MUSCLE

Network of Excellence

Multimedia Understanding through Semantics, Computation and Learning

Project no. FP6-507752

Deliverable D.5.1

Single Modality Processing

State-of-the-Art Report

Due date of deliverable: 01.09.2004
Actual submission date: 01.09.2004

Start date of Project: 1 March 2004 Duration: 48 Months

Name of responsible editor(s):

- Nozha Boujemaa (nozha.boujemaa@inria.fr), Valérie Gouet-Brunet

Revision: 1.0

	Project co-funded by the European Commission **within the Sixth Framework Programme (2002-2006)**	
	Dissemination Level	
PU	Public	X
PP	Restricted to other programme participants (including Commission Services)	
RE	Restricted to a group specified by the consortium (including Commission Services)	
CO	Confidential, only for members of the consortium (including Commission Services)	

Keyword List:

WP5
Single Modality Processing

State of the Art Report

Edited by Nozha Boujemaa[1], Valérie Gouet-Brunet[1]

Acknowledgements
Mathieu Coutaud[1] and Hichem Houissa[1]

[1]INRIA - Imedia, France
www-rocq.inria.fr/imedia/Muscle/WP5

Contents

List of authors

Part I : State of the Art on Image and Video Processing

Editors : Valérie Gouet-Brunet, Nozha Boujemaa

FT R&D, France	
Nathalie Laurent	nathalie.laurent@francetelecom.com
Christophe Laurent	christophe2.laurent@francetelecom.com
Mariette Maurizot	mariette.maurizot@wanadoo.fr
ENST, France	
Béatrice Pescquet-Popescu	beatrice.pesquet@enst.fr
Gemma Pielle	piella@tsi.enst.fr
INRIA-Vista, France	
Patrick Bouthemy	Patrick.Bouthemy@irisa.fr
Frederic Cao	Frederic.Cao@irisa.fr
Gwenaelle Piriou	Gwenaelle.Piriou@irisa.fr
Thomas Veit	Thomas.Veit@irisa.fr
Patrick Perez	Patrick.Perez@irisa.fr
ISTI-CNR Group, Italy	
Emanuele Salerno	emanuele.salerno@isti.cnr.it
Ovidio Salvetti	Ovidio.Salvetti@isti.cnr.it
UCL, Uk	
Fred Stentiford	f.stentiford@ee.ucl.ac.uk
TAU-VISUAL, Israel	
Nahum Kiryati	nk@eng.tau.ac.il
Nir Sochen	sochen@math.tau.ac.il
UFR, Germany	
Nikos Canterakis	canterakis@informatik.uni-freiburg.de
UPC, Spain	
Miriam Leon	mleon@gps.tsc.upc.es
Joan Llach	Joan.Llach@thomson.net
Montse Pardàs	montse@gps.tsc.upc.es
Philippe Salembier	philippe@gps.tsc.upc.es
AUTH, Greece	

Costas Cotsaces	cotsaces@aiia.csd.auth.gr
Irene Kotsia	ekotsia@aiia.csd.auth.gr
Nikos Nikolaidis	nikolaid@aiia.csd.auth.gr
Ioannis Pitas	pitas@aiia.csd.auth.gr
Vasilios Solachidis	vasilis@zeus.csd.auth.gr

ICCS-NTUA, Greece

Petros Maragos	maragos@cs.ntua.gr
Anastasia Sofou	sofou@cs.ntua.gr
Iasonas Kokkinos	jkokkin@cs.ntua.gr
Rapantzikos Konstantinos	rap@image.ntua.gr
Raouzaiou Amaryllis	araouz@image.ntua.gr
Ioannou Spyros	sivann@image.ntua.gr
Karpouzis Konstantinos	karpou@image.ntua.gr

INRIA-IMEDIA, France

Nozha Boujemaa	Nozha.Boujemaa@inria.fr
Valérie Gouet	Valerie.Gouet@inria.fr
Hichem Sahbi	Hichem.Sahbi@inria.fr
Anne Verroust	Anne.Verroust@inria.fr
Sabri Boughorbel	Sabri.Boughorbel@inria.fr

Part II : State of the art in selected areas of Speech and Audio Processing

Editor : Khalid Daoudi

ICCS-NTUA, Greece

A. Katsamanis	nkatsam@cs.ntua.gr
V. Pitsikalis	vpitsik@cs.ntua.gr
P. Maragos	maragos@cs.ntua.gr

AIIA Lab, Aristotle Univ. of Thessaloniki, Greece

E. Benetos	
M. Kotti	
C. Kotropoulos	costas@zeus.csd.auth.gr

VUT, Austria

A. Rauber	rauber@ifs.tuwien.ac.at
T. Lidy	lidy@ifs.tuwien.ac.at

Part III : State of the Art on Natural Language Processing

Editor : Gregory Grefenstette

AIIA Lab, Aristotle Univ. of Thessaloniki, Greece	
Nikoletta Bassiou	nbassiou@zeus.csd.auth.gr
Constantine Kotropoulos	costas@zeus.csd.auth.gr
George Almpanidis	galba@zeus.csd.auth.gr
Athanasios Papaioannou	apapaion@csd.auth.gr
CEA LIST, France	
Gregory Grefenstette	Gregory.Grefenstette@cea.fr
Patrick Hede	Patrick.Hede@cea.fr
Svitlana Zinger	zinger@enst.fr
IRISA, Rennes, France	
Pascale Sebillot	Pascale.Sebillot@irisa.fr
CNRS, France	
Fathi Debili	fathi.debili@ehess.fr

Introduction

The aim of the WP5 is to improve performance of each single modality processing (still images and video, speech and audio, text) in the perspective of multimedia understanding. The choice of Muscle was to keep all these modalities in the same WP even this involves different technical communities. This choice is motivated by the fact that we are convinced that these communities will benefit from putting together their technical expertise and have the opportunity to know how much progress is achieved in the other different modalities in the field of information retrieval by content. The natural language processing community is one of the earliest to deal with information retrieval. Visual content retrieval community learns a lot form NLP community in this field. More recently, audio and speech community have investigated this topic.

This document is the first deliverable for WP5 activities that consists on the state of the art for each single modality. It is structured in three major parts as follows:

- Part I: State of the art for Image and Video Processing
- Part II: State of the art in selected area of Speech and Audio Processing
- Part III: State of the art in Natural Language Processing

Part I

State of the art for Image and video processing

Chapter 1

Introduction

This report provides an overview about recent methods and algorithms for processing video and image data. The report concentrates on techniques and methods that have been used in the broad context of multimedia understanding. It addresses the whole spectrum of required methologies from visual features (static and dynamic), organization of visual features, segmentation, saliency and attention, and object detection and recognition.

Chapter 2

Visual Features for CBR

Content-based retrieval (CBR) from image and multimedia databases is one key application for multi-modal interfaces and search in multimedia databases. This chapter addresses the issue of feature extraction for modeling the visual appearance of images and their use for indexing the database. A variety of feature extraction algorithms exists, including the most classical ones: color, shape and texture. This state of the art will go beyond this simple "low-level" features and extract also more semantic information from the images using structural properties of images and patterns in terms of high-level features. In particular, we present in section 2.1 un local description of the image, which is based on points of interest and that is relevant for queries on parts of images or objects. In section 2.2, 3D shape descriptors are investigated. Finally, the chapter ends in section 2.3 with an overview on visual saliency approaches, that play a crucial role in analyzing the fast amount of visual information.

2.1 Local descriptors for content-based image retrieval

When considering sub-image retrieval or object recognition, local image characterization approaches provide better results than global characterization approaches classically based on color, texture And shape. It allows to gain robustness against occlusions and cluttering since only a local description of the patch of interest is involved. Such approaches are said to be *salient features-based*, because the information extracted from the image is condensed into limited but salient sites. Many kind of salient features can be envisaged : they can be represented by salient regions resulting from an image segmentation step [CBGM02, JJG01], or by non-connected image zones resulting from the construction of saliency maps as proposed by Itti *et al.* in [LCE98]. They can also be represented by edges [GS00, ST02], junctions, or special points [BS96]. This last case is the most encountered one and conducts to the extraction of points of interest (often called key points or salient points).Using points of interest is mainly motivated by observing that they provide the most compact representation of the image content by limiting the correlation and redundancy between the detected features. Indeed, contrary to an edge that exhibits a grey-level regularity along the contour curve, or a region in which all pixels fulfill some homogeneity criterion, there does not exist evident correlation nor dependency between non-connected and isolated salient points. Moreover, due to their sparse spatial distribution, object or image recognition based on salient points is much more robust to occlusions.

It is also important to note that the use of points of interest are often inspired from many psychovisual works [Mar82, Gor97] that have shown that the sensitivity of the HVS (*Human Visual System*) is not uniformly distributed across the image content. This observation leads to saliency-based image indexing approaches [AMS$^+$00] in which the first step consists in condensating the global information contained in the image into a limited number of feature values. Consequently, these salient features are to be selected with precision and in accordance with the properties of the HVS since the computation of the signature will uniquely depend on them. Ideally, salient features should be robust to geometric transforms, slight changes of viewpoint and variations of imaging conditions. The robustness regarding to coding schemes is also of prime importance for indexing purpose as underlined in [JB99]. Finally, most of the global image content should be contained in the extracted salient features or in a close neighborhood.

Image retrieval based on local descriptors follows a classical computation flow : first a robust salient feature detector must be designed. Such a detector determines the location of salient points in the image, plus the support regions around them where the local descriptor is to be computed. Second, a rich and compact descriptor is computed for each salient point, by analyzing the image information within the support region exhibited. Finally, a similarity measure must be found to compare two salient point signatures.

This survey is organized as follows : in section 2.1.1, we remind and briefly describe the point detectors we have encountered in the literature of Computer Vision. We revisit the classical Harris and Stephens point detector before presenting derived approaches that involve different support regions according to the involved image transformations. Other different techniques are also listed. In section 2.1.2, we revisit the local descriptors usually associated to such points of interest and support regions.

2.1.1 Point of interest detection

In the context of content-based images retrieval, the extracted points must have the following characteristics : they represent single points located in image area where the information is considered as perceptually relevant. Ideally, the point detector should have a good repeatability, i.e. should be able to repeat the extracted points from an image to another whatever the photometric/geometric transforms involved (translation, rotation, changes of scale, of viewpoint, of illumination, etc).

Since the early work of Moravec [Mor77] for stereo matching, many point extractors have been proposed in the literature of Computer Vision. Few comparison studies has been done for these approaches. See for example the ones of 2000 for grey value images [SMB00] and color images [GMDP00].

The most popular one is probably the *Harris and Stephens detector* [HS88], which has been used first for stereo purposes and then for image retrieval. It is based on the computation of the local auto-correlation function M of the signal at each pixel location. Large eigenvalues of this function indicate large curvatures in the two directions, which denotes the presence of a salient point. The eigenvalues computation is replaced by computing local maxima of the function $Det(M) - k.Trace(M)^2$. A modified version of this detector has been proposed in [SM97] by improving the computation of the spatial derivatives with precise gaussian derivatives. This precise version allows to gain in repeatability, as demonstrated in [SMB00]. The precise version of the Harris detector has been also extended to deal with color images in [MGD98], where it gets a better repeatability [GMDP00]; see for example the figure 2.1.

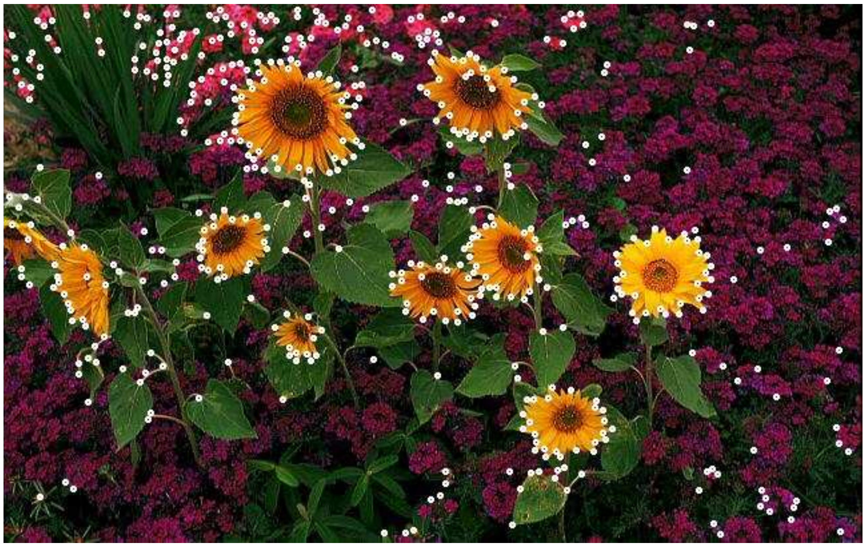


Figure 2.1: Automatic Harris Color Points extraction (500 points).

To obtain robustness to changes of viewing conditions, point of interest detectors should be invariant to image transformations. We revisit the more recent works that have been done for some usual image transformations. The encountered approaches provide specific point of interest detectors associated to specific support regions.

Scale invariance

The Harris detector (and its precise or color version) is invariant to rotation but is very sensitive to changes in image scale. Recent works proposed derived or new detectors to achieve scale invariance. The problem of identifying an appropriate scale for feature detection has been studied by Lindeberg who has described it as a problem of selection of a characteristic scale [Lin93, Lin94]. From these considerations, several works on scale invariance have been proposed for local features. They are described in the next paragraphs.

In [Lin98], Lindeberg has found a stable keypoint location in scale space by searching for 3D maxima of a function based on the Laplacian normalized with the scale. On the other hand, Lowe considered the local extrema in scale-space of Differences of Gaussian images [Low99]. Such points of interest are often called *DoG points*. As demonstrated by Lowe and evaluated later in [MS01], the DoG approach represents a close approximation of the Laplacian one, which is successfully compared to other functions (the Gradient and the standard Harris functions).

The paper [MS01] also presents an extension of the Harris precise detector. Here, the *Harris-Laplace detector* is introduced. It consists in extracting Harris points at a characteristic scale : each point is localized in 2D with the Harris function and then in scale-space where the Laplacian attains a maximum over scales. The authors demonstrate that this detector provides the best repeatability according to other scale-space detectors like the Laplacian one, the DoG one, the Gradient one and the standard Harris.

All these approaches provide points of interest that are invariant to rotation and scale changes. The selected scale determines the size of the support region to consider around the point (usually a uniform gaussian kernel) during the local description step.

Affine transformations

Some recent approaches have been proposed to adapt the support region to affine transformations (skew and stretch). In [Bau00], Baumberg extracts interest points at several scales and then adapts the shape of the support region to the local image structure using an iterative procedure based on the second order moment matrix. The uniform gaussian kernel which usually defines the support region is replaced by an ellipsoidal one.

In [MS02], the scale and the location of the points are directly extracted in an affine invariant way. The Harris-Laplace detector is extended to cope with such a transformation. The key idea is that the eigenvalues of the second order moment matrix computed in a point can be used to normalize the region according to affine transforms. The properties of the second order matrix were also explored in [SZ01], but their goal was to obtain an affine invariant texture descriptor.

Other techniques : spatial-frequency approaches

Different approaches have been proposed to extract points on sharp region boundaries instead of on corners. They are usually based on wavelet salient features detection. In [LSBJ00a], the authors propose a multi-resolution approach where salient points are associated to highest wavelet coefficient values. Paper [LLV03a] presents a similar approach that is independent of the wavelet filter size and that does not favor any contour direction. This approach also proposes a method for automatically determining the optimal number of salient points to extract.

2.1.2 Local image description

The second step consists in designing salient signatures that describe the support regions associated to the extracted points. Here the main task is to bridge the gap between image semantics and pixels. As underlined in [LNMT04], a saliency-based image representation paradigm can be defined by observing that two different approaches can be used to describe an image from a limited number of points of interest : a *global salient approach* and a *local salient approach*. In the former case, the technique consists in extracting a unique signature by considering globally the information contained in the support regions. This method can be considered as a tradeoff between classical global approaches working on the entire image and purely saliency-based approaches. Indexing proposals belonging to this class of methods include [STLH00, NM01, WJKB00]. In the latter case, the approach consists in considering independently the information contained in each support region, resulting in the computation of a local signature per salient point. In this document, we are mainly focusing on this second kind of technique.

The description is a function of the photometric and geometric information around the point and be the most compact possible. Many techniques have been developed, they depend on the considered application and more precisely on the image transformations involved. Roughly speaking, two kind of local signatures can be envisaged : the signatures coming from the global image indexing approaches (color histograms, Gabor jets, etc.) and the purely salient signatures that are dedicated to the use of a salient point detector.

The simplest local description that can be associated to a point of interest has been proposed for stereo matching purposes [ZDFL95] : it is the vector of the pixels located in a window around the point. A cross-correlation measure serves as similarity measure between two points. But the high dimensionality of such a descriptor makes it unapplicable for content-based images retrieval purposes. More compact point descriptors must be designed.

The differential invariants family

A set of image derivatives computed up to a given order approximates a point of interest neighborhood. Such a set is usually referred to *local jet* and the local description obtained is invariant to image translation. A stable estimation of the derivatives can be obtained by convolution with Gaussian derivatives. From the work of Koenderink [KVD87] and Florack [FtHRKV91, FtHRKV94] on the properties of local derivatives, a lot of work has been done on differential descriptors. Basically, the components of the local jet can be combined to obtain invariance to image rotation, providing differential quantities which are invariant under SO(2) group of similitudes. Such an approach has been used in [SM97, GB01] for image retrieval purposes and in [MGD98] for stereo matching purposes. When considering grey value images, they need to be computed up to order 3 [SM97], providing a set of 9 invariants. A generalization to color images has been proposed in [MGD98]. The use of the color information allows to reduce the description to the first order invariants, providing 8 color invariants less sensitive to noise than high order derivatives [GB02].

Another technique, referred as *steerable filters* [FA91], consists in computing the derivatives of the local jet in a particular direction. For instance, steering derivatives in the direction of the gradient makes them invariant to rotation.

Some similar approaches of local description are based on *complex filters*. In [Bau00] and[SZ02] such filters are derived from the family $K(x, y, \theta) = f(x, y)e^{i\theta}$, where θ is the orientation. [Bau00] uses Gaussian derivatives for $f(x, y)$ whereas in [SZ02] a polynomial is applied. These filters differ from the Gaussian derivatives by a linear coordinates change in filter response space. When dealing with rotation invariance, it is necessary to consider the modulus of each response filter (the phase is sensitive to rotation).

All these descriptors are invariant to 2D translation and rotation. To adapt them to scale changes, a solution consists first in computing the derivatives by using a Gaussian kernel parameterized by the characteristic scale founded during the salient point detection (see section 2.1.1) and second in normalizing such derivatives by this scale. See for example [MS01] where steerable filters are used as local descriptors. Another solution consists in directly normalizing the support regions associated to the points computed in scale-space : all supports are mapped to a circular region of constant radius. As a result, descriptors computed on such supports become invariant to scale changes. The same technique can be applied for affine transformations by mapping to a circular window the ellipsoidal support associated to point of interest extracted under affine transformations. The invariants obtained from this normalized support are invariant to affine transformations. Such an approach is employed in [MS02, MS03].

When considering illumination changes, two approaches are possible to normalize the invariants to affine illumination changes [Gro00, GM02]: first, it is possible to normalize the support region. Second, the derivatives can be normalized by dividing them with the appropriate power of the gradient magnitude. Such a normalization eliminates the linear factor. The differentiation operation naturally eliminates the offset one.

The SIFT descriptors

Based on the DoG detector [Low99], Lowe has proposed the *Scale Invariant Feature Transform approach* (SIFT) for describing the local neighborhood of such points. Here, the local neighborhood of the salient point is described with multiple images that are orientation planes representing a number of gradient orientations. This descriptor can be divided by the square root of the sum of squared components to obtain illumination invariance. This approach provides robustness against localization errors and small geometric distortions. Its main drawback is the high dimension of the feature space involved (128 gradient orientations). A complete use of The SIFT approach has been presented in [Low04] for reliable point matching between different views of an object or scene.

A recent performance evaluation of local descriptors [MS03] has shown that the SIFT descriptor performs best. Steerable filters come second. As noticed by the authors of this evaluation, this signature can represent a good choice given the low dimensionality.

Other techniques : spatial-frequency approaches

All other approaches are generally picked in the global image indexing literature. Many techniques which describe the frequency content of the images have been developed. The well known Fourier transform decomposes the image content into basis functions. But this representation makes not explicit the spatial relations between points and the basis functions are infinite, therefore difficult to adapt to a local approach. Frequency approaches based on the Gabor transform [Gab46] allow to take spatial relations between points into account, see for example [WJKB00]. Approaches based on wavelets have been also explored [Vet95, STLH00]. These classes of techniques are usually employed in the context of texture classification. Finally, the classical use of the color information have also been proposed [LLV03a, STLH00]. However, all the obtained signatures need to be high dimensional to give a precise signature.

Similarity measures

Except for the SIFT descriptor where the distance measure is L2 [Low99], the similarity between descriptors is usually computed with the Mahalanobis distance. The involved covariance matrix takes the different magnitudes, possible correlations and variability of the feature components into account. Point descriptors are subject to different kinds of noises : in practice, they are sensitive to image acquisition (sensors and sampling errors), to numerical errors, to points of interest delocalization, etc.

These considerations show the importance of the similarity measure which must be carefully chosen for the considered descriptor to achieve best performances. An optimal similarity measure is directly related to the shape of their variability. When considering the Mahalanobis distance, this noise can be integrated in the similarity measure via the covariance matrix Λ. When a model of noise of the components cannot be specified, the way to estimate the covariance of the components comes down to different empiric solutions :

- Estimating Λ from the whole available data. This simple solution generates weights that are not discriminant, since representing a rough model of noise. Even so, this is the most common solution encountered to compare features with the Mahalanobis distance;

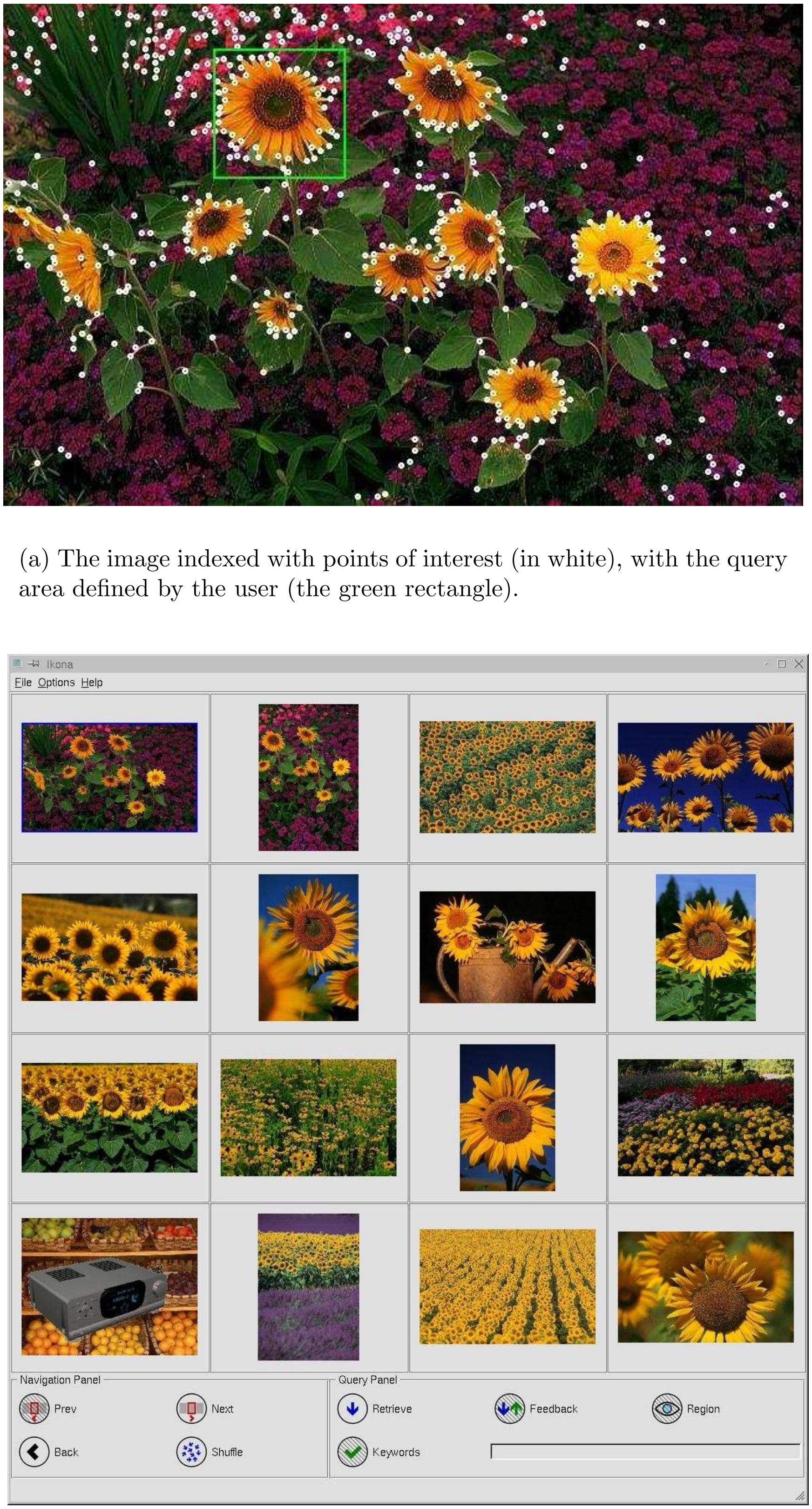
- Estimating Λ from points of interest whose local neighborhood is submitted to synthetic photometric and geometric transformations and perturbations that usually apply to images;

- Estimating Λ from training sequences of real images. Several points on different images with representative perturbations are tracked and a combination of the covariance matrices obtained can be used as the model of noise of point characterization. This solution has been adopted for example in [SM97] for image retrieval and in [GL04] for object tracking in video sequences.

Adding geometrical constraints between points of interest

Semi-local geometric constraints that consider the spatial relations between neighbor salient points of the same image can be added to enrich the local point description [SM97, GB01, BFGS04]. Obviously, they depend on the image transformations involved by the application. We present in figure 2.2 an example of object retrieval with local descriptors, with the CBIR system IKONA[1] developed by INRIA.

[1] http://www-rocq.inria.fr/imedia/ikona.html



(a) The image indexed with points of interest (in white), with the query area defined by the user (the green rectangle).

(b) 16 first results of this sub-image retrieval, presented by decreasing order of similarity.

Figure 2.2: An example of object retrieval with local descriptors. The query is a particular object (here a sunflower).

2.2 3D shape matching methods for content-based retrieval

A survey [ZP04] has been done recently on this subject and two papers have reported the result of comparative studies of the efficiency of several shape descriptors [PSF04, TV04].

2.2.1 Specificity of 3D shapes

Shape representation

In most of the 3D shape retrieval methods, the 3D shapes are represented by their boundary as a 2 dimensional surface such as a polyhedral surface, but they can also be voxelized.

Existing shape benchmarks (Princeton University [2], Konstanz University [3] and Utrecht University [4]) propose databases of polyhedral models but not give a guaranty on the way the surface is described (well defined or a soup of faces). Then some shape matching approaches will need a pre-processing phase to obtain a manifold surface.

Pose normalization

The 3D models may have arbitrary scale and position. As most of the dissimilarity measures are sensitive to translations and rotations, it is necessary to put the models in a canonical coordinate system. Most of the approaches use a PCA transform (the method of [DVR01] is used in several approaches).
The PCA transforms do not give the axes orientation (as noticed in [ROT02, ZP02, MSO00]) and may lead to ambiguities in the choice of the coordinate axes when the eigenvalues are similar. The dissimilarity measures of [ZP02, MSO00] take into account the coordinate systems obtained by switching the principal axes, which leads to a real invariance in rotation.

2.2.2 3D Shape matching methods

The approaches will be grouped into four categories :

- the methods based on a spatial decomposition of the 3D object,
- the methods working on the surface of the objects,
- the graph based methods,

[2] Princeton 3D models search engine : http://shape.cs.princeton.edu/search.html.

[3] Konstanz 3D models similarity search engine :http://merkur01.inf.uni-konstanz.de/cccc/.

[4] Utrecht 3D shape retrieval engine :http://www.cs.uu.nl/centers/give/imaging/3drecog/3dmatching.html.

- the 2D visual similarity based methods.

Spatial decomposition

Most of the methods grouped in this section put the 3D shape in a spatial grid formed either by : voxels, concentric spheres, tetrahedra, etc... and use this spatial decomposition to compute the shape descriptor :
- **Voxels :**

in [EPT00], a shape descriptor based on wavelets is proposed.

in [TV03], the descriptor is a set of weighted points, each weighted point being a "salient" point belonging to a voxel intersecting the surface. The dissimilarity measure uses the proportional transportation distance [GV02] derived from the Earth Mover's distance.

in [KCD$^+$03], a reflective symmetry descriptor is built measuring the reflective symmetry of the object w.r.t. every plane passing by the centroid.
- **Subdivision of a sphere** [MAS99] : the 3D object is put into a decomposition of a sphere in angular sectors and concentric spheres. A histogram based on this decomposition is built and a quadratic form distance measure is used to compare two objects.
- **Concentric spheres :** two approaches are based on spherical harmonics [VS02],[KF02]. Another is based on Zernike descriptors [NK03b, NK03a] and seems to obtain better results.
- **Tetrahedra** [ZC01] : the descriptor is computed using Fourier transforms on an approximation of the 3D object into a set of tetrahedra.
- **2D slices** [NK01] : the similarity is evaluated by computing 2D distance fields between the successive slices of the voxelized shapes and transforming them into a 3D distance between the two objects. This computation must be made for each couple of 3D objects and then it can be long.
- **principal axes of inertia** [ROT02] : the 3D models are parameterized along the principal axes of inertia and three histograms (moment of inertia about the axis, average distance of the surface about the axis and variance of the distance from the surface to the axis) are computed.

Working on the surface of the objects

These two approaches work only on well defined surfaces :
- **Shape index histogram (SF3D)** [ZP02] : a histogram of the shape index (function of the two principal curvatures on continuous surfaces) is adapted for polyhedral surfaces.
- **Curvature map** [JAP03] : the shape is represented by its curvature map, transforming the 3D problem into an image indexing problem. This method only works on genus 0 surfaces.

The other approaches work on the facets and do not need to have a manifold surface.
- **Complex EGI** [KI93] : a histogram is built on the Gaussian sphere. In this representation, the weight associated with each outward surface normal depends on the facet's area and on the distance from the origin. This method is sensitive to change in facets orientations.
- **Moment-based** [EPT00, MEA01] : these two approaches calculate the 3D statistical moments and compare them. The second approach adapts the similarity computation with SVM to obtain an interactive method.
- **Cord-based** [EPT00] : three histograms describing the distribution of angles between the cords and the first (respectively the second) principal axis and the distribution of the length of the cords are built. The Hamming distance is used when comparing the histograms.
- **Shape distribution** [ROD02, CYIR02] : these approaches compute a shape descriptor by combining probability distributions on shape functions representing the geometric properties of the 3D shape (angle between three points of the surface, distance between the centroid and a point, distance between two points of the surface, etc...). These methods achieve scale and affine invariance.
- **Hough transform** [ZP02] : a shape descriptor based on a 3D Hough transform is computed. The ambiguities of the PCA algorithm are taken into account in their method to obtain a rotation invariant shape descriptor. The 3D Hough transform descriptor was evaluated on the Princeton database (907 models). Some retrieval results (Human biped & Fighter jet airplane) are presented in figure 2.3

Graph based approaches

The main point of the graph based approaches is to extract the topological information which may be lost in the other approaches. These methods are affine invariant but most of them only work on well defined polyhedral models. The similarity measure between shapes consists in a recursive graph comparison.
The method of [HSD03] works on a voxelization of the shape and use a thinning algorithm to compute the skeletal nodes.
The skeletal graphs of [MHK01, DBS03, TS04] use Reeb graphs based on the computation of a geodesic distance on the surface.

2D visual similarity-based methods

When the polyhedral shape is ill-defined, the similarity can be computed comparing their appearance :
- in [CK01], each object is associated to 72 viewpoints computed by rotating the camera along an axis. The views are then structured in a shock graph. The comparison of the view of an object with the views of the database models uses

a shock graph matching.
- in [ROT03], after a normalization process, they compute depth images from 42 viewpoints and compare the shape feature vectors using Zhang's Fourier descriptors [ZL02].
- in [DCO03], one hundred orthogonal projections are encoded by Zernike moments and Fourier descriptors for retrieval.

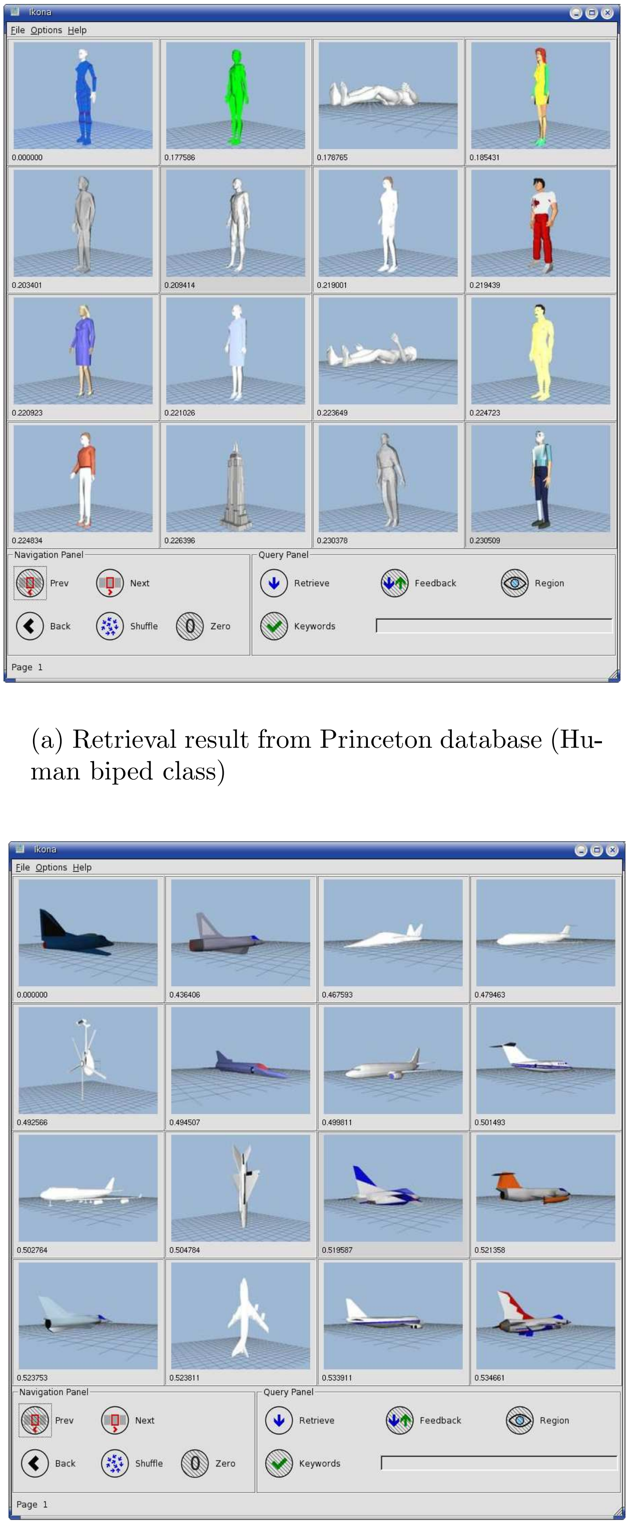
2.3 Saliency & Visual Feature Organization

Knowing what is important in an image or video is the first vital step in determining the associations between multimedia objects and hence capturing the relationships and ultimately the semantic content of those objects. Recent research on modeling human visual attention is showing distinct promise in this area. Unimportant background regions in an image may be compressed without significantly affecting the overall perceptual quality of the image. Therefore visual saliency operators play a crucial role in analyzing the fast amount of visual information.

2.3.1 A state of the art

Visual systems that have evolved in nature appear to exercise a mechanism that places emphasis upon areas in a scene without necessarily recognising objects that lie in those areas. Organisms having the benefit of vision are thereby able to sense danger and direct attention rapidly towards the unusual without having to tolerate the initial delay of a recall from memory. Treisman and Gelade [TG80] in their feature-integration theory make the distinction between scenes that require relatively slow focused attention to analyse and those which can be processed more rapidly during a pre-attentive stage. Evidence shows that it is relatively easy to spot a target ”O” that pops out amongst a background of ”N”s and ”T”s, but time consuming to locate one's offspring in a school photograph. They posed, as others have done since, the question why features that distinguish a target from the background in pre-attentive vision when applied separately often do not when they appear in conjunction. Wolfe [Wol98] emphasises that there is no clear distinction between slow serial and fast parallel mechanisms in visual search and that the evidence shows a continuum of search results in which both mechanisms perhaps play a part.

Desimone and Duncan [DD95] in their review confirm that strengthening the perceived grouping between targets and background objects makes the background harder to ignore. Furthermore they suggest that there is little evidence that there are separate representations for different features such as orientation and colour in the cortex stating that cells that respond to a single type of stimulus have yet to be found. They conclude that pre-attentive vision is an emergent



(a) Retrieval result from Princeton database (Human biped class)

(b) Retrieval result from Princeton database (Fighter jet airplane)

Figure 2.3: An evaluation of the 3D Hough transform on the Princeton database (907 models). The screenshots belong to the IKONA platform developed at INRIA. They show the performance and robustness of the approach in the 3D-model retrieval process. Here, the system returns a list of outputs ranking on the degree of similarity to the object query (First position). In these tables, we just show the top sixteen matches.

property of competitive interactions acting in parallel across the visual field and not the binding together of a set of separate feature measures.

Nothdurft [Not00] has shown that the salience of targets in human vision is nearly always increased if multiple contrasts in orientation, luminance and motion are present. The addition was mostly nonlinear, which indicated that the underlying mechanisms were not independent and not separable as Desimone and Duncan suggest.

Experiments by Reinagel and Zador [RZ99] using eye trackers show that subjects are attracted by image regions possessing high contrast and also by neighbourhoods in which pixel correlations drop off rapidly with distance. They observe that this strategy increases the entropy of the effective visual input and is in accord with measures of informativeness and cognitive surprise.

Early computational models by Koch et al [KU85] of attention generate maps that encode the visual environment for different elementary features such as orientation of edges and colour contrast and combine these into an overall saliency map. The most conspicuous neighbourhoods are taken to be those that give rise to the most activity in the saliency map as a result of activity in corresponding feature maps. Many authors have put emphasis upon identifying specific features that are normally associated with saliency and combining these to produce such maps. Milanese et al [MGP95] used five feature maps in their analysis of static scenes. After applying filters and passing the maps through a nonlinear relaxation process, they are averaged and thresholded to produce the saliency map. Itti et al [LCE98] have defined a system which models visual search in primates. 42 features based upon linear filters and centre surround structures encoding intensity, orientation and colour, are used to construct a saliency map that reflects areas of high attention. Supervised learning is suggested as a strategy to bias the relative weights of the features in order to tune the system towards specific target detection tasks. They observed that salient objects appearing strongly in only a few dimensions may be masked by noise present in a larger number of dimensions. Han et al [25, 29] use the Itti model to determine the best positions to seed a region growing algorithm for object extraction.

Osberger and Maeder [OM98] identified perceptually important regions by first segmenting images into homogeneous regions and then scoring each area using a number of intuitively selected measures. The approach was limited by the success of the segmentation techniques used. Luo and Singhal [LS00] also devised a set of intuitive saliency features and weights and used them to segment images to depict regions of interest. The integration of the features was not attempted. Marichal et al [MVDM96] used fuzzy logic to segment object boundaries before assigning levels of interest based upon a number of criteria. Zhao et al [ZSO$^+$96] employed features reflecting size, distance from the centre of the image, boundary length, compactness and colour to determine region importance.

Reisfeld et al [RWY95] detect symmetries in grey level images as a means of identifying certain image locations that are worthy of attention. The work relies heavily upon edge extraction and was extended to colour images by Heidemann [Hei04].

Walker et al [WCT98] suggested that object features that best expose saliency are those which have a low probability of being mis-classified with any other feature. Mudge et al [MTV87] also considered the saliency of a configuration of object components to be inversely related to the frequency that those components occur elsewhere. Oliva et al [OATH03] again use the idea that frequent features in images are more likely to belong to the background. They use orientation, scale and texture features to calculate saliency likelihoods and also incorporated contextual information.

Several models of visual attention [LCE98, GPW03, MTF$^+$04] have their counterparts in surround suppression in primate V1 [16]. Grigorescu et al. [GPW03] use this model and confirm qualitative explanations of visual pop-out effects. They obtain their results using pre-selected orientation sensitive Gabor energy filters, and apply their ideas to contour detection. Whereas Grigorescu et al obtain their results using pre-selected orientation sensitive Gabor energy filters, Stentiford [Ste03, Ste01] takes an approach that is not so restricted and generates features appropriate to the image in question. In this way features that determine levels of attention, which may or may not be orientation dependent, are not excluded from consideration.

By observing that visual receptive fields are sensitive to orientations, scales and intensity variations, some wavelet-based salient point detectors have also been proposed [LSBJ00b, LLV03b] leading to a description of the image content by a limited number of single points that are considered as perceptually relevant.

2.3.2 Applications

There is no doubt that the selection of salient image regions is a key problem in many fields of image processing. These cover object recognition [HS93], content-based image retrieval [Ste03, PG98, TSL$^+$01], image compression [Ste01, PS00, HXCM04], natural image classification [RL04], and medical image registration [MV98]. Specific examples include :

- Fan et al [FXM$^+$03] use attentional heuristics to identify sub-pictures likely to be of interest and more easily displayed while browsing on the small screen of a PDA.
- Models of visual attention have been used to extend the JPEG 2000 compression standard with some success [NCSP03].
- A blurring strategy based on visually salient regions of video frames is used to compress video signals without substantially interfering with human fixations [Itt04].

- Measures of attention have been shown to enhance the performance of object recognisers by focusing analysis on the regions likely to contain relevant objects [HXCM04].

2.3.3 Discussion

Almost all researchers in the field agree that saliency plays a major part in the recognition processes that take place in the human visual system, but the exact relationship is unclear. There is also agreement that local salient features certainly include those that occur relatively rarely in a scene. Indeed it is hard to see how a feature that was present throughout an image could be visually attentive.

Many attention models make use of plausible features such as orientation, intensity, colour and texture as a first step in the identification of regions of interest. Concepts of centre surround suppression are often applied that emphasise the importance of local differences and lead to an attention map that displays the relative saliency of various regions in the image.

However, there is no evidence that visual systems make use of a small number of predefined features that we might intuitively believe to be important in attention mechanisms. The favoured few certainly characterise saliency in many images, but there is an infinitude of other possible feature combinations to choose from and there will always therefore be images whose saliency is not captured. Salient features are most likely to be different in different images and the diverse possibilities emerge only at the time the images are viewed. This means that implementations of attention mechanisms that use any form of pre-defined feature measurements may preclude solutions in the search space and be unable to handle unseen material.

The same problem arises again in recognition where we are seeking features in common between two or more patterns. There is no guarantee that a pre-selected feature set and associated representations will encompass the inter-pattern similarities necessary to obtain a satisfactory performance, although sensible feature design based on prior knowledge will always yield good results. Indeed it is a criticism that techniques in feature extraction and pattern classification are treated separately with the result that features are not constructed according to classification performance [DHS01].

One of the most difficult problems in saliency-based image representation is to design efficient local descriptors that aim at describing the image content in the neighborhood of each salient area. Usually, these descriptors are designed by hand and are inspired from intuitive considerations. However, there is no evidence that the resulting descriptor provides a sufficiently good image representation and/or a good discriminative power for an image recognition task. Finally, when an image is represented by a set of salient zones, the natural order between pixel values is lost and the image can no longer be seen as a vector, increasing thus the complexity of the matching step for assessing the similarity between images or objects of interest. Classically, image registration approaches are used but in these cases, salient zones are considered independently of each others. However, human eyes are able to recognize a scene from a set of focus of attention and saccadic eye movements, showing that the information is accumulated during these movements. A registration approach is not able to reach this property since the local information contained in each salient feature ignores the local information localized in other places of the image. To simulate the image categorization from saccadic movements, a solution has been proposed in [RL04] that consists in linking the detected salient features to obtain an attributed string of local descriptors. An order between salient features is thus recovered and the matching step can be performed thanks to a string-edit distance.

Bibliography

[AMS$^+$00] A.W.M.Smeulders, M.Worring, S.Santini, A.Gupta, and R.Jain. Content-Based Image Retrieval at the End of the Early Years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, December 2000.

[Bau00] A. Baumberg. Reliable feature matching across widely separated views. In IEEE *Conference on Computer Vision and Pattern Recognition*, pages 774–781. 2000.

[BFGS04] N. Boujemaa, F. Fleuret, V. Gouet, and H. Sahbi. Automatic textual annotation of video news based on semantic visual object extraction. In *IS&T/SPIE Conference on Storage and Retrieval Methods and Applications for Multimedia*. 2004.

[BS96] C. Bauckhage and C. Schmid. Evaluation of keypoint detectors. Technical report, INRIA, 1996.

[CBGM02] C. Carson, S. Belongie, H. Greespan, and J. Malik. Blobworld: Image Segmentation Using Expectation-Maximization and Its Application to Image Querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8):1026–1038, August 2002.

[CK01] C. Cyr and B. Kimia. 3D object recognition using shape similarity-based aspect graph. In *International Conference on Computer Vision (ICCV)*, pages 254–261. 2001.

[CYIR02] L. S. C. Yiu Ip, D. Lapadat and W. C. Regli. Using shape distributions to compare solid models. In *7th ACM Symposium on Solid Modeling and Applications*. June 2002.

[DBS03] W. C. R. D. Bespalov, A. Shokoufandeh and W. Sun. Scale-space representation of 3d models and topological matching. In *Proceedings of the eighth ACM symposium on Solid modeling and applications*, pages 208–215. ACM Press, 2003.

[DCO03] Y. S. D.Y. Chen, X.P. Tian and M. Ouhyoung. On visual similarity based 3D model retrieval. *Computer graphics forum*, 22(3):223–232, 2003.

[DD95] R. Desimone and J. Duncan. Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.*, 18:193–222, 1995.

[DHS01] R. O. Duda, P. E. Hart., and D. G. Stork. *Pattern Classification.* John Wiley, 2001.

[DVR01] D. S. D. Vranic and J. Richter. Tools for 3D-object retrieval : Karhune-Loeve transform and spherical harmonics. In *2001 Workshop Multimedia Signal Processing.* Cannes, France, October 2001.

[EPT00] T. N. E. Paquet, M. Rioux A. Murching and A. Tabatabai. Description of shape information for 2-D and 3-D objects. *Signal Processing : Image Communication*, 16:103–122, 2000.

[FA91] W. Freeman and E. Adelson. The design and use of steerable filters. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 13(9):891–906, 1991.

[FtHRKV91] L. Florack, B. ter Haar Romeny, J. Koenderink, and M. Viergever. General intensity transformations and second order invariants. In *7th Scandinavian Conference on Image Analysis*, pages 338–345. Aalborg, Denmark, 1991.

[FtHRKV94] L. Florack, B. ter Haar Romeny, J. Koenderink, and M. Viergever. General intensity transformations and differential invariants. *Journal of Mathematical Imaging and Vision*, 4(2):171–187, 1994.

[FXM$^+$03] X. Fan, X. Xie, W. Ma, H. Zhang, and H. Zhou. Visual attention based image browsing on mobile devices. In *ICME*. 2003.

[Gab46] D. Gabor. Theory of communication. *Journal of the Inst. Elec. Eng.*, 93(26):429–457, 1946.

[GB01] V. Gouet and N. Boujemaa. Object-based queries using color points of interest. In *IEEE Workshop on Content-Based Access of Image and Video Libraries (CBAIVL 2001)*, pages 30–36. Kauai, Hawaii, USA, 2001.

[GB02] V. Gouet and N. Boujemaa. On the robustness of color points of interest for image retrieval. In IEEE *International Conference on Image Processing (ICIP'2002).* Rochester, New York, USA, September 2002.

[GL04] V. Gouet and B. Lameyre. SAP: a robust approach to track objects in video streams with snakes and points. In *To appear in the British Machine Vision Conference*. Kingston University, London, UK, September 2004.

[GM02] V. Gouet and P. Montesinos. Normalisation des images en couleur face aux changements d'illumination. In *13me Congrs Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle*, volume II, pages 415–424. Angers, France, 2002.

[GMDP00] V. Gouet, P. Montesinos, R. Deriche, and D. Pelé. Evaluation de détecteurs de points d'intérêt pour la couleur. In *Reconnaissance des Formes et Intelligence Artificielle*, volume II, pages 257–266. Paris, France, 2000.

[Gor97] Gordon I.E. *Theories of Visual Perception*. Wiley, second edition edition, 1997.

[GPW03] C. Grigorescu, N. Petkov, and M. Westenberg. Contour detection based on nonclassical receptive field inhibition. In *IEEE Trans on Image Processing*, volume 12, pages 729–739. 2003.

[Gro00] P. Gros. Color illumination models for image matching and indexing. In *Proceedings of* 15^{th} *International Conference on Pattern Recognition*, pages 580–583. Barcelona, Spain, September 2000.

[GS00] Gevers T. and Smeulders W.M. PicToSeek: Combining Color and Shape Invariant Features for Image Retrieval. *IEEE Transactions on Image Processing*, 9(1):102–119, January 2000.

[GV02] P. Giannopoulos and R. Veltkamp. A pseudo-metric for weighted point sets. In *European Conference on Computer Vision (ECCV 2002)*, pages 715–730. 2002.

[Hei04] G. Heidemann. Focus-of-attention from local color symmetries. In *IEEE Trans Pattern Analysis and Machine Intelligence*, volume 26. jul 2004.

[HS88] C. Harris and M. Stephens. A combined corner and edge detector. *Proceedings of the* 4^{th} *Alvey Vision Conference*, pages 147–151, 1988.

[HS93] R. M. Haralick and L. G. Shapiro. *Computer and Robot Vision*, volume 2. Addison-Wesley, 1993.

[HSD03] N. G. H. Sundar, D. Silver and S. Dickinson. Skeleton based shape matching and retrieval. In *Shape Modeling and Applications Conference, SMI 2003*. Seoul, Korea, May 2003.

[HXCM04] Y. Hu, X. Xie, Z. Chen, and W. Ma. Attention model based progressive image transmission. In *ICME*. 2004.

[Itt04] L. Itti. Automatic attention-based prioritization of unconstrained video for compression. In *Proc. SPIE Human Vision and Electronic Imaging IX (HVEI04)*, volume 5292, pages 272–283. San Jose, California, jan 2004.

[JAP03] A. D. B. J. Assfalg and P. Pala. Retrieval of 3D objects using curvature maps and weighted walkthroughs. In *International Conference on Image Analysis and Processing (ICIAP 03)*. Mantova, Italy, September 2003.

[JB99] J.-M. Jolion and S. Bres. Influence du codage jpeg sur des descripteurs d'images. *Traitement du signal*, 15(4):309–320, 1999.

[JJG01] J.Z.Wang, J.Li, and G.Wiederhold. SIMPLicity: Semantics-Sensitive Integrated Matching for Picture Libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9):947–963, September 2001.

[KCD$^+$03] M. Kazhdan, B. Chazelle, D. Dobkin, T. Funkhouser, and S. Rusinkiewicz. A reflective symmetry descriptor for 3D models. *Algorithmica*, 38(2):201–225, November 2003.

[KF02] M. Kazhdan and T. Funkhouser. Harmonic 3D shape matching. In *SIGGRAPH 2002 Technical Sketch*. 2002.

[KI93] S. B. Kang and K. Ikeuchi. The complex EGI : a new representation for 3D pose determination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(7):707–721, July 1993.

[KU85] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurbiology*, 4:219–227, 1985.

[KVD87] J. Koenderink and A. Van Doorn. Representation of local geometry in the visual system. *Biological Cybernetics*, 55:367–375, 1987.

[LCE98] L.Itti, C.Koch, and E.Niebur. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, November 1998.

[Lin93] T. Lindeberg. On scale selection for differential operators. In *Proceedings of 8^{th} Scandinavian Conference on Image Analysis, Tromso, Norway*, pages 857–866. May 1993.

[Lin94] T. Lindeberg. Scale-space theory: A basic tool for analysing structures at different scales. *Journal od Applied Statistics*, 21(2):224–270, 1994.

[Lin98] T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):79–116, 1998.

[LLV03a] C. Laurent, N. Laurent, and M. Visani. Color image retrieval based on wavelet salient features detection. In *Intl. Workshop on Content-Based Multimedia Indexing*. 2003.

[LLV03b] C. Laurent, N. Laurent, and M. Visani. Color image retrieval based on wavelet salient features detection. In *Third International Workshop on Content-Based Multimedia Indexing*, pages 327–334. Rennes, sep 2003.

[LNMT04] C. Laurent, N.Laurent, M.Maurizot, and T.Dorval. In Depth Analysis and Evaluation of Saliency-based Color Image Indexing Methods using Wavelet Salient Features. *to appear in Multimedia Tools and Applications*, 2004.

[Low99] D. G. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, pages 1150–1157. Corfu, Greece, 1999.

[Low04] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Accepted for publication int the International Journal of Computer Vision*, 2004.

[LS00] J. Luo and A. Singhal. On measuring low-level saliency in photographic images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR-00)*, pages 84–89. Los Alamitos, June 13–15 2000.

[LSBJ00a] E. Loupias, N. Sebe, S. Bres, and J. Jolion. Wavelet-based salient points for image retrieval. In *IEEE Int. Conf. On Image Processing*. Vancouver, September 2000.

[LSBJ00b] E. Loupias, N. Sebe, S. Bres, and J.-M. Jolion. Wavelet-based salient points for image retrieval. In *ICIP*. Vancouver, sep 2000.

[Mar82] D. Marr. *Vision*. W.H. Freeman and Company, 1982.

[MAS99] H.-P. K. M. Ankerst, G. Kastenmller and T. Seidl. 3D shape histograms for similarity search and classification in spatial databases. In Springer, editor, *6th International Symposium on Large Spatial Databases (SSD'99)*, volume 1651, pages 207–226. Hong Kong, China, 1999.

[MEA01] A. T. M. Elad and S. Ar. Content based retrieval of VRML objects : an iterative and interactive approach. In *EG Multimedia*, pages 97–108. September 2001.

[MGD98] P. Montesinos, V. Gouet, and R. Deriche. Differential Invariants for Color Images. In *Proceedings of* 14^{th} *International Conference on Pattern Recognition*. Brisbane, Australia, 1998.

[MGP95] R. Milanese, S. Gil, and T. Pun. Attentive mechanisms for dynamic and static scene analysis. *Optical Engineering*, 34(8):2420–2434, aug 1995.

[MHK01] T. K. M. Hilaga, Y. Shinagawa and T. L. Kunii. Topology matching for fully automatic similarity estimation of 3D shapes. In *SIGGRAPH 2001 Conference Proceedings*, pages 203–212. Los Angeles, August 2001.

[Mor77] H. Moravec. Towards automatic visual obstacle avoidance. In *Proceedings of the* 5^{th} *International Joint Conference on Artificial Intelligence*, page 584. Cambridge, Massachusetts, Etats-Unis, August 1977.

[MS01] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *International Conference on Computer Vision*. Vancouver, Canada, July 2001.

[MS02] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *European Conference on Computer Vision*, volume 1, pages 128–142. 2002.

[MS03] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *Intl. Computer Vision and Pattern Recognition*, 2003.

[MSO00] T. K. M.T. Suzuki and N. Otsu. A similarity retrieval of 3D polygonal models using rotation invariant shape descriptors. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC2000)*, pages 2946–2952. Nashville, Tennessee, October 2000.

[MTF$^+$04] O. L. Meur, D. Thoreau, E. Francois, D. Barba, and P. L. Collet. From low-level perception to high-perception level: a coherent approach for va modelling. In *Proc. SPIE Human Vision and Electronic Imaging IX (HVEI04)*, volume 5292, pages 284–295. San Jose, California, jan 2004.

[MTV87] T. Mudge, J. Turney, and R. A. Volz. Automatic generation of salient features for the recognition of partially occluded parts. *Robotica*, 5:117–127, 1987.

[MV98] J. Maintz and M. Viergever. A survey of medical image registration. *Medical Image Analysis*, 2(1):1–36, 1998.

[MVDM96] X. Marichal, C. D. Vleeschouwer, T. Delmot, and B. Macq. Automatic detection of interest areas of an image or of a sequence of images. In *ICIP'96 - Third IEEE International Conference on Image Processing*, volume 3, pages 371–374. Lausanne, Suisse, sep 1996.

[NCSP03] A. Nguyen, V. Chandran, S. Sridharan, and R. Prandolini. Guidelines to using region of interest coding in jpeg 2000. In *Proc. Int. Symposium on Digital Signal Processing and Communication Systems (DSPCS)*. Gold Coast, Australia, dec 2003.

[NK01] M. Novotni and R. Klein. A geometric approach to 3D object comparison. In *International Conference on Shape Modeling and Applications*, pages 167–175. Genova, May 2001.

[NK03a] M. Novotni and R. Klein. Shape retrieval using 3D Zernike descriptors. *Computer-Aided Design*, 36(11):1047–1062, September 2003. Special issue on solid modeling theory and applications.

[NK03b] M. Novotni and R. Klein. Zernike descriptors for content based shape retrieval. In *Proceedings of the eighth ACM symposium on Solid modeling and applications*, pages 216–225. 2003.

[NM01] N.Sebe and M.S.Lew. Salient Points for Content-based Retrieval. In *Proc. of the British Machine Vision Conference*. Manchester-UK, 2001.

[Not00] H. C. Nothdurft. Salience from feature contrast: additivity across dimensions. *Vision Research*, 40:1183–1201, 2000.

[OATH03] A. Oliva, M. C. A. Torralba, and J. Henderson. Top-down control of visual attention in object detection. In *ICIP*. 2003.

[OM98] W. Osberger and A. J. Maeder. Automatic identification of perceptually important regions in an image using a model of the human visual system. In *International Conference on Pattern Recognition*. Brisbane, Australia, August 1998.

[PG98] E. Pauwels and G.Frederix. Finding salient regions in images: Nonparametric clustering for image segmentation and grouping. In *Computer Vision and Image Understanding*, pages 73–85. aug 1998.

[PS00] C. M. Privitera and L. W. Stark. Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Trans Pattern Analysis and Machine Intelligence*, 22(9):970–982, 2000.

[PSF04] M. K. P. Shilane, P. Min and T. Funkhouser. The Princeton shape benchmark. In *Shape Modeling and Applications Conference, SMI'2004*, pages 167–178. Genova, Italy, June 2004.

[RL04] J. Ros and C. Laurent. Natural image classification using foveal strings. In *The International Workshop on Image, Video, and Audio Retrieval and Mining*. Sherbrooke, oct 2004.

[ROD02] B. C. R. Osada, T. Funkhouser and D. Dobkin. Shape distributions. *ACM Transactions on Graphics*, 21(4):807–832, 2002.

[ROT02] M. I. R. Ohbuchi, T. Otagiri and T. Takei. Shape-similarity search of three-dimensional models using parameterized statistics. In *Pacific Graphics 2002*. Beijing, October 2002.

[ROT03] M. N. R. Ohbuchi and T. Takei. Retrieving 3D shapes based on their appearance. In *5th ACM SIGMM Workshop on Multimedia Information Retrieval (MIR 2003)*. Berkeley, Californica, USA, November 2003.

[RWY95] D. Reisfeld, H. Wolfson, and Y. Yeshurun. Context-free attentional operators: the generalised symmetry transform. *Int. J. Computer Vision*, 14:119–30, 1995.

[RZ99] P. Reinagel and A. M. Zador. Natural scene statistics at the centre of gaze. *Network: Computation in Neural Systems*, 10(4):341–350, November 1999.

[SM97] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–534, May 1997.

[SMB00] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *International Journal of Computer Vision*, 37(2):151–172, 2000.

[ST02] S. Shim and T.Choi. Edge Color Histogram for Image Retrieval. In *IEEE International Conference on Image Processing*, volume 3, pages 957–960. Rochester (NY), September 2002.

[Ste01] F. W. Stentiford. An estimator for visual attention through competitive novelty with application to image compression, April 2001.

[Ste03] F. W. Stentiford. An attention based similarity measure with application to content-based information retrieval. In *Storage and Retrieval for Media Databases*, volume 5021. Santa Clara, jan 2003.

[STLH00] N. Sebe, Q. Tian, M. S. Lew, and T. Huang. Color indexing using wavelet-based salient points. In IEEE *Workshop on Content-based Access of Image and Video Libraries*, pages 15–19. 2000.

[SZ01] F. Schaffalitzky and A. Zisserman. Viewpoint invariant texture matching and wide baseline stereo. In *International Conference on Computer Vision*, pages 636–643. Vancouver, Canada, 2001.

[SZ02] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets. In *European Conference on Computer Vision*, pages 414–431. 2002.

[TG80] A. M. Treisman and G. Gelade. A feature-integration theory of attention. *ACM Computing Surveys Cognitive Psychology*, 12:97–136, 1980.

[TS04] T. Tung and F. Schmitt. Augmented reeb graphs for content-based retrieval of 3D mesh models. In *Shape Modeling and Applications Conference, SMI 2004*, pages 157–166. IEEE, Genova, Italy,, June 2004.

[TSL$^+$01] Q. Tian, N. Sebe, M. Lew, E. Loupias, and T. Huang. Image retrieval using wavelet-based salient points. *Journal of Electronic Imaging, Special Issue on Storage and Retrieval of Digital Media*, 10(4):835–849, oct 2001.

[TV03] J. Tangelder and R. Veltkamp. Polyhedral model retrieval using weighted point sets. In *Shape Modeling and Applications Conference, SMI 2003*. IEEE, Seoul, Korea, May 2003.

[TV04] J. Tangelder and R. Veltkamp. A survey of content based 3D shape retrieval methods. In *Shape Modeling and Applications Conference, SMI 2004*, pages 145–156. IEEE, Genova, Italy, June 2004.

[Vet95] J. Vetterli. *Wavelets and Subband Coding.* Prentice Hall, 1995.

[VS02] D. V. Vranic and D. Saupe. Description of 3D-shape using a complex function on the sphere. In *IEEE International Conference on Multimedia and Expo (ICME 2002)*, pages 177–180. Lausanne, August 2002.

[WCT98] K. N. Walker, T. F. Cootes, and C. J. Taylor. Locating salient object features. In *British Machine Vision Conference*. 1998.

[WJKB00] C. Wolf, J.-M. Jolion, W. Kropatsch, and H. Bishof. Content based image retrieval using interest points and texture features. In IAPR *International Conference on Pattern Recognition*, volume 4, pages 234–237. Barcelona, Spain, 2000.

[Wol98] J. M. Wolfe. *Visual Search in Attention*. Psychology Press, 1998.

[ZC01] C. Zhang and T. Chen. Efficient feature extraction for 2D/3D objects in mesh representation. In *ICIP 2001*. Thessaloniki, 2001.

[ZDFL95] Z. Zhang, R. Deriche, O. Faugeras, and Q.-T. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence*, 78:87–119, December 1995.

[ZL02] D. S. Zhang and G. Lu. Shape-based image retrieval using generic Fourier descriptor. *Signal Processing : Image Communication*, 17(10):825–848, November 2002.

[ZP02] T. Zaharia and F. Prteux. Shape-based retrieval of 3D mesh models. In *2002 IEEE International Conference on Multimedia and Expo (ICME'2002)*. Lausanne, August 2002.

[ZP04] T. Zaharia and F. Prêteux. 3D versus 2D/3D shape descriptors : A comparative study. In *Proceedings SPIE Conference on Image Processing : Algorithms and Systems III - IS & T / SPIE Symposium on Electronic Imaging, Science and Technology '03*, volume 5298. San Jose, CA, January 2004.

[ZSO$^+$96] J. Zhao, Y. Shimazu, K. Ohta, R. Hayasaka, and Y. Matsushita. An outstandingness oriented image segmentation and its application. In *Int. Symposium on Signal Processing and its Applications*, volume 1, pages 45–48. aug 1996.

Chapter 3

Object Detection and Recognition

Computer vision was born with the aim at building machines that "can see" [EC01]. This program has a twofold implication, since it may be intended either as a tool to understand the underlying properties and mechanisms of human vision, or as the theoretical basis for a set of applications aiming at extracting high level perceptual information from the analysis of natural scenes. The first aspect is mailnly related to similar applications of artificial intelligence (in the sense of simulating intelligent behavior); the second aspect is related to emulating intelligent responses to external stimuli, and has been studying all the perceptual cues that are useful to extract structure information from image data. A distinction exists between low-level vision and high-level vision, based on the increasing structuredness of the specific cues treated as primary data for processing. Attempts have been made to integrate different processing levels and different visual cues.

An extended notion of image analysis and understanding, however, needs extended concepts of "images" and "objects" to be introduced, which are not necessarily "natural" : any n-dimensional map of a scalar or vector quantity is referred to as an image.

The list of different applications of object detection and recognition would be practically infinite. We can safely say that almost every technological area is affected by this discipline. Medical diagnosis [BNPS03], industrial nondestructive testing [CAC$^+$00], remote sensing [LDZ00], optical character recognition [VBT02], virtual/ augmented reality [ZCHS03], automatic image and video indexing [BS02], visual databases [FPZ04], security/ surveillance systems [CLK00] are just a few examples. In particular, remote sensing and security issues are assuming an increasing importance in facing the present natural or human-made threats.

Computer vision subfields are now very mature disciplines, in the sense that they have sound foundations in different branches of mathematics, statistics and information theory. Another pointer to maturity is the very large number of groups and researchers involved in this area, and the wide range of applications, especially in such fields as face recognition [ZCPR03], where decades of basic and applied research have produced efficient commercial off-the-shelf systems.

Very robust techniques have been developed and applied successfully in many real-world applications. Also, this has been made possible by the increasing capabilities of general-purpose and specialized hardware (ultra-fast CPUs, digital signal processors, etc.), which brought many tasks to quasi-real-time performances. To exemplify, pattern recognition techniques are being increasigly employed in medical imaging. Pattern segmentation and recognition are the artificial vision tools that permit to build robust systems in face and fingerprint recognition, in security/surveillance applications, and optical character recognition in automatic text processing. Another emerging field where object recognition has a relevance is in protection of the intellectual property. Besides common practice applications, a number of real-world problems remain unsolved for the lack of robust solutions. Normally, the existing algorithms are effective in at least partially controlled environments (indoor, structured. . .), but their performance degrades in more general cases. Both new insights in image/scene perception theory and new computational paradigms are now being exploited to solve problems that lack efficient solutions.

Object detection and recognition, just as all image science, is a highly interdisciplinary area. Since its early days, the views of engineers, mathematicians, physicists, psychologists and computer scientists were put together to find solutions at different levels of imaging and vision problems. All the recent algorithmic advancements are now being used to improve current solutions or to find solutions to new problems. Besides the classical tools of functional artificial intelligence (neural networks and algorithms [CU93]), the fields of statistical image and signal processing [CiA02] are gaining new importance, also due to the availability of new probabilistic models and large image databases. New algorithmic approaches, such as nature-inspired computation [Fre02] [ea00], complex systems and chaos [BDT99], etc., are also used, often showing both promising results and new deep insights.

As mentioned above, just a few detection/recognition procedures work robustly when faced with an unstructured (and dynamically changing) environment. As an example, we mention again the case of remote sensing and surveillance, with all the related subfields and critical applications, such as pollution control, disaster management, anti-terrorist measures. On one hand, this depends on the environment [NS00], but not totally on that [EC01]. The extraction of visual information can also depend on the attitude of the agent [Mog97]. This fact has already been recognized in biological visual systems, where the identification of the different classes in a scene depends on the pre-attentive or attentive mood and on the specific task being performed. Further theoretical developments and interaction among different disciplines are needed to attack this problem, whereas classifying and recognizing all the objects in a scene still remains impossible. To find empirically effective and robust applications on the basis of current theoretical developments is thus the first challenge for the near future. This is a nontrivial task when truly real-world problems are to be solved. One of the proposed solutions is to rely on active vision systems [BPPP98], which are able to enrich the data space by capturing suitably chosen additional images.

Benchmarking

A crucial problem when working recognition systems are to be developed is benchmarking. Indeed, some standardized procedure to evaluate different products is necessary, along with a common test database, in order to allow a significant comparison to be made. To this end, very large image databases must be available, whose entries should be as similar as possible to the actual test images. This means that evaluation procedures and databases should be specifically application-oriented [ZCPR03]. Some work has been done for optical character recognition and fingerprint classification [BBT02] [BCG$^+$94], and for face recognition, where, among others, the different releases of the FERET face database and evaluation procedure are available [PWJR98] [PMRR00]. Benchmarking is particularly important in the mentioned applications, where many commercial off-the-shelf systems are now available. Refining the evaluation procedures would further help monitoring the performances of the different algorithms in their evolutions, and identifying the most urgent and/or promising research areas. At present, for example [ZCPR03], two major problems in face recognition are recognition under illumination and pose variations.

This chapter is organized as follows: section 3.1 details approaches for statistical data processing whereas section 3.2 presents a typology of approaches for object recognition based on the processed data. Then, the two last sections deal with specific approaches of detection and recognition for particular classes of objects: faces in section 3.3 and text in section 3.4.

3.1 Statistical data processing

The basic cue to accomplish object detection, classification and recognition is the feature. A feature can be any attribute of an image in any relevant space, such as the original pixel space or whatever other representation space (Fourier, wavelet…). Any image can thus be described by means of a feature vector.

With "appearance-based vision", we mean an approach to detect, classify, and recognize objects on the basis of sets of 2D images, or views. The concept of appearance-based vision is opposed to the one of "geometry-based vision", which relies on matching geometric primitives or templates. In the case of detection and classification of objects within complicated backgrounds, the appearance-based approach has proved to achieve better results than the approaches based of feature or template matching [ZCPR03].

The "features" on which appearance-based vision has to rely are not predetermined primitives or templates (possibily invariant to particular transformations), but are implicitly defined from the sample images chosen, from which they are to be learned [FFFP03] [FPZ03]. This implies that large sets of training images should be available. Typical tasks are identification (to determine which component in a database, if any, is represented in the input image), verification (to determine whether or not an input image actually corresponds to a particular object in a database) [ZCPR03], and detection [Ros02] (to determine presence and location of one or more objects of interest in an image). The choice of the analysis strategy depends on the final goal.

For appearance-based vision, the pixel (scalar or vector) values can simply be assumed as image features. This results, in general, in a very high-dimensional feature space, thus implying the need for dimensionality reduction [BPPP98] [Mar01] [Nel98]. In any case (whether pixel-based or not), appearance matching can be performed by global (holistic) or local methods. Among global mehods, we find, for example, histogram matching and eigenspace representation. The first approach exploits suitable distances between histograms to establish a similarity between an input pattern and the items in a database. The second approach projects the input image onto the subspace spanned by the training set, in order to find the minimum-distance element and be able to label the input pattern.

The relevant image space can be built, for example, through principal component analysis (PCA [CiA02]), or linear discriminant analysis (LDA [BHK97b] [Mar01]). In the PCA approach, a basis of orthogonal "eigenimages" is found [Kir90] [PT91], from which it is possible to reconstruct any element of the original dataset to a fixed level of accuracy. Projecting an input image onto this basis and finding the minimum-distance element of the original dataset amounts to classify (or recognize) the image. The effectiveness of this strategy is based on data variance, that is, the eigenimages computed are generalized vectors oriented in the same direction as the eigenvectors of the data covariance matrix. This is not always a valid criterion to discriminate between different classes. In the LDA approach (or *Fisher analysis*), indeed, the selection criterion is based on maximizing the between-class variances over the within-class variances, instead of maximizing the data variances.

Both PCA and LDA rely on second-order statistics to reduce the dimensionality of the image space. However, important information may be encoded in higher-order statistics, so that higher-order methods should outperform second-order methods [BMS02]. In other words, the common redundancy that results from sensed data is thought of as originated by the combination of independent components. This is the rationale of the independent component analysis (ICA) approaches. In [BMS02], two different ICA approaches and a hybrid approach are shown, with application to face recognition. One possibility is to derive a set of independent images on which the training data are projected, as is done by the PCA approach. In this case, the basis faces are not only uncorrelated, but independent of each other. Another possibility is to evaluate a number of components whose combinations can give the individual faces in the database via independent coefficients. In this way, a factorial code for face images is found.

Statistical data processing has also revealed its usefulness in tasks that are only indirectly related to object detection and recognition. As ICA is a typical strategy to perform blind source separation [Hyv00], it can be used to extract or remove interfering patterns from the raw data prior to proper recognition. This has been done, by ICA or similar techniques, in astrophysical image processing [KBP+03] and in document image analysis [TBS04]. In the former case, blind source separation is used as a sort of object detection tool, aiming at separating statistically distinct diffuse patterns superimposed to one another. In the latter, the aim is to reduce or cancel various degradations typically affecting ancient documents, in order to improve the error rate of optical character recognition systems or the legibility by a human reader. This can be done by both ICA and different color decorrelation approaches [TSMB04], among which PCA, applied to color or multispectral images of the document pages.

When their basic assumptions are not verified, the ICA-based approaches can fail their goal. Research is being done to extend the conditions of separability in these cases. Moreover, flexible algorithms are being developed to possibly take prior information into account, when the problem at hand is not totally blind (for example, when the statistical distribution of the sources is known, or when prior knowledge on the data model is available).

In particular, there is an additional hypothesis besides mutual independence that requires nongaussian distributed sources (except at most one). It has been shown that, for nonstationary processes, separation can be achieved even if more than one sources are Gaussian. If the sources are known to possess temporal structures, then the independence assumption can be dropped. The resulting separation techniques are commonly denoted as *dependent component analysis* (DCA) [Bar00] [Bar01] [BBB+03] [DCP03].

Another hypothesis that does not need to be verified is the linearity of the source mixture. Especially when there is a sensible within-class variability (see [BMS02] for face recognition), this hypothesis cannot be considered true. To overcome this difficulty is not an easy task, and no general solution has been found so far. Research is active in determining the kind of nonlinearities that allow the problem to be efficiently solved [CiA02] [YAC98] [Bac02].

3.2 A typology of methods for object recognition

Object recognition is the field of automatic description and classification of patterns [DHS00]. The recognition task may be 1) supervised where the object class is known, and some *a priori* informations are given or 2) unsupervised where the object is assigned to unknown class. Interest in the area of object recognition has been renewed due to the emerging applications such as handwritten identification, human/face recognition, image databases browsing, etc. Major approaches are presented in the following.

3.2.1 Template Matching

Template matching is one of the earliest approaches to object recognition. It is also known by the name of model-based object recognition [Agg95, Lon98, CF01, PS00]. The object to be recognized is matched against the template by taking into account various poses (translation and rotation) and scale changes. The template and the similarity measure (correlation) are optimized based on the available training set. Template matching is computationally consuming, but the availability of faster processors has now made this approach more feasible. The rigid template matching mentioned above, while effective in some application domains, has a number of disadvantages. For instance, it would fail if images presents distortion due to some processing or viewpoint change. Deformable template models [Gre93, BK89] can be used for object recognition when the deformation cannot be directly modelled.

3.2.2 Statistical Approach

The goal of the statistical approach is to establish decision boundaries given a set of training images [HAMJ01]. This boundaries should split feature space such that different image categories occupy compact and disjoint regions in a d-dimensional feature space. The effectiveness of the representation space is determined by how well images from different classes can be separated. In the statistical decision theoretic approach, the decision boundaries are determined by the probability distributions of the images belonging to each class, which must either be specified or learned [DGL96, DHS00, Vap98]. One can also take a discriminant analysis-based approach to classification : First a parametric form of the decision boundary (e.g., linear or quadratic) is specified; then the best decision boundary of the specified form is found based on the classification of training images. Non parametric approaches based on maximizing the margin between the decision boundaries and the training samples leads to the SVM classifiers [Vap98, Bur98, CHV99] (see figure 3.1).

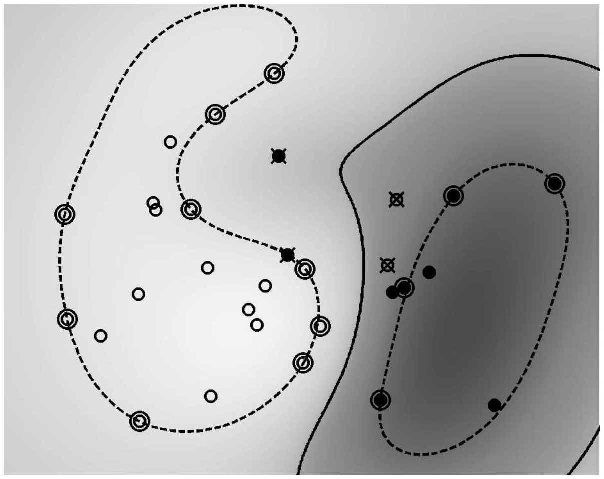


Figure 3.1: Support Vector Machine is becoming the state of the art of object recognition using statistical learning approach. It consists in finding a decision boundary that maximizes the margin between positive and negative training examples. In this 2D toy problem, positive example are represented with white circles and negative example are represented with black filled circles. The SVM maps data into a high dimensional space via kernels where a linear decision boundary is constructed. Such boundary corresponds to a non-linear one in the input space which is depicted with a solid line in the figure. One of the advantage of SVM is that only few training examples (surrounded examples in the figure) called support vectors are involved in th construction of the classifer. Dotted lines depict the edge of the margin. The SVM allows also to handle outliers (crossed examples in the figure) using soft-margin versions.

3.2.3 Syntactic Approach

In many recognition tasks, hierarchical representation is more appropriate to adopt. The image is viewed as being composed of simple sub-images which are themselves built from yet simpler sub-images [Fu82, Pav77] called primitives. In syntactic object recognition, an analogy can be drawn between the structure of images and the syntax of a language. The images are viewed as sentences belonging to a language, primitives are viewed as the alphabet of the language, and the sentences are generated according to a grammar [FB86a, FB86b]. Thus, a large collection of complex images can be described by a small number of primitives and grammatical rules. The grammar for each image class must be inferred from the available training samples. The implementation of a syntactic approach, however, leads to many difficulties which primarily have to do with the segmentation of noisy images (to detect the primitives) and the inference of the grammar from training data.

3.2.4 Neural Networks

Neural networks can be viewed as parallel computing systems with a large number of simple processors having many interconnections [Tho96, Zha00, Sme95]. Neural networks have the ability to learn complex nonlinear input-output relationships [EPdRH02], use sequential training procedures, and adapt themselves to the data. In spite of the seemingly different underlying principles, most of the well known neural network models are implicitly equivalent or similar to classical statistical object recognition methods. In [Rip93] a discussion is given on this relationship between neural networks and statistical object recognition.

3.3 Faces

3.3.1 Face detection

Many methods for face detection are discussed in the literature, including artificial neural networks [RBK98] [Sun96], support vector machines [OFG97][EPPP00][RTSB01], Bayesian inference [CWT00], deformable templates [MYW$^+$99], graph-matching [LBP95], skin color learning [HAMJ01][SB00] and coarse-to-fine processing [FG01][VJ01]. Distinguishing factors include whether they can solve the face detection problem with real complex backgrounds and the run-time cost.

Finding faces on simple background

Color provides a computationally efficient method which is robust under rotations in depth, partial occlusions and can be used to model and filter skin color efficiently from a training set using standard classifiers [SB02, HAMJ01, YLW98, CG99]. Face detection can also be achieved using motion. The general idea is to detect differences between the current and the previous frame in a video sequence. If the difference between pixel values is greater than a given threshold, the movement is considered to be significant and the pixel is set to be of interest. Many prior knowledge (cf. below) can be used to decide whether a pixel of interest belong to a face or not [EJ95, ST98].

It is also advantageous to use prior knowledge. For face detection, we know that the head is located at the top of the body, that a human normally walks upright, etc [SP96, YC98]. For example, blinking patterns in an image sequence is an easy and a reliable mean to detect the presence of a face since blinking provides a space-time signal which is easily detected and unique to faces [GBGB01]. The fact that both eyes blink together provides a redundancy which makes it possible to discriminate faces from other motion in the scene. Furthermore, symmetry and the fixed separation between the eyes provides a way to estimate the size and the orientation of the head.

Finding faces in a complex background

This is the most interesting, challenging, and practical case, since it serves many applications. The generic approach is based on modeling the face appearance using an a priori geometric or learned model. Many methods have been proposed to perform face detection in a complex background using machine learning techniques such as neural network classifiers [RBK98][SP98], Bayesian inference [CWT00], support vector machines [OFG97][RTSB01] and eigenfaces [MP95]. Other techniques are based on the analysis of facial structures, and include : graph matching [LBP95], geometrical hashing [LSW98], edge counting and coarse-to-fine processing [FG01], together with AdaBoost [VJ01] and FloatBoost [LZZ$^+$02]. All these methods operate by extracting windows at different locations and scales; pre-processing subimages using normalization techniques and encoding them using an appropriate structure or a feature space. The underlying extracted information is classified as face/non face using a suitable classifier and a search strategy. Again, techniques differ in the training model used, the amount of training data necessary to capture the complexity of the decision boundary, the facial representation, and mainly the search strategy used to reduce computation.

3.3.2 Face recognition

Public transactions based on passwords, cards, etc., show only that the information provided by the user is valid, not that their rightful owner is present. Face recognition is *a science of identification*, which substitutes a key or a password by the facial characteristics which are unique for a given individual. This science measures the statistical variability of faces using a human population and requires a preliminary step of face localization.

At this time it is not yet clear whether face recognition is *holistic* or a *local task*, i.e., if it depends on the whole face characteristics or some of them. In the human visual cortex, face recognition is a dedicated task different from object recognition as a dedicated part in the brain is responsible for this task. The psychophysics aspects have been largely studied by biologists in order to understand face recognition in the human visual cortex [YE89][BHB93] and this will help developing original and effective face recognition algorithms.

In the remainder of this abstract, we will review the main representative state of the art techniques for face recognition based on modeling the face appearance. Existing techniques can be classified either as local or holistic [BP93, LLN00], they use intensity or infrared images [WPJW96] [SWNE01], their query mode can be face images or sketches [UL96][Kon96], and the nature of the task can be identification or verification [ZCR00]. We will illustrate the methods from the holistic or local processing point-of-views.

Holistic versus local methods

Holistic methods consider the whole face image in the recognition process. The basic method is correlation [BP93] which declares two faces as similar if their inner product is relatively high. This method is simple but sensitive to changes in the viewing conditions and the lighting effects. More sophisticated face recognition methods have been introduced among them the well studied Eigenfaces and Fisherfaces [SK87, MN95, PMS94, TP91]. The eigenface method finds the principal axes, in an Euclidean space, which maximize the variance of the faces through a training set. Then, faces are projected into a subspace spanned by the principal axes and the coefficients of projection will be used as a face description. In contrast to eigenface, fisherface finds the principal axes which make the between-class variance large while maintaining faces related to the same individual with a small within-class variance [BHK97a]. Active appearance models [CWT00, LW99] combine both shape and texture characteristics in order to train a projection matrix which is used to infer the parameters of face identity, pose and shape variations. Other methods for face detection, based on statistics and geometry, have been introduced including optical flow [LCK02], neural networks [CF90, HCZZ00], support vector machines [JKYL00, GLC00, LGL00], Hidden Markov models [Rig01, SY94, Nyf95, AB96, Eic02] and linear subspaces [Sha92, OvdM02, BJ03, BH01]. The drawback of holistic methods resides mainly in their sensitivity to partial occlusion, variations of the contrast and the pose. Local methods have been introduced to overcome these drawbacks and they usually proceed by extracting some facial components (mouth, nose, etc.) [GHL71, KK72, CB65, Ble66] [LHLM02, RY92, Hal91, Kan73] and by combining these components using appropriate classifiers [GLC00, HHP01, XPL02, HHP01, GZ01, TV00, CSS00].

Several methods for face component extraction and analysis have been introduced. A generalized symmetry operator is used in [RY92] to find the eyes and mouth locations. This approach stems from the almost symmetric nature of the face about a vertical line through the nose and a particular symmetry measure is maximized. [Hal91] used a template-based approach in order to locate the facial components. [Kan73] introduced a technique based on the use of a Laplacian operator in order to obtain a binary image and the facial components. [BP93] used instead a gradient operator in order to derive their edge map, which is partitioned into vertical and horizontal image gradients. The horizontal image gradient is used to extract the left and the right boundaries of face and the nose while the vertical image gradient is used to detect the head-top, the eyes, the nose base and the mouth. Face outline detection was performed using dynamic programming principle as the outline of a face can be followed by a line, so the order can be defined without ambiguity. The issues of the accuracy of face component extraction has been addressed in [Zha99]. In many systems, good recognition results are dependent on accurate component extraction and registration, so the performance may degrade when these components are not determined accurately.

[WFKvdM97] introduced a face recognition algorithm based on graphs. A face is seen as a graph where its nodes correspond the facial components and the arc-labels are the distances between these components. The facial components are extracted by maximizing a correlation between a bunch of Gabor filters related to a prototype graph. Afterward, each node in the graph is moved locally in order to maximize further this correlation and to localize the underlying facial component with higher precision. Face recognition is based on graph matching. In [GHL71, KK72, CB65, Ble66], the overall configuration of a face can be represented by a vector of numerical data representing the relative position and the size of the main facial components and the face outline coordinates. In [LHLM02], the authors use an augmented Gabor feature vector as a concatenation of multiple responses of Gabor filters at different orientation, scales and locations.

Statistical methods have also been used to recognize faces using their local components among them support vector machines [GLC00, HHP01, XPL02] and Ada-boost [GZ01]. For instance [HHP01] trained many SVM classifiers which are used as filters in order to locate the facial components (eyes, nose, tips of the nose, corners of the mouth, etc). The extracted components are used as input to a multi-class SVM which returns the identity of the aggregate face components. [GZ01] adapt the Ada-boost method [CSS00] for face recognition. Ada-boost is a method to combine a collection of weak classifiers (the weak learners are related to the facial components) to form a strong classifier after many iterations.

Hybrid methods

Hybrid methods and sensor fusion have received significant attention in the last decade. Sensor fusion has already been used in the combination of speech and face recognition for person identification. In the work of [GHS^{+}95], the conclusion was that the future in person identification lies in hybrid recognition systems.

The work of [Gor95] was the first attempt to conceive such a hybrid recognition system combining two template based face recognition systems for both frontal and profile faces. [AB96] introduced a parallel hybrid recognition system based on three face classifiers, namely the profile approach by [YJB95], an HMM similar to the one in [SY94], and the EigenFace approach by [TP91].

3.3.3 Emotions detection

There is a long history of interest in the problem of recognizing emotion from facial expressions [EF78], and extensive studies on face perception during the last twenty years [Ekm73] [DM75][SK84]. The salient issues in emotion recognition from faces are parallel in some respects to the issues associated with voices, but divergent in others.

As in speech, a long established tradition attempts to define the facial expression of emotion in terms of qualitative targets, i.e. static positions capable of being displayed in a still photograph. The still image usually captures the apex of the expression, i.e. the instant at which the indicators of emotion are most marked. More recently emphasis, has switched towards descriptions that emphasize gestures, i.e. significant movements of facial features.

In the context of faces, the task has almost always been to classify examples of archetypal emotions. That may well reflect the influence of Ekman and his colleagues, who have argued robustly that the facial expression of emotion is inherently categorical. More recently, morphing techniques have been used to probe states that are intermediate between archetypal expressions. They do reveal effects that are consistent with a degree of categorical structure in the domain of facial expression, but they are not particularly large, and there may be alternative ways of explaining them - notably by considering how category terms and facial parameters map onto activation-evaluation space [KK00].

Analysis of the emotional expression of a human face requires a number of preprocessing steps which attempt to detect or track the face, to locate characteristic facial regions such as eyes, mouth and nose on it, to extract and follow the movement of facial features, such as characteristic points in these regions, or model facial gestures using anatomic information about the face.

Facial features can be viewed [EF75] as either static (such as skin color), or slowly varying (such as permanent wrinkles), or rapidly varying (such as raising the eyebrows) with respect to time evolution. Detection of the position and shape of the mouth, eyes, particularly eyelids, wrinkles and extraction of features related to them are the targets of techniques applied to still images of humans. It has, however, been shown [Bas78], that facial expressions can be more accurately recognized from image sequences, than from a single still image. His experiments used point-light conditions, i.e. subjects viewed image sequences in which only white dots on a darkened surface of the face were visible. Expressions were recognized at above chance levels when based on image sequences, whereas only happiness and sadness were recognized at above chance levels when based on still images. Techniques which attempt to identify facial gestures for emotional expression characterization face the problems of locating or extracting the facial regions or features, computing the spatio-temporal motion of the face through optical flow estimation, and introducing geometric or physical muscle models describing the facial structure or gestures.

3.4 Text extraction methods in images and video sequences

The volume of data available nowadays in video format makes it necessary to create tools which allow to extract information from these video sequences in order to be classified (video indexing) or to be analyzed (video analysis) without human supervision. However, these tasks are very complex and remain an open problem. Therefore, the combination of different techniques could be of great interest to help solving this problem. Text extraction can be a key feature because it is usually synchronized and related to the scene in the sequence, obtaining extra information about the scene.

The contents in a video sequence can be perceptive, such as colour, shapes, textures, frequency changes, or semantic, such as objects or events and its relationships. The perceptive ones are easier to analyse automatically and the semantic ones are easier to handle linguistically. Therefore, since the computational cost of text extraction is lower than the cost of recognising objects, events or people, a good option is to detect and recognise text in a sequence.

At the moment there are many algorithms that are able to extract text from document files (OCR Optical Character Recognition) providing acceptable results. The main differences between OCR techniques and text recognition in video sequences are that document files are usually binary images where, of course, letters are static and the background is white or black, while in video sequences background is complex and characters can move throughout time. OCR will be the last step in text recognition in video sequences.
Two kinds of text can be found in video sequences :

Scene text : It is composed by text that is integrated in the scene and is captured directly by the camera. It is more difficult to recognise because it can appear in any tilt or perspective, different illumination, on both planar and ridged surfaces, complete or partially hidden.

Caption text : It is the text that is introduced artificially over the frame. It can be static or in motion, but in any case its aim is to be readable and comprehensive by a viewer. In some references it is also called artificial text or graphic text.

Mainly all the different methods try to detect and recognize caption text, but some of them show also tools to deal with both texts. A minority is targeting only scene text. The aim of most methods is to find out the content in the video sequence, and their results are usually utilised for indexation. In this state of the art we are focusing on caption text methods.

3.4.1 Text features

Some of the main caption text features are the following.

Contrast between text and backgroud

Contrast is an important feature since in most images text must be readable, it cannot be blurred or occluded. Very often a high contrast is required as well as a steady brightness. One of the main problems localizing text is unsurprisingly the low contrast and the complex background. When characters have similar hue and brightness to the background their detection is almost impossible. In these cases, some enhancement tools must be employed as a pre-processing step in order to improve the contrast. Some standards for subtitling recommend using a black background. If coloured background is used then a legible text colour should be chosen. The most legible text colours on a black background are white, yellow, cyan and green, but it should be avoided using magenta, red and blue. Moreover, it has been checked empirically that for caption text there is no rule for choosing a colour. As a consequence, to avoid the possibility of characters blurring with background, they are highlighted by increasing the contour contrast.

Spatial cohesion

TYPOGRAPHY : Typography refers to the type of font used. Indeed, some of them are useful because their structure avoids confusions among the different letters. This property takes into account not only able people, but physically handicapped too, like deaf viewers ? This feature doesn't help directly for the localization but for character recognition system (OCR).

SIZE : An important fact is that text has to be readable. On one hand, researchers on vision have investigated human eye resolution, concluding that when letters are smaller than a minimal size, people cannot differentiate them. The minimal high and width sizes are approximately 15 and 7 respectively, but these values differ a little in each article. On the other hand, a maximal size is also bounded. Other important value is the ratio between height and width of a letter, which takes values around 0.9. Another size constraint is called word length or sentence length. Characters have similar heights and spacing within a text string. Therefore, this feature allows separating words.

COMPACTNESS (or FILLFACTOR) : If a bounding box containing the letter is build, compactness is the relation between pixels belonging to the letter and those belonging to the background. These features can be applied for both a simple character and an entire word. Allowed values used to be included in the interval [0.1, 1] [LE98], but it depends on the authors' criteria.

HORIZONTAL GEOMETRY : Text possesses certain orientation. To make text more easily readable it is usually displayed horizontally, in languages that are written horizontally.

Textured appearance :

The two previous features, contrast and spatial cohesion, can cause that the text search turns into texture segmentation. There are some papers which describe quite well what 'textured appearance' means, such us : "For example, by looking at the comic page of a news paper a few feet away, one can probably tell quickly where the text is without actually recognizing individual characters.", see [WM99]. Therefore, considering text as a whole entity, it has enough features to be detectable as a texture. Problems can appear when image textures and text features are very similar, like the leaves of a tree or grass in a field. Normally both textures and edges are fine structures and a level-pixel processing is required in order to localize and extract them.

Colour homogeneity :
Some papers take colour homogeneity as the main feature, because of the fact that colour segmentation preserves characters contour better than for example contrast segmentation, which may blur some edges [LE98]. Colour homogeneity could be analyzed in two different ways. On one hand, assuming that all the characters have the same colour, a whole word or line could be detected. On the other hand, each character might be written in different colour, but the letter colour itself remains homogenous. Both possibilities can be solved with the same algorithms. Summing up, monochrome characters are found more frequently than polychrome. Both of them can be detected with colour segmentation, but perhaps it could be assumed that polychrome characters are related to some artistic more than an informative purpose, so that some authors tend to discard this kind of characters.

Strokes thickness :
Another feature that contributes to the textured appearance of the text is strokes thickness, because stroke is almost ever uniform. Thickness usually remains constant, except for some typography. In [RDD03] the different features between Roman languages and non-roman languages can be found. One of them is precisely the stroke density : in non-roman languages density varies in the character itself, whereas it remains constant in roman languages. Another attribute related to stroke is its number in the character, which differs in both languages too. In Non-roman languages it fluctuates from 1 to 20, in Roman ones the variation is lower, from 1 to 4.

Temporal uniformity and redundancy :
People need time to read a sentence. This means that if every second 25 frames are displayed, the same caption text will be overlapped in so many frames as needed in order to make the sentence readable. Comprehension must be achieved by the reader at the time of viewing. Vision research has determined that humans need between two and three seconds in order to understand or process a complex image. Temporal uniformity affects not only the visualization time, but also the variation of the text size or their movement throughout the video sequence. These parameters can not sharply change from frame to frame. In this way this variation could be detected.

Mouvement on the frame :
This characteristic is related to the previous feature, but in this case it describes the most common text behaviour on the screen along the time :

- Static text. Characters present no movement, for example, when the name of a person appears on the screen, subtitles in a film, the scoreboard in a match, etc.

- Scrolling and crawling text. Normally it is linearly moving either horizontally from right to left (crawling), or vertically from bottom to top (scrolling). For example, the opening and closing credits on a film, last news, share prices information, etc. In roman languages to obtain the same level of comprehension, scrolling speed should be slower than crawling speed, because its movement is working against the reader's natural reading strategy. However, TV programs time is often limited and this additional information is not put on the screen to be fully readable.

- Flying text. This is the less common kind of movement. For example, it could be found in some TV advertisement and its movement is free, not predictable. As some papers stress, see [LE98], caption text aim doesn't fit with the effect achieved with flying text. In other words, flying text is not intended to give more information, but to attract attention. Its detection is also possible, but computational cost increases and the advantages of both temporary uniformity and movement cannot be exploited. Moreover, this kind of text is more artistic and does not always have the same features, such as same character size or colour.

Due to the fact that flying text is quite improbable and usually not very interesting, when text candidates present neither static nor linear motion, candidate text blocks can be discarded. On the other hand, velocity is another discarding element. When candidate regions are moving faster than a threshold [pixels/frame], the regions are not designed in order to be read.

Position in the frame :
The main aim in this case is to avoid obscuring any important part or activity in the frame. Normally caption text is superimposed on the same area, which means caption text is normally found centred at the screen bottom. But it can be placed in a more appropriate place (e.g. football match scoreboard : top left/right corner). Sometimes, when the caption text in a frame is replaced for another one in a consecutive frame, the position of this new information differs a little with regard to the previous one.

In the literature, algorithms can be divided in two big groups, those that work in the compressed domain and those that work in the spatial domain. Many of them use the following parameters in order to evaluate its behaviour :

Recall $Recall = \frac{Correct}{(Correct+FalseNegatives)}$

$Precision\ Precision = \frac{Correct}{(Correct+FalsePositives)}$

where False Negatives are those characters, which have been discarded, and False Positives are those objects belonging to the background, which have been taken as characters.

3.4.2 Compressed domain techniques

In this group we include algorithms that are both in the compressed and in the semi-compressed domain.

Compressed domain : The only reference has been found in [ADS02]. It is based on the localization of static characters over background in motion taking into account the macro-blocks belonging to P frames (MPEG-4). Moreover it assumes that text has horizontal geometry, that it does not occupy the whole frame and that it has to appear at least in three frames. These three features allow the algorithm to isolate macro-blocks and to determinate if the macro-blocks are candidates to contain text. Both recall and precision are high in those sequences with moving background and static text, like sports sequence (e.g. score in a football match). But it cannot be used in sequences containing moving text or static background.

Semi-compressed domain : This section is called semi-compressed domain, because algorithms don't work directly with macro-blocks but analysing the DCT (Discrete Cosinus Transformation) components [ADS02], [COB04] and [ZZJ00]. DCT plus motion compensation are utilised in the MPEG standard video compression in order to reduce spatial redundancy in a frame and temporal redundancy in consecutive frames, respectively. DCT coefficients represent spatial and directional periodicity. Thus, low level features can be directly extracted from compressed images. AC coefficients from horizontal harmonics show horizontal intensity variations; therefore, they would be high in case of having a text line. On the other hand, AC coefficients from vertical harmonics show vertical intensity variations; they would be high in case of having more than one single text line. In [SS96] a detailed explanation about the DCT coefficient meaning can be found. The DCT block size, the character size and their ratio are important. For instance, if each letter is bigger than the block size we will be evaluating a single letter stroke intensity variation, but not text intensity variation relative to the background. In the same way if the letter size is too small any texture could be analogous to text and easily confused (e.g. grass field). In [ADS02] the algorithms are classified in edge-based method and correlation methods. The Edge-based methods take into account contrast between text and background, this is the reason why as a first step they calculate the Horizontal Intensity Variation in the DCT coefficients. The Correlation method [GAK98] is applied only when a shot changes, so a shot detection must be previously done. In order to detect if the new shot contents text the intra-coded blocks increment is calculated in the B- and P-frames. Once the candidate blocks are chosen some text features are applied, such as that characters are made of strokes and its colour homogeneity and horizontal geometry, in order to localize text. In [ZZJ00] this method is commented and discarded due to its vulnerability to scene changes.

Some other transformations could be used like the DWT (Discrete Wavelet Transform). This transformation gives more information than the previous one because spatial information is not lost with the transformation. Therefore, those areas with high values in high frequencies can be more easily found. Both [LDK00] and [LDK98] suggest wavelet transformation because of its capability to preserve spatial information. The text boxes are found through a hybrid : wavelet transformation and neural network. From the WT output some relevant statistical features can be chosen. In particular the more discriminant are the mean, the second and third order level calculated from the HL, LH and HH sub-bands. These vectors are used as input in a neural network.
In [CG96] some other transformations such as DHT (Discrete Haar Transform), DFT (Discrete Fourier Transform) and WHT (Walsh-Hadamard Transform), as well as the DWT (Discrete Wavelet Transform) are explained.
As a pre-processing tool, in [ZRC02] this kind of algorithms is used in a first step to localize candidate areas.

3.4.3 Spatial domain techniques

Those methods that work with the pixel values and positions are called methods in the spatial domain and they can be classified according to the following image features :

- Edge-based [AD99],[ADS02],[HCWZ01] and [AK97]. Methods in this group are focused in the search of those areas that have a high contrast between text and background. In this way, edges from letters are identified and merged. Once these regions are recognised, spatial cohesion features are applied in order to discard false positives.

- Connected Components-based [LS96] and [Lie98]. These methods use a bottom-up approach by grouping small components into successively larger components until all regions are identified in the image. Also in this case spatial cohesion features are applied.
  Both edge-based and Connected Components-based methods could be included under the same group, region-based methods.

- Texture-based [JB92],[JY98],[LDK00],[WM99] and [WMR97]. In this we can include many of the existing methods : they use the property that text in images have distinct textural properties that distinguish them from the background. Example are those which use the Gabor filter [JB92], Gaussian filter [WMR97] or those based on the colour and shape of the regions [WM99] and [JY98]. If we had not classified into spatial and compressed domain, those methods based on wavelet or FFT would accomplish this textural properties.

- Correlation-based [WC03]. These methods can be summarized as those that use any kind of correlation in order to decide if a pixel belongs to a character or not.

- Others [Tan02]. All the methods that have been mentioned don't use temporal information or use it as a complementary tool. In [Tan02], temporal information is the main feature. After applying a shot detection technique, a vector through time for each pixel is calculated in a set of frames. The authors prove that, computing the PCA for each vector, feature vectors related to the background can be separated from those related to text. The main problem of this method is that it only can be applied when the sequence has static text and a moving background.

Bibliography

[AB96] B. Achermann and H. Buncke. Combination of face classifiers for person indentification. *Proceedings of the 13th IAPR international conference on pattern recognition (ICPR)*, 3:416–420, 1996.

[AD99] L. Agnihotri and N. Dimitrova. Text detection for video analysis. In *IEEE Workshop on CBAIVL*, pages 109–13. 1999.

[ADS02] L. Agnihotri, N. Dimitrova, and M. Soletic. Multi-layered videotext extraction method. In *IEEE International Conference on Multimedia and Expo (ICME)*. Lausanne, Switzerland, August 26-29 2002.

[Agg95] J. K. Aggarwal. A model-based object recognition in dense range image. *ACM Computing Survey*, 25(1):5–43, 1995.

[AK97] M. A.Smith and T. Kanade. Video skimming and characterization through the combination and language understanding techniques. In *IEEE Computer Vision and Pattern Recognition*, pages 775–781. 1997.

[Bac02] M. I. J. F. R. Bach. Kernel independent component analysis. *J. Mach. Learn. Res.*, 3:1–48, 2002.

[Bar00] A. K. Barros. Dependent component analysis. In M. Girolami, editor, *Advances in Independent Component Analysis*, page 63. Springer Series Perspectives in Neural Computing, 2000.

[Bar01] A. C. A. K. Barros. Extraction of specific signals with temporal structure. *Neural Comput.*, 13(9):1995–2003, September 2001.

[Bas78] J. Bassili. Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face. *Journal of Personality and Social Psychology*, 37:2049–2059, 1978.

[BBB$^+$03] L. Bedini, S. Bottini, C. Baccigalupi, P. Ballatore, E. E. K. D. Herranz, E. Salerno, and A. Tonazzini. A semi-blind approach for statistical source separation in astrophysical maps. In *ISTI-CNR*. Pisa, Italy, October 2003. ISTI–TR-35.

[BBT02] F. S. Brundick, A. E. M. Brodeen, and M. S. Taylor. A statistical approach to the generation of a database for evaluating ocr software. *Int. J. on Docum. Anal. & Recogn.*, 4(3):170–176, March 2002.

[BCG$^+$94] J. L. Blue, G. T. Candela, P. J. Grother, R. Chellappa, and C. L. Wilson. Evaluation of pattern classifiers for fingerprint and ocr applications. *Pattern Recognition*, 27(4):485–501, April 1994.

[BDT99] E. Bonabeau, M. Dorigo, and G. Theraulaz. *Swarm Intelligence: From Natural to Artificial Systems.* Oxford University Press, New York, 1999.

[BH01] A. U. Batur and M. H. Hayes. Linear subspaces for illumination robust face recognition. *In proceedings of IEEE conference on computer vision and pattern recognition*, 2001.

[BHB93] V. Bruce, P. Hancock, and A. Burton. Human face perception and identification. *in face recognition: from theory to application (H. Wechsler, P.J. Phillips, V. Bruce, F.F. Soulie and T.S. Huang editors. Berlin Springer verlag*, pages 51–72, 1993.

[BHK97a] P. Belhumeur, J. P. Hespanha, and D. Kriegman. Eigenfaces vs fisherfaces: Recognition using class specific linear projection. *IEEE transactions on pattern analysis and machine intelligence*, 19(7):711–720, 1997.

[BHK97b] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces : Recognition using class specific linear projection. *IEEE Trans. on PAMI*, 19(7):711–720, July 1997.

[BJ03] R. Basri and D. W. Jacobs. Lambertian reflectance and linear subspaces. *In Proceedings on pattern analysis and machine intelligence*, 25(2):218–233, 2003.

[BK89] R. Bajscy and S. Kovacic. Multiresolution elastic matching. *Computer Vision Graphics Image Processing*, 46:1–21, 1989.

[Ble66] W. Bledsoe. Man machine facial recognition. *Technical report Rep. PRI:22 Panoramic research Inc, Paolo Alto, Cal*, 1966.

[BMS02] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski. Face recognition by independent component analysis. *IEEE Trans. on Neural Networks*, 13(6):1450–1464, November 2002.

[BNPS03] S. D. Bona, H. Niemann, G. Pieri, and O. Salvetti. Brain volumes characterisation using hierarchical neural networks. *Artificial Intelligence in Medicine*, 28(3):307–22, July 2003.

[BP93] R. Brunelli and T. Poggio. Face recognition: Features versus templates. *In Pattern Analysis and Machine Intelligence.*, 15(10):1042–1052, 1993.

[BPPP98] H. Borotschnig, L. Paletta, M. Prantl, and A. Pinz. Active object recognition in parametric eigenspace. In *Proc. Brit. Mach. Vis. Conf.*, volume 2, pages 629–638. 1998.

[BS02] N. Boujemaa and B. L. Saux. Unsupervised categorization for image database overview. *VISUAL Lecture Notes in Computer Science*, 2314:163–174, 2002.

[Bur98] C. Burges. A tutorial on support vector machines for pattern recognition. In *Data Mining and Knowledge Discovery*, volume 2, pages 121–167. 1998.

[CAC$^+$00] G. Cavaccini, M. Agresti, M. Chimenti, E. Bozzi, and O. Salvetti. An evaluation approach to ndt ultrasound processes by wavelet transform. In *15th World Conference on Nondestructive Testing*, pages 15–21. Roma, Italy, October 2000. Nondestructive Testing.

[CB65] H. Chan and W. W. Bledsoe. A man machine facial recognition system: some preliminary results. *Technical report, panoramic research, Inc, cal*, 1965.

[CF90] G. W. Cottrell and M. Fleming. Face recognition using unsupervised feature extraction. *In proceedings international Neural network conference*, 1:322–325, 1990.

[CF01] R. Campbell and P. Flynn. A survey of free-form object representation and recognition techniques. *Computer Vision and Image Understanding*, 81(2):166–210, 2001.

[CG96] N. Chaddha and A. Gupta. Text segmentation using linear transforms. In *Proceedings of Asilomar Conf. Circuits and Computers International*, pages 422–427. 1996.

[CG99] J. Cai and A. Goshtasby. Detecting human faces in color images. *Image and Vision Computing.*, 18(1):63–75, 1999.

[CHV99] O. Chapelle, P. Haffner, and V. Vapnik. Svms for histogram-based image classification. *IEEE Transactions on Neural Networks*, 1999.

[CiA02] A. Cichocki and S. ichi Amari. *Adaptive Blind Signal and Image Processing : Learning Algorithms and Applications.* Wiley, April 2002.

[CLK00] R. T. Collins, A. J. Lipton, and T. Kanade. Introduction to the special section on video surveillance. *IEEE Trans. on PAMI*, 22(8):745–746, August 2000.

[COB04] D. Chen, J.-M. Odobez, and H. Bourlard. Text detection, recognition in images and video frames. *Pattern Recognition*, 37(3):595–608, 2004.

[CSS00] M. Collins, R. E. Schapire, and Y. Singer. Logistic regression, adaboost and bregman distances. In *Computational Learing Theory*, pages 158–169. 2000.

[CU93] A. Cichocki and R. Unbehauen. *Neural Networks for Optimization and Signal Processing.* Wiley, 1993.

[CWT00] T. Cootes, K. Walker, and C. Taylor. View-based active appearance models. *In IEEE International Conference on Face and Gesture Recognition.*, pages 227–232, 2000.

[DCP03] J. Delabrouille, J.-F. Cardoso, and G. Patanchon. Multidetector multicomponent spectral matching and applications for cosmic microwave background data analysis. *Mon. Not. Roy. Astr. Soc.*, 346(4):1089–1102, December 2003.

[DGL96] L. Devroye, L. Gyrfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer, 1996.

[DHS00] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification.* Hardcover, 2000.

[DM75] C. H. Davis M. *Recognition of Facial Expressions.* Arno Press, New York, 1975.

[ea00] R. T. et al. Nature-inspired computation : Towards novel and radical computing. *BT Technol. J.*, 18(1):73–75, January 2000.

[EC01] J.-O. Eklundh and H. I. Christensen. Computer vision: Past and future. In *Informatics - 10 Years Back. 10 Years Ahead.*, pages 328–340. Springer-Verlag, London, UK, 2001.

[EF75] P. Ekman and W. V. Friesen. *Unmasking the Face.* Prentice-Hall, 1975.

[EF78] P. Ekman and W. V. Friesen. The facial action coding system. *Consulting Psychologists Press, San Francisco, CA.*, 1978.

[Eic02] S. Eickeler. Face database retrieval using pseudo 2d hidden markov models. *In proceedings of IEEE conference on Face and Gesture*, 2002.

[EJ95] A. Eleftheriadis and A. Jacquin. Automatic face location and tracking for model assisted coding of video teleconferencing sequences at low bit rates. *Signal Processing: Image Communications.*, 7(3):231–248, 1995.

[Ekm73] P. Ekman. *Darwin and Facial Expressions.* Academic Press, 1973.

[EPdRH02] M. Egmont-Petersen, D. de Ridder, and H. Handels. Image processing using neural networks - a review. *Pattern Recognition*, 35(10):279–301, 2002.

[EPPP00] T. Evgeniou, M. Pontil, C. Papageorgiou, and T. Poggio. Image representations for object detection using kernel classifiers. *In Asian Conference on Computer Vision*, pages 687–692, 2000.

[FB86a] K. Fu and T. Booth. Grammatical inference: Introduction and survey: Part i. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(3):343–360, 1986.

[FB86b] K. Fu and T. Booth. Grammatical inference: Introduction and survey: Part ii. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(3):360–376, 1986.

[FFFP03] L. Fei-Fei, R. Fergus, and P. Perona. A bayesian approach to unsupervised one-shot learning of object categories. In *Proc. 9th Int. Conf. on Comp. Vis.*, pages 1134–1141. Nice, France, October 2003.

[FG01] F. Fleuret and D. Geman. Coarse-to-fine visual selection. *In International Journal of Computer Vision*, 41(2):85–107, 2001.

[FPZ03] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc. IEEE Conf. Comp. Vis. Patt. Rec.*, pages 264–271. October 2003.

[FPZ04] R. Fergus, P. Perona, and A. Zisserman. A visual category filter for google images. In *Proc. 8th Europ. Conf. Comp. Vis.* May 2004.

[Fre02] A. A. Freitas. *Data Mining and Knowledge Discovery with Evolutionary Algorithms.* Springer-Verlag, 2002.

[Fu82] K. Fu. *Syntactical Pattern Recognition and Applications.* Prentice-Hall, 1982.

[GAK98] U. Gargi, S. Antani, and R. Kasturi. Indexing text events in digital video databases. In *Fourteenth International Conference on Pattern Recognition*, volume 1, pages 916–918. Brisbane, Australia, aug 1998.

[GBGB01] K. Grauman, M. Betke, J. Gips, and G. Bradski. Communication via eye blinks detection and duration analysis in real time. *In Proceeding of Computer Vision and Pattern Recognition*, 2001.

[GHL71] A. J. Goldstein, L. Harmon, and A. B. Lesk. identification of human faces. *In proceeding IEEE*, 59:748, 1971.

[GHS$^+$95] S. Gutta, J. Huang, D. Singh, I. Shah, B. Takacs, and H. Wechsler. Benchmark studies on face recognition. In *Proceedings of International Workshop on Automatic,Face - and Gesture Recognition (IWAFGR).* 1995.

[GLC00] G. Guo, S. Li, and K. Chan. Face recognition by support vector machines. *In proceedings on the international conference on Face and Gesture recognition*, pages 196–201, 2000.

[Gor95] G. Gordon. Face recognition from frontal and profil views. *Proceedings of the international workshop on automatic face and gesture recognition*, pages 47–52, 1995.

[Gre93] U. Grenander. General pattern theory. *Oxford University Press*, 1993.

[GZ01] G.-D. Guo and H.-J. Zhang. Boosting for face recognition. In *IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems (RATFG-RTS'01).* 2001.

[Hal91] P. Hallinan. Recognizing human eyes. *In spie proceedings: Geometric methods in computer vision*, 1570:214–226, 1991.

[HAMJ01] R. Hsu, M. Abdel-Mottaleb, and A. K. Jain. Face detection in color images. *In Proceedings of the IEEE International Conference on Image Processing*, pages 1046–1049, 2001.

[HCWZ01] X.-S. Hua, X.-R. Chen, L. Wenyin, and H.-J. Zhang. Automatic location of text in video frames. In *MULTIMEDIA '01: Proceedings of the 2001 ACM workshops on Multimedia*, pages 24–27. ACM Press, New York, NY, USA, 2001. ISBN 1-58113-395-2.

[HCZZ00] F. J. Huang, T. Chen, Z. Zhou, and H.-J. Zhang. Pose invariant face recognition. *In proceedings of IEEE conference on Face and Gesture*, 2000.

[HHP01] B. Heisele, P. Ho, and T. Poggio. Face recognition with support vector machines : Global versus component-based approach. *In proceedings of International conference on computer vision*, 2001.

[Hyv00] E. O. A. Hyvärinen. Independent component analysis : Algorithms and applications. *Neural Networks*, 13:411–430, 2000.

[JB92] A. K. Jain and S. Bhattacharjee. Text segmentation using gabor filters for automatic document processing. *Matching Vision and Application*, 5:169–184, 1992.

[JKYL00] K. Jonsson, J. Kittler, and J. M. Y.P. Li. Learning support vectors for face verification and recognition. *In proceedings of IEEE conference on Face and Gesture*, 2000.

[JY98] A. K. Jain and B. Yu. Efficient automatic text location method and content-based indexing and structuring of video database. *Journal of Visual Communication and Image Representation*, 7(4):336–344, dec 1998.

[Kan73] T. Kanade. Picture processing by computer complex and recognition of human faces. *Technical report, kyoto university departement of computer science*, 1973.

[KBP$^+$03] E. E. Kuruoğlu, L. Bedini, M. T. Paratore, E. Salerno, and A. Tonazzini. Source separation in astrophysical maps using independent factor analysis. *Neural Networks*, 16(3-4):479–491., April-May 2003.

[Kir90] L. S. M. Kirby. Application of the karhunen-loeve procedure for the characterization of human faces. *IEEE Trans. on PAMI*, 12(1):103–108, January 1990.

[KK72] Y. Kaya and K. Kobayashi. A basic study on human face recognition. *In S. watanabe editors, frontiers of pattern recognition*, page 265, 1972.

[KK00] K. S. Karpouzis K., Tsapatsoulis N. Moving to continuous facial expression space using the mpeg-4 facial definition parameter (fdp) set. In *Proc. of SPIE Electronic Imaging 2000, San Jose, CA, USA*. 2000.

[Kon96] W. Konen. Comparing facial line drawings with gray-level images: A case study on phantomas. *in proceedings, international conference on Artifical Neural neutworks*, pages 727–734, 1996.

[LBP95] T. Leung, M. Burl, and P. Perona. Finding faces in cluttered scenes using random labelled graph matching. *In Proceedings of the International Conference on Computer Vision*, pages 637–644, 1995.

[LCK02] X. Liu, T. Chen, and B. V. Kumar. On modeling variations for face authentication. *In proceedings of IEEE conference on Face and Gesture*, 2002.

[LDK98] H. Li, D. Doermann, and O. Kia. Automatic text detection and tracking in digital video. Technical Report LAMP-TR-028, CAR-TR-900, University of Maryland, College Park, 1998.

[LDK00] H. Li, D. Doermann, and O. Kia. Automatic text detection and tracking in digital video. In *IEEE Trans. on Image Processing*, volume 9, pages 147–155. jan 2000.

[LDZ00] A. Lorette, X. Descombes, and J. Zerubia. Texture analysis through a markovian modelling and fuzzy classification : Application to urban area extraction from satellite images. *Int. J. Comp. Vis.*, 36(3):221–236, 2000.

[LE98] R. Lienhart and W. Effelsberg. Automatic Text Segmentation and Text Recognition for Video Indexing. Technical Report TR-98-009, Department for Mathematics and Computer Science, University of Mannheim, may 1998.

[LGL00] Y. Li, S. Gong, and H. Liddell. Support vector regression and classification based multi-view face detection and recognition. *In proceedings of IEEE conference on Face and Gesture*, 2000.

[LHLM02] Q. Liu, R. Huang, H. Lu, and S. Ma. Face recognition using kernel based fisher discriminant analysis. *In proceedings of IEEE conference on Face and Gesture*, 2002.

[Lie98] R. Lienhart. Text segmentation and text recognition in digital videos, may 1998. MoCA Project.

[LLN00] R. Liao, S. Z. Li, and Nanyang. Face recognition based on multiple facial features. *In proceedings of IEEE conference on Face and Gesture*, 2000.

[Lon98] S. Loncaeic. A survey of shape analysis techniques. *Pattern Recognition*, 31(8):983–1001, 1998.

[LS96] R. Lienhart and F. Stuber. Automatic text recognition in digital videos. In *Proceedings of SPIE Image and Video Processing IV 2666*, pages 180–188. 1996.

[LSW98] Y. Lamdan, J. Shwartz, and H. J. Wolfson. Object recognition by affine invariant matching. *In Proceedings of Computer Vision and Pattern Recognition*, pages 335–344, 1998.

[LW99] C. Liu and H. Wechsler. Face recognition using shape and texture. *In proceedings of IEEE conference on computer vision and pattern recognition*, 1999.

[LZZ$^+$02] S. Li, L. Zhu, Z. Zhang, A. Blake, H. Zhang, and H.-Y. Shum. Statistical learning of multi-view face detection. *In Proceedings of the European Conference on Computer Vision.*, pages 67–81, 2002.

[Mar01] A. C. K. A. M. Martínez. Pca versus lda. *IEEE Trans. on PAMI*, 23(2):228–233, February 2001.

[MN95] H. Murase and S. K. Nayar. Visual learning and recognition of 3D objects from appearance. *ijcv*, 14(5):5–24, 1995.

[Mog97] A. P. B. Moghaddam. Probabilistic visual learning for object representation. *IEEE Trans. on PAMI*, 19(7):696–710, July 1997.

[MP95] B. Moghaddam and A. Pentland. Probabilistic visual learning for object detection. *In Proceedings of the International Conference on Computer Vision*, pages 786–793, 1995.

[MYW$^+$99] J. Miao, B. Yin, K. Wang, L. Shen, and X. Chen. A hierarchical multiscale and multiangle system for human face detection in complex background using gravity center template. *In Pattern Recognition*, 32(7):1237–1248, 1999.

[Nel98] A. S. R. Nelson. A cubist approach to object recognition. In *Int. Conf. Comp. Vis.*, pages 614–621. Bombay, India, January 1998.

[NS00] R. C. Nelson and A. Selinger. Improving appearance-based object recognition in cluttered background. In *Proc. Int. Conf. Patt. Rec.*, pages 1–8. Barcelona, Spain, September 2000.

[Nyf95] C. Nyffengger. Gesichterkennung mit hmm. *Masters thesis IAM*, 1995.

[OFG97] E. Osuna, R. Freund, and F. Girosi. Training support vector machines: an application to face detection. *In Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 130–136, 1997.

[OvdM02] K. Okada and C. von der Malsburg. Pose-invariant face recognition with parametric linear subspaces. *In proceedings of IEEE conference on Face and Gesture*, 2002.

[Pav77] T. Pavlidis. *Structural Pattern Recognition.* Springer-Verlag, 1977.

[PMRR00] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The feret evaluation methodology for face recognition algorithms. *IEEE Trans. on PAMI*, 10(10):1090–1104, October 2000.

[PMS94] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspace for face recognition. *iccv*, pages 84–91, 1994.

[PS00] R. Plamondon and S. Srihari. Online and off-line handwriting recognition: a comprehensive survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):63–84, 2000.

[PT91] A. Pentland and M. Turk. Eigenfaces for recognition. *J. Cogn. Neurosci.*, 3:72–86, 1991.

[PWJR98] P. J. Phillips, H. Wechsler, J. Juang, and P. J. Rauss. The feret database and evaluation procedure for face-recognition algorithms. *Image Vision Computing J.*, 16(5):295–306, 1998.

[RBK98] H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, 1998.

[RDD03] A. Rosenfeld, D. Doermann, and D. DeMenthon. *Video Mining.* Kluwer Academic Publishers, Norwell, USA, 2003.

[Rig01] G. Rigoll. Hidden markov models for pattern recognition and man-machine-communication. *Pre-Conference Tutorial at 23. Annual Symposium for Pattern Recognition (DAGM 2001)*, 2001.

[Rip93] B. Ripley. *Statistical Aspects of Neural Networks.* U. Bornndorff-Neilsen, J. Jensen, and W. Kendal, eds., Chapman and Hall, 1993.

[Ros02] M. H. C. Rosenberg. Training object detection models with weakly labeled data. In *Proc. Brit. Mach. Vis. Conf.*, pages 577–586. 2002.

[RTSB01] S. Romdhani, P. Torr, B. Schlkopf, and A. Blake. Computationally efficient face detection. *iccv*, pages 695–700, 2001.

[RY92] D. Reisfeld and Y. Yeshurun. Robust detection of facial features by generalized symmetry. *In proceedings , International conference on pattern recognition*, pages 117–120, 1992.

[SB00] H. Sahbi and N. Boujemaa. From coarse-to-fine skin and face detection. *ACM International Conference on Multimedia.*, pages 432–434, 2000.

[SB02] H. Sahbi and N. Boujemaa. Robust face recognition using dynamic space warping. *In Proceedings of Springer Verlag Lecture Notes In Computer Science. ECCV's Workshop on Biometric Authentication.*, pages 121–132, 2002.

[Sha92] A. Shashua. Geometry and photometry in 3d visual recognition. *PhD thesis, Massachussets Institute of Technology*, 1992.

[SK84] E. P. Scherer K. *Approaches to Emotion.* Lawrence Erlbaum Associates, 1984.

[SK87] L. Sirovich and M. Kirby. Low dimensional procedure for the characterization of human faces. *In journal of opt soc. Am. A*, 4(3):519–524, 1987.

[Sme95] Y. Smetanin. Neural networks as systems for pattern recognition: a review. *Pattern Recognition and Image Analysis*, 5(2):254–293, 1995.

[SP96] K. Sobottka and I. Pitas. Segmentation and tracking of faces in color images. In *FG*, pages 236–241. 1996.

[SP98] K. K. Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):39–51, 1998.

[SS96] B. Shen and I. K. Sethi. Direct feature extraction from compressed images. In *Storage and Retrieval for Image and Video Databases (SPIE)*, pages 404–414. 1996.

[ST98] E. Saber and A. Tekalp. Frontal-view face detection and facial feature extraction using color, shape and symmetry based cost functions. *Pattern Recognition Letters*, 19(8):669–680, 1998.

[Sun96] K.-K. Sung. Learning and example selection for object and pattern detection,. *Ph.D. Thesis, Massachusetts Institute of Technology, Electrical Engineering and Computer Science*, 1996.

[SWNE01] D. A. Socolinsky, L. B. Wolff, J. D. Neuheisel, and C. K. Eveland. Illumination invariant face recognition using thermal infrared imagery. *In proceedings of IEEE conference on computer vision and pattern recognition*, 2001.

[SY94] F. Samaria and S. Young. Hmm based architecture for face identification. *Image and vision computing*, 12(8):537–543, 1994.

[Tan02] X. e. a. Tang. Video text extraction using temporal feature vectors. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*. Lausanne, Switzerland, aug 2002.

[TBS04] A. Tonazzini, L. Bedini, and E. Salerno. Independent component analysis for document restoration. *Int. J. on Docum. Anal. & Recogn.*, 7(1):17–27, 2004.

[Tho96] H. Thodberg. Review of bayesian neural networks with an application to near infrared spectroscopy. *IEEE Transactions on Neural Networks*, 7(1):56–72, 1996.

[TP91] M. Turk and A. Pentland. Eigenfaces for recognition. *In Journal of Cognitive Neuroscience*, 3(1):72–86, 1991.

[TSMB04] A. Tonazzini, E. Salerno, M. Mochi, and L. Bedini. Bleed-through removal from degraded documents using a color decorrelation method. Technical report, ISTI-CNR, ISTI–TR-04, Pisa, February 2004.

[TV00] K. Tieu and P. Viola. Boosting image retrieval. *cvpr*, pages 228–235, 2000.

[UL96] R. Uhl and N. Lobo. A framework for recognizing a facial image from a police sketch. *In proceedings IEEE conference on computer vision and pattern recognition*, pages 586–593, 1996.

[Vap98] V. N. Vapnik. *Statistical learning theory.* J. Wiley and sons, 1998.

[VBT02] S. Vezzosi, L. Bedini, and A. Tonazzini. An integrated system for the analysis and the recognition of characters in ancient documents. Technical report, ISTI-CNR, 2002.

[VJ01] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *In Second International Workshop On Statistical and Computational Theories of Vision-Modeling, Learning, Computing and Sampling*, 2001.

[WC03] E. K. Wong and M. Chen. A new robust algorithm for video text extraction. *Pattern Recognition*, 36(6):1397–1406, 2003.

[WFKvdM97] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *Pattern Analysis and Machine Intelligence*, 19(7):775–779, 1997.

[WM99] V. Wu and R. Manmatha. Textfinder: An automatic system to detect and recognize text in images. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 21(11):1224–1229, nov 1999.

[WMR97] V. Wu, R. Manmatha, and E. M. Riseman. Automatic text detection and recognition. In *Proceedings of Image Understanding Workshop*, Images and Multimedia, pages 707–712. 1997.

[WPJW96] J. Wilder, P. J. Phillips, C. Jiang, and S. Wiener. Comparison of visible and infra-red imagery for face recognition. *Proceedings of the international conference on Automatic face and gesture recognition*, pages 182–187, 1996.

[XPL02] D. Xi, I. T. Podolak, and S.-W. Lee. Facial component extraction and face recognition with support vector machines. *In proceedings of IEEE conference on Face and Gesture*, 2002.

[YAC98] H. H. Yang, S. I. Amari, and A. Cichocki. Information-theoretic approach to blind separation of sources in nonlinear mixture. *Signal Processing*, 64(3):291–300, 1998.

[YC98] K. C. Yow and R. Cipolla. Enhancing human face detection using motion and active contours. *accv*, pages 515–522, 1998.

[YE89] A. Young and H. Ellis. *Handbook of Research on Face Processing.* Elsevier, North Holland, 1989.

[YJB95] K. Yu, X. Jiang, and H. Bunke. Face recognition by facial profile analysis. *Proceedings of the international workshop on automatic face and gesture recognition*, pages 208–213, 1995.

[YLW98] J. Yang, W. Lu, and A. Waibel. Skin-color modeling and adaptation. *accv*, pages 687–694, 1998.

[ZCHS03] L. Zhang, B. Curless, A. Hertzmann, and S. M. Seitz. Shape and motion under varying illumination : Unifying structure from motion, photometric stereo, and multi-view stereo. In *9th IEEE Int. Conf. Comp. Vis.* Nice, France, October 2003.

[ZCPR03] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition : A literature survey. *ACM Computing Surveys*, 35(4):399–458, December 2003.

[ZCR00] W. Zhao, R. Chellappa, and A. Rosenfeld. Face recogntion: A litterature survey. *Technical report University of Maryland*, 2000.

[Zha99] W. Zhao. Improving the robustness of face recognition. *In proceedings of International conference on audio and video based person authentification*, pages 78–83, 1999.

[Zha00] G. Zhang. Neural networks for classification: a survey. *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews*, 30(4):451–462, 2000.

[ZRC02] D.-Q. Zhang, R. K. Rajendran, and S.-F. Chang. General and domain-specific techniques for detecting and recognizing superimposed text in video. In *IEEE International Conference in Image Processing (ICIP).* Rochester, NY, USA, September 22-25th 2002.

[ZZJ00] Y. Zhong, H. Zhang, and A. Jain. Automatic caption localization in compressed video. *Institute of Electrical and Electronics Engineers on Pattern Analysis and Machine Intelligence (PAMI)*, 22(4):385–392, April 2000.

Chapter 4

Spatial relations and geometrical configuration

This chapter presents more fondamental concepts of signal processing, with the idea to exhibit relevant image representations that can be applied or adapted to content-based retrieval. In particular, a state of the art on multiresolution representations of the signal is proposed in section 4.1. The paradigm of multiresolution is introduced, while the study mainly focuses on pyramid and wavelet representations. In section 4.2, the problem of image segmentation is investigated, by presenting the different existing methodologies according to classical mathematical concepts.

4.1 Multiresolution and content-driven representations

4.1.1 Classical approaches for multiresolution representations

There are several ways to transform one representation of a given signal into another one. The most classical example is the Fourier transform, where a signal is decomposed into sinusoidal waves. Such a decomposition gives the intensity of the fluctuations (frequencies) in the signal which is often of great importance. However, due to the infinite extent of the sinusoidal functions, any local signal characteristics (i.e., an abrupt change in the signal) are spread over the entire representation, thus making them 'invisible'. This is a serious drawback since singularities and irregular structures often carry the most important information in signals. For instance, in images, discontinuities in the intensity may provide the location of the object contours which are particularly meaningful for recognition purposes. For many other types of signals such as electro-cardiograms or radar signals, the interesting information is given by transients such as local extrema. Furthermore, such singularities usually occur with different location and localization (i.e., range, scale) in time and frequency. Consequently, transform methods that represent the signal at multiple scales are better suited for extracting information than methods that represent the signal at a single scale.

Over the last century, scientists in different fields struggled to overcome limitations of the Fourier transform and to build representations of signals that are able to adapt themselves to the nature of the signal. On the one hand, to 'pick up' the transients without giving up the frequency information, the signal should be decomposed over functions which are well localized in time (or space) and frequency. This leads to so-called *time-frequency representations*. On the other hand, since signal structure depends on the scale at which the signal is being perceived, it should be analyzed at different scales or levels of resolution. This results in so-called *multiresolution representations* which, besides a time parameter, also contain a scale parameter.

4.1.2 Multiresolution approaches

MR methods span a very broad array of concepts and approaches. In this report, we will mainly focus on pyramid and wavelet representations. There are, however, many other MR techniques such as quadtrees, multigrid methods and scale-space representations, to name a few. They have the 'multiresolution paradigm' in common, but apart from that they differ in many respects, both in theory and in practice.

Pyramids

Pyramids have been recognized early as an interesting tool for computer vision and image coding [BA83]. A classical pyramid scheme consists of three steps : (*i*) deriving a coarse approximation of an input image, (*ii*) predicting this image based on the coarse version, and (*iii*) taking their difference as the prediction error. This defines the analysis part. At synthesis, the prediction error is added back to the prediction from the coarse version, guaranteeing perfect reconstruction. Iteration of the analysis part over the coarse approximation yields a pyramid representation of the original image as an approximation image at the lowest resolution and a set of detail images at successive higher resolutions.

Wavelets

Wavelets[Mal89] are functions that are well localized in time and frequency and that can be used to decompose a signal into different frequency bands with different time resolutions. This leads to the wavelet transform. Of particular interest is the discrete wavelet transform, which applies a two-channel filter bank (with downsampling) iteratively to the low-pass band (initially the original signal). The wavelet representation consists then of the low-pass band at the lowest resolution and the high-pass bands obtained at each step. This transform is invertible and non-redundant. As such, the corresponding decomposition differs from various other MR decompositions such as pyramids, which are redundant, and scale-spaces, which are non-invertible in general. Both aforementioned properties, i.e., invertibility and non-redundancy, turn the discrete wavelet transform into a highly efficient and applicable representation for a broad range of signal and image processing tasks such as denoising and, particularly, compression.

The wavelet decomposition (and reconstruction) of a discrete signal from a resolution to the next one is implemented by a two-channel perfect reconstruction filter bank such as in Fig. 4.1.

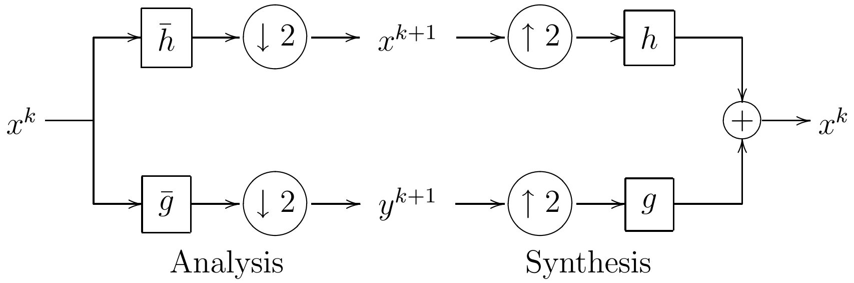


Figure 4.1: *Perfect reconstruction filter bank with low-pass and high-pass analysis filters $\bar{h}$, $\bar{g}$, respectively, and low-pass and high-pass synthesis filters h, g, respectively.*

There exists a great variety of wavelet families depending on the choice of the prototype wavelet or, alternatively, the filter's coefficients. However, imposing additional requirements such as orthogonality, symmetry, compactness of support, rapid decay and smoothness limits our choice. The 'optimal' choice of the wavelet basis will depend on the application at hand, and therein lies part of the difficulty of building a suitable wavelet representation.

Most wavelet (and pyramid) transforms have been designed in the one-dimensional case. By successive application of such one-dimensional transforms on the rows and the columns (or vice versa) of an image, one obtains a so-called *separable* two-dimensional transform. Non-separable transforms can also be constructed [HG00, KV92]. Although they provide decompositions with more general properties, they have been used less often in image applications due to the lack of general tools for their design.

4.1.3 The need for adaptive wavelets

Wavelets have had a tremendous impact on signal processing, both because of their unifying role and their success in several applications. The applicability of the wavelet transform (as well as for other MR decompositions) is somewhat limited, however, by the linearity assumption. Coarsening a signal by means of linear operators may not be compatible with a natural coarsening of some signal attribute of interest (e.g., the shape of an object), and hence the use of linear procedures may be inconsistent in such applications. In general, linear filters smear the singularities of a signal and displaces their locations, causing undesirable effects.

Moreover, standard wavelets are often not suited for higher dimensional signals because they are not adapted to the 'geometry' of higher dimensional signal singularities. For example, an image comprises smooth regions separated by piecewise regular curves. Wavelets, however, are good at isolating the discontinuity across the curve, but they do not 'see' the smoothness along the curve. These observations indicate the need for MR representations which are data-dependent.

The importance of such 'data-driven' representations has led to a wealth of new directions in multiresolution approaches such as bandelets [LM00], ridgelets [Don98], curvelets [Can01], morphological wavelets [HG00], etc., which go beyond standard wavelet theory.

We will look for adaptive (i.e., content-driven) transforms which retain the desirable properties of the standard wavelet transform (e.g., non-redundancy and invertibility) while exploiting, in a simple way, the geometrical information of the underlying signal. This will allow for a better localization and representation of the singularities, as well as for sharper (perceptually better) approximations at lower resolutions.

In [PH02, PPPH02, HPPP03, HPPP04] we have presented the construction of *adaptive wavelets* by means of an extension of the lifting scheme. The basic idea is to choose the update filters according to some decision criterion which depends on the local characteristics of the input signal. In this way, only homogeneous regions are smoothed while discontinuities are preserved. An interesting aspect of our approach is that it is neither causal nor redundant, i.e., it does not require any bookkeeping to enable perfect reconstruction. We show that these adaptive schemes yield lower entropies than schemes with fixed update filters, a property that is highly relevant in the context of *compression*. Despite all these attractive properties, a number of open theoretical and practical questions need to be addressed before such schemes become useful in signal processing and analysis applications. For example, we need to get a better understanding how to design update and prediction operators that lead to adaptive wavelet decompositions that satisfy properties key to a given application at hand. In our work, we have focused on binary decision maps and as a result, the adaptive scheme can only discriminate between two 'geometric events' (e.g., edge region or homogeneous region). In order to deal with the great richness of real-world signals and images, one must be able to incorporate the geometrical structure of the signals, for example, by using multiple criteria.

Another issue that needs to be addressed is the stability of the scheme. In particular, the behavior of the adaptive scheme under quantization needs a more thorough investigation. Stability of decompositions is of utmost importance when they are being used in lossy compression schemes for image or video coding.

4.2 Image Segmentation

Image segmentation can be described as the process of partitioning the image into disjoint regions, each one being homogeneous and connected with respect to some cue(s), such as image intensity, texture, color, etc. Image segmentation methods can be classified as *boundary-based*, which generate an edge image which delineates the segments of the image, or *region-based*, which group image pixels based on the homogeneity of spatially localized features.

The cues that lead the segmentation process include most commonly image intensity, and usually additional features, like color and texture features, or motion and depth estimates. The incorporation of multiple-cue information in a segmentation process increases the potential of producing better results, when the corresponding features are treated appropriately.

Computer vision researchers have used different methodologies to attack the segmentation problem. The discrimination of different methodologies will be done according to the mathematical methodology they employ; the segmentation categories that will be presented are the following :

- Variational methods
- Statistical methods
- Graph-based methods
- Morphological methods

In the rest of this report, we will refer to recent advances in these four domains and present how they deal with multiple cues.

4.2.1 Variational Methods

In the variational framework the solution to the segmentation problem is expressed as the minimization of some energy functional J defined on partitions $\mathcal{R} = R_{1\ldots N}$ of the image domain Ω; J is commonly expressed in terms of the contours ∂R_i and the interiors R_i of the image segments :

$$J(\mathcal{R}) = \sum_{i=1}^{N} \int_{R_i} f(x, R_i) dx + \sum_{i=1}^{N} \oint_{\partial R_i} g(s) ds \tag{4.1}$$

In the above formula f quantifies the homogeneity of the features at location x with those of region R_i and $g(s)$ is a decreasing function of edge strength.

Active Contours/Snakes : In the setting of Active Contours a closed curve or set of curves enclosing some initial region(s) is given, and we wish to modify them until they come into alignment with the contours of prominent image objects. This is accomplished by evolving the contour in the direction of steepest descent of the energy functional and regarding as a solution a location where the contour no longer changes with time. This is accomplished mathematically by deriving a partial differential equation (PDE) from the Euler-Lagrange equations corresponding to the energy functional and then solving numerically this PDE.

This idea was first introduced in the purely boundary-based method of Snakes in [KWT87], which were implemented using splines to parametrize the evolving curves; the fact that the number of objects detected could not dynamically change limited their applicability since a good initialization is almost always essential.

Geodesic Active Contours : One of the primary developments in the field of active contour models in the last decade has been the reformulation in [CKS97] of the Snakes model as the computation of geodesics, with respect to a metric defined using the image. This gave rise to the Geodesic Active Contour model (GAC), which is parametrization free, and where the geodesic curve computation is reduced to the numerical solution of a geometric flow. This geometrical flow can be efficiently solved numerically using level set methods [Set96], where the contour is represented as the zero level-set of a 3D function. Thereby the evolving contours naturally split and merge, allowing the simultaneous detection of several objects. Still, a specific initialization step is necessary, where the initial curve should lie completely exterior or interior to the object boundaries.

Active Regions : Along a different line, based primarily on the (region based) Mumford-Shah functional [MS89a], variational methods have been designed to deal with multiple cues [ZY96] and have been later rephrased [PD02] in terms of geometric curve evolution schemes implemented using level set methods [Set96]. Specifically, in [PD02] the boundary and region based terms have been brought together, so that the curve evolution is driven by both statistical terms as in [ZY96], and geometric terms as in [CKS97]. In [CV01a] a special case of the Mumford-Shah functional is used to give rise to purely region-based curve evolution equations which are implemented using an efficient variation of level set methods.

Unsupervised Methods, Prior Information : In [RBD03] an unsupervised textured image segmentation method is proposed, which extends and simplifies the methods of [PD02]. This scheme simultaneously updates the regions and the parameters of the distributions, thereby dynamically learning the model of each region. Another active field deals with the incorporation of prior knowledge about the desired segmentation, like shape knowledge [RP02, CSS04].

Multiple Cues : Multiple cues are dealt with using a statistical model to express the feature similarity term within each region; intensity, texture, motion, etc. features are merged in a vector of random variables which is assumed to follow a region-specific multivariate distribution. The only modification of the evolution equations is then in the region-based term, which uses a high dimensional vector to determine the corresponding forces, as in [ZY96, PD02, RBD03, BRDW03].

4.2.2 Statistical Methods

Another approach to segmentation considers the segmentation labels as a random field, and recasts the segmentation problem as a problem of deriving estimates from this field, like Maximum-A-Posteriori (MAP), or Minimum Mean Squared Error (MMSE) estimates.

Markov/Gibbs Random Fields : By assuming that each segmentation label conditioned on its neighbors is independent of the rest of the field leads to the Markov Random Field (MRF) model. These dependencies between the labels x_i, x_j at neighboring nodes in an MRF are encoded in the *clique potentials* $\Psi(x_i, x_j)$, while the influence of the observed data is encoded in the *observation potentials* $\Phi_i(x_i)$. The energy of a configuration $X = \{x_1, \dots, x_N\}$ is then given by

$$E(X) = \prod_i \Phi_i(x_i) \prod_{x_j \in \mathcal{N}_i} \Psi(x_i, x_j) \tag{4.2}$$

The Gibbsian probability of X is related to its energy by Boltzmann's law :

$$P(X) = \frac{1}{Z} e^{-E(X)/T} \tag{4.3}$$

Efficient Inference Algorithms for MRFs : Despite their flexibility and their robustness compared to variational algorithms, MRF-based segmentation algorithms had fallen into dismay during the previous decade, since the stochastic relaxation-based algorithms that dominated research in the 1980's [GG84] were not fast enough for practical applications. During the recent years interest in MRFs has resurged due to the introduction of efficient algorithms for inference, based on deterministic, local computations [YFW01]. The Belief Propagation (BP) algorithm has been used since the 1980's for inference on random fields, where the dependencies between the random variables can be expressed in terms of a graph without loops. The results obtained in this case are exact and variations of the same algorithm can be used for deriving MAP estimates and estimating the marginal distributions at nodes. Applying the BP algorithm on graphs with loops (like MRFs for low-level vision) gives rise to the Loopy Belief Propagation (LBP) algorithm which, even though not guaranteed to give the exact estimates, gives very good results in practice. Theoretical justifications for the application of the LBP algorithm have established links with the Bethe approximation from statistical physics. In [FH04] specially designed LBP algorithms for computer vision are proposed that are shown to result in substantial speedups.

In [BZ03] a sampling method from Statistical Physics, Swendsen Wang Cuts, has been modified to be applicable to MRFs for computer vision. Samples are drawn from the posterior probability of the MRF's configurations efficiently, by flipping clusters of labels which significantly speeds up inference. A closely related algorithm devised for inference on MRFs is the Graph-Cuts algorithm which is described in the following section.

Generative Models : Another active field of research in statistical methods for image segmentation is the field of generative models-based segmentation (and vision in general). The idea is to construct statistical models of the appearance of each region, and then assign the observations in the image to the region that explains them best. Most important contributions in this field include [ZY96, TZ02, TCYZ03], which have given rise to impressive segmentation results. In this work, image segmentation is posed as parameter estimation, where the parameters include a) the boundaries of the regions, b) the type of each region (e.g. texture, intensity, face regions) c) the parameters of the generative model describing the features within each region and d) the number of the regions. The observed data is the image itself, and image segmentation and parsing is interpreted as sampling from the posterior on these parameters. In case the the only regularity enforced on the segmentation labels is that of minimum boundary length, the inference of the region labels given all the other parameters can be speeded up using a variational algorithm like active regions which is then interpreted as a greedy (non-stochastic) search in the posterior distribution of labels.

Multiple Cues : The conventional statistical approach, which has also been adopted by variational methods, groups the features from all cues in a vector of random variables and has been described in the previous section. In [TZ02] an alternative method is used to allow each region to be either a texture, a color region or any other categorical type of regions, by stochastically performing *jump moves* in the posterior distribution of the segmentation parameters. This makes it possible to choose among the cues that are used to perform segmentation in an elegant and theoretically sound way.

4.2.3 Graph-based Algorithms

Graph Theoretic Algorithms represent the image as a weighted graph, where the nodes are image pixels and the edge weights encode the information available about the desired segmentation : this can be either in the form of pairwise weights, like in the case of MRFs, or bottom-up knowledge, like the output of some edge-detector, which indicates that two nodes (pixels) should belong to different segments. The segmentation is determined, using techniques from graph theory, by finding a partition (cut) of the graph such that a criterion is (min-) maximized.

Graph Cuts : Graph Cuts [BVZ01] have been devised to perform efficient inference of the MAP estimate on MRFs, and they are guaranteed to converge to a solution that is at least equal to a fixed fraction of the global optimum. By a combination of swap moves the segmentation labels are changed in clusters which results in important speedups in the convergence rates. Specifically, even though the inference of the MAP estimate is NP hard, based on classical computer science algorithms for graphs, the complexity of estimating a suboptimal solution is reduced to polynomial in the product of the number labels and pixels. This method has been also applied to minimize the cost function of the geodesic active contour model [BK03].

Normalized Cuts : Based on a different approach, in the Normalized Cuts algorithm [SM01] a graph is constructed where all the pixels-nodes in the image are connected, with weights on the graph determined by the affinity (in feature space and in location) of the pixels. The problem of segmentation is then phrased as determining a partition of the graph, where the affinity of the nodes 'lost' by separating the graph (the cut) is minimal *when divided by* the total affinities of each cluster to the whole graph (the normalization). Thereby, even though separating a single node into a cluster may result in a low cut value, when normalized this will be close to the maximum normalized cut value. The partitions are calculated by considering initially the labels of the nodes continuous, solving an eigenvector problem and then discretizing the eigenvectors. This algorithm gives state of the art results and has been extended to deal with motion segmentation and object-based segmentation.

Minimal Path Finding Methods : In [MM99] the image is considered as a graph where the nodes are either the pixels of the image or regions (tiles) of the mosaic (fine segmentation) of the image and the segmentation is equivalent to finding the shortest distance on the graph or the minimum spanning tree, where the distance is defined as the length of some path. Depending on the way that the path is defined, e.g shortest path, cheapest path, easiest path, different segmentation schemes occur. In [CK97] the problem of segmentation is investigated using a minimum path approach : the global minimum attainable by an active contour is detected between two end points, by modifying the model energy to include an internal regularization term in addition to the external potential term. In a similar manner, the proposed approach in [AC04] relies on some kind of energy partition of the image domain, where the energy is defined by measuring a pseudometric-based distance to a source point; thus, the choice of an energy and a set of sources determines the tessellation of the domain.

Multiple Cues : Multiple cues are integrated by expressing the weights between nodes as the product (or maximum) of the weights corresponding to each cue. For example in the normalized cuts method, affinity terms based on texture features are multiplied with affinity terms based on pixel proximity as well as edge-based affinities etc.

4.2.4 Morphological Methods

Mathematical morphology is a nonlinear image analysis methodology that is primarily based on set- and lattice- theoretic approaches whose goals are to quantify the geometrical structure of images. Among the morphological methodologies that have been developed for various image analysis tasks, the most prominent segmentation methodology proposed is the *watershed transform*.

The idea of the watershed-based morphological segmentation can be summarized in three stages. At a first stage the image is simplified and processed in such a way that the presence of noise is reduced and useless (redundant) information is removed, thus producing an image that consists mostly of flat and large regions. The second stage involves the region-feature extraction, where the goal is to extract some special features such as small homogeneous regions, called markers, which will be used as the starting points for the flooding process. The selection of the markers is probably the most difficult task and the strategies for finding them are diverse and problem dependent; many case studies are listed in [MB90]. The last stage is the application of the watershed transform, where a gradient image is constructed and its topographic surface is flooded by sources placed at the position of the markers. The watershed construction grows the markers until the exact contours of the objects are found. Catchment basins without sources are flooded by already flooded neighboring catchment basins. In order to avoid the merging of lakes produced by different sources, dams are erected to keep them separated. The set of dams constitutes the boundaries of the resulting segmentation.

This model of segmentation has been extendedly studied and used during the past years. Compared to the other segmentation methods, the watershed has several advantages, including the proper handling of gaps and the placement of boundaries at the most significant edges. Despite its success it has reached its limits as new segmentation needs have appeared, which have triggered the development of a rich and coherent framework able to deal with a large variety of segmentation tasks. Today, mathematical morphology proposes a series of tools, to be used in a sequence or combined with the classical watershed model, in order to satisfy the present segmentation needs :

Multiscale segmentation : The use of multiple scales appears to be an essential part for all types of morphological segmentation whether they are automatic or interactive. A multiscale segmentation can be produced by the progressive merging of regions, and extraneous knowledge can be incorporated in this process [MM99]. Useful partitions can be extracted from multiscale representations of the image, with a degree of coarseness that can be determined through interaction with the user.

Generalized floodings : In [Mey99, Mey00] F. Meyer presented several particular modes of flooding, in order to properly segment different types of images. Synchronous flooding is investigated, where all lakes share some common criterion, such as altitude, depth, area or volume. He also established a continuum between multiscale segmentation and segmentation with markers by using fuzzy markers : sources are placed, whose flood is slowed down by a factor associated to each marker. In [MV02], in order to introduce geometrical regularization constraints in a morphological segmentation scheme and make it comparable to energy-driven methods, the computation of the watershed on a smooth relief by implementing viscous flooding has been proposed.

Levelings : Without prior filtering an image contains numerous tiny and meaningless homogeneous zones. For this reason, filtering before segmenting is often mandatory. The levelings [MM00] are novel object-oriented morphological filters, which enable the separation of homogeneous zones from the transition zones. Homogeneous zones can thus be enlarged, without blurring or displacing the contours : for this reason its is possible to directly segment the filtered image, with no need to go back to the original image.

PDE formulation of watershed transform : Maragos and Butt [MB00] explore the common theoretical concepts, tools, and numerical algorithms used in differential morphology and curve evolution. They focus on morphological operator representations for curve evolution as well as evolution laws for various morphological curve operations and distance transforms and consider them as the major route to connect differential morphology and curve evolution to the eikonal PDE. A minimum distance algorithm is derived for the watershed; in a continuous formulation, this is modelled via the eikonal PDE, which can be solved using curve evolution algorithms. They introduce the implementation of the watershed transform via the eikonal PDE. Another approach to the PDE formulation of the watershed transform was proposed in [NWB03], where watershed segmentation is represented as an energy minimization problem, using the distance-based definition of the watershed line. A priori considerations about smoothness are then be imposed by adding the contour length to the energy function. This leads to a segmentation method called *Watersnakes*, that integrates the strengths of watershed segmentation and energy based segmentation, in a PDE formulation.

Lattice-theoretic segmentation and Connectivity : In [Ser00] Serra investigated the problem of segmentation from the viewpoint of the algebraic lattice of image partitions. His approach is based on a new concept of connectivity defined on algebraic complete sup-generated lattices.

Multiple Cues : Multiple Cues come into play when estimating the gradient image, by using edge detectors that combine multiple cues. As an alternative, in [VPS03] a hierarchical method is proposed where a scale-space edge detection scheme is derived by the vector-valued diffusion of multiple cues.

4.2.5 Benchmarks and comparatives

Only recently has a common benchmark been made available for public use : in the Berkeley Segmentation Dataset (4.2.6) thousands of manually segmented images have been made available, so as to systematize the comparison among image segmentation methods.

4.2.6 Useful URLs

Variational Methods :

The Odyssee team (PDEs for segmentation) :
`http://www-sop.inria.fr/odyssee/research/1/index.en.html`
Level-set methods for computer vision :
`http://www.math.ucla.edu/~imagers/`

Statistical Methods :
The UCLA team's work on statistical methods for segmentation :
`http://civs.stat.ucla.edu/Segmentation/Segment.htm`
The Pattern Theory Group :
`http://www.dam.brown.edu/ptg`

Graph based Methods :
The Berkeley vision group segmentation project & normalized cuts :
`http://www.cs.berkeley.edu/projects/vision/Grouping/overview.html`
`http://www.cis.upenn.edu/~jshi/`
Graph cuts and belief propagation for vision :
`http://www.cs.cornell.edu/vision/`

Benchmarks :
The Berkeley segmentation benchmark :
`http://www.cs.berkeley.edu/projects/vision/grouping/segbench/`

4.2.7 Prior-based Segmentation in the Variational Framework

From Occam's Razor to Mumford-Shah

Image segmentation is a key element in the extraction of semantic content from images. It transforms the voluminous and intricate raw pixel data to a compact set of potentially meaningful segments. Classical methods for object segmentation and boundary determination rely on intrinsic local image features such as gray level values, texture features or image gradients. However, when the image to segment is noisy or taken under less than ideal illumination conditions, the use of a-priori knowledge is essential.

A-priori knowledge regarding image structure can be obtained from various sources. We distinguish between generic prior knowledge, that applies to most images, to application-specific prior knowledge. Notable examples for generic prior knowledge are the principle of parsimony (Occam's Razor) [Hey97] and the Gestalt principles of perception [Kof63]. In contrast, shape priors are generally application specific.

The incorporation of a-priori knowledge in image segmentation is not trivial.

The challenge is to represent the a-priori knowledge in mathematical terms that can be linked to a segmentation algorithm. For example, the generic principle of parsimony is the basis of the Minimum Description Length (MDL) [BRY98] approach. However, the seemingly appealing straightforward representation of parsimony using Kolmogorov complexity [LV97] can lead to computationally intractable algorithms.

In their seminal paper, Mumford and Shah [MS89b] casted the segmentation problem as a functional minimization problem. The functional consists of a *fidelity term*, that quantifies the discrepancy between the segmentation output and the input image data, and two additional terms that represent, in fact, the principle of parsimony. One measures the non-uniformity within segments, the other evaluates the total length of boundaries segment boundaries.

From Mumford-Shah to Chan-Vese

Formally, Mumford and Shah [MS89b] proposed to segment an input image $f\colon \Omega \to \mathbb{R}$ by minimizing the functional

$$E(u,C) = \frac{1}{2}\int_\Omega (f-u)^2 dxdy + \lambda\frac{1}{2}\int_{\Omega-C} |\nabla u|^2 dxdy + \nu|C|\ , \tag{4.4}$$

simultaneously with respect to the segmenting boundary C and the piecewise smooth approximation u, of the input image f. The functional minimization problem is approached via the calculus of variations. Straightforward minimization is not trivial, due to the dependence of the integration domains on the unknown segmentation. This technical difficulty can be alleviated using, e.g., the Γ-convergence framework [AT90].

When the weight λ of the smoothness term tends to infinity, u becomes a piecewise constant approximation, $u = \{u_i\}$, of f. The functional can now be expressed as

$$E(u,C) = \frac{1}{2}\sum_i \int_{\Omega_i} (f-u_i)^2 dxdy + \nu|C| \qquad \cup_i\Omega_i = \Omega, \quad \Omega_i \cap \Omega_j = \emptyset \tag{4.5}$$

In the two phase case, Chan and Vese [CV01b] used a level-set function $\phi \in \mathbb{R}^3$ to embed the contour $C = \{x \in \Omega |\ \phi(x) = 0\}$ [OS88], and introduced the Heaviside function $H(\phi)$ into the energy functional:

$$E_{CV}(\phi, u_+, u_-) = \int_\Omega \left[(f-u_+)2H(\phi) + (f-u_-)2(1-H(\phi)) + \nu|\nabla H(\phi)|\right] dxdy \tag{4.6}$$

where

$$H(\phi) = \begin{cases} 1 & \phi \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.7}$$

Using Euler-Lagrange equations for the functional (4.6), the following gradient descent equation for the evolution of ϕ is obtained :

$$\frac{\partial \phi}{\partial t} = \delta(\phi)\left[\nu \text{ div } (\frac{\nabla \phi}{|\nabla \phi|}) - (f - u_+)2 + (f - u_-)2\right] \ . \tag{4.8}$$

A smooth approximation of $H(\phi)$ (and $\delta(\phi)$) must be used in practice [CV01b]. The scalars u_+ and u_- are updated in alternation with the level set evolution to take the mean value of the input image f in the regions $\phi \geq 0$ and $\phi < 0$, respectively :

$$u_+ = \frac{\int f(x,y)H(\phi)dxdy}{\int H(\phi)dxdy} \qquad u_- = \frac{\int f(x,y)(1-H(\phi))dxdy}{\int (1-H(\phi))dxdy} \tag{4.9}$$

The bi-level Chan-Vese functional (4.6), provides the framework for modern prior-based segmentation techniques, that use application-specific shape priors in addition to the generic principle of parsimony. The case of polygonal contours was considered in [UYK04].

From generic priors to shape priors

In the presence of occlusion, shadows and low image contrast, generic prior knowledge is insufficient by itself. Prior knowledge on the shape of interest is then necessary [Ull96]. The recovered object boundary should then be compatible with the expected contour, in addition to being constrained by length, smoothness and fidelity to the observed image. To achieve this, the energetic formulation (4.6) can be extended by adding a prior shape term [CSS03]:

$$E(\phi, u_+, u_-) = E_{CV}(\phi, u_+, u_-) + \mu E_{shape}(\phi), \qquad \mu \geq 0. \tag{4.10}$$

The main difficulty in the integration of prior information into the variational segmentation process is the need to account for possible pose transformations between the known contour of the given object instance and the boundary in the image to be segmented. Many algorithms [CTT$^+$01, CKS02, CKS01, LGF00, TYW$^+$01, LFGW00, RP51] use a comprehensive training set to account for small deformations. These methods employ various statistical approaches to characterize the probability distribution of the shapes. They then measure the similarity between the evolving object boundary (or level set function) and representatives of the training data. The performance of these methods depends on the size and coverage of the training set.

Selected recent contributions of MUSCLE partners

The incorporation of prior knowledge in image segmentation using the variational framework is a hot research topic. Here a examples of recent contributions of MUSCLE partners in this field.

- **INRIA-ARIANA :** M. Rochery, I. H. Jermyn and J. Zerubia have recently developed high order active contours, and applied them to the detection of line networks in satellite imagery [RJZ03].
- **TAU-VISUAL :** T. Riklin-Raviv, N. Kiryati and N. Sochen brought together concepts from variational segmentation and vision geometry. Their method is deterministic, and accounts for significant projective transformations between a *single* prior image and the image to be segmented [RRKS04].

Bibliography

[AC04] P. A. Arbelaez and L. Cohen. Energy partitions and image segmentation. *Journal of Mathematical Imaging and Vision*, 20:4357, 2004.

[AT90] L. Ambrosio and V. Tortorelli. Approximation of functionals depending on jumps by elliptic functionals via γ-convergence. *Communications on Pure and Applied Mathematics*, 1053:999–1036, 1990.

[BA83] P. J. Burt and E. H. Adelson. The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540, April 1983.

[BK03] Y. Boykov and V. Kolmogorov. Computing geodesics and minimal surfaces via graph cuts. In *Proc. International Conference on Computer Vision*. 2003.

[BRDW03] T. Brox, M. Roussonl, R. Deriche, and J. Weickert. Unsupervised segmentation incorporating colour, texture, and motion. Technical Report RR-4760, INRIA, 2003. `ftp://ftp.inria.fr/INRIA/publication/publi-pdf/RR/RR-4760.pdf`.

[BRY98] A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Trans. Information Theory*, 44:2743–2760, 1998.

[BVZ01] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.

[BZ03] A. Barbu and S. Zhu. Graph partition by Swendsen-Wang cuts. In *Proc. International Conference on Computer Vision*. 2003.

[Can01] E. J. Candès. The curvelet transform for image denoising. In *Proceedings of the IEEE International Conference on Image Processing*. Thessaloniki, Greece, October 7-10 2001.

[CK97] L. D. Cohen and R. Kimmel. Global minimum for active contour models: A minimal path approach. *International Journal of Computer Vision*, 24(1):57–78, 1997.

[CKS97] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. *International Journal of Computer Vision (IJCV)*, 22(1):61–79, 1997.

[CKS01] D. Cremers, T. Kohlberger, and C. Schnorr. Nonlinear shape statistics via kernel spaces. In *Proc. DAGM01*, pages 269–276. 2001.

[CKS02] D. Cremers, T. Kohlberger, and C. Schnorr. Nonlinear shape statistics in mumford-shah based segmentation. In *Proc. ECCV02*, volume 2, pages 93–108. 2002.

[CSS03] D. Cremers, N. Sochen, and D. Schnorr. Towards recognition-based variational segmentation using shape priors and dynamic labeling. In *Proc. Intl. Conf. on Scale-Space Theories in Computer Vision*, pages 388–400. 2003.

[CSS04] D. Cremers, N. Sochen, and C. Schnorr. Multiphase dynamic labeling for variational recognition-driven image segmentation. In *Proc. European Conference on Computer Vision*. 2004.

[CTT$^+$01] Y. Chen, S. Thiruvenkadam, H. Tagare, F. Huang, and D. Wilson. On the incorporation of shape priors into geometric active contours. In *Proc. VLSM01*, pages 145–152. 2001.

[CV01a] T. Chan and L. Vese. Active contours without edges. *IEEE Tans. on Image Processing*, 10(2), 2001.

[CV01b] T. Chan and L. Vese. Active contours without edges. *IEEE Trans. Image Processing*, 10:266–277, 2001.

[Don98] D. L. Donoho. Orthonormal ridgelets and linear singularities. Technical report, Statistics Department, Stanford University, California, 1998.

[FH04] P. Felzenszwalb and D. Huttenlocher. Efficient belief propagation for early vision. In *Proc. International Conference on Computer Vision & Pattern. Recognition*. 2004.

[GG84] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

[Hey97] F. Heylighen. Occam's razor. *Principia Cybernetica Web (Principia Cybernetica, Brussels)*, 1997.

[HG00] H. J. A. M. Heijmans and J. Goutsias. Nonlinear multiresolution signal decomposition schemes. Part II: Morphological wavelets. *IEEE Transactions on Image Processing*, 9(11):1897–1913, 2000.

[HPPP03] H. J. A. M. Heijmans, B. Pesquet-Popescu, and G. Piella. Building nonredundant adaptive wavelets by update lifting, 2003. Submitted to *Applied and Computational Harmonic Analysis*, preliminary work: Research Report PNA-R0212, CWI, Amsterdam.

[HPPP04] H. J. A. M. Heijmans, G. Piella, and B. Pesquet-Popescu. Adaptive wavelets for image compression using update lifting: Quantisation and error analysis, 2004. Submitted to *International Journal of Wavelets, Multiresolution and Information Processing* (IJWMIP).

[Kof63] K. Koffka. *Principles of Gestalt Psychology.* Harcourt Brace, New York, 1963.

[KV92] J. Kovačević and M. Vetterli. Nonseparable multidimensional perfect reconstruction filter banks and wavelet bases for $\mathbb{R}^n$. *IEEE Transactions on Information Theory*, 38:533–555, 1992.

[KWT87] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. In *Proc. International Conference on Computer Vision.* 1987.

[LFGW00] M. Leventon, O. Faugeras, W. Grimson, and W. Wells. Level set based segmentation with intensity and curvature priors. In *Proc. Workshop on Mathematical Methods in Biomedical Image Analysis*, pages 4–11. 2000.

[LGF00] E. Leventon, W. Grimson, and O. Faugeras. Statistical shape influence in geodesic active contours. In *Proc. CVPR00*, volume 1, pages 316–323. 2000.

[LM00] E. Le Pennec and S. G. Mallat. Image compression with geometrical wavelets. In *Proceedings of the IEEE International Conference on Image Processing.* Vancouver, Canada, September 10-13 2000.

[LV97] M. Li and P. Vitanyi. *An Introduction to Kolmogorov Complexity and Its Applications.* Springer Verlag, 2nd edition, 1997.

[Mal89] S. G. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:674–693, 1989.

[MB90] F. Meyer and S. Beucher. Morphological segmentation. *Journal of Visual Communication and Image Representation*, 1(1):21 – 45, 1990.

[MB00] P. Maragos and M. A. Butt. Curve evolution, differential morphology, and distance transforms applied to multiscale and eikonal problems. *Fundamenta Informaticae*, 41:91 – 129, 2000.

[Mey99] F. Meyer. Morphological multiscale and interactive segmentation. In *Proc. IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing*. 1999.

[Mey00] F. Meyer. Flooding and segmentation. In *Proc. International Symposium on Mathematical Morphology*. 2000.

[MM99] F. Meyer and P. Maragos. Multiscale morphological segmentations based on watershed, flooding, and eikonal PDE. In *Proc. of Scale-Space*, volume 1682 of *LNCS*, pages 351–362. 1999.

[MM00] F. Meyer and P. Maragos. Nonlinear scale-space representation with morphological levelings. *Journal of Visual Communication and Image Representation*, 11:245–265, 2000.

[MS89a] D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational-problems. *Communications on Pure and Applied Mathematics*, 42(5):577–685, 1989.

[MS89b] D. Mumford and J. Shahem. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on Pure and Applied Mathematics*, 42:577–684, 1989.

[MV02] F. Meyer and C. Vachier. Image segmentation based on viscuous flooding. In *Proc. International Symposium on Mathematical Morphology*. 2002.

[NWB03] H. T. Nguyen, M. Worring, and R. Boomgaard. Watersnakes: energy-driven watershed segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(3):330–342, 2003.

[OS88] S. Osher and J. Sethian. Fronts propagating with curvature-dependent speed: Algorithms based on hamilton-jacobi formulations. *Journal of Computational Physics*, 79:12–49, 1988.

[PD02] N. Paragios and R. Deriche. Geodesic active contours and level set methods for supervised texture segmentation. *Internation Journal of Computer Vision*, 2002.

[PH02] G. Piella and H. J. A. M. Heijmans. Adaptive lifting schemes with perfect reconstruction. *IEEE Transactions on Signal Processing*, 50(7):1620–1630, July 2002.

[PPPH02] G. Piella, B. Pesquet-Popescu, and H. J. A. M. Heijmans. Adaptive update lifting with a decision rule based on derivative filters. *IEEE Signal Processing Letters*, 9(10):329–332, October 2002.

[RBD03] M. Rousson, T. Brox, and R. Deriche. Active unsupervised texture segmentation on a diffusion based space. In *Proc. International Conference on Computer Vision & Pattern Recognition*. 2003.

[RJZ03] M. Rochery, I. H. Jermyn, and J. Zerubia. Higher order active contours and their application to the detection of line networks in satellite imagery. In *Proc. IEEE Workshop Variational, Geometric and Level Set Methods in Computer Vision, (ICCV)*. Nice, France, October 2003.

[RP02] M. Rousson and N. Paragios. Shape priors for level set representations. In *Proc. European Conference in Computer Vision*. 2002.

[RP51] M. Rousson and N. Paragios. Shape priors for level set representations. *Proc. ECCV 2002*, 2002:78–92, LNCS 2351.

[RRKS04] T. Riklin-Raviv, N. Kiryati, and N. Sochen. Unlevel-sets : Geometry and prior-based segmentation. In *Proc. 8th European Conference on Computer Vision (ECCV'2004)*, volume Part IV: *Lecture Notes in Coputer Science* #3024, pages 50–61. Springer, Prague, Czech Republic, May 2004.

[Ser00] J. Serra. Connections for sets and functions. *Fundamenta Informaticae*, 41, 2000.

[Set96] J. Sethian. *Level Set Methods*. Cambridge University Press, 1996.

[SM01] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2001.

[TCYZ03] Z. Tu, X. Chen, A. Yuille, and S. Zhu. Image parsing: Unifying segmentation, detection, and recognition. In *Proc. International Conference on Computer Vision*. 2003.

[TYW$^+$01] A. Tsai, A. Yezzi, W. Wells, C. Tempany, D. Tucker, A. Fan, W. Grimson, and A. Willsky. Model-based curve evolution technique for image segmentation. *Proc. CVPR01*, 1:463–468, 2001.

[TZ02] Z. Tu and S. Zhu. Image segmentation by data-driven MCMC. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(5):657–673, 2002.

[Ull96] S. Ullman. *High-Level Vision: Object Recognition and Visual Cognition*, chapter 8. MIT Press, 1996.

[UYK04] G. B. Unal, A. J. Yezzi, and H. Krim. Information-theoretic active polygons for unsupervised texture segmentation. *International Journal of Computer Vision*, 62(3):199–220, 2004.

[VPS03] I. Vanhamel, I. Pratikakis, and H. Sahli. Multiscale gradient watersheds of color images. *IEEE Trans. on Image Processing*, 12(6):617–626, 2003.

[YFW01] J. Yedidia, W. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. In *Proc. Uncertainty in Artificial Intelligence*. 2001. `http://www.merl.com/reports/docs/TR2001-22.pdf`.

[ZY96] S. Zhu and A. Yuille. Region competition: Unifying snakes, region growing and Bayes/MDL for multiband image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(9):884–900, 1996.

Chapter 5

Image Sequence Features

Motion is seamlessly perceived by human beings when directly observing a day-life scene, but also when watching films, videos or TV programs, or even various domain-specific image sequences such as meteorological or heart ultrasound ones. However, motion information is hidden in the image sequences supplied by image sensors. It has to be recovered from the observations formed by the image intensities in the successive frames of the sequence.

Assumptions (i.e., *data models*) must be formulated to relate the observed image intensities with motion. When dealing with video, the commonly used data model is the brightness constancy constraint which states that the intensity does not change along the trajectory of the moving point in the image plane (at least, to a short time extent). The motion constraint equation can then be expressed in a differential form that relates the 2D velocity vector, the spatial image gradient and the temporal intensity derivative at any point p in the image. Nevertheless, this enables to locally retrieve one component of the velocity vector only, the so-called normal flow, which corresponds to the aperture problem. Then, other constraints (i.e., *motion models*) must be added. They are supposed to formalize known, expected or learned properties of the motion field, and this implies to somehow introduce spatial coherence or more generally contextual information.

Visual motion information can involve different kinds of mathematical variables. First, we can deal with *continuous variables* to represent the motion field : velocity vectors $\mathbf{w}(p)$ with $\mathbf{w}(p) \in \mathbb{R}^2$, or parametric motion models with parameters $\theta \in \mathbb{R}^d$ with d denoting the number of parameters. Let us note that the latter can be equivalently represented by the model flow vectors $\{\mathbf{w}_\theta(p)\}$ with $\mathbf{w}_\theta(p) \in \mathbb{R}^2$. Second, we can consider *discrete values or symbolic labels* to code motion detection output : binary values $\{0, 1\}$, or motion segmentation output : number n of the motion region or layer with $n \in \{1, ..., N\}$.

Spatial coherence can be formalized by conditional densities defined on local neighborhoods as in Markov Random Field (MRF) models, or equivalently by potentials on cliques as in Gibbs distributions. Another way is to first segment each image into spatial regions according to a given criterion (grey level, colour, texture) and to analyse the motion information over these regions. Perceptual grouping schemes can also be envisaged.

5.1 Local motion measurements

The brightness constancy assumption along the trajectory of a moving point $p(t)$ in the image plane, with $p(t) = (x(t), y(t))$, can be expressed as $dI(x(t), y(t), t)/dt = 0$, with I denoting the image intensity function. By applying the chain rule, we get the well-known motion constraint equation [HS81, Nag87] :

$$r(p,t) = \mathbf{w}(p,t).\nabla I(p,t) + I_t(p,t) = 0 \ , \tag{5.1}$$

where ∇I denotes the spatial gradient of the intensity, with $\nabla I = (I_x, I_y)$, and I_t its partial temporal derivative. The above equation can be straightforwardly extended to the case where a parametric motion model is considered, and we can write :

$$r_\theta(p,t) = \mathbf{w}_\theta(p,t).\nabla I(p,t) + I_t(p,t) = 0 \ , \tag{5.2}$$

where θ denotes the vector of motion model parameters. It can be easily derived from equation (5.1) that the motion information which can be locally recovered at a pixel p is contained in the *normal flow* given by :

$$\nu(p,t) = \frac{-I_t(p,t)}{\|\nabla I(p,t)\|} \ . \tag{5.3}$$

It can also be written in a vectorial form: $\boldsymbol{\nu}(p,t) = \frac{-I_t(p,t)}{\|\nabla I(p,t)\|}\boldsymbol{\omega}_{\nabla I}(p,t)$, where $\boldsymbol{\omega}_{\nabla I}$ denotes the unit vector parallel to the intensity spatial gradient. However, it should be clear that the orientation of the normal flow vector does not convey any information on the motion direction, but implicitly on the object texture (for inner points) or on the object shape (for points on the object border). Besides, the normal flow can be computed at the right scale to enforce reliability as explained in [FB03].

In case of a moving camera and assuming that the dominant image motion is due to the camera motion and can be correctly described by a 2D parametric motion model, we can exhibit the *residual normal flow* given by :

$$\nu_{res}(p,t) = \frac{-DFD_{\hat{\theta}}(p,t)}{\|\nabla I(p,t)\|} \ , \tag{5.4}$$

where $DFD_{\hat{\theta}}(p,t) = I(p + \mathbf{w}_{\hat{\theta}}, t+1) - I(p,t)$ is the displaced frame difference corresponding to the compensation of the dominant motion described by the estimated motion model parameters $\hat{\theta}$.

Since the computation of intensity derivatives is usually affected by noise and can be unreliable in nearly uniform areas, it may be preferable to consider the local mean of the absolute magnitude of normal residual flows weighted by the square of the norm of the spatial intensity gradient (as proposed in [IRP94, OB97]) :

$$\bar{\nu}_{res}(p,t) = \frac{\sum_{q\in\mathcal{F}(p)} \|\nabla I(q,t)\|.|DFD_{\hat{\theta}_t}(q)|}{\max\left(\eta^2, \sum_{q\in\mathcal{F}(p)} \|\nabla I(q,t)\|^2\right)} \quad , \tag{5.5}$$

where $\mathcal{F}(p)$ is a local spatial window centered in pixel p (typically a 3×3 window), and η^2 is a predetermined constant related to the noise level. An interesting property of the local motion quantity $\bar{\nu}_{res}(p)$ is that the reliability of the conveyed motion information can be locally evaluated. Given the lowest motion magnitude δ to be detected, we can derive two bounds, $l_\delta(p)$ and $L_\delta(p)$, verifying the following properties [OB97]. If $\bar{\nu}_{res}(p) < l_\delta(p)$, the magnitude of the (unknown) true velocity vector $\mathbf{w}(p)$ is necessarily lower than δ. Conversely, if $\bar{\nu}_{res}(p) > L_\delta(p)$, $\|\mathbf{w}(p)\|$ is necessarily greater than δ. The two bounds l_δ and L_δ can be directly computed from the spatial derivatives of the intensity function within the window $\mathcal{F}(p)$.

By defining the motion quantity $\bar{\nu}_{res}(p)$, we already advocate the interest of considering spatial coherence to compute motion information. Here, it simply amounts to a weighted averaging over a small spatial support and it only concerns the data model. In the same vein, more information can be locally extracted by considering small spatio-temporal supports, either through spatio-temporal (frequency-based) velocity-tuned filters as in [FJ90] or using 3D orientation tensors [BGW91, NG98]. On the other hand, more benefit can be gained by introducing contextual information through the motion models.

Figure 5.1 presents some results of estimated dominant image motion and maps of residual normal flow magnitudes [PBY04].

5.2 Motion detection and segmentation

Previous approaches to motion detection can be split into two categories: methods based on motion segmentation and methods thresholding image differences. A synopsis of methods from both categories is proposed in [MB96].

Motion segmentation methods require an accurate estimation of the 2D apparent motion in the image. This is not trivial since computing motion estimation on the various image supports arising from objects in the scene is definitely demanding. However, some specific problems cannot be solved without information on the orientation of motion. For example, apparent motion estimation enables to conclude that the motion of a static background and a static foreground, although different in their 2D projection, are induced by the same camera motion and should therefore not be detected as independently moving. This can be done using parallax and rigidity constraints [IA99], [Nel91], [TP90], [TLS93] and [CB71].

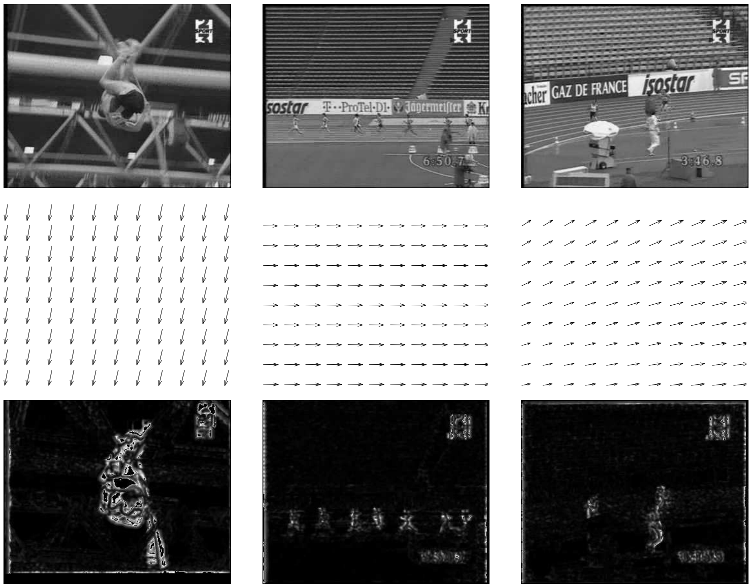


Figure 5.1: Track-and-field video: Top row: three images at different time instants of a track-and-field sport TV program involving respectively an upward-tilt camera motion, a left panning one, and a zoom combined with a panning motion. Middle row: the velocity vector fields corresponding to the estimated dominant image motion (due to camera motion). Bottom row: maps of residual normal flow magnitudes (zero-value in black) [Piriou04].

Thresholding methods are applied to temporal image differences. The temporal image differences of static regions and those of moving objects can be statistically modeled. Within the class of thresholding methods, different kinds of image structures can be considered. They may range from decision on single pixels to spatial regularization using MRFs or active contours. The highest level of structure is representing moving objects by their spatio-temporal envelopes. A different form of structure is to consider only edges and to detect those that are moving.

A review of basic methods is given in [Kon00]. Likelihood tests for motion detection were early introduced in [HNR84]. More complex methods are proposed in [Ros02] for modeling the spatial distribution of either noise or signal and selecting an appropriate threshold. Markov Random Fields are a natural extension in order to introduce contextual information into the detection scheme [AK95]. Previous to image differencing, camera-induced motion can also be estimated and compensated. To this end, a 2D parametric motion model is used in [OB97] along with hierarchical MRFs. In [DKA04] wavelet analysis and robust techniques are introduced to estimate dominant motion and hierarchical MRFs are again exploited.

A higher level of spatial structure can be reached using active contour and the level set theory. For instance the approach described in [PD00] applies active contours for detection and tracking of moving objects. It relies on the gray level intensities for providing object boundaries. Two methods using level sets implementation are introduced in [MK03]: one purely based on motion and another enforcing correspondence between motion boundaries and intensity boundaries. Besides, they can distinguish between different moving objects.

Bayesian approaches like MRF and variational methods both rely on energy minimization techniques. In the Bayesian framework, the aim is to maximize the posterior probability of a model given the observations. Variational techniques minimize an energy functional yielding a contour evolving according to some constraints. The formulation of energies for both approaches is similar. It consists of a regularization term and a term to fit observations. However, it is not possible to assess the validity of the extremum. In other words, it is not possible to interpret the value of the extremum. Thus, the best state is reached given the model and the observations, but the quality of this state cannot be assessed.

Temporal integration improves quality and stability of the detection. Accumulating motion information over time makes moving objects more salient with regard to noise. This enables to detect small slowly moving objects. Directionally consistent flow is accumulated over time in [Wix00]. A graph to represent moving objects is exploited in [CM99], and object trajectories are optimal paths in this graph. Spatio-temporal image intensity gradients to create mosaics of the background are used in [PBA00]. Residual motion is then propagated and accumulated without optical flow computation. A threshold allows one to balance between false alarms and minimal detectable motion. It is not clear how to set this threshold, unless empirically.

As mentioned above, motion information accumulates on edges. The work in [SDC04] concentrates on these highly contrasted features to detect multiple layered motions. In [JKM04] contour fragments are matched. The different transformations issued from those matchings are clustered into background and moving objects. This work is closer to shape matching issues than to motion detection. Both methods rely on a Canny-type edge detector.

Some work has also been devoted to distinguish between motion and changes. Changes are variations of image intensities which do not correspond to a real moving object : shadows and reflections but also aliasing [BFY00].

The use of perceptual criteria for change detection appears in [LM03] and [SMK02]. In [LM03] an *a contrario* framework is applied to detect changes in satellite images of urban landscapes. The considered local change information is the image gradient orientation. In [SMK02], the use of perceptual organization is considered to build the spatio-temporal envelopes of moving objects. The approach is essentially applied to human undergoing fronto-parallel motion in front of a static camera. The envelopes localize the motion information but do not provide shape information. In [VCB04], camera-induced motion is first compen-

sated using 2D parametric motion models. The perceptual grouping principle allows the computation of automatic detection thresholds. Detection operates on three frames only. Boundaries of moving objects are retrieved through an image segmentation based on meaningful intensity level lines. Moreover, a confidence level for each detected region is derived through the so-called number of false alarms evaluated according to the *a contrario* model. The lower this number, the more reliable the detected event.
Figures 5.2, 5.3, 5.4 and 5.5 illustrate the approach with examples of objects detection on different sequences.

Let us now consider motion segmentation meant as the competitive partitioning of the image into motion-based homogeneous regions. One important step ahead in solving the motion segmentation problem was to formulate the motion segmentation problem as a statistical contextual labeling problem or in other words as a discrete Bayesian inference problem. Segmenting the moving objects is then equivalent to assigning the proper (symbolic) label (i.e., the region number) to each pixel in the image, while estimating 2D parametric (usually affine) motion models over the region supports (which is obviously a chicken-and-egg problem). The advantages of this class of methods are mainly two-fold. Determining the support of each region is then implicit and easy to handle: it merely results from extracting the connected components of pixels with the same label. Introducing spatial coherence can be straightforwardly (and locally) expressed by exploiting MRF models. This formulation can also encompass the determination of motion layers by assuming that the regions of same label are not necessarily connected [SA96].

Specifying (simple) MRF models at a pixel level (i.e., sites are pixels and a 4- or 8-neighbour system is considered) is efficient, but remains limited to express more sophisticated properties on region geometry (e.g., more global shape information [CKS02]) or to handle extended spatial interaction. Multigrid MRF models [HPB94] (as used in [OB97, OB98]) is a means to address somewhat the second concern (and also to speed up the minimization process while usually supplying better results). An alternative is to first segment the image into spatial regions (based on grey level, colour or texture) and to specify a MRF model on the resulting graph of adjacent regions as done in [GB00]. The motion region labels are then assigned to the nodes of the graph (which are the sites considered in that case). This enables to exploit more elaborated and less local *a priori* information on the geometry of the regions and their motion. However, the spatial segmentation stage is often time consuming, and getting an effective improvement on the final motion segmentation accuracy remains questionable. Using the level-set framework is another way to precisely locate region boundaries while dealing with topology changes [PD00], but handling a competitive motion partioning of the image (with the number of regions *a priori* unknown) remains an open issue in that context even if recent attempts have been reported [CS03, MK03].

Let us also mention other recent work on Bayesian motion segmentation, exploring the use of edge motion [SDC04], offering extension to spatio-temporal models [CS03], or introducing (two-step) hidden Markov measure field (HMMF) models [MSB03]. Tensor voting could also be considered as an implicit way to enforce spatial coherence [NM03].

5.3 Optical flow computation

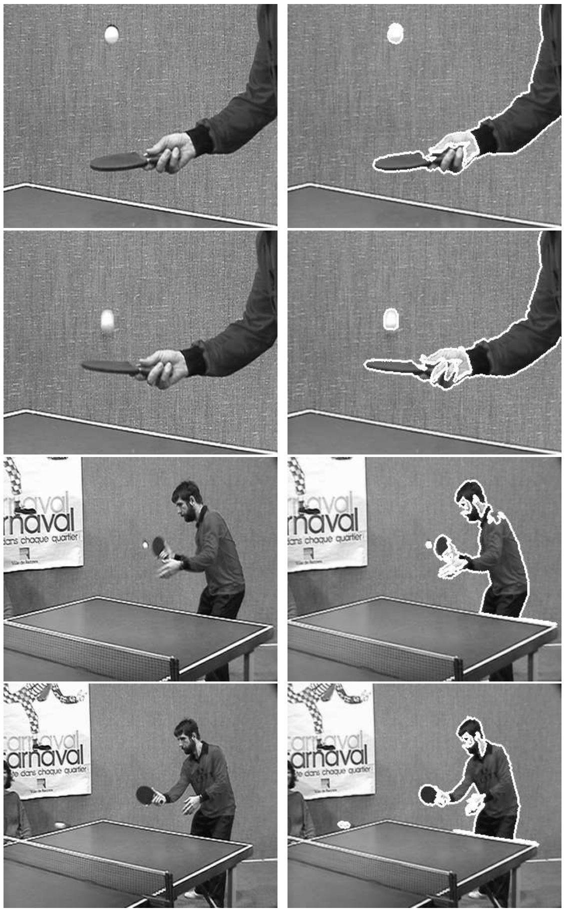
By definition, the velocity field formed by continuous vector variables is a complete representation of the motion information. Computing optical flow based on the data model of equation (5.1) requires to add a motion model enforcing the expected spatial properties of the motion field, that is, to resort to a regularization method. Such properties of spatial coherence (more specifically, piecewise continuity of the motion field) can be expressed on local spatial neighborhoods. First methods to estimate discontinuous optical flow fields were based on MRF models associated with Bayesian inference [HB93, MB96, SK99] (i.e., minimization of a discretized energy function). Then, continuous-domain models were designed based on PDE formalism [AWS00, CH99, KDA99, WBPB03]. Spatial coherence can also be explicitly formulated by first segmenting the image in spatial regions forming the delimited domains where motion models, either dense or parametric ones, can be defined and estimated [BJ96, GB00].

A general formulation of the global (discretized) energy function to be minimized to estimate the velocity field $\mathbf{w}$ can be given by :

$$E(\mathbf{w},\zeta) = \sum_{p\in S} \rho_1[r(p)] \; + \sum_{p\sim q} \rho_2[\|\mathbf{w}(p)-\mathbf{w}(q)\|, \zeta(p'_{p\sim q})] + \sum_{A\in\chi} \rho_3(\zeta_A) \;\; , \qquad (5.6)$$

where S designates the set of pixel sites, $r(p)$ is defined in (5.1), $S' = \{p'\}$ the set of discontinuity sites located midway between the pixel sites and χ is the set of cliques associated with the neighborhood system chosen on S'. In [HB93], quadratic functions were used and the motion discontinuities were handled by introducing a binary line process ζ. Then, robust estimators were popularized [BA96, MP98] leading to the introduction of so-called auxiliary variables ζ now taking their values in $[0, 1]$. Depending on the followed approach, the third term of the energy $E(\mathbf{w}, \zeta)$ can be optional. Multigrid MRF are moreover involved in the scheme developed in [MP98]. Besides, multiresolution incremental schemes are required to compute optical flow in case of large displacements. Dense optical flow and parametric motion models can also be jointly considered and estimated, which enables to supply a segmented velocity field as designed in [MP02].

Recent advances have dealt with the computation of fluid motion fields involving the definition of a new data model (derived from the continuity equation of the fluid mechanics) and of a motion model preserving the underlying physics of the visualized fluid flows (2^{nd} order div-curl constraint) as defined in [CMP02]. A comprehensive investigation of physics-based data models is described in [HF01].



(a) Original images (b) Boundaries of detected moving regions (in white)

Figure 5.2: Table tennis sequence. Original images in the left column. The boundaries of the detected moving regions appear in white in the right column. Both the player and the ball are accurately detected in the four images. In the first part of this video (two upper rows of Fig. 5.2), the camera is almost static and only the forearm of the player appears. The ball, the arm and the racket are correctly detected. The detected moving regions closely fit the contours of the moving object. The regions detected on the ball are associated with NFAs of 10^{-20}. Let us recall that the lower the NFA, the higher the confidence in the detection. NFAs on the arm are even much lower, about 10^{-150}. This reflects that the motion of the arm is perceptually much more meaningful. In the second part of the video (two lower rows of Fig. 5.2), the camera is zooming out. The global dominant motion estimation performs well , although the assumption of a planar background is violated by the table. Both the ball and the body of the player are accurately detected. The NFA on the ball is about 10^{-1} in the bottom left image. This means that the residual motion observations on the ball hardly distinguish from noise. Indeed, the velocity of the ball reaches a minimum before being hit by the racket. On top of that, the size of the ball is only about thirty pixels. It is hardly possible to gather enough motion evidence on such a small region. In the bottom right image, the velocity of the ball increased dramatically. This is reflected by an increase of the confidence in the detection and the NFA of the region corresponding to the ball decreases to 10^{-15}. The velocity of the player varies inversely to that of the ball. It is maximal just before hitting the ball. The NFA is about 10^{-80} on the player in the lower left image. The velocity of the player decreases afterward and the associated NFA raises to 10^{-30} in the bottom right image.

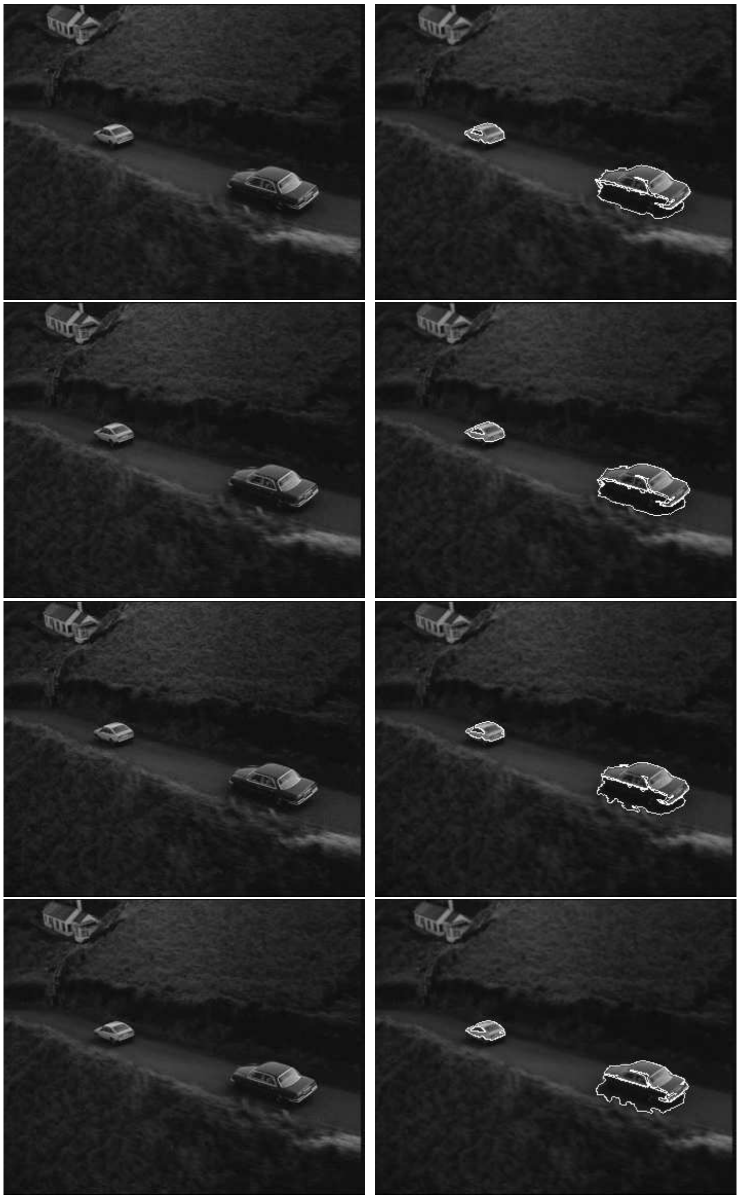


Figure 5.3: Four consecutive images of the road sequence. Original images in the left column. Outline of detected regions in white in the right column. NFAs are about 10^{-10} on the white car on the left of the scene. Two regions are detected on the darker car on the right. The upper region corresponding to the more contrasted part has an NFA about 10^{-70} in the four images. The lower region has an NFA of 10^{-15}. The higher NFA (lower confidence) is explained by the saturated gray levels on the lower region.

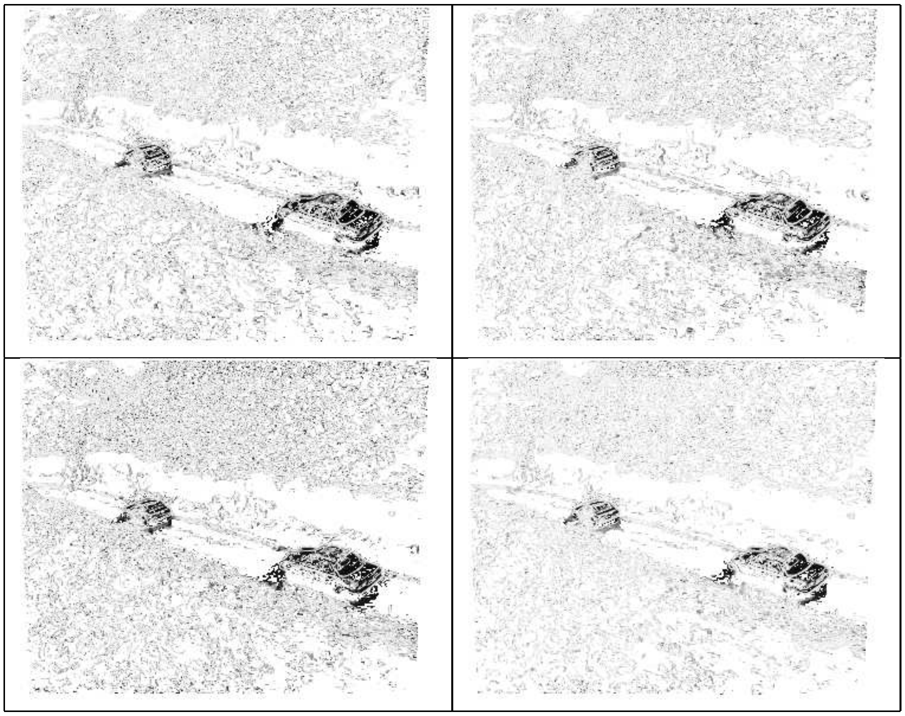


Figure 5.4: Road sequence. Maps of residual motion for the four successive images displayed in Fig. 5.3. On the lower part of the right car, the low contrast prevents from extracting motion measures. The *a contrario* approach consists in contradicting the fact that the value of residual motion are independently distributed. The number of false alarms (NFA) of a region, is related to the probability to observe by chance values of residual motion that are as high as the ones which are actually observed.

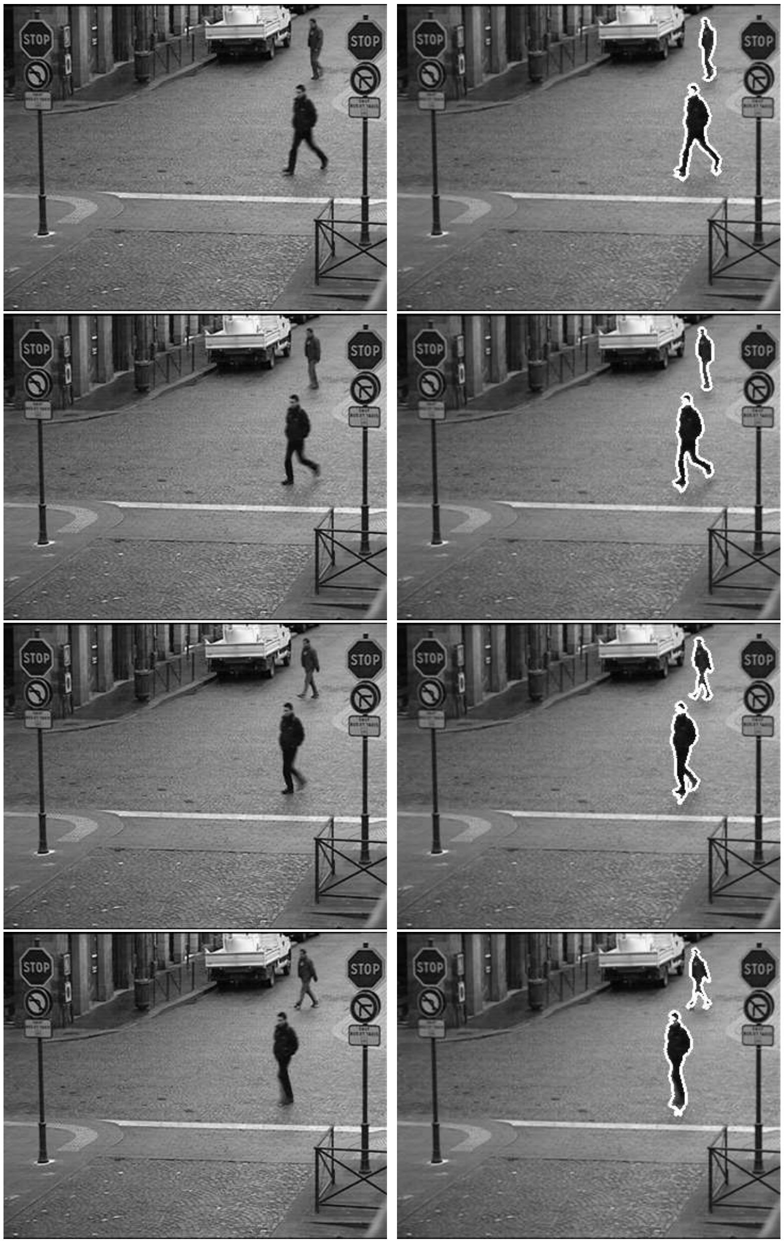


Figure 5.5: Four images of the street sequence. Original images on the left. Detected regions are outlined in white on the right. NFAs are about 10^{-80} on the pedestrian closer to the camera and about 10^{-18} for the pedestrian in the background

5.4 Motion recognition

Exploiting the tremendous amount of multimedia data, and specifically video data, requires to develop methods able to extract information at a higher level (more semantic) level. Video summarization, video retrieval or video surveillance are examples of applications. Inferring concepts from low-level video features is a highly challenging problem. The characteristics of a semantic event have to be expressed in terms of video primitives (color, texture, motion, shape ...) sufficiently discriminant w.r.t. content. This remains an open problem at the source of active research activities.

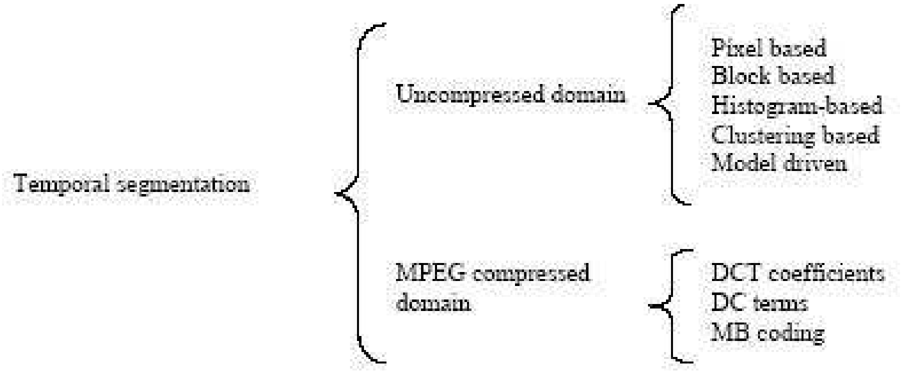
In [VL00], statistical models for components of the video structure are introduced to classify video sequences into different genres. The analysis of image motion is widely exploited for the segmentation of videos into meaningful units or for event recognition. Efficient motion characterization can be derived from the optical flow, as in [RA00] for human action change detection. In [ZMI01], the authors use very simple local spatio-temporal measurements, i.e., histograms of the spatial and temporal intensity gradients, to cluster temporal dynamic events. In [YB98], a principal component representation of activity parameters (such as translation, rotation ...) learnt from a set of examples is introduced. The considered application was the recognition of particular human motions, assuming an initial segmentation of the body.

In [DRP02], video abstraction relies on a measure of fidelity of a set of key-frames and a measure of summarizability derived from MPEG-7 descriptors. In [NPZ02], spatio-temporal slices extracted in the volume formed by the image sequence are exploited for clustering and retrieving tasks. Sport videos are receiving specific attention due to the economical importance of sport TV programs and to future services to be designed in that context. Different approaches have been recently investigated to detect highlights in sports videos [ETM03].

In [PBY04], a statistical approach is proposed involving modeling, (supervised) learning and classification issues, to deal with concepts related to events in videos, more precisely, to dynamic content. Focus is put on motion information. Since no analytical motion models are available to account for the variety of dynamic contents to be found in videos, motion models have to be specified and learned from the image data. To this end, new mixed-state probabilistic motion models are introduced. Furthermore, such a probabilistic modelling allows the derivation of a parsimonious motion representation while coping with errors in the motion measurements and with variability in a given kind of motion content. Scene motion (i.e., the residual image motion) and camera motion (i.e., the dominant image motion) are handled in a distinct way. Indeed, these two sources of motion bring important, different but complementary, information which have to be explicitly taken into account for event detection. As for motion measurements, on one hand, 2D parametric motion models are considered to capture the camera motion, and on the other hand, low-level local motion features to account for the scene motion. They convey more information than those used in [ZMI01], while still easily computable contrary to optic flow.

5.5 Video shot detection algorithms

During the last years several papers describing the state of the art in video shot detection have been published, like [BR96], [DAE95] or [KC01]. We would like to provide an extended review of current techniques describing the most relevant approaches in view of our applications in the context of video analysis for information retrieval. A well-accepted classification of shot boundary detection algorithms is based on the format of the input sequence : uncompressed (raw) images, or an MPEG bit-stream. Under these main categories, we group the different techniques in the state of the art, as represented in the figure below.



5.5.1 Segmentation in the compressed domain

Most of existing algorithms work in the uncompressed domain. They use as input data a sequence of frames, being the color of the pixels the only information directly available. Commonly, transitions are detected by means of a threshold over a similarity measure. According to the metric used to compute the distance between frames and the technique used in the segmentation process, current algorithms can be classified into five sub-groups [KC01].

Pixel comparison :
This type of transition detection algorithms, also known as template matching, relies on the computation of pixel-wise difference between consecutive frames. One of the earliest methods was implemented by Kikukawa and Kawafuchi in [KK92]. They compute the difference between images as the absolute sum of pixel differences. For gray-level images, the frame difference curve FD(t) is the absolute sum of pixel differences between frames at t and t + 1. By extension, for color images, the frame difference curve is computed as the mean of the FD

in each color component. The proposed algorithm detects a transition when the value of the frame difference exceeds a certain threshold. The main drawback of this approach is that the system is not able to distinguish between a small change of the whole image and a large change in a small portion of the image (due to object motion, for example) resulting in a high rate of false positives. An improvement for this technique is presented by Zhang et al. [ZKS93]. In [HJT94], Hampapur et al. compute what they call chromatic images by dividing the change in gray level of corresponding pixels from two consecutive images by the value of that pixel in the last image. Then, fades should correspond to uniform chromatic images. But this approach is also very sensitive to camera and object motion.

Campisi et al. [CNS02] present a classic but robust method. A difference image sequence is computed from the video. Then each difference image is segmented into blocks. The similarity between consecutive collocated blocks is computed and the sum of the similarities for the whole image, and within a temporal window, is compared with a dynamic threshold to detect a fade or a dissolve.

An approach based on the variance of pixel intensities was proposed by Alattar [Ala97]. Fades were detected first by recording all negative spikes in the time series of the second order difference of the pixel intensity variance, and then by ensuring that the first order difference of the mean of the video sequence was relatively constant next to the negative spike. A combination of both approaches is described in Truong et al. [TDV00]. These methods have a relatively high false detection rate.

One algorithm for cuts and one for fades and dissolves have been developed by Huang and Liao [HL01]. The first one checks if the value of the DC image difference has a local maximum (within a window) at the current frame transition. Then the ratio of this maximum to the next highest maximum in the window is compared with two thresholds. If it is higher than the highest threshold, a cut is detected. If it is between the two thresholds, a heuristic algorithm that considers the histogram difference and the static or dynamic nature of the appearing and disappearing shots (based on edge differences) is used. In the algorithm for gradual transitions, the image difference from a specific frame is differentiated, its zero-crossings are computed and then low-pass filtered. Then the transition is the interval where there are few resultant zero-crossings, meaning that the difference is monotonically increasing.

Lienhart [Lie99] proposed first to locate all monochromatic frames in the video as potential start/end points of fades. Monochromatic frames were identified as frames with standard deviation of pixel intensities close to zero. Fades were then detected by starting to search in both directions for a linear increase in the standard deviation of pixel intensity/color.An average hit rate of 87% was reported at false alarm rate of 30%.

Block-based pixel comparison :

The previous techniques are based on global image characteristics, which makes them very sensitive to camera and object motion. In order to increase the robustness, block-based approaches use local characteristics of the image. Under this approach the frame difference is computed comparing the B blocks in which the image is divided. Following this strategy, different techniques can be derived depending on the computation of the difference between blocks. Kasturi and Jain [KJ91] compute the similarity between blocks using a likelihood ratio based on the mean and the variance of the luminance values in each block. Blocks are then counted as changed if and only if the likelihood ratio exceeds a certain threshold. When the percentage of changed blocks is above a second threshold, a cut is declared. This method is more robust in front of slow camera and object motion, but it still presents a high ratio of false positives. Moreover, because statistical values must be computed, its computational burden is higher than for template matching approaches. In order to improve the robustness in front of motion, Shahraray [Sha95] proposes a method where several matches are considered for each block. A new algorithm called net comparison is presented by Xiong et al. in [XLI95]. As previous techniques, it declares a cut according to the percentage of changed blocks between consecutive frames, but in order to increase the processing speed only a part of the image is analyzed. As an extension, in [XL98] the same authors propose to sub-sample in both time and space. The last technique we present based on block differences was developed by Demarty [9]. In the first step, her approach computes the difference between consecutive frames (based on pixel value differences) and applies a block-based criterion to determine the amount of change in each block, creating what the author calls the transition mask. Then, a global criterion is applied on the transition mask in order to determine the amount of change between consecutive frames. By applying the same procedure all over the sequence, a monodimensional curve is created. This curve is filtered using morphological operators (a modified version of the top hat) so that transitions can be detected by imposing a threshold. While this approach has proven to be robust for the detection of scene cuts, it is less capable to detect gradual transitions because of the thresholding of the last step. Moreover, according to the author, it is sensible to abrupt changes on the images, causing that flashes or rapid motion yield false alarms.

Histogram comparison :

The main idea behind the use of the histogram is to further reduce the sensibility to the camera motion. Since histograms represent a distribution of colors, they are robust in front of rotations or changes in the viewing angle [Swa93]. Although in theory it is possible that completely different images present identical histograms, this is a very unlikely situation in real scenarios. As we have seen in the case of pixel comparisons, histogram-based approaches can also operate over the whole image (global histogram) or by regions (local histograms).

Global histogram comparison :

First techniques based on histogram comparisons [NT92], [Ton91], [ZKS93], [YYWL95] used the same approaches we have previously described for pixel-based algorithms. However, instead of using a distance based on pixel values, they define a new metric based on histogram comparisons. The histogram difference between two gray-level images at t and t + 1 can be formulated as :

$$HD(t) = \frac{1}{M} \sum m = 0^{M-1} \| h_{t+1}(m) - h_t(m) \| \tag{5.7}$$

where M is the total number of bins of the histogram. On that basis, a cut is declared when the value of the histogram difference $HD(t)$ exceeds a certain threshold. In spite of its simplicity, such approaches provided promising results. In order to improve the performance of these algorithms, Ueda et al. [UMY91] and Zhang et al. [ZKS93] take into account color information by means of color histograms. An important factor to take into account when histograms are compared is the color space used to represent colors. Both Smith [Smi97] and Gargi et al. [GOK$^+$95] analyze among others, the following color spaces : RGB, HSV, YUV. Up to now, the histogram-based methods are only suited to detect abrupt transitions (cuts) by using a single threshold. The algorithm proposed in [ZKS93] by Zhang et al. implements a technique called twin comparison. This algorithm also relies on the difference between color histograms but it uses two thresholds : one to detect cuts, and another to detect gradual transitions. Tests conducted by Borezcky [BR96] show that this approach significantly reduces the number of false positives. However, it slightly reduces the number of detected transitions, too.

Local histogram comparison :
The use of local histograms (computed over a portion of the image) intends to recover part of the spatial information lost from the pixel domain. The idea is to take profit of the robustness of the histograms in front of movement while introducing some spatial information to improve the performance. A block-based approach is presented by Nagasaka and Tanaka in [NT92]. Images are split into 16 non-overlapping blocks. A color histogram is computed for each block and compared to that of the corresponding block. The largest differences are discarded to be more tolerant to object and camera motion. Then, as we have seen in the case of pixel-based block comparisons, a two-threshold approach is used to detect the transitions. The first threshold is used to decide if a block is counted as changed between consecutive frames. The second threshold specifies the minimum number of changed blocks to declare a transition. This technique reduces the number of missed transitions compared to global histogram approaches at the expense of a higher rate of false alarms. Another technique based on local histograms, named selective HSV histograms, is presented by Lee and Ip in [LI94].

Clustering-based temporal video segmentation :
Thus far, the algorithms we have reviewed rely on the thresholding of a similarity measure between consecutive frames. In the following, we discuss some techniques based on different approaches. In [GFT98], Gnsel et al. present an

unsupervised clustering technique based on the K-Means algorithm. According to this technique, the segmentation is viewed as a 2-class clustering problem, where each transition between consecutive frames must be classified as shot change or no shot change using the K-Means algorithm. Observe that although frame transitions are labelled individually, gradual shot changes can be naturally detected : when several successive frame transitions are marked as shot change, the whole set corresponds to a gradual change. The information used to measure similarities is based on color histograms both in the YUV and RGB color spaces. Observe that this technique does not classify the different types of gradual transitions. However it overcomes one of the main drawbacks of previous techniques : the dependence of the threshold on the type of sequence. The main advantage of the clustering-based segmentation is that it allows multiple features to be simultaneously used. In [FT98], histogram and pixel-based features are jointly used.
The method of Qi et al. [QHL03] uses two binary classifiers trained with manually labelled data : one for separating cut form non-cut frames and one for separating abrupt from gradual cuts. These are either k-nearest-neighbor classifiers, naive bayesian networks or support vector machines. The gradual/abrupt classifier input is preprocessed with wavelet smoothing. Classifier input is a vector composed of whole-frame and blockwise histogram differences from 30 neighbor frames, camera motion estimations derived from MPEG2 features and blackness of the frame. The value of the work lies in the comparison of the different types of classifiers for the task, and the fact that the experimental results are on the standard TRECVID 2001 data. The optimal F-score of 0.94 for cuts and 0.70 for gradual transitions is obtained from the k-nearest-neighbor classifiers.

A novel idea was developed by Lienhart [Lie01] where dissolves (and *only* dissolves) are detected by a learning classifier (specifically a neural network). The classifier detects possible dissolves at multiple temporal scales, and merges the results using a winner-take all strategy. The interesting part is that the classifier is trained using a *dissolve synthesizer* which creates artificial dissolves from any available set of video sequences. Although the videos used for experimental verification are non-standard, performance is shown to be superior to the simple edge-based methods commonly used for dissolve detection.

Feature-based temporal video segmentation :
Another category of techniques, which are not based on the thresholding of a frame similarity curve, bases the segmentation process on the analysis of the evolution of a certain feature. In [BBB^{+}98], [BG96], [GB98], Bouthemy et al. present several algorithms based on motion analysis (global or local). According to their reasoning, in a real video sequence the camera and object motion are continuous in the shots. However, each transition represents a discontinuity in such motion. This approach declares a transition each time there is a change in the motion model. While it is quite robust in front of movement, it is only able to detect abrupt transitions. Relaying on the same principle of motion continuity, in [ZMM99], [ZMM95] Zabih et al. present a technique based on the analysis of

contour pixels. A cut is declared when most of the edges change; a progressive change of the edges should correspond to gradual transitions.

Heng and Ngan [HN01] present an advanced edge-based approach. They extract the edges from the video and refine them. Then they match them across frames using a complex technique which includes, among other things, exhaustive matching, various constraints, motion estimating through clustering, dilation of edges, and edge joint tracking. Matched edges transfer their identity to edges of the next frame, and thus it is possible to determine the backward lifetime (age) of each edge. Each frame's majority object lifetime is the near-maximum lifetime of edges in the frame. If it is below a threshold, a gradual transition is detected. If it is near zero, a cut is detected. In addition, in order to detect the type of gradual transition the frame is segmented into blocks and the forward and backward lifetimes of the edges in each block are checked to determine if they belong to the appearing or disappearing shot.

Li et al. citeLi2002 define the Joint Probability Image of two frames as a matrix where the element $[i, j]$ contains the probability that a pixel with color i in the first image has color j in the second image. They also derive the Joint Probability Projection Vector (JPPV) and the Joint Probability Projection Centroid (JPPC) as 1-dimensional and scalar statistics of the JPI. Specific types of transition (dissolve, fade, dither) are observed to have specific patterns of JPI. To detect shot changes, possible locations are first detected by using the JPPC, and they are then refined and validated by matching the JPPV with patterns of specific transitions.

Zhao and Grosky [ZG03] split each frame into blocks and compute the average hue and saturation for each. For every hue and saturation value, the angles of the Delaunay triangulation of the blocks that have this value is found and the angles of the triangulation are histogrammed. The feature vector for each frame is 10 hue values $\times$ 36 angle bins + 10 saturation values $\times$ 36 angle bins = 720 values. This is compared to the standard HS histogram method with 10 (H) $\times$ 10 (S) = 100 bins. Optionally, latent semantic indexing is applied. This is essentially a SVD dimensionality reduction of the frame vectors which is followed by normalization and frequency based weighting. A shot change is detected if the magnitude of the difference of the feature vectors (histogram, anglogram, or LSI of either) is above a threshold T_2. If it is between two thresholds T_1 and T_2 it is verified semi-automatically. The biggest drawback of the work, however, is the necessity of user interaction for the detection of gradual transitions.

The algorithm of Zhang et al. [ZCS03], called *PixSO*, begins by calculating the classic sum of pixel differences. If this above a threshold, a cut is detected. If it is between a high and a low threshold, each frame is segmented into two classes (foreground and background) by an unsupervised segmentation algorithm and the change between these classes is computed. If it is above a threshold, objects are defined as connected components of foreground. The object correspondence between frames is computed based on distance of centroids. If overlap between corresponding objects is above a threshold, a cut is detected. The authors claim 0.96 recall and 0.92 precision.

Model-driven temporal video segmentation :
All the techniques we have surveyed so far extract some kind of data from the video sequence and then analyze it in order to detect some pattern which corresponds to shot transitions. These techniques are known as data-driven. Alternatively, model-driven approaches tackle the problem of video segmentation from a different point of view. They characterize the different types of transition by means of mathematical models. Hampapur et al. [HJW95] create models for different types of transition. Their method takes into account the editing process of video sequences to characterize the transitions. For instance, dissolves are modeled as the weighted average of frames from the shots before and after the transition. The segmentation algorithm tries to locate those frames of the sequence that fit the model to declare a transition. As other previously discussed techniques, it is quite sensitive to motion. In [AJ94] Aigrain and Joly introduce what they call differential model of motion picture. Instead of modeling the value of image pixels across transitions, they define a model to characterize the difference between them. The technique presented by Yu et al. [YBH97] is a mixture between data- and model-driven approaches. At the first step, an histogram-based algorithm is applied to detect cuts. Then, fades and dissolves are detected by inspection of the frames in-between. Their model is based on the evolution of the number of edge pixels across gradual transitions. A different technique is discussed in [BW98], where Boreczky and Wilcox use hidden Markov models (HMM) to detect shot transitions. They generate a three component vector for each pair of consecutive frames, based on color (gray-level histogram difference), motion (estimate of object motion) and audio information. They use the sequence of vectors to feed a six state HMM (shot, cut, fade, dissolve, pan and zoom). The main advantage of this algorithm is that thanks to the training phase (inherent to any HMM) there is no need for thresholds.

Shot detection is performed by Janvier et al. [JBMMP03] in three independent steps. First the appearance of a cut is detected by estimating the probability that a boundary exists between two frames. The various probability functions are derived experimentally and their parameters are computed from training data and the derived “shots” are further segmented into pieces that exhibit greater continuity using a linear model and dynamic programming with minimum message length criterion to fit this model. In order to give acceptable results in the task of shot detection, these segments are then merged using a 2-class K-means. This operates on color histograms and mutual information [CNP02] of frames within a window around the segment change and classifies them into “shot change” and “no shot change” categories.

A thorough analysis of the shot detection problem is presented by Hanjalic

[Han02], and a probabilistically based algorithm is described. A discontinuity value between frames is defined as the average difference between average YUV components of matching blocks within these frames. For detecting abrupt transitions adjacent frames are compared, while for gradual ones frames that are apart by the minimum shot length are compared. The a priori likelihood functions of the discontinuity values are obtained by experiments on manually labelled data. Their choice is largely heuristic. The a priori probability of shot detection conditional on time from last shot detection is modelled as a Poisson function, which was observed to offer good results. The shot detection probability is refined with a heuristic derived from the pattern of the discontinuity values (for cuts and fades) and in-frame variances (for dissolves) in a window around the current frame. Effectively a different detector is implemented for each transition type.

Sánchez and Binefa [SB03] model each shot as a Coupled Markov Chain (CMC), which encodes the probabilities of temporal change of a set of features. Here, hue and motion of a 16x16 image block are used. For each frame, two CMCs are computed, one from the shot up to and including that frame, and one from that frame and the next one. Then the Kullback-Leibler divergence between the two distributions is compared with a threshold adaptively (but heuristically) computed from the mean and standard deviation of the changes encoded in the first CMC. The fact that the dynamics of the disappearing shot are inherently taken into consideration for detecting shot changes is what gives this method its strength and elegance.

Subspace-based video segmentation :
A feature-independent method is presented by Liu and Chen [LC02]. They use a modified PCA algorithm on each shot to extract its principal eigenspace. For this, two novel algorithms are presented to facilitate the gradual computation of the eigenspace, and also to place greater weight on the most recent frames (by heuristic weights). A cut is detected when the difference between a frame and its projection into the eigenspace is above a static threshold.

In order to make the feature space more discriminant, Cernekova et al. [CKP03] propose performing SVD is on the color histograms of each frame to produce reduced dimension feature vectors. Shot change is detected by comparing the angle between the feature vector of each frame and the average of the feature vectors of the current shot. Shots whose feature vectors exhibit high sparseness are considered to depict a gradual transition between two shots. The main problem with this approach is the static threshold which is applied on the angle between vectors to detect shot change, which may be problematic in the case of large variation in intra-frame content and small variations in inter-frame content.

5.5.2 Segmentation in the MPEG compressed domain

Due to the increasing amount of material stored in MPEG format, performing temporal segmentation in the compressed domain presents several advantages.

By avoiding to decode the analyzed sequence, the processing speed is increased and the storage requirements are greatly reduced. Moreover, operations should be faster because the lower gross data rate of compressed video. Another important advantage is that MPEG bit-streams contain several features that are not present in the raw sequence, such as motion vectors or average gray intensities. In the following sections, we review some of these techniques grouped by the type of information they use: DCT coefficients, DC terms or macro-block (MB) coding type.

Temporal segmentation based on DCT coefficients :
The main goal of the first approaches working on the compressed domain was speed. In [AHC93], Arman et al. introduce a cut detection algorithm based on the comparison of the DCT coefficients of corresponding blocks from consecutive I frames by means of a normalized inner product. In order to increase the processing speed, only a subset of coefficients from a subset of blocks is taken into account. The technique presented by Zhang et al. in [ZLGS94] is an adaptation of the well-known techniques that apply on the pixel domain. Here, on compressed video, the pixel-wise comparison is used to compute the difference between DCT coefficients of corresponding blocks. Both of the above techniques only take into account I frames because they are the only ones for which the DCT coefficients are directly available. On the one hand, this allows a great processing speed. On the other hand, however, it greatly reduces the performance of the algorithms. The loss of temporal resolution makes them even more sensible to object and camera motion, increasing the number of false positives compared to the techniques working on the uncompressed domain. Besides, gradual transitions cannot be detected.

Temporal segmentation based on DC terms :
A large set of techniques working on the compressed domain are based on the processing of DC terms. These values correspond to the mean luminance of each block of the coded image. The full set of DC terms constitutes a low resolution version of the original image, called DC-image. These techniques use basically the same approaches based on pixel [YL95] and histogram [PS97] differences we have seen in the previous sections, applied over the DC-images. The main problem here is how to efficiently extract these images from the input bit-stream. I frames are easy to handle because DCT coefficients are directly available. However, DC images for P and B frames are very costly to compute since motion vectors must be taken into account. In [SD95], Shen and Delp propose a fast reconstruction of color DC-images using an approximated approach. In [TD98], Taskiran and Delp present an extension of the previous technique based on a generalized sequence trace, which is the evolution of the difference between the feature vector of consecutive frames. In order to reduce the complexity of the DC-image creation, Sethi and Patel [SP95] present an algorithm where only I frames are processed. Ardizzone et al. [AGCM96] also propose an approach relying only on I frames.

An advanced statistical approach is presented by Lelescu and Schonfeld [LS03]. This additionally benefits from operating on MPEG-compressed videos. They extract luminance and chrominance for each block in every I and P frame, and then perform PCA on the resulting vectors. It should be noted that the eigenvectors are computed based only on the M first frames of each shot. The resulting vectors are modelled by a gaussian random distribution. Then the mean and covariance matrix of the distribution are computed, also in the M first frames of each shot. The originality of the approach is that a change statistic is computed for each new frame by maximum likelihood methodology (the Generalized Likelihood Ratio algorithm) and if it exceeds an experimentally determined threshold a new shot is started. The estimation can be either additively or non-additively based. The algorithm is tested on a number of videos originating from news, music clips, sports and other types where gradual transitions are common. The best results of 0.92 recall and 0.72 precision were recorded for the additively-based algorithm.

Temporal segmentation using MB coding mode and other features : Another feature from the compressed domain used in several segmentation algorithms is the coding mode of the macro-blocks (MB). For I images all the MB are coded in intra mode, while for P and B images they may be coded either intra- or inter-mode. For those portions of the sequence with small changes, most of the blocks are coded in inter-mode because it is possible to find similar blocks in the preceding frames. However, across a transition there is a lot of change and hence the number of MB coded in intra-mode should be much higher. Notice that this only applies for P and B frames. Different techniques make use of this information, usually combined with other features, as for instance [MJC95], [ZLS95], [KC98] or [FLH96]. The approaches described in Little et al. and Deardoff et al. [KC98] use as input data a sequence encoded using the M-JPEG format and searches for differences in the size of the encoded image to detect cuts. More precisely, a cut is detected when there is a sudden change on the size of the coded image. One of the main drawbacks of compressed domain techniques is the dependence of the performance on the input bit-stream itself. First, it implies a lack of generality of the algorithms because the information contained in the bit-stream depends on the coding method. Second, even when the same standard is used, sequences encoded with different encoders may lead to significant differences in performance [PS97]. Additionally, the performance achieved using compressed domain techniques is lower than that achieved using uncompressed domain approaches, especially for gradual transitions.

Bibliography

[AGCM96] E. Ardizzone, G. Gioiello, M. L. Cascia, and D. Molinell. A real-time neural approach to scene cut detection. In I. I. Sethi and e. R.C. Jain, editors, *S&T/SPIE Conference on Storage and Retrieval for Image and Video Databases IV, San Jose, CA, USA*. 1996.

[AHC93] F. Arman, A. Hsu, and M.-Y. Chiu. Image processing on compressed data for large video databases. In *First ACM Internation Conference on Multimedia*, pages 267–272. 1993.

[AJ94] P. Aigrain and P. Joly. The automatic real-time analysis of film editing and transition effects and its applications. In *Computers & Graphics*, volume 18, pages 93–103. January-February 1994.

[AK95] T. Aach and A. Kaup. Bayesian algorithms for change detection in image sequences using Markov random fields. *Signal Processing : Image Communication*, 7:147–160, 1995.

[Ala97] A. M. Alattar. Detecting fade regions in uncompressed video sequences. In *Proc. 1997, IEEE Int. Conf. Acoustics, Speech and Signal Processing*, pages 3025–3028. 1997.

[AWS00] L. Alvarez, J. Weickert, and J. Sánchez. Reliable estimation of dense optical flow fields with large displacements. *International Journal of Computer Vision*, 39(1):41–56, 2000.

[BA96] M. Black and P. Anandan. The robust estimation of multiple motions : Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63:75–104, 1996.

[BBB$^+$98] S. Benayoun, H. Bernard, P. Bertolino, P. Bouthemy, M. Gelgon, R. Mohr, C. Schmid, and F. Spindler. Structuration de video pour des interfaces de consultation avancees. In *4mes Journes d'Etudes et d'Echanges Compression et Reprsentation des Signaux Audio-Visuels, CORESA'98*. June 1998.

[BFY00] M. J. Black, D. J. Fleet, and Y. Yacoob. Blackfleetyacoob. *Computer Vision and Image Understanding*, 78:8–31, 2000.

[BG96] P. Bouthemy and F. Ganansia. Video partitioning and camera motion characterization for content-based video indexing. In *the IEEE International Conference on Image Processing, ICIP'96*, volume 1, pages 905–908. Lausanne, Switzerland, september 1996.

[BGW91] J. Bign, G. Granlund, and J. Wiklund. Multidimensional orientation estimation with applications to texture analysis and optical flow. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 13(8):775–790, August 1991.

[BJ96] M. Black and A. Jepson. Estimating optical flow in segmented images using variable-order parametric models with local deformations. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(10):972–986, October 1996.

[BR96] J. Boreczky and L. Rowe. Comparison of video shot boundary detection techniques. In I. I. Sethi and e. R.C. Jain, editors, *S&T/SPIE Conference on Storage and Retrieval for Image and Video Databases IV, San Jose, CA, USA*, pages 170–179. 1996.

[BW98] J. Boreczky and L. Wilcox. A hidden markov model framework for video segmentation using audio and image features. In *In IEEE Proceedings of the the International Conference on Acoustics, Speech and Signal Processing, ICASSP'98*, volume 6, pages 3741–3744. Seattle, WA, USA, May 1998.

[CB71] G. Csurka and P. Bouthemy. Direct identification of moving objects and background from 2D motion models. In *7th International Conference on Computer Vision*. Kerkyra, Greece, 566–571.

[CH99] I. Cohen and I. Herlin. Non uniform multiresolution method for optical flow and phase portrait models : Environmental application. In *IJCV*, volume 33, pages 29–49. September 1999.

[CKP03] Z. Cernekova, C. Kotropoulos, and I. Pitas. Video shot segmentation using singular value decomposition. In *International Conference on Multimedia and Expo*. Jul 2003.

[CKS02] D. Cremers, T. Kohlberger, and C. Schnrr. Nonlinear shape statistics in mumford-shah based segmentation. In *7th European Conference on Computer Vision, ECCV'2002*, volume LNCS 2351. Springer Verlag, Copenhagen, 2002.

[CM99] I. Cohen and G. Medioni. Detecting and tracking moving objects for video surveillance. In *IEEE Conf. Computer Vision and Pattern Recognition*. Fort Collins CO, June 1999.

[CMP02] T. Corpetti, E. Mmin, and P. Prez. Dense estimation of fluid flows. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(3):365–380, 2002.

[CNP02] Z. Cernekova, C. Nikou, and I. Pitas. Shot detection in video sequences using entropy-based metrics. In *International Conference on Image Processing*, pages 0 – 0. Oct 2002.

[CNS02] P. Campisi, A. Neri, and L. Sorgi. Automatic dissolve and fade detection for video sequences. In *International Conference on Digital Signal Processing*. Jul 2002.

[CS03] D. Cremers and S. Soatto. Variational space-time motion segmentation. In *Proc. 9th IEEE Int. Conf. on Computer Vision, ICCV'2003*. Nice, October 2003.

[DAE95] A. Dailianas, R. B. Allen, and P. England. Comparison of automatic video segmentation algorithms. In *SPIE Proceedings, Integration Issues in Large Commercial Media Delivery Systems*, pages 2–16. Philadelphia, PA, USA, October 1995.

[DKA04] C. Demonceaux and D. Kachi-Akkouche. Motion detection using wavelet analysis and hierarchical Markov models. In *First International Workshop on Spatial Coherence for Visual Motion Analysis*. Prague, May 2004.

[DRP02] A. Divakaran, R. Radhakrishnan, and K. Peker. Motion activity-based extraction of key-frame from video shots. In *ICIP'02*. Rochester, September 2002.

[ETM03] A. Ekin, A. Tekalp, and R. Mehrotra. Automatic soccer video analysis and summarization. *IEEE Int. Trans. on Image Processing*, 12(7):796–807, 2003.

[FB03] R. Fablet and P. Bouthemy. Motion recognition using non parametric image motion models estimated from temporal and multiscale cooccurrence statistics. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(12):1619–1624, December 2003.

[FJ90] D. Fleet and A. Jepson. Computation of component image velocity from local phase information. *International Journal of Computer Vision*, 5(1):77–104, 1990.

[FLH96] J. Feng, K.-T. Lo, and M. H. Scene change detection algorithm for mpeg video sequence. In *the IEEE International Conference on Image Processing, ICIP'96*. Lausanne, Switzerland, september 1996.

[FT98] A. Ferman and A. Tekalp. Efficient filtering and clustering for temporal video segmentation and visual summarization. *Journal on Visual Communications and Image Representation*, 9(4), 1998.

[GB98] M. Gelgon and P. Bouthemy. Determining a structured spatio-temporal representation of video content for efficient visualization and indexing. In *5th European Conference on Computer Vision, ECCV'98*, volume 1, pages 595–609. Freiburg, Germany, June 1998.

[GB00] M. Gelgon and P. Bouthemy. A region-level motion-based graph representation and labeling for tracking a spatial image partition. *Pattern Recognition*, 33(4):725–740, April 2000.

[GFT98] B. Gunsel, A. Ferman, and A. Tekalp. Temporal video segmentation algorithm using unsupervised clustering and semantic object tracking. *Journal on Electronic Imaging*, 7(3):592–604, 1998.

[GOK$^+$95] U. Gargi, S. Oswald, D. Kosiba, S. Devadiga, and R. Kasturi. Evaluation of video sequence indexing and hierarchical video indexing. In I. I. Sethi and e. R.C. Jain, editors, *S&T/SPIE Conference on Storage and Retrieval for Image and Video Databases, San Jose, CA, USA*, pages 1522–1530. 1995.

[Han02] A. Hanjalic. Shot-boundary detection: Unraveled and resolved? *IEEE Transactions on Circuits and Systems for Video Technology*, 12(3), Feb 2002.

[HB93] F. Heitz and P. Bouthemy. Multimodal estimation of discontinuous optical flow using markov random fields. *IEEE Trans. on PAMI*, 15(12):1217–1232, 1993.

[HF01] H. Haussecker and D. Fleet. Estimating optical flow with physical models of brightness variation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(6):661–673, 2001.

[HJT94] A. Hampapur, R. Jain, and T.E.Weymouth. Digital video segmentation. In *In Proceedings of the ACM Multimedia Conference and Exposition (verificar)*, pages 357–364. San Francisco, CA, USA, October 1994.

[HJW95] A. Hampapur, R. Jain, and T. Weymouth. Production model based digital video segmentation. *Journal of Multimedia Tools and Applications*, 1(1):9–46, March 1995.

[HL01] C.-L. Huang and B.-Y. Liao. A robust scene-change detection method for video segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(13):1281 – 1288, Dec 2001.

[HN01] W. Heng and K. Ngan. An object-based shot boundary detection using edge tracing and tracking. *Journal of Visual Communication and Image Representation*, 12(4):217 – 239, Sep 2001.

[HNR84] Y. Z. Hsu, H.-H. Nagel, and G. Rekers. New likelihood test methods for change detection in image sequences. *Computer Vision, Graphics and Image Processing*, 26:73–106, 1984.

[HPB94] F. Heitz, P. Prez, and P. Bouthemy. Multiscale minimization of global energy functions in some visual recovery problems. In *CVGIP : Image Understanding*, volume 59, pages 125–134. January 1994.

[HS81] B. Horn and B. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[IA99] M. Irani and P. Anandan. A unified approach to moving object detection in 2D and 3D scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(6):577–589, June 1999.

[IRP94] M. Irani, B. Rousso, and S. Peleg. Computing occluding and transparent motion. *International Journal of Computer Vision*, 12(1):5–16, 1994.

[JBMMP03] B. Janvier, E. Bruno, S. Marchand-Maillet, and T. Pun. Information-theoretic framework for the joint temporal partitioning and representation of video data. In *International Workshop on Content-Based Multimedia Indexing*. Sep 2003.

[JKM04] V. Jain, B. B. Kimia, and J. L. Mundy. Segregation of moving objects using elastic matching. In *Workshop on Spatial Coherence for Visual Motion Analysis*. Prague, May 2004.

[KC98] I. Koprinska and S. Carrato. Detecting and classifying video shot booundaries in mpeg compressed sequences. In *In Proceedings of IX European Signal Processing Conference, EUSIPCO' 98*, pages 1729–1732. Rodhes, Grece, 1998.

[KC01] I. Koprinska and S. Carrato. Temporal video segmentation : A survey. *Signal Processing : Image communications*, 16(5):477–500, January 2001.

[KDA99] P. Kornprobst, R. Deriche, and G. Aubert. Image sequence analysis via partial differential equations. *Journal of Mathematical Imaging and Vision*, 11(1):5–26, 1999.

[KJ91] R. Kasturi and R. Jain. *Dynamic vision*, chapter Computer Vision : Principles, pages 469–480. IEEE Computer Society Press, Whashington DC, USA, 1991.

[KK92] T. Kikukawa and S. Kawafuchi. Development of an automatic summary editing system for the audio-visual resources. *Transactions on Electronic Information. J75-A*, pages 204–212, 1992.

[Kon00] J. Konrad. *Handbook of Image and Video Processing*. Academic Press, 2000. Motion detection and estimation.

[LC02] X. Liu and T. Chen. Shot boundary detection using temporal statistics modeling. In *International Conference on Acoustics, Speech and Signal Processing*. May 2002.

[LI94] C.-M. Lee and D. M.-C. Ip. A robust approach for camera break detection in color video sequences. In *IAPR Workshop Machine Vision Applications*, pages 502–505. Kawasaki, Japan, 1994.

[Lie99] R. Lienhart. Comparison of automatic shot boundary detection algorithms. In *Proc. of SPIE Storage and Retrieval for Image and Video Databases VII, San Jose, CA, U.S.A.*, volume 3656, pages 290–301. January 1999.

[Lie01] R. Lienhart. Reliable dissolve detection. In *SPIE Storage and Retrieval for Media Databases*. Jan 2001.

[LM03] J. L. Lisani and J.-M. Morel. Detection of major changes in satellite images. In *IEEE International Conference on Image Processing*. Barcelona, September 2003.

[LS03] D. Lelescu and D. Schonfeld. Statistical sequential analysis for real-time video scene change detection on compressed multimedia bitstream. *IEEE Transactions on Multimedia*, 5(2), Mar 2003.

[MB96] A. Mitiche and P. Bouthemy. Computation and analysis of image motion : A synopsis of current problems and methods. *International Journal of Computer Vision*, 19:29–55, 1996.

[MJC95] J. Meng, Y. Juan, and S.-F. Chang. Scene change detection in a mpeg compressed video sequence. In I. I. Sethi and e. R.C. Jain, editors, *S&T/SPIE Conference on Storage and Retrieval for Image and Video Databases, San Jose, CA, USA*, volume 2417, pages 14–25. 1995.

[MK03] A.-R. Mansouri and J. Konrad. Multiple motion segmentation with level sets. *IEEE Transactions on Image Processing*, 12(2):201–220, 2003.

[MP98] E. Mmin and P. Prez. Optical flow estimation and object-based segmentation with robust techniques. *IEEE Trans. on Image Processing*, 7(5):703–719, May 1998.

[MP02] E. Mmin and P. Prez. Hierarchical estimation and segmentation of dense motion fields. *Int. Journal of Computer Vision*, 46(2):129–155, February 2002.

[MSB03] J.-L. Marroquin, E. Santana, and S. Botello. Hidden markov measure field models for image segmentation. *IEEE Trans. on PAMI*, 25(11):1380–1387, November 2003.

[Nag87] H.-H. Nagel. On the estimation of optic flow : Relations between different approaches and some new results. *Artificial Intelligence*, 33:299–324, 1987.

[Nel91] R. C. Nelson. Qualitative detection of motion by a moving observer. *International Journal of Computer Vision*, 7(1):33–46, 1991.

[NG98] H.-H. Nagel and A. Gehrke. Spatiotemporally adaptive estimation and segmentation of of-fields. In *5th Eur. Conf. on Comp. Vis., ECCV'98*, volume LNCS 1407. Springer, Freiburg, 1998.

[NM03] M. Nicolescu and G. Medioni. Layered 4d representation and voting for grouping from motion. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(4):492–501, April 2003.

[NPZ02] C.-W. Ngo, T.-C. Pong, and H.-J. Zhang. On clustering and retrieval of video shots through temporal slices analysis. *IEEE Trans. Multimedia*, 4(4):446–458, December 2002.

[NT92] A. Nagasaka and Y. Tanaka. *Automatic video indexing and full-video search for object appeareances*, chapter Visual Database Systems II, pages 113–127. Elsevier, 1992.

[OB97] J. Odobez and P. Bouthemy. *Video Data Compression for Multimedia Computing*, chapter Chapter 8 : Separation of moving regions from background in an image sequence acquired with a mobile camera, pages 283–311. Kluwer Academic, 1997.

[OB98] J.-M. Odobez and P. Bouthemy. Direct incremental model-based image motion segmentation for video analysis. *Signal Processing*, 6(2):143–155, 1998.

[PBA00] R. Pless, T. Brodksy, and Y. Aloimonos. Detecting independent motion : The statistics of temporal continuity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):768–773, August 2000.

[PBY04] G. Piriou, P. Bouthemy, and J.-F. Yao. Extraction of semantic dynamic content from videos with probabilistic motion models. In *Proc. European Conf. on Computer Vision, ECCV'04*. Prague, May 2004.

[PD00] N. Paragios and R. Deriche. Geodesic active contour and level sets for the detection and tracking of moving objects. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(3):266–280, March 2000.

[PS97] N. Patel and I. Sethi. Video shot detection and characterization for video databases. *Pattern Recognition*, 30:583–592, 1997.

[QHL03] Y. Qi, A. Hauptmann, and T. Liu. Supervised classification for video shot segmentation. In *International Conference on Multimedia and Expo*. Jul 2003.

[RA00] Y. Rui and P. Anandan. Segmenting visual actions based on spatio-temporal motion patterns. In *CVPR'2000*. 2000.

[Ros02] P. L. Rosin. Thresholding for change detection. In *ICCV*, pages 79–95. May 2002.

[SA96] H. S. Sawhney and S. Ayer. Compact representations of videos through dominant and multiple motion estimation. *IEEE Trans. on PAMI*, 18(8):814–830, August 1996.

[SB03] J. Sánchez and X. Binefa. Shot segmentation using a coupled Markov chains representation of video contents. In *Iberian Conference on Pattern Recognition and Image Analysis*. Jun 2003.

[SD95] K. Shen and E. Delp. A fast algorithm for video parsing using mpeg compressed sequences. In *Proceedings of the IEEE International Conference on Image Processing, ICIP'95*, pages 252–255. Washington DC, USA, October 1995.

[SDC04] P. Smith, T. Drummond, and R. Cipolla. Layered motion segmentation and depth ordering by tracking edges. *IEEE Transactions on Pattern Analysis and Machine Intellingence*, 26(4):479–494, April 2004.

[Sha95] B. Shahraray. Scene change detection and content-based sampling of video sequences. In R. S. A. Rodriguez and e. E. Delp, editors, *S&T/SPIE Conference on on Digital Video Compression : Algorithms and Technologies, San Jose, CA, USA*, volume 2419, pages 2–13. 1995.

[SK99] C. Stiller and J. Konrad. Estimating motion in image sequences (a tutorial on modeling and computation of 2d motion). *IEEE Signal Processing Magazine*, 16:70–91, July 1999.

[Smi97] J. Smith. *Integrated Spatial and Feature Image Systems : Retrieval, Analysis and Compression*. Phd thesis, Columbia University, New York City, NY, USA, 1997.

[SMK02] S. Sarkar, D. Majchrzak, and K. Korimilli. Perceptual organization based computational model for robust segmentation of moving objects. *Computer Vision and Image Understanding*, 86:141–170, 2002.

[SP95] I. Sethi and N. Patel. A statistical approach to scene change detection. In I. I. Sethi and e. R.C. Jain, editors, *S&T/SPIE Conference on Storage and Retrieval for Image and Video Databases, San Jose, CA, USA*, volume 2420, pages 2–11. 1995.

[Swa93] M. Swain. Interactive indexing into image databases. In I. I. Sethi and e. R.C. Jain, editors, *S&T/SPIE Conference on Storage and Retrieval for Image and Video Databases, San Jose, CA, USA*, pages 173–187. 1993.

[TD98] C. Taskiran and E. Delp. Video scene change detecion using the generalized sequence trace. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98*, pages 2961–2964. Seattle, WA, USA, 1998.

[TDV00] B. T. Truong, C. Dorai, and S. Venkatesh. New enhancements to cut, fade and dissolve detection processes in video segmentation. In *ACM Multimedia 2000*, pages 219–227. November 2000.

[TLS93] W. B. Thompson, P. Lechleider, and E. R. Stuck. Detecting moving objects using the rigidity constraint. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15(2):162–166, 1993.

[Ton91] Y. Tonomura. Video handling based on structures information for hypermedia systems. In *Proceedings of the ACM International Conference on Multimedia Information Systems*, pages 333–344. 1991.

[TP90] W. B. Thompson and T.-C. Pong. Detecting moving objects. *International Journal of Computer Vision*, 4:39–57, 1990.

[UMY91] H. Ueda, T. Miyatake, and S. Yoshizawa. Impact : An alteractive natural-motion-picture dedicated multimedia authoring system. In *Proceedings of the ACM Conference on Human Interfaces (veficicar)*, pages 343–350. New Orleans, Louisiana, USA, April-May 1991.

[VCB04] T. Veit, F. Cao, and P. Bouthemy. Probabilistic parameter-free motion detection. In *Proc. Conf. Computer Vision and Pattern Recognition, CVPR'04*. Washington DC, June 2004.

[VL00] N. Vasconcelos and A. Lippman. Statistical models of video structure for content analysis and characterization. *IEEE Trans. on IP*, 9(1):3–19, 2000.

[WBPB03] J. Weickert, A. Bruhn, N. Papenberg, and T. Brox. Variational optic flow computation : From continuous models to algorithms. In L. Alvarez, editor, *International Workshop on Computer Vision and Image Analysis (IWCVIA'03)*. Las Palmas de Gran Canaria, December 2003.

[Wix00] L. Wixson. Detecting salient motion by accumulating directionally-consistent flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):774–780, August 2000.

[XL98] W. Xiong and J.-M. Lee. Efficient scene change detection and camera motion annotation for video classification. *Computer Vision and Image Understanding*, 71(2):166–181, 1998.

[XLI95] W. Xiong, J.-M. Lee, and D. M.-C. Ip. Net comparison : a fast and effective method for classifying image sequences. In I. I. Sethi and e. R.C. Jain, editors, *S&T/SPIE Conference on Storage and Retrieval for Image and Video Databases, San Jose, CA, USA*, volume 2420, pages 318–328. 1995.

[YB98] Y. Yacoob and J. Black. Parametrized modeling and recognition of activities. In *Sixth IEEE Int. Conf. on Computer Vision*. 1998.

[YBH97] H. Yu, G. Bozdagi, and S. Harrington. Featured-based hierarchical video segmentation. In *Proceedings of the IEEE International Conference on Image Processing, ICIP'97*, pages 497–501. Santa Barbara, CA, USA, October 1997.

[YL95] B. Yeo and B. Liu. Rapid scene analysis on compressed video. *IEEE Transactions on Circuits and Systems and Video Technology*, 5(6):533–544, 1995.

[YYWL95] M. Yeung, B. Yeo, W. Wolf, and B. Liu. Video browsing using clustering and scene transitions on compressed sequencess. In *S&T/SPIE Conference on Multimedia Computing and Networking, San Jose, CA, USA*, volume 2417. 1995.

[ZCS03] C. Zhang, S.-C. Chen, and M.-L. Shyu. Pixso: A system for video shot detection. In *Pacific-Rim Conference On Multimedia*. Dec 2003.

[ZG03] R. Zhao and W. I. Grosky. Video shot detection using color anglogram and latent semantic indexing: From contents to semantics. In *Handbook of Video Databases: Design and Applications*. CRC Press, Jan 2003.

[ZKS93] H. Zhang, A. KanKanhalli, and S. Smoliar. Automatic partitioning of full motion video. *Multimedia Systems*, 1(1):10–28, 1993.

[ZLGS94] H. Zhang, C. Low, Y. Gong, and S. Smoliar. Video parsing using compressed data. In *Proceedings of SPIE Conference on Image and Video Processing II*, pages 142–149. 1994.

[ZLS95] H. Zhang, C. Low, and S. Smoliar. Video parsing and browsing using compressed data. *Multimedia Tools and Applications*, 1:89–111, 1995.

[ZMI01] L. Zelnik-Manor and M. Irani. Event-based video analysis. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*. Kauai, Hawaii, December 2001.

[ZMM95] R. Zabih, J. Miller, and K. Mai. A feature-based algorithm for detecting and classifying scene breaks. In *ACM International Conference on Multimedia*. San Francisco, CA, USA, November 1995.

[ZMM99] R. Zabih, J. Miler, and K. Mai. A feature-based algorithm for detecting and classifying production effects. *Multimedia Systems*, 7:119–128, 1999.

Part II

State of the art in some areas of speech and audio processing

Chapter 6

Content-based description of Audio

6.1 Introduction

Scientific and technological progress nowadays facilitate the continuous accumulation of data. This fact yields unprecedented opportunities for the transition of our society to the prospect of a knowledge-driven society. Semantic manipulation of the deluge of information that is being accrued is a basic prerequisite towards this direction. Diversity and multimodality of this information pose additional challenges.

Within this framework, tackling automatic content analysis of audio data is of major importance. Audio cues, either alone or integrated with information extracted from other modalities, may contribute significantly to the overall semantic interpretation of data.

Event detection in audio streams is an aspect of the aforementioned analysis. The concept of "*event*" corresponds to a noteworthy happening and is application dependent. For example, applause, chanting, laughter or alterations in the speech rate of the sport-caster may be regarded as events in a sports-video. Speaker changes, changes between various speech quality levels, between speech and silence, or speech and music are common events, e.g. in the case of broadcast news.

Event detection in audio streams aims at delineating audio as a sequence of homogeneous parts each one identified as a member of a predefined audio class. *Determination of such audio classes* (e.g. speech, music, silence, laughter, noisy speech etc.) is the first step in the design of an event detection algorithm. Obviously, these classes are inferred from the specificity of the application. The *selection and extraction of appropriate audio-features* ensues and is probably the most significant phase. It is understood that the correct choice of these features favors the successful completion of the following phases. Proper exploitation of the extracted attributes leads to the *segmentation* of the audio stream and *classification* of the resulting parts. Classificiation can be done by discirmination of the parts, e.g. into speech or music and subsequently into specific sub-classes of a category, e.g. various forms of speech, musical instruments or musical genres. Furthermore, the extracted attributes allow automatic organization of audio based on its content.

In this report we will attempt to discuss the state of the art in all these aspect, namely selection of audio features, segmentation, classification and organization. Research in these areas has been quite intense and indicates that the problem is an arduous one. Relevant overview articles have been written by Foote [18] and Downie [9]. The remainder of this Chapter is structured as follows: Section 6.2 describes the range of features currently employed in audio analysis, focusing on spectral, temporal, and specific high-level features. Section 6.3 describes current work in a task that is fundamental to many subsequent tasks in audio analysis, namely audio segmentation. this is followed by a description of the current state in differents classification tasks, briefly listing the various classification approaches, followed by the two core audio classification tasks, namelymusical instruments and genre classification, in Section 6.4. This will be rounded off by a short description of issues pertaining to user interfaces, i.e. how the approaches covered in this report may be put to use to assis users in interacting with audio collections, in Section 6.5, before summing up and pointing at core challenges in Section 6.6

6.2 Audio Feature Selection

During the short life of research in the area of general audio segmentation various types of features have been proposed. Various audio descriptive features are used for classification purposes in sound and speech recognition applications and many of them are also widely used in music analysis, although there exist also specially designed features that try to account for specific characteristics of each audio type, or that can discriminate among them. Another case of specially designed features are the ones that discriminate between two classes of audio data, e.g. common discrimination tasks include speech vs. music or speech vs. silence.

6.2.1 Spectral features

Spectral descriptors are most frequently used in general audio recognition applications, but many extracted features are also used in musical instrument and genre classification applications [46] [47] [7] [31] [11] [51]. The most frequently used are:

- *Mel frequency cepstrum coefficients*: They are a widely used variation of the cepstral coefficients that follow the Mel-frequency scale as proposed by psychoacoustics and are the standarad features used in Automatic Speech Recognition (ASR). They have been adopted in many approaches, most of which afterwards use model-based (e.g. posterior probabilities) classification methods [54, 32, 6, 14, 50] and/or are designated to be part of larger ASR systems.

- *Zero-Crossing Rate (ZCR)*: related to the mean frequency of a segment. It provides a noise measure for the given signal [57], [37], [22], [36], [30]. A variation is *high zero-crossing rate ratio* defined as the ratio of the number of frames whose ZCR is above 1.5 times the average ZCR.

- *short-time energy (STE)*: provides a convenient representation of the amplitude. A variation is *low short-time energy ratio* which is proportional to the ratio of the number of frames whose STE is less than 0.5 time of average STE.

- *Silence crossing rate* is the number of times that the energy falls below some silence level criterion.

- *Spectral Flux (Delta Spectrum)*: It is defined as the average variation value of the spectrum between two adjacent frames.

- *Spectral Rolloff*: Measures the spectral shape and is defined as the frequency below which a percentage of the magnitude distribution is concentrated.

- *Spectral centroid*: Sometimes it is referred to as "Brightness" of the signal, it provides a measure of spectral shape where higher values correspond to brighter textures.

- *Harmonicity*: It measures the deviation of the signal spectrum from a harmonic spectrum.

Even the root mean square amplitude along with the ZCR has been used as feature in [37] resulting in quite high classification rates.

Other spectral features used by Eronen [11] are the mean and standard deviation of the spectral centroid, the fundamental frequency along with its mean and standard deviation, the normalized and maximum of the normalized spectral centroid. In addition, the MPEG-7 Audio group has proposed a variety of spectral descriptors, consisting of the following groups: *basic spectral*, *signal parameters*, *timbral spectral* and *spectral basis representations*. MPEG-7 Harmonicity and Fundamental Frequency features were used by Slezak [51] in musical instrument classification experiments.

Among the features that have been proposed recently especially for the specific problem of audio segmentation and classification, the following should be mentioned: *average probability dynamism*, *mean per frame entropy* [52], [2], [6], *background-label energy ratio*, *phoneme distribution match* [52], *Teager energy* and related *modulation features* [13].

In more detail, the use of *average probability dynamism* as an audio-feature is based on the observation that the posterior phoneme probability estimates (according to an acoustic model) for speech segments change frequently whereas in non-speech segments they change much less frequently and more gradually. *Mean per-frame entropy* exhibits the suitability of an acoustic model for an audio segment. *Background-label energy ratio* compares the expected non-speech energy of a segment to the expected speech energy. *Phoneme distribution match* shows how close the phoneme probability distribution for a whole segment is to the corresponding distribution estimated from a training set of known speech segments. Finally, *Teager energy* and related *modulation features*, namely instantaneous amplitude and frequency, are based on the observation that speech resonance signals may be modelled as amplitude and frequency modulated signals.

6.2.2 Temporal features

Temporal descriptors are widely used in instrument recognition applications, mainly because the timbral characteristics that differentiate musical instruments are not related with their spectral features. In their experiments featuring musical instrument classification, Eronen [11] and Slezak [51] have used several temporal features, which are outlined below:

- *Length*: Signal length.
- *Rise time (Attack)*: Relative length of the attack (till reaching 75% of maximal amplitude).
- Steady time: Relative length after the end of the attack (till the final fall under 75% of maximal amplitude)
- *Decay time*: Decay time (the rest of the signal).
- *Maximum*: Moment of reaching maximal amplitude.
- *Crest factor*: Defined as max/rms of signal amplitude.

It should be noted that the audio description framework as defined in the MPEG-7 Audio protocol [24] contains two groups of temporal descriptors: *basic* and *timbral temporal*. The former contains two descriptors: instantaneous waveform and power values, while the latter contains descriptors for log attack time and temporal centroid.

6.2.3 Specialized high-level features

There are few cases that the features selected might be characterized by unnecessary detail and they are not so general as they should be. As an example, in [42], various over-detailed quantities are proposed as features, like cepstrum resynthesis residual magnitude and spectral roll-off point.

Features that discriminate between speech and music should be mentioned explicitly, since they have been the object of specialized research. Most sounds generated by musical instruments as sources share the common characteristic of being harmonic. This means that they contain superimposed a fundamental frequency tone plus its integer multiples. On the other hand, speech is a mixed harmonic/non-harmonic sound with voice/unvoiced segments correspondingly [57]. Although fundamental frequency estimation (referred as pitch) is a research field on its own, various features try to detect harmonic related properties.

Features describing rhythmic content can become useful in musical genre classification experiments. In addition, the analysis of the rhythmic structure of music by using tempo tracking algorithms can be used in music indexing and retrieval applications. Usually, tempo detection algorithms consist of a filterbank decomposition, an envelope extraction step and a periodicity detection algorithm. The features calculated for describing rhythmic content are based on the wavelet transform (WT), which provides high time resolution and low-frequency resolution for high frequencies and low time and high-frequency resolution for low frequencies. Moreover, the discrete wavelet transform (DWT) can provide a compact signal representation in the temporal and spectral domain, that can be used efficiently in a filterbank algorithm. More information on the tempo detection algorithm and the extraction of the rhythmic features using a beat histogram can be found in [47].

Up till now only few approaches in the area of content-based audio analysis have utilized the framework of psychoacoustics. Psychoacoustics deals with the relationship of physical sounds and the human brain's interpretation of them, cf. [59]. One of the first exceptions was [15], using psychoacoustic models to describe the similarity of instrumental sounds. The SOMeJB system [40] uses Rhythm Patterns as features, which describe time-invariant loudness modulation amplitudes per modulation frequency (i.e. energy or rhythm variation) on several frequency regions. The modulation amplitudes are calculated from a Sonogram, i.e. a spectrogram representing the specific loudness sensation according to the human psycho-acoustics. Specifically, the audio data is decomposed into frequency bands, which are then grouped according to the Bark critical-band scale. Then, loudness levels are calculated, referred to as phon using the equal-loudness contour matrix, which is subsequently transformed into the specific loudness sensation per critical band, referred to as sone. To obtain a time-invariant representation, recurring patterns in the individual critical bands are extracted in the second stage of the feature extraction process. These are weighted according to the fluctuation strength model, followed by the application of a final gradient filter and gaussian smoothing.

Recent experiments confirmed the importance of psycho-acoustic models for audio classification [28], also introducing two further high-level descriptors for audio characteristics, namely statistical Spectrum Descriptors and Rhythm histograms: The spectrum transformed into Bark scale in step represents rhythmic characteristics within the specific frequency range of a critical band. According to the occurence of beats or other rhythmic variation of energy on a specific band, statistical measures are able to describe the audio content. The following statistical moments on the values of each of the 24 critical bands are calculated to form statistical Spectrum Descriptors: mean, median, variance, skewness, kurtosis, min- and max-value, resulting in 168 attributes.

Rhythm Histogram features are another descriptor for general rhythmics in an audio document. Contrary to the Rhythm Patterns and the Statistical Spectrum Descriptor, information is not stored per critical band. Rather, the magnitudes of each modulation frequency bin of all 24 critical bands are summed up, to form a histogram of “rhythmic energy” per modulation frequency. The histogram contains 60 bins which reflect modulation frequency between 0 and 10 Hz. For a given piece of audio, the Rhythm Histogram feature set is calculated by taking the median of the histograms of every 6 second segment processed, resulting in a 60-dimensional feature space.

6.3 Segmentation

Audio segmentation is fundamental to subsequent tasks such as classification of audio. Many different approaches have been proposed in the literature for the segmentation of an audio stream into homogeneous parts.

In *rule-based* approaches, segmentation is based on rules applied to the set of features that have been extracted from the stream. *Energy* is the most common feature used. Energy based approaches [48], [25], [22] have been widely used and are particularly easy to implement. Silence periods in the input signal are detected as low energy sections of the signal. It is assumed that segment boundaries exist in theseRhythmic periods if a number of additional constraints is satisfied such as minimum length of the silence period. Others hypothesize segment boundaries when abrupt changes in the values of the features between subsequent moving frames are detected [55], [57].

In *metric-based* approaches segment boundaries are placed at local maxima or minima of a special distance calculated between neighboring sliding windows. One metric is the Kullback-Leibler divergence that was first used by Siegler et al [45] as an alternative to the Generalized Likelihood Ratio proposed in [20], in the case of speaker segmentation. The Bayesian Information Criterion (BIC) was applied as a metric by Chen and Gopalakrishnan [8] and exhibits improved stability and robustness at a high computational cost, however. Many have proposed variations in the application of the BIC in order to optimize efficiency [58], [23]. The VQ distortion measure, on the other hand, proposed in [33], is an alternative reported to have improved results.

In *decoder-guided* approaches the speech recognition system is used for segmentation. The stream is first decoded and then the segments are cut between long silence intervals [26], [53]. However, silence is not directly related to acoustic changes in the stream so usually a second stage segmentation follows usually based on rules.

In many cases, segmentation is performed explicitly at pre-specified time intervals. A post processing step, after classification, follows in order to concatenate neighboring segments of the same class.

6.4 Classification

Equally important for event detection in audio streams is the classification of the various audio segments in predetermined classes. Classification may take place in subsequent stages. Definition of the classes depends on the application. Several studies and overviews related to content-based audio signal clas-sification are available, e.g. [29].

6.4.1 Classification approaches

Rule-based methods follow a hierarchical heuristic scheme to achieve classification. Based on the properties of the various audio classes in the feature space, simple rules are devised and form a decision tree aiming at the proper classification of the audio segments [57], [22]. These methods usually lack robustness because they are threshold dependent, but no training phase is necessary and they can work in real-time.

In most of the *model-based* methods, segmentation and classification are performed at the same time. Models such as Gaussian Mixture Models and Hidden Markov Models are trained for each audio class and classification is achieved by Maximum Likelihood or Maximum a Posteriori selection over a sliding window [25], [5], [42], [39], [4], [1], [52]. These methods may yield quite good results but they cannot easily generalize, they do not work in real-time, since they usually involve a number of iterations, and data is needed for training.

Classical Pattern Analysis techniques cope with the classification issue as a case of pattern recognition. So, various well known methods of this area are applied, such as neural networks and Nearest Neighbor (NN) methods. Maleh et al [10] apply either a quadratic Gaussian classifier or an NN classifier. Shao et al [44] apply a multilayer perceptron combined with a genetic algorithm to achieve 16-

class classification. Lu et al [30] apply an algorithm based on K-nearest neighbor classifier and Linear Spectral Pairs Vector Quantization to determine speech / non-speech segments. Foote [19] uses a tree structure quantizer for speech/music classification. More modern approaches have also been tested like Nearest Feature Line Method [27] which performs better than simple NN approaches, and Support Vector Machines [21], [32]. Results are quite satisfactory.

State of the art in the classification methods for event detection is not completely described by the preceding categorization. *Hybrid* approaches that combine the aforementioned ideas are equally significant. The classifier proposed by Lu et al [30] is such an example. In the first stage, a variation of classical pattern analysis algorithms is applied to discriminate between speech and non-speech segments, then a finer classification is achieved by a rule-based schema and finally speaker clustering is performed by model-based analysis.

6.4.2 Musical instrument classification

The problem consists of a system being able to identify the instruments performing over an audio signal for a specific time frame. Whilst the problem of automatic instrument identification in polyphonic music by using the whole spectra of orchestral instruments still remains unsolved, the automatic classification of musical instruments in monophonic music has produced interesting results.

As far as monophonic audio streams are concerned, Brown [7] attempted to automatically classify sound segments consisting 4 classes of woodwind instruments (oboe, saxophone, clarinet and flute) using files derived from the UIOWA database. The features used were cepstral coefficients, bin-to-bin differences, constant Q-coefficients and autocorrelation coefficients, with correct identification results were at 79%-84%, comparable to human perception experiments. Eronen [11] attempted to classify isolated instrument tones of 30 instrument classes taken from the MUMS database (1498 samples total). By using a total of 43 features, ranging from temporal to spectral and a hierarchical classification framework utilizing the k-NN algorithm, he achieved 80.6% correct classification rate for individual instruments. Likewise, Marques [31] classified instrument tones (of 0.2 sec duration) for 8 musical instrument classes using Gaussian Mixture Models (GMMs) and Support Vector Machines (SVMs) for classification. 1024 training segments and 100 test segments were utilized for the experiments, taken from various recordings, and the features extracted were Linear Prediction Coefficients (LPCs), cepstral and mel-cepstral coefficients. The correct classification rate was at 70%. More recently, Slezak [51] applied temporal and spectral descriptors to sound data taken from the MUMS collection, consisting of 18 instrument classes. The features extracted were mainly taken from the MPEG-7 Audio framework: length, attack, decay, steady, harmonicity, fundamental frequency and brightness. Rough set decision rules and k-NN algorithms were used for classification, which at best reached 68.4% correct classification rate.

In an early work [15] the authors use psychoacoustic models to describe the similarity of instrumental sounds and organize a collection of instrument sounds using a Self-Organizing Map. For each instrument a 300 milliseconds sound was analyzed, extracting steady state sounds with a duration of 6 milliseconds. These steady state sounds can be regarded as the smallest possible building blocks of music. Contrary to pure classification tasks, the goal of this work was to form a basis for exploratory analysis of instrument sounds, offering a possibility to browse through groups of similar sounds.

At the present time, research has been focusing more on the area of instrument classification for polyphonic music. The first step in such a method is naturally to separate the audio stream into the different monophonic instrument sounds, and afterwards to classify the various instruments using the same methods as proposed before. Essid [12] has proposed a method for polyphonic instrument recognition, making possible to recognize up to 4 instruments playing concurrently. The system associates a hierarchical classification tree with a class-pairwise feature selection technique and GMMs to discriminate the possible instrument combinations. The features used are mel-frequency cepstrum coefficients, zero-crossing rate, autocorrelation coefficients, along with spectral features such as spectral centroid, spectral width, spectral asymmetry and the MPEG-7 spectrum flatness descriptor. With this proposed method, an average success rate of 90.88% is achieved, using samples from a jazz database. Wang [49] proposed a different method for polyphonic music separation, which is based on non-negative matrix factorization (NMF) techniques. By performing the NMF algorithm on the magnitude spectrogram of the audio stream, the signal is decomposed into temporal and spectral components. Two different experiments onto two audio stream were performed, but whilst the decomposition was highly accurate, the recovered components were manually classified. Research that could be performed in the future could focus on an automatic grouping method of the different components.

6.4.3 Musical genre classification

The problem of musical genre classification is referred as automatically classifying pieces of music to a hierarchy of musical genres and due to its subjective nature (the definitions and taxonomies vary) is an ill-defined problem. Some attempts have been made for a classification of basic musical genres, but the area has not yet been fully explored. A thorough survey on the various approaches on genre classification is conducted on [3], which confirms the existence of three different musical genre extraction approaches: a manual classification of titles, supervised classification techniques and data mining techniques such as co-occurrence analysis.

As far as pattern recognition approaches are concerned, Tzanetakis [47] proposed a system for classification of 20 musical genres, using spectral and rhythmic features. 2000 audio files were used in total and a classifier based on GMMs, while the genre classification accuracy for ten genres reached 61%, a result comparable to human musical genre classification.

Scheirer [43] presented a model of human perceptual behavior and briefly discussed how his model can be applied to classifying music into genre categories and performing music similarity-matching. However, he has not applied his model to large scale music collections. The collection he uses consisted of 75 songs from each of which he selected two 5-second sequences.

A more recent study by Xu [56] uses as extracted features the beat spectrum, the LPC-derived cepstrum, zero-crossing rate, spectrum power and MFCCs. Four different genre classes are taken into consideration, consisting of 100 total samples. The main classifier developed is based on SVMs and the results show 6.86% error rate, while the GMM classifier achieves 12.31%. Generally, the pattern recognition approach for genre classification is limited by inconsistencies of the built-in taxonomy and the assumption that genre can be assessed from signal attributes.

In [28] the authors use a combined approach of 3 feature sets for music genre classification on 3 databases: Rhythm Patterns describe loudness modulation amplitude per modulation frequency based on a critical bands accummulated spectrum and including psycho-acoustics. The statistical spectrum descriptor contains statistics about the spectral features. The Rhythm Histogram summarizes modulation energy per modulation frequency. Together, the features achieve up to 84.24 % classification accuracy (698 audio files, 8 genres) using SVMs.

Further state-of-the-art approaches have participated in the 2005 MIREX (Music Information Retrieval Evaluation eXchange).

6.5 Content-based Organization and Browsing

While actually being addressed by a specific, separate Workpackage within MUSCLE, we briefly would like to point to some recent work that utilized the concepts of the work described so far in order to build user interfaces and access methods for audio repositories. While being far from complete, the main intention of this section is to provide a brief flavour of activities herein, bridging the gap between content based description of audio and its potential utilization.

One of the core tasks of content-based description of audio is to support retrieval tasks. An overview of a range of music retrieval systems has recently been presented by Typke et al. [34] A system supporting browsing audio streams is the Sonic Browser [17]. In [16] this line of research is continued evaluating the browsing of everyday sounds. The investigation is directed at comparing browsing single versus multiple stream audio.

A different approach is taken by the SOMeJB system, i.e. the SOM-enhanced Jukebox [40]. The goal is to automatically create an organization of music archives following their perceived sound similarity. More specifically, characteristics of frequency spectra are extracted and transformed according to psychoacoustic models. The resulting psychoacoustic Rhythm Patterns are further organized using the Growing Hierarchical Self-Organizing Map (GHSOM), an unsupervised neural networkbased on the Self-Organizing Map. On top of this advanced visualizations including Islands of Music (IoM) and Weather Charts [38] offer an interface for interactive exploration of large music repositories. Interactive zooming, browsing and playlist generation are supported by the PlaySOM application [35]. Other interfaces building on the same or similar concepts are currently gaining larg epopularity, such as the Databionic Music Miner [41] using very large and sparsely populated SOMs.

Of course, ranges of other approaches and representations are being devised, integrating metadata, or utilizing symbolic music representation rather than pure audio as covered in this report.

6.6 Conclusions

In this report, we have presented the state of the art in the area of content-based description of audio. From the preceding discussion it becomes obvious that research in this field has been rather active in recent years. Although the achievements have been important, there is a number of key issues that still remain open. Generalization, robustness and real-time operation constitute challenging problems that ongoing research has to face. It seems that a widely accepted solution in the case of a general audio stream has not yet been proposed.

Furthermore, the much-proclaimed bridging of the semantic gap, i.e. understanding which concepts are actually burried withing a given set of audio data, and how to extract and convey this, is still far from being finished. The individual sub-disciplines touched upon in this report are in many cases still far from a commonly accepted model of the precise description and evaluation of their endeavours. For example, the concept of "genre" is still under hefty debate within the community aiming at genre classification, and its conceptual differences from, say, artist classification, are subject to ongoing discussion.

Feature extraction is still progressing steadily towards a better capture of specific characteristics of audio, particularly suitable to specific tasks of semantic annotation, whereas some of the main progresse obtained at this years' MIREX audio description contest have been obtained by significant improvements in optimized machine learning algorithms.

Furthermore, first works integrating, for example, textual and audio data to obtain more robust extraction and representation of semantic concepts, are beginning to appear, hinting at the potential of cross-modal analysis.

Further investigation of these areas is clearly warranted. Semantic interpretation of multimedia data will surely benefit from any possible improvements in the field of event detection in audio streams.

6.7 Related URLs

Related research projects:
http://research.microsoft.com/users/llu/Audioprojects.aspx

Audio search technologies:
http://www.musclefish.com

Audio mining technologies:
http://www.nexidia.com/

http://www.bbn.com/speech/am.html

Audio search using speech recognition:
http://speechbot.research.compaq.com/

Search the web for sounds:
http://www.findsounds.com/index.html

Musical audio mining:
http://www.ipem.rug.ac.be/MAMI/

Music Information Retrieval Evaluation eXchange
http://www.music-ir.org/mirexwiki/index.php/Main_Page

SOMeJB: The SOM-enhanced jukebox.
http://www.ifs.tuwien.ac.at/~andi/somejb

RISM. Repertoire international des sources musicales.
http://rism.stub.uni-frankfurt.de

Musica. the international database of choral repertoire.
http://www.musicanet.org

Marsyas: A software framework for research in computer audition.
http://www.cs.princeton.edu/~gtzan/wmarsyas.html

Harmonica. accompagnying action on music information in libraries.
http://projects/fnb/nl/harmonica

Cantate. computer access to notation and text in music libraries.
http://projects/fnb/nl/cantate

Bibliography

[1] J. Ajmera, I. McCowan, and H. Bourlard. Robust HMM-based speech/music segmentation. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP-02)*, pages 1746–1749, Orlando, Florida, 2002.

[2] J. Ajmera, I. McCowan, and H. Bourlard. Speech/music segmentation using entropy and dynamism features in a HMM classification framework. *Speech Communication*, 40:351–363, 2003.

[3] J.Aucouturier and F. Pachet, "Representing musical genre: a state of the art," *J. New Music Research*, vol. 32, no. 1, pp. 83-93, 2003.

[4] M. Baillie and J.M. Jose. An audio-based sports video segmentation and event detection algorithm. In *Proc. 2nd IEEE Workshop on Event Mining 2004, Detection and Recognition of Events in video in association with IEEE Computer Vision and Pattern Recognition (CVPR2004)*, Washington DC, USA, July 2004.

[5] R. Bakis, S. Schen, P. Gopalakrishnan, R. Gopinath, S. Maes, and L. Polymenakos. Transcription of broadcast news - system robustness issues and adaptation techniques. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP-97)*, volume 2, pages 711–714, 1997.

[6] A. Berenzweig and D. Ellis. Locating singing voice segments within music signals. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-01)*, 2001.

[7] J. C. Brown, O. Houix, and S. McAdams, "Feature dependence in the automatic identification of musical woodwind instruments," *Journal Acoustical Society of America*, vol. 109, no. 3, pp. 1064-1072, March 2001.

[8] S. Chen and P. S. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the Bayesian information criterion. In *Broadcast News Transcription and Understanding Workshop*, pages 127–132, Lansdowne, Virginia, February 1998.

[9] J.S. Downie. *Annual Review of Information Science and Technology*, volume 37, chapter Music information retrieval, pages 295–340. Information Today, Medford, NJ, 2003. `http://music-ir.org/downie_mir_arist37.pdf`.

[10] K. El-Maleh, M. Klein, G. Petrucci, and P. Kabal. Speech/music discrimination for multimedia applications. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP-00)*, pages 2445–2448, June 2000.

[11] A. Eronen and A. Klapuri, "Musical instrument recognition using cepstral coefficients and temporal features," in Proc. *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 2, pp. 753-756, June 2000.

[12] S. Essid, G. Richard and B. David, "Instrument recognition in polyphonic music," in Proc. *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 3, pp. 245-248, March 2005.

[13] G. Evangelopoulos and P. Maragos. Multiband energy tracking and demodulation towards noisy speech endpoint detection. IEEE Trans. Acoust., Speech, Signal Processing, 2004. submitted.

[14] H. Ezzaidi and J. Rouat. Speech, music and songs discrimination in the context of handsets variability. In *Proc. International Conference on Speech and Language Processing, (ICSLP-02)*, September 2002.

[15] B. Feiten and S. Günzel. Automatic indexing of a sound database using self-organizing neural nets. *Computer Music Journal*, 18(3):53–65, 1994.

[16] M. Fernström and E. Brazil. An auditory tool for multimedia asset management. In *Proceedings of the International Conference on Auditory Display (ICAD 2001)*, Espoo, Finland, July 29 - August 1 2001.

[17] M. Fernström and C. McNamara. After direct manipulation - direct sonification. In *Proceedings of the International Conference on Auditory Display (ICAD 98)*, Glasgow, UK, 1998.

[18] J. Foote. An overview of audio information retrieval. *Multimedia Systems, special issue on audio and multimedia*, 7(1):2–10, January 1999.

[19] J. T. Foote. Content-based retrieval of music and audio. In C.-C. J. Kuo et al., editor, *Multimedia Storage and Archiving Systems II, Proc. of SPIE*, volume 3229, pages 138–147, 1997.

[20] H. Gish and M. Schmidt. Text-independent speaker identification. *IEEE Signal Processing Mag.*, pages 18–32, October 1994.

[21] G. D. Guo and S. Z. Li. Content-based audio classification and retrieval by Support Vector Machines. *IEEE Trans. Neural Networks*, 14(1):209–215, January 2003.

[22] H. Harb, L. Chen, and J.-Y. Auloge. Speech/music/silence and gender detection algorithm. In *Proceedings of the 7th International conference on Distributed Multimedia Systems DMS01*, pages 257–262, September 2001.

[23] R. Huang and J.H.L. Hansen. Advances in unsupervised audio segmentation for the broadcast news and NGSW corpora. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP-04)*, 2004.

[24] ISO/IEC 15938-4:2001, "Multimedia Content Description Interface - Part 4: Audio", Version 1.0.

[25] T. Kemp, M. Schmidt, M. Westphal, and A. Waibel. Strategies for automatic segmentation of audio data. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP-00)*, 2000.

[26] F. Kubala, H. Jin, S. Matsoukas, L. Nguyen, R. Schwartz, and J. Makhoul. BBN byblos Hub-4 transcription system. In *Proc. of the 1997 DARPA Speech Recognition Workshop*, Chantilly, Virginia, 1997.

[27] S. Z. Li. Content-based audio classification and retrieval using the Nearest Feature Line Method. *IEEE Trans. Acoust., Speech, Signal Processing*, 8(5), September 2000.

[28] T. Lidy and A. Rauber Evaluation Of Feature Extractors And Psycho-acoustic Transformations For Music Genre Classification In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, London, UK, September 2005.

[29] M. Liu and C. Wan. A study of content-based classification and retrieval of audio database. In *Proceedings of the 5th International Database Engineering and Applications Symposium (IDEAS 2001)*, Grenoble, France, 2001. IEEE.

[30] L. Lu, H.-J. Zhang, and H. Jiang. Content analysis for audio classification and segmentation. *IEEE Trans. Acoust., Speech, Signal Processing*, (7):504–516, October 2002.

[31] J. Marques and P. J. Moreno, "A study of musical instrument classification using gaussian mixture models and support vector machines," *Technical Report Series, Cambridge Research Laboratory*, vol. 4, pp. 1-16, June 1999.

[32] P. J. Moreno and R. Rifkin. Using the Fisher Kernel Method for web audio classification. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP-00)*, 2000.

[33] S. Nakagawa and K. Mori. Speaker change detection and speaker clustering using VQ distortion measure. *Systems and Computers in Japan*, 34(13):25–35, 2003.

[34] R. Typke, F. Wiering and R.C. Veltkamp. A Survey Of Music Information Retrieval Systems. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, London, UK, September 2005.

[35] R. Neumayer, M. Dittenbach and A. Rauber. PlaySOM and PocketSOM-Player, Alternative Interfaces to Large Music Collections In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, London, UK, September 2005.

[36] N. Nitanda, M. Haseyama, and H. Kitajima. An audio signal segmentation and classification using fuzzy c-means clustering. In *Proc. 2nd International Conference on Information Technology for Application (ICITA-2004)*, 2004.

[37] C. Panagiotakis and G. Tziritas. A speech/music discriminator based on RMS and zero-crossings. *IEEE Trans. Multimedia*, 2004 (accepted).

[38] E. Pampalk, A. Rauber, and D. Merkl. Content-based organization and visualization of music archives. In *Proceedings of ACM Multimedia 2002*, pages 570–579, Juan-les-Pins, France, December 1-6 2002. ACM. `http://www.ifs.tuwien.ac.at/ifs/research/publications.html`.

[39] J. Pinquier, J.-L. Rouas, and R. Andre'-Obrecht. Robust speech / music classification in audio documents. In *Proc. International Conference on Speech and Language Processing, (ICSLP-02)*, pages 2005–2008, Vol. 3, Denver, USA, 16-20 septembre 2002. Causal Productions Pty Ltd.

[40] A. Rauber, E. Pampalk, and D. Merkl. The SOM-enhanced JukeBox: Organization and visualization of music collections based on perceptual models. *Journal of New Music Research*, 32(2):193–210, June 2003. `http://www.extenza-eps.com/extenza/loadHTML?objectIDValue=16745&type=ab%stract`.

[41] F. Mörchen, A. Ultsch, M. Nöcker and C. Stamm. Databionic Visualization Of Music Collections According To Perceptual Distance. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, London, UK, September 2005.

[42] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP-97)*, 1997.

[43] E. Scheirer. Music-Listening Systems. PhD Thesis, MIT Media Laboratory, April 2000.

[44] X. Shao, C. Xu, and M. S. Kankanhalli. Applying neural network on content based audio classification. In *IEEE Pacific-Rim Conference On Multimedia (PCM-03), Singapore*, 2003.

[45] M. A. Siegler, U. Jain, B. Raj, and R. M. Stern. Automatic segmentation, classification and clustering of broadcast news audio. In *Proceedings of the Ninth Spoken Language Systems Technology Workshop*, 1997.

[46] G. Tzanetakis and P. Cook, "Multifeature audio segmentation for browsing and annotation," in Proc. *IEEE Workshop Applications of Signal Processing to Audio and Acoustics*, pp. 103-106, 1999.

[47] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 5, pp. 293-302, July 2002.

[48] H. Wactlar, A. Hauptmann, and M. Witbrock. Informedia: News-on-demand experiments in speech recognition. In *Proceedings of ARPA Speech Recognition Workshop*, Harriman, NY, February 1996. Arden House.

[49] B. Wang and M. D. Plumbley, "Musical audio stream seperation by non-negative matrix factorization," in Proc. *DMRN Summer Conf.*, pp. 14-17, July 2005.

[50] Steven Wegmann, Francesco Scattone, Ira Carp, Larry Gillick, Robert Roth, and Jon Yamron. Dragon Systems' 1997 broadcast news transcription system. In *Proc. of the 1998 DARPA Broadcast News Workshop*, Landsowne, Virginia, 1998.

[51] A. Wieczorkowska, J. Wroblewski, P. Synak, and D. Slezak, "Application of temporal descriptors to musical instrument sound recognition," *Journal Intelligent Information Systems*, vol. 21, no. 1, pp. 71-93, July 2003.

[52] G. Williams and D. Ellis. Speech/music discrimination based on posterior probability features. In *Proc. Eurospeech99*, Budapest, September 1999.

[53] P. C. Woodland. The development of the HTK broadcast news transcription system: An overview. *Speech Communication*, 37:47–67, May 2002.

[54] P.C. Woodland, T. Hain, G.L. Moore, T.R. Niesler, D. Povey, A. Tuerk, and E.W.D. Whittaker. The 1998 HTK broadcast news transcription system: Development and results. In *Proc. of the DARPA Broadcast News Workshop*, Herndon, Virginia, 1999.

[55] Wang Y., Liu Z., and Huang J. Multimedia content analysis using audio and visual information. *IEEE Signal Processing Mag.*, 17(6):12–36, November 2000. Invited Paper.

[56] C. Xu, N. C. Maddage, X. Shao, F. Cao, and Q. Tian, "Musical genre classification using support vector machines," in Proc. *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 5, pp. 429-432, April 2003.

[57] T. Zhang and C.-C. J. Kuo. Audio content analysis for online audiovisual data segmentation and classification. *IEEE Trans. Speech Audio Processing*, 9(4):441–457, May 2001.

[58] B. Zhou and J. H. L. Hansen. Unsupervised audio stream segmentation and clustering via the Bayesian information criterion. In *Proc. International Conference on Speech and Language Processing, (ICSLP-00)*, pages 714–717, October 2000.

[59] E. Zwicker and H. Fastl. *Psychoacoustics, Facts and Models*, volume 22 of *Series of Information Sciences*. Springer, Berlin, 2 edition, 1999.

Chapter 7

Robust Features for Automatic Speech Recognition Systems

7.1 Introduction

Since the early scientific or fictitious envisagements of intelligent machines at Bell Labs, Lincoln Labs, or Clarke and Kubrick's Space Odyssey, computer systems have become ubiquitous, storing huge quantities of multimodal data (combinations of speech, audio, text and video). Managing this information is essential for the creation of a knowledge-driven society. Towards this direction, Automatic Speech Recognition (ASR) seems to be one of the most important tasks that should be successfully dealt with.

ASR technology has developed rapidly. The pioneering work of the first research years (Filterbanks, Spectrogram, Linear Prediction Coding) has been followed by several fundamental achievements (Dynamic Time Warping, Mel-Cepstral Coefficients, Hidden Markov Models). Although significant contributions have been made, ASR systems haven't yet reached the desirable standards of functionality in ordinary (everyday) conditions. Robustness is one of the main attributes that ASR systems lack. This could be tackled by the application of speech enhancement techniques, extraction of robust features for speech representation, or, finally, by model compensation.

As far as feature extraction is concerned, the main research areas cannot be easily classified in completely distinct categories, since the cross-fertilization of ideas has triggered approaches that combine ideas from various fields. *Filterbank analysis* is an inherent component of many techniques for robust feature extraction. It is inspired by the physiological processing of speech sounds in separate frequency bands that is performed by the auditory system. *Auditory processing* has developed into a separate research field and has been the origin of important ideas, related to physiologically and perceptually inspired features [16, 27, 51],[19, 21]. Equally important is the research field based on concepts relevant to speech resonance (short-term) *modulations*. Both physical observations and theoretical advances support the existence of modulations during speech production [52, 30],[37]. Other approaches are related to the long-term *modulation spectrum* [26, 11] and the features derived from that [23],[21], [22], which could be perceptually based or variants of noise robust features. Finally, special attention should be paid to the techniques [34, 38, 40, 4] that attempt to model nonlinear phenomena of the speech production system [52, 30]. These may quantify aerodynamic phenomena like turbulence and/or modulations, that the linear source-filter model cannot take into consideration.

7.2 Filterbanks

The corresponding group of ASR features is based on the idea of decomposing speech along the frequency domain using several overlapping bandpass filters. The filterbank scheme is motivated by observations made by Allen [2] and Fletcher [13]. They have provided evidences that the human auditory system processes speech in separate frequency bands and extracts their spectral content. The human cognitive system classifies the speech events accordingly. The most common features for ASR tasks are the time-localized energies of the different frequency bands, [39]. These features map the spectral subband energies to the appropriate acoustic events (phonemes). A common practice is to train separate recognizers to process each one of the band components.

For the definition of a filterbank certain parameters are required. These parameters are the number of filters, their placing (the center frequencies), their bandwidths and the type of filters used. The number of filters cannot be too small otherwise the ability to resolve the speech spectrum could be impaired. Also, their number cannot be too large because the filter bandwidth would be too small and some bands would have very low speech energy. The most common range for the number of filters is between 6 and 32 [48]. The filter placing can be linear where the center frequencies are spaced uniformly to span the whole frequency range of the speech signals. An alternative to uniform filterbanks is to space the filters uniformly along a logarithmic frequency scale (e.g. the Mel scale). Such a logarithmic frequency scale is motivated by the human auditory perception process. Finally, a common non-uniform filterbank placing is the critical band scale (Bark scale). The spacing of the filters along the critical band scale is based on perception studies and is intended to choose bands that give equal contribution to speech articulation [48]. The third parameter is the filters' bandwidths, which depend on the placing, the number and the desired overlap of the filters. It is common practice that the filter bandwidths are not equal along the frequency axis. Finally, many different types of bandpass filters have been proposed during the past few years depending on the analysis/recognition tasks. For instance, gammatone filters are popular for auditory speech analysis [28]. An alternative option are Gabor filters [10, 45] which have optimal time-frequency resolution.

Mel Frequency Cepstral Coefficients–MFCC: The MFCC are the most commonly used feature set for ASR applications. They were introduced by Davis and Mermelstein [9]. Cepstrum analysis can enable the separation of convolved signals. In the linear source-filter model such convolved signals are the source excitation signal and the impulse response of the vocal tract filter. So, the vocal tract's distortions to the speech signal can be removed.

The wide-spread use of the MFCC is due to the low complexity of the estimation algorithm and their efficiency in ASR tasks. In detail the algorithm consists of the following steps. The magnitude-squared of the Fourier transform is computed and triangular frequency weights are applied. These weights represent the effects of the peripheral auditory frequency resolution. Then, the logarithmic outputs of the filterbank are used for the estimation of the cepstrum of the signal. Finally, the feature vectors are estimated with the discrete cosine transform in order to reduce the dimensionality and decorrelate the vector components. These features are smoothed out by dropping the higher-order cepstra coefficients.

Even though MFCC are the most common features for ASR tasks, they appear to have major disadvantages. At first, as most HMMs use Gaussian distributions with diagonal covariance matrices, they cannot benefit from a cepstral liftering, since any multiplying factor that is applied to the observations does not affect the exponent calculation. Second, MFCC are easily affected by common frequency-localized random perturbations which have hardly any effect on human speech communication. Finally, robust feature extraction should process at least about syllable-length (around 200-500 ms) spans of the speech signal [23], in order to extract reliable information for classification of the phonemes.

Another different filterbank approach was proposed by Hermansky [24] and Boulard [7]. They examined Fletcher's proposal [13] to divide the speech spectrum into a number of frequency subbands and extracted spectral features from each of these. However, the recognition/classification task is done independently in each one of the bands by estimating the conditional probabilities for each band. Then, these estimates are merged in order to give the final output feature set. The merging is done by a multi-layer perceptron (MLP) trained on the same training data as the HMM-based classifiers. The input feature set is the power spectrum values obtained after the PLP critical band filtering, the compression by a cubic-root function and loudness equalization. These features showed a relative improvement of about 50% (compared to the MFCC) in the presence of frequency selective additive noises which corrupted only some of the frequency bands. On the other hand, they were ineffective for noises that corrupted the whole speech spectrum.

Subband Spectral Centroids: These features have been introduced by Paliwal et al, [15]. They can be considered as histograms of the spectrum energies distributed among nonlinearly-placed bins. They show properties similar to those of the distributions of the formant frequencies and they appear to be quite robust to noise. They can be used as a supplementary feature set to cepstral features. Note that the conventional feature sets like the MFCC utilize only amplitude information from the speech power spectrum while the proposed features utilize frequency information, too. It should be stated, though, that these features failed to show significant improvement at the recognition rates when compared to the MFCC.

7.3 Modulations

7.3.1 Short-term Modulations

The linear model of speech makes the assumption that resonances (and the center frequencies of the formants) remain constant for relatively short amounts of time. A nonlinear model proposes that these resonances are not constant but can fluctuate around their center frequency and can be modeled as a sum of AM-FM signals [37]. Short-term modulation features attempt to quantify these fluctuations and capture the temporal and dynamic nature of the speech resonances. The improvement of the recognition rates supports the accuracy of such a nonlinear model [10]. These features are used to enhance the classic cepstrum-based features as an augmented feature set for ASR applications and they show robustness in noisy speech signals due to the use of the filterbank and the Energy Separation Algorithm (ESA).

Alternatively, other algorithms have been proposed to obtain the instantaneous amplitude and frequency signals from the bandpassed speech. Such approaches use Kalman filtering [41] and the Hilbert transform [44]. The first approach has high computational complexity and cannot be used for real-time applications. The latter one has poor temporal resolution and rapid changes are smoothed out.

Finally, some experimentation has been done with merging the source-filter model with the nonlinear model of the resonances for the estimation of MFCC-like features. More specifically, the square amplitude of the bandpassed speech signals is replaced by the nonlinear Teager energy [29] in the standard estimation algorithm of the MFCC feature set. The correct phoneme recognition rates have shown marginal differences for clean speech signals when compared to the MFCC corresponding rates. This is due to the post-processing smoothing effect of the MFCC (i.e. cutting off the higher-order cepstra coefficients). However, the Teager energy-based MFCC features seem to be smoother and more robust for noisy signals and yield improved results.

Frequency Modulation Features: A mel-spaced Gabor filterbank of 6-filters is used in order to bandpass the speech signals. These signals are demodulated using the ESA and the instantaneous frequency and amplitude signals are yielded. Then, the 1^{st} and 2^{nd} moments of the instantaneous frequencys are estimated. The *Frequency Modulation Percentages* (FMP) are the ratio of the second over the first moment of these signals [10]. These spectral moments have been tested as input feature sets for various ASR tasks yielding improved results. For the TIMIT phoneme recognition task the correct phoneme accuracy rates are 40% for the FMPs compared to 55% for the MFCC using though half the vector length of the MFCC. In the AURORA-3 database word recognition task, the relative improvement is 16% compared with the auditory features [8] and 50% when compared with the MFCC [10].

Amplitude Modulation Features: Other nonlinear feature sets have been proposed taking under consideration the amplitude modulated (AM) part of the nonlinear speech model. The algorithm described above (the filterbank and the demodulation algorithm) has been used to estimate the instantaneous amplitude signals (absolute envelopes) of the bandpassed speech signals. These envelopes are modulated by lowpass signals containing the linguistic information, [10]. The proposed feature set parametrizes the modulating signals and yields their statistics (their 1^{st} and 2^{nd} spectral moments). This feature set shows a statistically important improvement compared to the baseline accuracies of the MFCC. Recent experiments on these features indicate that they are noise invariant, mainly due to their lowpass nature. Namely the instantaneous amplitudes are lowpass signals and, concequently, more robust in noise. Their estimates appear to be very smooth, in terms of spikes and discontinuities, even at low SNR. It has been shown that they contain significant amount of information concerning both the speaker and the linguistic content of the speech signals [46].

The instantaneous amplitude signals, and their corresponding modulating signals, have a very slow temporal evolution. This property is exploited from another viewpoint by research in long-term modulations i.e. the *Modulation Spectrogram.* The short and long-term modulations are two different concepts of the speech production mechanism. The short-term modulations are studied in time-windows up to 10-30ms in order to capture the micro-details (very rapid changes) of the speech signals. On the contrary, long-term modulations examine the temporal evolution of the speech energy and the corresponding time-windows are in the range of 200-500ms.

7.3.2 Long-term Modulations

Early experiments [12, 11, 26] on the perceptual ability of the human auditory system have shown that slow temporal modulations differ as far as their relative importance on different frequencies is concerned. In detail, speech intelligibility is not affected by low-pass filtering below 16 Hz, or high-pass filtering above 4 Hz. Furthermore, intelligibility in noise depends on the integrity of the modulation spectrum in the range between 2 and 8 Hz, on the global shape of the spectral envelope and not so much on its fine details [31]. Finally, the duration of the dominant component (around 4 Hz) is related to the average duration of syllables.

Typical short-time feature extraction methods (filterbank energies, LPC, Cepstrum, MFCC, PLP) form a representation of the spectral envelope of the signal framewise. This has the drawback of being sensitive to background noise. For example at particular frequency components, part of the signal that lies 100ms outside a certain phonetic labeled segment may still carry information relevant to the classification of the given phoneme [55].

The relative importance of the different frequencies of the modulation spectrum is supported in terms of recognition experiments by the different contributions of spectrum components. When the lowest frequency band is removed (cutoff frequency of 1Hz) the accuracy is increased to 93.6% compared to 86% for the unfiltered modulation spectrum. The relative contributions of the various bands of the modulation spectrum, as far as different features are concerned (MFCC, PLP), do not show major differences. The subband that has the maximum contribution is in the range of 2-4Hz of the spectrum. For filterbank related features [31] the more important contribution is in the subband of 4-8Hz and gets more affected by convolutional noise than MFCC and PLP.

The *Dynamic Cepstral Coefficients* method [14] attepmts to incorporate long-term temporal information. These coefficients are computed by first- and second-order orthogonal polynomial expansions of feature time trajectories, referred to as "delta and acceleration coefficients". They have become a standard method followed by every ASR system and are robust to slowly varying convolution distortions. Alternatively, in the method of *Cepstral Mean Normalization* the long-term average is substracted from the logarithmic speech spectrum and convolutive noise is suppressed.

An alternative to the DC component removal (i.e. Cepstral Mean Normalization), is to use a high-pass filter. In Relative Spectral Processing (RASTA) [21, 22] the modulation frequency components that do not belong to the range from 1 to 12 Hz are filtered out. Thus, this method suppresses the slowly varying convolutive distortions and attenuates the spectral components that vary more rapidly than the typical rate of change of speech.

Relative Spectra Processing–RASTA: RASTA processing has fundamental relations to both the temporal properties of hearing and the equalization of speech [21]. It achieves a broader, than the delta features, pass-band by adding a spectral pole, and allows the preservation of the linguistic content. RASTA band-pass filtering is applied either on the logarithmic spectrum or on a nonlinearly compressed spectrum and consists of filters with a sharp spectral zero at the zero modulation frequency. The moving average (MA) part of the RASTA filters is derived from the delta features. The spectral pole of the autoregressive (AR) part is obtained through experimentation and determines the high-pass cut-off frequency. The RASTA algorithm consists of the following steps. At first, the critical-band power spectrum is computed. Then, the spectral amplitude is transformed through a compressing static nonlinearity, and the time trajectories of each transformed spectral component are filtered. The filtered speech representation is transformed through an expanding static nonlinearity and is multiplied by the equal loudness curve raised to the power 0.33 in order to simulate the power law of hearing. Finally, an all-pole model of the resulting spectrum is computed.

Several variations have been proposed using a different nonlinear spectral domain than the logarithmic one like the J-RASTA and the Lin-Log RASTA algorithms. The use of such variations improves the correct recognition rates because they simulate more realistically the physiological hearing processes.

Many experiments have been performed in order to examine and compare the RASTA features to other analysis schemes such as the PLP and the PLP+Cepstral Mean Removal. Both logarithmic RASTA and cepstral mean removal improve the recognition rates for convolutional noise. However, PLP, logarithmic RASTA and cepstral mean removal all degrade severely in additive noise. Lin-Log RASTA with a linear mapping shows good robustness over both convolutional and additive noise. While cepstral mean subtraction performed better for purely convolutional noise, it was not as effective as the Lin-Log RASTA approach when additive noise was present.

Temporal Patterns–TRAP: This method was introduced by Hermansky et al. [23]. Conventional features in ASR describe the short-term speech properties. On the other hand, the TRAP features describe likelihoods of sub-word classes at a given time instant, derived from temporal trajectories of band-limited spectral densities in the vicinity of the given time instant.

Coding of linguistic information in a single short-term spectral frame of speech appears to be very complex. A single frame of such a short-term spectrum does not contain all the necessary information for the decoding scheme as the neighboring speech sounds influence the short-term spectrum of the current one. The mechanical inertia of human speech production organs results in spreading the linguistic information in time. At any given time at least 3-5 phonemes interact. This introduces high within-phoneme variability of the spectral envelope. ASR systems attempt to classify phonemes from individual slices of the short-term spectrum and need to deal with this within-class variability, even though experiments show that human listeners are not affected by such phenomena.

Such ASR systems expect feature vectors of uncorrelated and normally distributed features every 10 ms. So, a process is needed that is capable of examining long spans of speech within various frequency bands and deliver every 10 ms uncorrelated and normally distributed features. The proposed algorithm TRAP-TANDEM is such a module. The tandem submodule is an hierarchical tree-based structure that splits speech into different sound classes e.g. voiced, unvoiced, silence, etc.

This processing scheme is capable of examining relatively long spans of the speech signal within various frequency bands. It uses MLP (Multi-Layer Perceptron) to provide nonlinear mapping from temporal trajectories to phoneme likelihoods. The *TRAP* processing uses relatively long time windows (500-1000 ms) and frequency localized (1-3 Barks) overlapping time-frequency regions of the signal. The *TANDEM* algorithm refers to a way of converting the frequency-localized evidence to features for the HMM-based ASR systems.

The time-frequency spectral density plane is estimated using the front-end taken from the PLP analysis. It employs the short-time spectral analysis of the speech signal using a Bark-spaced filterbank. The input to the TRAP estimator consists of 1-3 time trajectories of critical-band energies. The individual trajectories are concatenated to form a longer input vector, and finally, PCA is introduced in order to reduce the vector's dimensionality.

TRAP estimator delivers vectors of posterior probabilities sub-word acoustic events, each estimated at an individual frequency band. The events targeted are the phonemes clustered into 6 broad phonetic classes, and separate estimators are trained for each frequency region of interest.

The TANDEM part derives a vector of posterior probabilities of sub-word speech events for every speech frame from the evidence presented to its input. An MLP is used in order to optimally cluster the input vectors. and estimate the posterior probabilities of the individual classes. These probabilities are post-processed by a static nonlinearity in order to match the gaussian probability distributions, and whitened by the KL transform derived by the training data.

The events targeted by the TRAP estimators do not need to be the same as those targeted by the TANDEM estimator. Also, TRAP estimators can be trained on different databases than the databases used in training the TANDEM estimator. Note that both the TRAP and TANDEM estimators are nonlinear feed-forward MLP discriminative classifiers.

So far, the TRAP-TANDEM features have been found more useful in combination with the conventional spectrum-based features like PLP and MFCC, where they brought more than 10% relative improvement (for the DARPA EARS program). Nowdays, the performance of the TRAP-TANDEM stand-alone features is becoming comparable with the traditional approaches. For example, for the OGI Numbers task they yield the same (5%) word error rate as the best system using the PLP+Delta+DDelta features. Finally, for the TIMIT database task the TRAP-based features gave 10% relative improvement in the phoneme error rates compared to the MFCC.

7.4 Auditory-based Features

The human auditory system is a biological apparatus with ideal performance, especially in noisy environments. Various ASR approaches incorporate characteristics of this system. The adaption of physiologically based methods for spectral analysis [16] is such an approach. The physiological model of the auditory system can be categorized into the areas of outer, middle and inner ear [48]. The cochlea and the basilar membrane, both located in the inner ear, can be modeled as a mechanical realization of a bank of filters. Along the basilar membrane are distributed the Inner Hair Cells (IHC), which sense mechanical vibrations and convert them into firing of the connected nerve fibers which in turn emit neural impulses to the auditory nerve.

Inspired by the above ideas, the **Ensemble Interval Histogram** (EIH) model is constructed by a bank of 'cochlear' filters followed by an array of level crossing detectors that model the motion to neural conversion. The probability distributions of the level crossings are summed up for each cochlear filter resulting on the ensemble interval histogram. Front-ends that use ideas of this approach have shown comparable recognition rates to common spectrum-based features. Moreover, they are characterized by increased noise resistance for lower SNR's [17].

Lateral inhibition is another characteristic that has been introduced into periphery models. This is defined as the suppression of the activity of nerve fibers on the basilar membrane caused by the activity of adjacent fibres. It accounts for the phenomenon caused when two tones of different amplitude are similar in frequency, leading to an inhibition in the perception of the weaker one. The aforementioned phenomenon has been used to improve noise robustness by convolving a frequency dependent lateral inhibition function with noisy speech [56]. Since narrowband SNR is higher on spectral peaks, by emphasising these areas and attenuating spectral valleys, the signal's SNR increases.

The **Joint Synchrony/Mean-Rate** model [50, 51] captures the essential features extracted by the cochlea in response to sound pressure waves. It includes parts that deal with peripheral transformations occurring in the early stages of the hearing process. These parts attempt to extract information relevant to perception, such as formants, and enhance sharpness of onset and offset of different speech segments. In detail, the speech signal is first pre-filtered through a set of four complex zero pairs to eliminate the very high and very low frequency components. Then it passes through a 40-channel critical-band linear filter bank whose single channels were designed in order to fit physiological data. Next, the hair cell synapse model is intended to capture prominent features of the transformation from basilar membrane vibration, represented by the outputs of the filter bank, to probabilistic response properties of auditory nerve fibers. The outputs of this stage represent the probability of firing as a function of time for a set of similar fibers acting as a group. The two output models that follow are the Generalized Synchrony Detector (GSD) and the Envelope Detector (ED). GSD which implements the known "phase-locking" property of nerve fibers, is designed with the aim of enhancing spectral peaks due to vocal tract resonances. ED computes the envelope of the signals at the output of the previous stage of the model and is important for capturing the very rapidly changing dynamic nature of speech.

An important type of filter that has been proposed for auditory processing is the **gammatone** function [28]. This has been shown to describe impulse-response data gathered physiologically from primary auditory filters in the cat. The gammachirp is constructed by adding a frequency modulation term to the gammatone function. This function has minimal uncertainty in joint time/scale representation. The gammachirp auditory filter is the real part of the analytic gammachirp function, has an asymmetric amplitude characteristic and provides an excellent fit to human masking data.

Auditory peripheral modeling is another area that incorporates auditory characteristics. These include critical band filtering, loudness curve properties, nonlinear energy compression, haircell modeling and short-time adaptation. By the use of such models there are improvements in the temporal localization and speech detectability in degraded environments, resulting in increased system robustness to noise.

Perceptual linear prediction (PLP) is a variant of Linear Prediction Coding (LPC) which incorporates auditory peripheral knowledge [19, 20]. The main characteristics for estimating the audible spectrum are realized by adding critical band integration, equal-loudness pre-emphasis and intensity to loudness compression. More specifically, the method considers the short-term power spectrum and convolves it with a critical-band masking pattern. Then, the critical band is resampled at about one Bark scale intervals. A pre-emphasis operation is performed with a fixed equal loudness curve and finally the resulting spectrum is compressed with a cubic root nonlinearity. The output low-order all-pole model is consistent with phenomena observed in human speech perception. It simulates the properties of the auditory system resulting in parameters compatible with LPC. The main advantage of PLP is the reduction of the order of the model (e.g. 5 coefficients vs. 15 for LPC).

7.5 Fractal-based Features

One of the latest approaches in speech analysis are the nonlinear/fractal methods. These diverge from the standard linear source-filter approach in order to explore nonlinear characteristics of the speech production system. They are based on tools that lie in the areas of fractals and dynamical systems. Their motivation stems from the observations of aerodynamic phenomena in speech production [52, 30]. Specifically, airflow separation, unstable air jet, oscillations between the walls and vortices are phenomena encountered in many speech sounds and lead to turbulent flow. Especially fricatives, plosives and vowels uttered with some speaker-dependent aspiration contain various amounts of turbulence. Moreover, the presence of vortices could result in additional acoustic sources. The initial significant contributions [52, 30] are further supported by acoustic and aerodynamic analysis of mechanical models [5, 25]. On the other hand it has been conjectured that geometrical structures in turbulence can be modeled using fractals [35]. Difference equation, oscillator and prediction nonlinear models were among the early works in the area [47, 33, 54]. Speech processing techniques that have been inspired by fractals have been introduced in [36, 38]. These measure the roughness of the signal in multiple scales as a quantification of the geometrical complexity of the underlying signal. Their application as short-time features in ASR experiments has shown significant improvement of 12%-18% error reduction in the tough recognition task over the E-set ISOLET database [38].

Recently various approaches [18, 34, 40, 53, 6, 49] apply such fractal-based measurements on reconstructed multidimensional phase spaces instead of the one-dimensional signal space. They argue that the multidimensional reconstructed space is closer to the true speech production dynamics compared to the one dimensional speech signal. The latter can be seen as a collapsed projection from a higher dimensional space. The analysis is carried out by computing invariants of the multidimensional signals, which are the fractal dimensions and Lyapunov exponents [32], [4], [43]. Generalized dimensions [42],[3] and multifractal spectrum [1] are alternative representations of the underlying geometrical complexity. Special cases include the standard fractal dimension (box-counting, Minkowski-Boulingand dimension), correlation dimension and information dimension. It should be noted that this field is not fully developped yet because the observed phenomena are neither completelly understood nor directly related to the various approaches and models reported. Moreover, a suitable way to integrate such analysis in ASR systems is not a simple task and only preliminary results have been reported [38].

7.6 Discussion

In this report we have presented briefly the main trends for robust feature extraction techniques in ASR systems. Feature extraction methods can be categorized into overlapping classes that share a number of common ideas. The most common ideas are related to filterbank processing, features inspired by the physiology of the auditory system, features utilizing perceptual knowledge, or inspired by phenomena that occur during speech production (e.g. modulations). A review of the proposed features for ASR systems indicates that cepstral analysis and the *MFCC* [9] features have become one of the most common approaches. A popular alternative are the *PLP* [20] or related features that are based on knowledge of the human auditory peripheral system. Finally, nonlinear speech processing techniques (e.g. modulations, fractals) have started to gain momentum. Many techniques share the concept of short-time processing. However, recently there have been introduced alternative methods e.g. *RASTA* [21] ,*TRAP* [23] that filter out parts of the modulation spectrum or process frames that span longer time intervals. There are not direct comparisons for every proposed feature set, but implicit conclusions may be assumed by considering their absolute recognition results.

Concluding, although research in this area has been active for many decades, robustness is still a key issue that should be considered. Thus more effort should be placed in order to accomplish satisfactory performance in adverse acoustic environments.

7.7 Useful URLs

Speech at Carnegie Mellon University:
`http://www.speech.cs.cmu.edu`

IDIAP Speech Processing Group:
`http://old-www.idiap.ch/speech/speechNF.html`

Center for spoken language understanding at Oregon Health and Science Univercity:
`http://cslu.cse.ogi.edu`

Center for Spoken language research at Univercity of Colorado:
`http://cslr.colorado.edu`

Spoken Language Systems at MIT Laboratory for Computer Science:
`http://www.sls.csail.mit.edu/sls/sls-blue-nospec.html`

The International Computer Science Institute Speech Group at Berkeley:
`http://www.icsi.berkeley.edu/Speech/`

Speech processing and Auditory perception laboratory at UCLA:
`http://www.icsl.ucla.edu/~spapl`

Various links of research groups:
`http://mambo.ucsc.edu/psl/speech.html`

Bibliography

[1] O. Adeyemi and F. G. Boudreaux-Bartels. Improved accuracy in the singularity spectrum of multifractal chaotic time series. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-97*, Munich, Germany, 1997.

[2] J. B. Allen. How do humans process and recognize speech? *IEEE Trans. Speech Audio Processing*, 4(2):567–577, 1994.

[3] Y. Ashkenazy. The use of generalized information dimension in measuring fractal dimension of time series. *Physica A*, 271(3-4):427–447, 1999.

[4] M. Banbrook, S. McLaughlin, and I. Mann. Speech characterization and synthesis by nonlinear methods. *IEEE Trans. Speech Audio Processing*, 7:1–17, 1999.

[5] A. Barney, C. H. Shadle, and P. O. A. L. Davies. Fluid flow in a dynamic mechanical model of the vocal folds and tract. i. measurements and theory. *J. Acoust. Soc. Am.*, 105(1):444–455, 1999.

[6] H.-P. Bernhard and G. Kubin. Speech production and chaos. In *XIIth International Congress of Phonetic Sciences*, Aix-en-Provence, France, Aug 19-24, 1991.

[7] H. Bourlard and S. Dupont. A new asr approach based on independent processing and re-combination of partial frequency bands. In *Proc. International Conference on Speech and Language Processing, ICSLP-96*, pages 426–429, Philadelphia, USA, 1996.

[8] J. Chen, D. Dimitriadis, H. Jiang, Q. Li, T. A. Myrvoll, O. Siohan, and F. K. Soong. Bell labs approach to aurora evaluation on connected digit recognition. In *Proc. International Conference on Speech and Language Processing, ICSLP-02*, pages 462–465, Denver, CO, USA, September 2002.

[9] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust., Speech, Signal Processing*, 28(4):357–366, 1980.

[10] D. Dimitriadis and P. Maragos. Robust energy demodulation based on continuous models with application to speech recognition. In *Proc. European Conference on Speech Communication and Technology, Eurospeech-03*, Geneva, Switzerland, September 2003.

[11] R. Drullman, J. Festen, and R. Plomp. Effect of reducing slow temporal modulations on speech reception. *J. Acoust. Soc. Am.*, 95(5):2670–2680, 1994.

[12] R. Drullman, J. Festen, and R. Plomp. Effect of temporal smearing on speech reception. *J. Acoust. Soc. Am.*, 95(2), 1994.

[13] H. Fletcher. *Speech and Hearing in Communication.* Krieger, New York, 1953.

[14] S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Trans. Acoust., Speech, Signal Processing*, 29(2):254–272, 1981.

[15] B. Gajic and K. K. Paliwal. Robust feature extraction using subband spectral centroid histograms. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-01*, volume 1, pages 85–88, 2001.

[16] O. Ghitza. Auditory nerve representation as a front-end in a noisy environment. *Computer Speech and Language*, 2(1):109–130, 1987.

[17] O. Ghitza. Auditory nerve representation as a basis for speech processing. In S. Furui and M. M. Sondhi, editors, *Advances in Speech Signal Processing*, pages 453–486. Marcel Dekker, New York, 1992.

[18] S. Haykin and J. Principe. Making sense of a complex world. *IEEE Signal Processing Magazine*, page 6681, May 1998.

[19] H. Hermansky. An efficient speaker independent automatic speech recognition by simulation of some properties of human auditory perception. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-87*, pages 1156–1162, 1987.

[20] H. Hermansky. Perceptual linear predictive (plp) analysis of speech. *J. Acoust. Soc. Am.*, 87(4):1738–1752, 1990.

[21] H. Hermansky and N. Morgan. Rasta processing of speech. *IEEE Trans. Speech Audio Processing*, 2(4):578–589, 1994.

[22] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn. Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP). In *Proc. European Conference on Speech Communication and Technology, Eurospeech-91*, pages 578–589, 1991.

[23] H. Hermansky and S. Sharma. Traps - classifiers of temporal patterns. In *Proc. International Conference on Speech and Language Processing, ICSLP-98*, 1998.

[24] H. Hermansky, S. Tibrewala, and M. Pavel. Towards asr on partially corrupted speech. In *Proc. International Conference on Speech and Language Processing, ICSLP-96*, pages 462–465, Philadelphia, USA, 1996.

[25] H. Herzel, D. Berry, I. Titze, and I. Steinecke. Nonlinear dynamics of the voice: Signal analysis and biomechanical modeling. *CHAOS*, 5(1):30–34, 1995.

[26] T. Houtgast and H. J. M. Steeneken. A review of the mtf concept in room acoustics and its use for estimating speech intelligibility in auditoria. *J. Acoust. Soc. Am.*, 77(3):1069–1077, 1985.

[27] M.J. Hunt and C. Lefebvre. Speech recognition using a cochlear model. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-86*, pages 1979–1982, 1986.

[28] T. Irino and R. D. Patterson. A time-domain, level-dependent auditory filter: The gammachirp. *J. Acoust. Soc. Am.*, 101:412–419, 1997.

[29] F. Jabloun, A. E. Cetin, and E. Erzin. Teager energy based feature parameters for speech recognition in car noise. *IEEE Signal Processing Lett.*, 6(10), 1999.

[30] J. F. Kaiser. Some observations on vocal tract operation from a fluid flow point of view. In I. R. Titze and R. C. Scherer, editors, *Vocal Fold Physiology: Biomechanics, Acoustics and Phonatory Control*, pages 358–386. Denver Center for Performing Arts, Denver, CO, 1983.

[31] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel. On the importance of various modulation frequencies for speech recognition. In *Proc. European Conference on Speech Communication and Technology, Eurospeech-97*, pages 1079–1082, Rhodes, Greece, 1997.

[32] I. Kokkinos and P. Maragos. Nonlinear speech analysis using models for chaotic systems. *IEEE Trans. Acoust., Speech, Signal Processing*, to be published.

[33] G. Kubin and W. B. Kleijn. Time-scale modification of speech based on a nonlinear oscillator model. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-94*, 1994.

[34] A. Kumar and S. K. Mullick. Nonlinear dynamical analysis of speech. *J. Acoust. Soc. Am.*, 100(1):615–629, 1996.

[35] B. Mandelbrot. *The Fractal Geometry of Nature*. Freeman, NY, 1982.

[36] P. Maragos. Fractal aspects of speech signals: Dimension and interpolation. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-91*, 1991.

[37] P. Maragos, J. F. Kaiser, and T. F. Quatieri. Energy separation in signal modulations with application to speech analysis. *IEEE Trans. on Signal Processing*, 41(10):3024–3051, October 1993.

[38] P. Maragos and A. Potamianos. Fractal dimensios of speech sounds: Computation and application to automatic speech recognition. *J. Acoust. Soc. Am.*, 105(3):1925–1932, 1999.

[39] C. Nadeu, D. Macho, and J. Hernardo. Time and frequency filtering of filter-bank energies for robust hmm speech recognition. *Speech Communication*, 34:93–114, 2001.

[40] S. Narayanan and A. Alwan. A nonlinear dynamical systems analysis of fricative consonants. *J. Acoust. Soc. Am.*, 97(4):2511–2524, 1995.

[41] W.-C. Pai and P. C. Doerschuk. Statistical am-fm models, extended kalman filter demodulation, cramer-rao bounds, and speech analysis. *IEEE Trans. on Signal Processing*, 48(8):2300–2313, August 2000.

[42] V. Pitsikalis, I. Kokkinos, and P. Maragos. Nonlinear analysis of speech signals: Generalized dimensions and lyapunov exponents. In *Proc. European Conference on Speech Communication and Technology, Eurospeech-03*, Geneva, Switzerland, September 2003.

[43] V. Pitsikalis and P. Maragos. Speech analysis and feature extraction using chaotic models. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-02*, Orlando, USA, May 2002.

[44] A. Potamianos and P. Maragos. A comparison of the energy operator and the hilbert transform approach to signal and speech demodulation. *Signal Processing*, 37:95–120, May 1994.

[45] A. Potamianos and P. Maragos. Speech formant frequency and bandwidth tracking using multiband energy demodulation. *J. Acoust. Soc. Am.*, 99(6):3795–3806, June 1996.

[46] A. Potamianos and P. Maragos. Speech analysis and synthesis using an am-fm modulation model. *Speech Communication*, 28:195–209, 1999.

[47] T. F. Quatieri and E. M. Hofstetter. Short-time signal representation by nonlinear difference equations. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'90, Albuquerque*, April 1990.

[48] L. R. Rabiner and B.H.Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[49] S. Sabanal and M. Nakagawa. The fractal properties of vocal sounds and their application in the speech recognition model. *Chaos, Solitons and Fractals*, 7 No11:1825–1843, 1996.

[50] S. Seneff. Pitch and spectral estimation of speech based on an auditory synchrony model. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-84*, pages 3621–3624, 1984.

[51] S. Seneff. A joint synchrony/mean-rate model of auditory speech processing. *Journal of Phonetics*, 16(1):57–76, 1988.

[52] H. M. Teager and S. M. Teager. Evidence for nonlinear sound production mechanisms in the vocal tract. In *Speech Production and Speech Modelling W.J. Hardcastle & Marchal, Eds., NATO ASI Series D*, volume 55, 1989.

[53] N. Tishby. A dynamical systems approach to speech processing. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-90*, 1990.

[54] B. Townshend. Nonlinear prediction of speech signals. *IEEE Trans. Acoust., Speech, Signal Processing*, 1990.

[55] H. H. Yang, S. J. Van Vuuren, and H. Hermansky. Relevancy of time-frequency features for phonetic classification measured by mutual information. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-99*, 1999.

[56] Y.M.Cheng and D.O'Shaughnessy. Speech enhancement based conceptually on auditory evidence. *IEEE Trans. Acoust., Speech, Signal Processing*, 39(9):1943–1954, 1991.

Chapter 8

Speech analysis

8.1 Description of the problem

Among the variety of acoustic signals the ear is exposed to, speech is one of the few for which an approximated production model is available. All the speech signals thus share common characteristics that are of interest for various areas in automatic speech processing: speech synthesis, automatic speech recognition, speech synthesis, hearing aids, telecommunications…

The production of a speech signal $s(n)$ can be approximated as the convolution of an excitation signal $e(n)$ by a filter corresponding to the vocal tract $h(n)$: $s(n) = e(n) * h(n)$. In the spectral domain this equation becomes $S(\omega) = E(\omega) \times H(\omega)$.

Speech analysis aims at finding contributions of the excitation signal (noise in the case of unvoiced sounds or periodic signal for voiced sounds) and of the vocal tract filter, and separating these two contributions.

The excitation signal is produced by either the vibration of vocal folds (periodic excitation) or the turbulent flow of air somewhere in the vocal tract (noise excitation).

The filter corresponding to the vocal tract depends its geometrical shape and thus depends on the position of speech articulators, i.e. the lower jaw, position and shape of the tongue body, position of the tongue apex, aperture and rounding of lips, larynx and velum position.

The three main areas of research in speech analysis are **(i)** spectral analysis, **(ii)** the determination of the fundamental frequency and **(iii)** automatic formant tracking.

8.2 State of the art

8.2.1 Spectral analysis

The main objective of spectral analysis is to get a relevant spectrum of the vocal tract filter. The main challenge is to get an information as independent as possible of the excitation. The spectrum of a voiced excitation is a spectrum of regularly spaced (the fundamental frequency F0) lines. The vocal tract spectrum is thus "sampled" by F0. The higher F0, the less the precision of the vocal tract spectrum. This means that the vocal tract spectrum is not well approximated for high F0 voices, i.e. for female speakers and children.

The first tool for spectral estimation is the well know Fourier transform. The size of the analysing window influences the frequency smoothing. When the time window is small (approximately 4 ms) compared to the F0 period the smoothing is strong and resonance frequencies of the vocal tract can be seen. However, the spectrum depends on the location of the window with respect to F0 periods and this solution is only used to display speech spectrograms used by phoneticians. When the time window is long (approximately 32 ms) compared to the F0 periods the smoothing is weak and thus, harmonics (multiples of F0) are visible. Fourier analysis is the basic tool for spectral analysis of speech. One of the difficulties is the choice of the size and position or the analysing window with respect to the periods of the fundamental frequency. There exists some reassignment methods that reduce the effect of the window location by moving the spectral energy where it should appear [11].

Beside Fourier transform there are two main families of spectral analyses in speech processing. The first is that of linear prediction methods that correspond to the assumption of an all pole model. The idea is to approximate the speech sample $s(n)$ by $\sum_{k=1}^{k=p} a_k s(n-k)$. Parameters a_k are obtained by minimising the squared error over an analysing window. The spectrum can be easily calculated from these coefficients. The advantage of this method is its low computation cost. Its main disadvantage is that the underlying all pole hypothesis is only valid for oral vowels and not for nasal vowels and consonants. There are a number of derived methods, the selective linear prediction for instance, that enable the analysis to be applied over a limited spectral region [10].

The second family is that of cepstral smoothing. The principle is to eliminate the contribution of the excitation in a Fourier spectrum calculated over a rather long window (approximately 32 ms). The underlying idea is to compute the inverse Fourier transform of the spectrum, called cepstrum, to isolate the contribution of harmonics. This contribution appears as an isolated peak that can be easily filtered. An extra Fourier transform gives the smoothed spectrum (see [12] for a more detailed presentation). Derived from this idea Davis and Mermelstein [2] proposed Mel cepstra that are calculated from the energy vector computed over a Mel[1] filter bank after the Fourier transform. This enables a more concise representation of speech spectra and removes speaker variability to some extent. These Mel cepstral coefficients are widely used as input spectral vectors in automatic speech recognition. There are interesting methods derived from cepstral analysis. The spectral envelope method [7] is an iterative version of the standard cepstral analysis that enables a good energy approximation in the vicinity of harmonics. This gives a better approximation and a more steady estimation of spectral peaks. The discrete cepstral analysis [3] approximates a number of spectral points by a sum of cosinusoids. The spectral points have to be chosen carefully to represent relevant spectral information. One therefore chooses harmonics or other spectral peaks.

Despite their theoretical interest wavelets are not often used.

8.3 Determination of the fundamental frequency

The fundamental frequency is the frequency of vocal fold vibration. When vocal folds vibrate, the vocal tract is excited by a periodic signal which gives rise to voiced sounds. The fundamental frequency, called F0, often improperly called pitch which is related to perception, plays a central role in speech analysis. Indeed, the fundamental frequency the is the prosody parameter that gives intonation and, as explained above, has a major impact of the shape of the speech spectrum. Its determination thus has received considerable attention. Furthermore, the fundamental frequency is very important within the framework of speech coding and synthesis. There are basically two kinds of determination method: (i) methods that operate in the time domain as the famous autocorrelation method [5, 12], and (ii) methods that operate in the frequency domain as the cepstrum or spectral comb methods. The difficulties lie in the false determination of double F0 or half F0 values, and in the voicing decision, i.e. how to decide whether a speech window corresponds to voiced or unvoiced speech. These problems and the processing of noisy signal are the current challenges.

8.4 Automatic formant tracking

As explained in the introduction one of the objectives in speech analysis is to find spectral information related to the vocal tract filter. Formants are spectral peaks that correspond to the resonance frequencies of the vocal tract. As formants directly derive from the geometrical shape of the vocal tract, they may be exploited to recover the place of articulation and thus identify sounds pronounced, especially vowels and other vocalic sounds. Formant tracks are utilised

[1] The Mel frequency is a non linear frequency scale that approximates the frequency resolution of the ear.

to pilot formant synthesisers [6], to study coarticulation effects, vowel perception, articulatory phenomena and in some rare cases to provide a speech recognition with additional data [4].

Given the potential interest of formant data numerous works have been dedicated to the design of automatic formant tracking algorithms. Nature and complexity of the problem explain the success of dynamic programming algorithms [13, 14]. The first steps of these algorithms is the extraction of formant candidates at each frame of the speech signal. The second stage is dynamic programming that utilises the evaluation of transition costs between two frames. Other algorithms aim at explaining the acoustic signal [1] or the spectrogram energy. In [9, 8] we showed how active curves could be used to track formants. The underlying idea is to deform initial rough estimates of formants under the influence of the spectrogram to get regular tracks close to lines of spectral maxima which are potential formants.

8.5 Perspectives

Despite of constant efforts, automatic formant tracking remains an open challenge. This challenge is all the more important since good formant estimates could be exploited in various areas of automatic speech processing: speech recognition, synthesis, speaker identification... As formants are closely related to speech production, analysis by synthesis methods are the most promising approach. Moreover, progress in speech production in terms of talking heads and speaker adaptation could provide additional constraints to improve results.

In the domain of F0 determination the robustness is still an open challenge, especially when performances are compared against those of human listeners who are able to detect speech even in a strong ambiant noise. Improvement of F0 determination techniques is probably closely linked to the development of new spectral analyses that achieve a better precision in the localization of energy. Another potential source of improvement is a better cooperation with psychoacoustics that focuses on the human perception of acoustic signals and investigates the reasons why the ear is far better than early processing of speech.

Bibliography

[1] I. Bazzi, A. Acero, and L. Deng. An expectation maximization approach for formant tracking using a parameter-free non-linear predictor. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Hong-Kong, May 2003.

[2] S. B. Davis and P. Mermelstein. Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-28(4):357–366, August 1980.

[3] T. Gallas and X. Rodet. Generalized functionnal approximation for source-filter system modelling. In *Proceedings of European Conference on Speech Technology*, Genova, Italy, September, 1991.

[4] Philip N. Garner and Wendy J. Holmes. On the robust incorporation of formant features into hidden makov mdels for automatic speech recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 1–4, Seattle, USA, May 1998.

[5] W. J. Hess. *Pitch Determination of Speech Signals - Algorithms and Devices.* Springer Berlin, 1983.

[6] J. N. Holmes. Formant synthetisers: cascade or parallel ? *Speech Communication*, 2:251–273, 1983.

[7] S. Imai and Y. Abe. Spectral envelope extraction by improved cepstral method. *Trans. IECE*, J62-A(4):217–223, 1979 (en japonais).

[8] Y. Laprie. A concurrent curve strategy for formant tracking. In *International Conf. on Spoken Language Processing - ICSLP2004, Jegu, Korea*, October 2004.

[9] Y. Laprie and M.-O. Berger. Cooperation of regularization and speech heuristics to control automatic formant tracking. *Speech Communication*, 19(4):255–270, October 1996.

[10] J.D. Markel and A.H. Gray. Automatic formant trajectory estimation. In *Linear Prediction of Speech*, chapter 7. Springer-Verlag, Berlin Heidelberg New York, 1976.

[11] F. Plante, G. Meyer, and W.A. Ainsworth. *IEEE Transactions on Speech and Audio Processing*, 6(3):282–287, 1998.

[12] L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals.* Prentice-Hall, Englewood Cliffs, N.J., 1978.

[13] D. Talkin. Speech formant trajectory estimation using dynamic programming with modulated transition costs. *Journal of the Acoustical Society of America*, S1:S55, March 1987.

[14] K. Xia and C. Epsy-Wilson. A New Strategy of Formant Tracking based on Dynamic Programming. In *International Conf. on Spoken Language Processing - ICSLP2000, Beijing, China*, October 2000.

Chapter 9

State of the art in acoustic-to-articulatory inversion

9.1 Description of the problem

The acoustic-to-articulatory inversion consists in recovering the vocal tract shape dynamics from the acoustical speech signal, that can possibly be completed by the knowledge of the speaker's face. Estimating the vocal tract shape from the speech signal has received considerable attention because it offers new perspectives for speech processing. Indeed, it would enable knowing how a speech signal has been articulated. This potential knowledge could give rise to a number of breakthroughs in automatic speech processing. For speech coding, this would allow spectral parameters to be replaced by a small number of articulatory parameters[8] that vary slowly with time. In the case of automatic speech recognition the location of critical articulators could be exploited[7] in a view of discarding some acoustical hypotheses. For language acquisition and second language learning this could offer articulatory feedbacks. Lastly, in the domain of phonetics, inversion would enable knowing how sounds were articulated without requiring medical imaging techniques.

Basically, the acoustic-to-articulatory inversion is an acoustical problem and data are therefore formant frequencies, i.e. the resonance frequencies of the vocal tract. However, formants cannot be extracted easily from the speech signal and most of the existing methods thus need to be generalized to accept standard spectral input data (for instance, Mel Frequency Cepstral Coefficients). This represents a first difficulty to solve the inverse problem.

The main difficulty is that acoustic-to-articulatory inversion is an ill-posed problem. There is no one-to-one mapping between acoustic and articulatory domains and there are thus an infinite number of vocal tract shapes that can produce the same formants and thus the same speech signal. Indeed, the problem is under-determined because there are more unknowns than data. Generally the first three formant frequencies are used as data and there are more than six articulatory parameters, for instance seven in the case of the famous Maeda's model [5]. One important issue is thus to add constraints that are both sufficiently restrictive and realistic from a phonetic point of view.

9.2 State of the art

Most of the acoustic-to-articulatory methods rest on an analysis-by-synthesis approach. Indeed, among the variety of acoustical signals the ear is exposed to, speech is one of the few an approximated production model is available for. The synthesis corresponds to the use of an articulatory synthesizer that computes speech spectra or formants from articulatory or geometrical parameters. Adjusting the faithfulness of the articulatory synthesizer with respect to the static and dynamic characteristics of the human vocal tract allows constraints to be put on the shape of inverse solutions and thus the number of inverse solutions to be reduced. The simplest articulatory models approximate the vocal tract shape geometrically as a set of concatenated uniform tubes (generally between 6 and 8 tubes). Their main weakness is that they are unable to render the vocal tract shape and that the total length of the vocal tract is an extrinsic parameter that has to be calculated independently. More faithful models can be built from medical imaging of the vocal tract. The 2D sagittal articulatory model proposed by Meada [5] was derived from X-ray images and describes the vocal tract through seven deformation modes. More recent models based on MRI images describe the 3D shape of the vocal tract [2] articulatory parameters. The strong noise and lying position imposed by a MRI machine create some discrepancies between normal and MRI modes of articulation that cannot be evaluated with precision. Even if they are more flexible than concatenation of uniform tubes they require prior adaptation before being used for any speaker.

The number of parameters of an articulatory model generally ranges from 6 to 9 and the solution space cannot be explored during the inversion. Inversion methods therefore exploit some explicit or implicit table lookup method to recover at each time frame the set of articulatory parameters. Explicit table lookup methods required efficient sampling and representation methods [6] to limit the table size. Implicit table lookup methods often use neural networks [9, 1] but cannot guarantee that a uniform acoustic resolution is achieved.

Once inverse solutions have been recovered at each time frame of the speech signal articulatory trajectories are then built from these local solutions by some optimal path search algorithm, generally dynamic programming [3]. Other methods exploit regularization techniques or physical constraints [4] to obtain smooth trajectories.

The evaluation of an acoustic-to-articulatory inversion procedure comprises two aspects. The first is the acoustical faithfulness and ensures that inverted data are able to reproduce a speech signal as close as possible to the original. The closeness is generally evaluated by measuring the distance between original and synthetic formant frequencies.

The second aspect is that of the articulatory faithfulness. Unlike domains where data can be acquired easily (automatic speech recognition for instance) acquisition here requires medical imaging techniques which are often expensive (MRI, X-ray, electromagnetic articulography), hazardous (X-ray), perturb articulation (noise produced by medical machine, MRI especially), not fast enough to capture continuous speech (MRI), not precise enough (electromagnetic articulography). This explains that very few data are available all the more because some are required to build or adapt the articulatory model.

Current inversion techniques mostly concern vowels and sequences of vowels for one speaker. This domains thus necessitates substantial efforts to provide a general purpose inversion framework.

Given these results perspectives concern the incorporation of additional constraints in order to reduce the under-determination of the inverse problem. These constraints could be static and provided by phonetics to penalize unrealistic vocal tract shapes given a formant 3-tuple. But they could be dynamic and provided by computer vision techniques used to track visible articulators, i.e. lips and lower jaw. The additional knowledge of visible articulators gives rise to two or three articulatory parameters (jaw position, lip aperture and protrusion) and therefore considerably reduces the under-determination of the problem. This corresponds to a multimodal audio-visual approach of the inverse problem.

The second perspective is the use of standard spectral parameters as input data and the development of a general inversion method for all the classes of speech sounds. This is a hard problem and requires to be able to derive the acoustical behavior of the articulatory model from a standard model whatever the geometrical characteristics of an arbitrary speaker. Furthermore, this implicitly requires the development of a general articulatory model that works for consonants (voiced or unvoiced) as well as for vowels.

Bibliography

[1] G. Bailly, C. Abry, R. Laboissièrre, P. Perrier, and J.-L. Schwartz. Inversion and speech recognition. In J. Vandewalle, R. Boite, M. Mooner, and A. Osterlinck, editors, *Signal processing VI: Theories and Applications*, volume 1, pages 159–164, Brussels, Belgium, August 1992. Elsevier.

[2] O. Engwall. Modelling of the vocal tract in three dimensions. In *Proceedings of the 6th European Conference on Speech Communication and Technology*, pages 113–116, Budapest, September, 1999.

[3] S. K. Gupta and J. Schroeter. Pitch-synchronous frame-by-frame and segment-based articulatory analysis by synthesis. *Journal of Acoustical Society of America*, 94(5):2517–2530, Nov 1993.

[4] Schoentgen J. and S Ciocea. Kinematic formant-to-area mapping. *Speech Communication*, 21:227–244, 1997.

[5] S. Maeda. Un modèle articulatoire de la langue avec des composantes linéaires. In *Actes 10èmes Journées d'Etude sur la Parole*, pages 152–162, Grenoble, Mai 1979.

[6] Slim Ouni and Yves Laprie. Exploring the Null Space of the Acoustic-to-Articulatory Inversion Using a Hypercube Codebook. In *Eurospeech, Aalborg, Danemark*, volume 1, pages 277–280, September 2001.

[7] R.C. Rose, J. Schroeter, and M.M. Sondhi. An investigation of the potential role of speech production models in automatic speech recognition. In *Proceedings of International Conference on Spoken Language Processing*, volume 2, pages 575–578, Yokohama, Japan, September, 1994.

[8] J. Schroeter and M. M. Sondhi. Speech coding based on physiological models of speech production. In S. Furui and M. M. Sondhi, editors, *Advances in Speech Signal Processing*, pages 231–267. Dekker, New York, 1992.

[9] K. Shirai and T. Kobayashi. Estimating articulatory motion from speech wave. *Speech Communication*, 5(2):159–170, 1986.

Part III

State of the art in Natural Language Processing

Chapter 10

Introduction

Of the three media covered in the MUSCLE network: image, sound and text, text is the easiest media to treat. The features to use in text, those related to meaning, i.e. words, are easy to manipulate. The main problem with understanding text derives from the great number of features (modern languages possess hundred of thousands of words forms) and ambiguity (different words can be used to express the same meaning). Added to these problems are the number of different languages that are used. For example, there are 20 languages available on the home page of the European community. The number of features make the problem of natural language processing important for understanding.

This section contains three parts concerning the state-of-the-art in Natural Language Processing for Multimedia Understanding. The first part covers building language models. Language models are important for both recovering signal in text and speech processing. Words do not appear in a random order in text. Understanding this order allows use to separate importance occurrences from unimportant (from the viewpoint of meaning). Language models help use also map words and structures between words onto meaning. After this section there is a section on information retrieval. Information retrieval concerns the problems involved in reducing the sequential order and variety of text to normalized units that can be efficiently stored and matched against other texts (queries). It is this reduction that allows us to explore meaning and operationally understand text, that is, we can say that this text corresponds to that text. When one of the texts represents some information need, then providing the relevant text in response to this need simulates understanding.

The third part of this section covers the use of natural language processing techniques for processing text associated with images. Considering the use of text in content based image retrieval (CBIR) systems, this last part describes the techniques drawn from natural language processing that are currently used by research teams participating in CBIR competitions such as ImageCLEF and TRECVID. It is hoped that this description will be useful for understanding the current state of the art in natural language processing particularly for applications involving multiple media (of which one of the media is text). General recommendations on the issues in natural language processing which still require attention and research conclude the section.

Chapter 11

Language Modelling

11.1 Introduction

Statistical natural language processing (NLP) attempts to do statistical inference for the field of natural language by taking some data and making some inferences about the distribution under which they were generated. The most classic task of statistical estimation is language modelling where the problem lies in predicting the next word given the previous words. In general, statistical language modelling (SLM) aims to estimate the probability distribution of various linguistic units, such as words, sentences, or even whole documents, in order to encode linguistic information in a way to be useful to systems which process human language.

Statistical language modelling have been applied in a wide range of natural language processing tasks including speech or optical character recognition, handwriting recognition, machine translation, spelling correction, information retrieval, and many more.

In more detail, in speech recognition systems, sophisticated statistical models which included linguistic knowledge were developed in order to transcribe spoken text into a written form [Jel98, Cla99]. In the field of machine translation, the proposed statistical models along with tagging and parsing techniques minimized the amount of ambiguity and variability of both the source and target languages which result from language-specific phenomena, such as idiomatic expressions, multiple sense words, word order constraints and others [BCP$^+$90, BF95]. The benefits from language modelling were also exploited in optical character recognition where the original text must be recovered from a potentially distorted image, and in spelling correction where the "correct" text is sought [Ros94]. In information retrieval, a language modelling approach was first proposed in [PC98, Hie98], and later described in terms of a “noisy channel” model in [BL99]. In the following years, successful applications of the LM approach to a number of retrieval tasks have also been reported [XWN01, LCC02, SJCO02], and research carried out by a number of groups has confirmed that the language modelling approach is a theoretically attractive and potentially very effective probabilistic framework for studying information retrieval problems [CL03].

Automatic word categorization is another important field of application in statistical natural language processing. Research in this area points out that it is possible to determine the structure of a natural language by examining the regularities of the statistics of the language [Fin93]. Various clustering algorithms which partition, in a hierarchical or non-hierarchical way, a set of objects into groups/clusters according to a predefined objective function have been induced in order to construct clusters of similar words. This attempt is important since clustering similar words may alleviate problems that arise in other fields of NLP, such as speech recognition and information retrieval.

The most commonly used language models are based on the n-gram language model [Jel98]. However, there are many improvements over this simple model, including skipping models, higher order n-grams, caching and clustering sentence-mixture models, all of which are presented in the following sections.

11.2 Language Modelling Techniques/Approaches

Given a sequence of M words $\mathbf{W} = w_1, w_2, \ldots, w_M$ a language model estimates the *a priori* probability $P(\mathbf{W}) = P(w_1, w_2, \ldots, w_M)$. This probability is typically broken down into its component probabilities using the Markov chain rule as follows:

$$P(\mathbf{W}) = P(w_1) \times P(w_2|w_1) \times \cdots \times P(w_M|w_1 \ldots w_{M-1})$$

or, in a more condensed form

$$P(\mathbf{W}) = P(w_1) \prod_{i=2}^{M} P(w_i|w_1, \ldots, w_{i-1}). \tag{11.1}$$

11.2.1 n-Gram Model

Since, the estimation of such a large set of probabilities from a finite set of training data is not feasible, the n-gram model is employed where the current word is predicted based on the preceding $n-1$ words expecting that most of the relevant syntactic information lies in the immediate past. The probability of the sequence of M words $P(\mathbf{W})$ is then expressed by

$$P(\mathbf{W}) \approx P(w_1) \prod_{i=2}^{M} P(w_i|w_{i-n}, \ldots, w_{i-1}). \tag{11.2}$$

The probabilities of the above equation are estimated by means of the relative frequency approach as follows

$$P(w_i|w_{i-n}, \ldots, w_{i-1}) \simeq \frac{c(w_{i-n}, \ldots, w_{i-1}, w_i)}{c(w_{i-n}, \ldots, w_{i-1})} \tag{11.3}$$

where $c(.)$ represents the number of occurrences of the corresponding event in the training corpus.

The number of parameters in n-gram models increases considerably as n increases, resulting in an increase in the size of the model and the data required for training. Thus, the computation of the *a priori* probability for large n is difficult imposing generally low values for n, usually 2 or 3. That is, it is assumed that the probability of a word depends on only the preceding word (bigram) or the two previous words (trigram).

For the trigram model which has been shown to work well in practice for most applications, $P(\mathbf{W})$ is given by:

$$P(\mathbf{W}) \approx P(w_1)P(w_2|w_1)\prod_{i=3}^{M} P(w_i|w_{i-2}, w_{i-1}). \tag{11.4}$$

The above trigram probability is estimated from their occurrences in the training corpus in a way similar to (11.3):

$$P(w_i|w_{i-2}w_{i-1}) \approx \frac{c(w_{i-2}w_{i-1}w_i)}{c(w_{i-2}w_{i-1})} \tag{11.5}$$

where $c(w_{i-2}w_{i-1}w_i)$ and $c(w_{i-2}w_{i-1})$ represent the number of occurrences of word sequences $< w_{i-2}w_{i-1}w_i >$ and $< w_{i-2}w_{i-1} >$ respectively.

Smoothing

It is obvious that the above approximation can be very noisy, since there are many trigrams that never occur in the training corpus. This problem is known as the sparse data problem or the zero frequency problem. To illustrate it, if we consider a vocabulary of size $|V| = 20000$, then we have $|V \times V| = 4 \times 10^8$ (400 million) possible word bigrams, but the training corpus consists rarely of more than 10×10^6 (10 million) words, that is only 2.5% of all bigrams can be observed. It is obvious that for trigram models the effect will be even more disastrous. To alleviate this problem various smoothing techniques are applied that take some probability away from some occurrences and redistribute it to other events. These techniques are further described below.

i. Additive Smoothing

   It is one of the simplest types of smoothing used in practice. In this method we pretend that each n-gram occurs δ times more than it does, where $0 < \delta \leq 1$ and w_{i-n+1}^{i} is the sequence of words $w_{i-n+1} \dots w_i$, i.e.,

$$p_{add}(w_i|w_{i-n+1}^{i-1}) = \frac{\delta + c(w_{i-n+1}^{i})}{\delta|V| + \sum_{w_1} c(w_{i-n+1}^{i})}. \tag{11.6}$$

   Authors in [GWC90, GWC94] have argued that this method generally performs poorly.

ii. Good-Turing Estimate

The Good-Turing estimate [Goo53] is central to many smoothing techniques. The Good- Turing estimate states that for any n-gram that occurs r times, we should pretend that it occurs r^* times where

$$r^* = (r+1) \cdot \frac{n_{r+1}}{n_r} \tag{11.7}$$

and n_r is the number of n-grams that occur exactly r times in the training data. To convert this count to a probability, we just normalize: For an n-gram a with r counts, we take

$$p_{GT}(a) = \frac{r^*}{\sum_{r=0}^{\infty} n_r r^*}. \tag{11.8}$$

The Good-Turing estimate cannot be used when $n_r = 0$. It is

generally necessary to "smooth" n_r, e.g., to adjust the n_r so that they are all above zero. In practice, the Good-Turing estimate is not used by itself for n-gram smoothing, because it does not include the combination of higher-order models with lower-order models necessary for good performance. However, it is used as a tool in several smoothing techniques.

iii. Jelinek-Mercer Smoothing

In this method [JFM80], we linearly interpolate n-gram models of high class and n-gram models of low class as follows:

$$\begin{aligned} p_{interp}(w_i|w_{i-n+1}^{i-1}) &= \lambda_{w_{i-n+1}^{i-1}} p_{ML}(w_i|w_{i-n+1}^{i-1}) + \\ &+ (1-\lambda_{w_{i-n+1}^{i-1}}) p_{interp}(w_i|w_{i-n+2}^{i-1}). \end{aligned} \tag{11.9}$$

That is, the nth-order smoothed model is defined recursively as a linear interpolation between the nth-order maximum likelihood model and the $(n-1)$th order smoothed model. To end the recursion, we can take the smoothed 1st-order model to be the maximum likelihood distribution, or we can take the smoothed 0th-order model to be the uniform distribution $p_{unif}(w_i) = \frac{1}{|V|}$. In order to estimate $\lambda_{w_{i-n+1}^{i-1}}$, the data that will be used should be different from the data used to calculate p_{ML}. The optimal $\lambda_{w_{i-n+1}^{i-1}}$ will be different for different histories w_{i-n+1}^{i-1}. However, training each parameter $\lambda_{w_{i-n+1}^{i-1}}$ for every w_{i-n+1}^{i-1} independently is not generally felicitous, so [Che96] suggests that bucketing $\lambda_{w_{i-n+1}^{i-1}}$ according to the average number of counts per nonzero element in a distribution according to $\frac{\sum_{w_i} c(w_{i-n+1}^{i})}{w_i : c(w_{i-n+1}^{i}) > 0}$.

iv. Katz Smoothing

Katz Smoothing [SK87] extends the intuitions of the Good-Turing estimate by adding the combination of higher-order models with lower-order models. For the bigram models we have the following equation

$$c_{katz}(w_{i-1}^{i}) = \begin{cases} d_r r & r > 0 \\ \alpha(w_{i-1}) p_{ML}(w_i) & r = 0 \end{cases} \tag{11.10}$$

with $r = c(w_{i-1}^{i})$. That is, all bigrams with a nonzero count r (except from some bigrams with $r > k$) are discounted according to a discount ratio d_r which is approximately $\frac{r^*}{r}$ where r^* is the discount predicted by the Good-Turing estimate. The counts subtracted from the nonzero counts are then distributed among the zero-count bigrams according to the next lower-order distribution, i.e., the unigram model.

v. Witten-Bell Smoothing

Witten-Bell Smoothing [BCCW90] was developed for the task of text compression, and can be considered to be an instance of Jelinek-Mercer smoothing. In particular, the nth-order smoothed model is defined recursively as a linear interpolation between the nth-order maximum likelihood model and the $(n-1)$th-order smoothed model

$$p_{WB}(w_i|w_{i-n+1}^{i-1}) = \lambda_{w_{i-n+1}^{i-1}} p_{ML}(w_i|w_{i-n+1}^{i-1}) + (1 - \lambda_{w_{i-n+1}^{i-1}}) p_{WB}(w_i|w_{i-n+2}^{i-1}). \tag{11.11}$$

According to Witten-Belley Smoothing to compute the parameters $\lambda_{w_{i-n+1}^{i-1}}$ we will need to use the number of unique words that follow the history w_{i-n+1}^{i-1} which is defined as

$$N_{1+}(w_{i-n+1}^{i-1}\bullet) = |\{w_i : c(w_{i-n+1}^{i-1}) > 0\}|. \tag{11.12}$$

So the parameters $\lambda_{w_{i-n+1}^{i-1}}$ are defined as

$$1 - \lambda_{w_{i-n+1}^{i-1}} = \frac{N_{1+}(w_{i-n+1}^{i-1}\bullet)}{N_{1+}(w_{i-n+1}^{i-1}\bullet) + \sum_{w_i} c(w_{i-n+1}^{i-1})}. \tag{11.13}$$

vi. Absolute Discounting

Absolute discounting [NEK94] involves the interpolation of higher and lower order models. In this method the higher-order distribution is created by subtracting a fixed discount $D \leq 1$ from each nonzero count. So we have

$$p_{abs}(w_i|w_{i-n+1}^{i-1}) = \frac{max\{c(w_{i-n+1}^{i}) - D, 0\}}{\sum_{w_i} c(w_{i-n+1}^{i})} + (1 - \lambda_{w_{i-n+1}^{i-1}}) p_{abs}(w_i|w_{i-n+2}^{i-1}). \tag{11.14}$$

vii. Kneser-Ney Smoothing

It is an extension of absolute discounting [KN95] where the lower-order distribution that one combines with a higher-order distribution is built in a novel manner. Specifically, for the bigram model the probability of unigram is not proportional to the number of occurrences of a word, but instead to the number of different words that it follows. So the probability of unigram equals to

$$p(w_i) = \frac{N_{1+}(\bullet w_i)}{N_{1+}(\bullet\bullet)} \tag{11.15}$$

where $N_{1+}(\bullet w_i) = |\{w_{i-1} : c(w_{i-1}w_i) > 0\}|$ is the number of different words w_{i-1} that precede w_i and $N_{1+}(\bullet\bullet) = \sum_{w_{i-1}} N_{1+}(w_{i-1}\bullet) = |\{(w_{i-1}, w_i) : c(w_{i-1}w_i) > 0\}| = \sum_{w_i} N_{1+}(\bullet w_i)$. Generalizing to higher-order models, we have that:

$$p(w_i|w_{i-n+2}^{i-1}) = \frac{N_{1+}(\bullet w_{i-n+2}^{i})}{N_{1+}(\bullet w_{i-n+2}^{i-1}\bullet)}. \tag{11.16}$$

viii. Modified Knesser-Ney Smoothing

Modified Knesser-Ney Smoothing [SG96] is an improved version of Knesser-Ney smoothing, where instead of using a single discount D for all nonzero counts as in Knesser-Ney Smoothing, we have three different parameters D_1 D_2 D_{3+} that are applied to n-grams with one, two or more counts, respectively. So we use the above formula

$$p(w_i|w_{i-n+1}^{i-1}) = \frac{c(w_{i-n+1}^{i}) - D(c(w_{i-n+1}^{i}))}{\sum_{w_i} c(w_{i-n+1}^{i})} + \gamma(w_{i-n+1}^{i-1})p_{KN}(w_i|w_{i-n+2}^{i-1}) \tag{11.17}$$

where

$$D(c) = \begin{cases} D_0 & c = 0 \\ D_1 & c = 1 \\ D_2 & c = 2 \\ D_{3+} & c \geq 3 \end{cases} \tag{11.18}$$

and

$$\gamma(w_{i-n+1}^{i-1}) = \frac{D_1 N_1(w_{i-n+1}^{i-1}\bullet) + D_2 N_2(w_{i-n+1}^{i-1}\bullet) + D_{3+} N_{3+}(w_{i-n+1}^{i-1}\bullet)}{\sum_{w_i} c(w_{i-n+1}^{i})}. \tag{11.19}$$

11.2.2 Higher-order n-Grams

The trigram assumption in many cases is inefficient and longer contexts are sometimes more helpful. For this reason, trigram models are extended to higher-order

N-Grams, such as 4-grams and 5-grams. That is, instead of computing probabilities of the form $P(w_i|w_{i-2}w_{i-1})$, probabilities of longer context are estimated, such that of the 5-gram model, $P(w_i|w_{i-4}w_{i-3}w_{i-2}w_{i-1})$.

Experiments with longer contexts showed little benefit since there are many cases where no sequence of the form $< w_{i-4}w_{i-3}w_{i-2}w_{i-1}w_i >$ is seen in the training data, forcing to backoff techniques or interpolation with lower-order n-grams, such as 4-grams, trigrams, bigrams, or even unigrams [Goo04].

In general, the efficiency of these models depends significantly on the smoothing technique employed. For example, the Interpolated Kneser-Ney smoothing works better with higher-order n-grams than with lower-order ones, but higher-order n-grams are often impractical due to memory limitations. The tradeoff between memory and performance typically requires heavy pruning of 4-grams and 5-grams, reducing the potential improvement from them.

11.2.3 Skipping Models

Skipping models [HAH+93, Ros94, NEK94, MHL+99, SO00] make use of the observation that when moving to higher-order n-grams, the chance of having seen the exact context before decreases while the chance of having seen a similar context, with the most of the words in it, increases.

In a 5-gram context, for instance, there are many subsets of the 5-gram to be considered, such as $P(w_i|w_{i-4}w_{i-3}w_{i-1})$ or $P(w_i|w_{i-4}w_{i-2}w_{i-1})$. These skipping 5-grams can be interpolated with a normal 5-gram, forming models such as

$$\lambda P(w_i|w_{i-4}w_{i-3}w_{i-2}w_{i-1})+\mu P(w_i|w_{i-4}w_{i-3}w_{i_1})+(1-\lambda-\mu)P(w_i|w_{i-4}w_{i-2}w_{i-1}) \tag{11.20}$$

where, as usual, $0 \leq \lambda \leq 1, 0 \leq \mu \leq 1$ and $0 \leq (1-\lambda-\mu) \leq 1$.

In another popular variation of the skipping model all component probabilities depend on two previous words, like a trigram, but the overall probability is N-gram-like. An example of such a model is the following

$$\lambda P(w_i|w_{i-2}w_{i-1}) + \mu P(w_i|w_{i-3}w_{i-1}) + (1-\lambda-\mu)P(w_i|w_{i-3}w_{i-2}) \tag{11.21}$$

where the overall probability is 4-gram like since it depends on $w_{i-3}w_{i-2}$ and w_{i-1}.

11.2.4 Caching Models

Caching Models ([KM90, KM92, JMRS91]) depend on the assumption that if a speaker uses a word, it is likely that he will use the same word again in the near future. In particular, in a unigram cache, a unigram model from the most recently spoken words is formed and is then linearly interpolated with a conventional n-gram.

In another approach depending on the context a smoothed bigram or trigram formed from the previous words is interpolated with the standard trigram [Goo04]:

$$\begin{aligned} P_{trigram-cache}(w|w_1 \dots w_{i-2}w_{i-1}) &= \lambda\, P_{smooth}(w|w_{i-2}w_{i-1}) \\ +(1-\lambda)\, P_{tricache}(w|w_1 \dots w_{i-1}) & \end{aligned} \tag{11.22}$$

where $P_{tricache}(w|w_1 \dots w_{i-1})$ is a simple interpolated trigram model, using counts from the preceding words in the same document.

In another technique conditional caching is used by weighting the trigram cache differently depending on whether or not the context has been previously seen or not, which means that the trigram cache $P_{tricache}(w|w_{i-2}w_{i-1})$ is interpolated only if at least w_{i-1} has been seen in the cache [Goo04].

11.2.5 Sentence Mixture Models

Sentence mixture models [IOR94, IO99] depend on the fact that within a corpus, there may be several different sentence types which can be grouped by topic, or style, or some other criterion. In these models, each sentence type is modelled separately. The probability of a sentence is computed once for each sentence type and then a weighted sum of the probabilities across sentence types is taken. In general, the sentence type is treated as a hidden variable.

Denoting with s_j the condition that the sentence under consideration is a sentence of type j, the probability of the sentence, given that it is of type j can be written as

$$\prod_{i=1}^{M} P(w_i|w_{i-2}w_{i-1}s_j). \tag{11.23}$$

Taking into consideration all sentence types, a global more efficient model is obtained. Assuming s_0 to be a special context that is always true, then it holds that $P(w_i|w_{i-2}w_{i-1}s_0) = P(w_i|w_{i-2}w_{i-1})$. Having S different sentence types (usually $4 \leq S \leq 8$) and assuming that $\sigma_0, \sigma_1, \dots \sigma_S$ are the sentence interpolation parameters optimized on held-out data subject to the constraint $\sum_{j=0}^{S} \sigma_j = 1$, the overall probability of a sentence $w_1 \dots w_M$ is equal to

$$\sum_{j=0}^{S} \sigma_j \prod_{i=1}^{M} P(w_i|w_{i-2}w_{i-1}s_j). \tag{11.24}$$

Since the probabilities $P(w_i|w_{i-2}w_{i-1}s_j)$ may suffer from data sparseness, they are often linearly interpolated with the global model $P(w_i|w_{i-2}w_{i-1})$, using interpolation weights optimized on held-out data.

Sentence mixture models can be used for combining a stochastic context-free grammar model with a bigram model, resulting in marginally better results than either model used separately, as proposed in [JWS$^+$95].

11.2.6 Clustering

This approach attempts to estimate the probability of word sequences by exploiting the similarities in meaning or syntactic function between words derived from clustering. Successful assignment of words to classes enable more reasonable predictions for previously unseen histories by assuming that they are similar to other seen histories.

Supposing that we partition a vocabulary of V words into L classes using a function which maps a word w_i into its corresponding class C_i where $i = 1, \ldots, L$, an n-gram class language model is an n-gram model for which it holds for $1 \leq k \leq n$ that

$$P(w_k|w_1, \ldots, w_{k-1}) = P(w_k|C_k)P(C_k|C_1 \ldots C_{k-1}). \tag{11.25}$$

The model described by the above equation has $L^n - 1 + V - C$ independent parameters which are always fewer than a general N-gram model.

The way to find the best clusters has been a great research topic recently. Previous research [BCP$^+$90, KN93, Bel97, YS99] has found that there are small differences between the different developed techniques for finding clusters.

11.3 APPENDIX

11.3.1 Project URLs

Some well known tools for language modelling are given below:

- **CMU**
  - Open Source Speech Software (for example CMU Statistical Language Modelling toolkit, Hephaestus, Sphinx)

    http://www.speech.cs.cmu.edu/
  - Statistical Language Modelling Toolkit

    http://mi.eng.cam.ac.uk/ prc14/toolkit.html

    The CMU-Cambridge Statistical Language Modelling toolkit facilitates the construction and testing of statistical language models.

- **SRI Speech Technology and Research Laboratory**

  SRILM - The SRI Language Modelling Toolkit

  http://www.speech.sri.com/projects/srilm/

  SRILM is a toolkit for building and applying statistical language models (LMs), primarily for use in speech recognition, statistical tagging and segmentation.

- **Bow: A Toolkit for Statistical Language Modelling, Text Retrieval, Classification and Clustering**

  http://www-2.cs.cmu.edu/ mccallum/bow/

  Bow is useful for writing statistical text analysis, language modelling and information retrieval programs. The current distribution includes the library, as well as front-ends for document classification (rainbow), document retrieval (arrow) and document clustering (crossbow).

- **Lemur Toolkit for Language Modelling and Information Retrieval**

  http://www.lemurproject.org/

  The Lemur Toolkit is designed to facilitate research in language modelling and information retrieval, where IR is broadly interpreted to include such technologies as ad hoc and distributed retrieval, cross-language IR, summarization, filtering, and classification. The toolkit supports indexing of large-scale text databases, the construction of simple language models for documents, queries, or subcollections, and the implementation of retrieval systems based on language models as well as a variety of other retrieval models.

- **Maximum Entropy Modelling**

  http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html

  Maximum Entropy Model is a general purpose machine learning framework that has proved to be highly expressive and powerful in statistical natural language processing, statistical physics, computer vision and many other fields

Bibliography

[BCCW90] Bell, T. C., J. G. Cleary, and I. H. Witten. *Text Compression.* Prentice Hall, Englewood Cliffs, 1990. N.J.

[BCP$^+$90] P. F. Brown, J. Cocke, S. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, 1990.

[Bel97] J. R. Bellegarda. A latent semantic analysis framework for large-span language modeling. In *Proc. Eurospeech'97*, volume 3, pages 1451–1454. September 1997.

[BF95] R. Brown and R. Frederking. Applying statistical english language modeling to symbolic machine translation. In *Proc. of the 6th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI95)*, pages 221–239. July 1995.

[BL99] A. Berger and J. Lafferty. Information retrieval as statistical translation. In *Proc. of the 22nd Annual International ACM SIGIR Conference*, pages 222–229. 1999.

[Che96] F. S. Chen. *Building probabilistic models for natural language.* Ph.D. thesis, Harvard University, June 1996.

[CL03] W. Croft and J. Lafferty, editors. *Language Modeling for Information Retrieval.* Kluwer, 2003.

[Cla99] P. R. Clarkson. *Adaptation of Statistical Language Models for Automatic Speech Recognition.* Ph.D. thesis, University of Cambridge, 1999.

[Fin93] S. Finch. *Finding Structure in language.* Ph.D. thesis, Centre for Cognitive Science, University of Edinburgh, 1993.

[Goo53] I. J. Good. The population frequencies of species and the estimation of population parameters. In *Biometrika*, pages 40(3 and 4):237–264. 1953.

[Goo04] J. Goodman. A bit of progress in language modeling. Technical Report MSR-TR-2001-72, Microsoft, July 2004.

[GWC90] Gale, A. William, and K. W. Church. Estimation procedures for language context:poor estimates are worse than none. In *COMPSTAT, Proceedings in Computational Statistics, 9th Symposium*, pages 69–74. 1990.

[GWC94] Gale, A. William, and K. W. Church. What's wrong with adding one? In *N. Oostdijk and P. de Haan, editors, Corpus-Based Research into Language*. Rodolpi, Amsterdam, 1994.

[HAH+93] X. Huang, F. Alleva, H. Hon, M. Hwang, K. Lee, and R. Rosenfeld. The SPHINX-II speech recognition sytem: An overview. *Computer Speech and Language*, 2:137–148, 1993.

[Hie98] D. Hiemstra. A linguistically motivated probabilistic model of information retrieval. In *Proc. European Conf. Digital Libraries*, pages 569–584. 1998.

[IO99] R. Iyer and M. Ostendorf. Modeling long distance dependence in language: Topic mixtures versus dynamic cache models. *IEEE Transactions on Acoustics, Speech and Audio Processing*, 7:30–39, January 1999.

[IOR94] R. Iyer, M. Ostendorf, and J. R. Rohlicek. Language modeling with sentence-level mixtures. *In DARPA-HLT*, pages 82–86, 1994.

[Jel98] F. Jelinek. *Statistical Methods for Speech Recognition.* MIT Press, Cambridge, Massachussetts, 1998.

[JFM80] Jelinek, Frederic, and R. L. Mercer. Interpolated estimation of markov source parameters from sparce data. In *Proceedings of the Workshop on Pattern Recognition in Practice*. Amsterdam, The Netherlands: North Holland, May 1980.

[JMRS91] F. Jelinek, B. Merialdo, S. Roukos, and M. Strauss. A dynamic lm for speech recognition. In *Proc. ARPA Workshop on Speech and Natural Language*, pages 293–295. 1991.

[JWS+95] D. Jurafsky, C. Wooters, J. Segal, A. Stolcke, E. Fosler, G. Tajchman, and N. Morgan. Using a stochastic context-free grammar as a language model for speech recognition. In *Proc. ICASSP '95*, pages 189–192. Detroit, MI, 1995.

[KM90] R. Kuhn and R. D. Mori. A cache-based natural language model for speech recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 12(6):570–583, June 1990.

[KM92] R. Kuhn and R. D. Mori. Correction to a cache-based natural language model for speech reproduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(6):691–692, 1992.

[KN93] R. Kneser and H. Ney. Improved clustering techniques for class-based statistical language modeling. In *Eurospeech 93*, volume 2, pages 973–976. 1993.

[KN95] R. Kneser and H. Ney. Improved backing–off for mgram language modeling. In *Proc. ICASSP*, volume 1, pages 181–184. May 1995.

[LCC02] V. Lavrenko, M. Choquette, and W. B. Croft. Cross-lingual relevance models. In *Proc. of the 25th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*. 2002.

[MHL$^+$99] S. Martin, C. Hamacher, J. Liermann, F. Wessel, and H. Ney. Assessment of smoothing methods and complex stochastic language modeling. In *6th European Conference on Speech Communication and Technology*, volume 5, pages 1939–1942. Budapest, Hungary, September 1999.

[NEK94] H. Ney, U. Essen, and R. Kneser. On structuring probabilistic dependences in stochastic language modelling. In *Computer Speech and Language*, pages 8:1–38. 1994.

[PC98] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval system. In *Proc. SIGIR 98*, pages 275–281. New York, 1998.

[Ros94] R. Rosenfeld. *Adaptive Statistical Language Modeling: A Maximum Entropy Approach*. Ph. D. Thesis, School of Computer Science, Pittsburgh, PA, April 1994.

[SG96] C. F. Stanley and J. Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*, pages 310–318. San Francisco, 1996.

[SJCO02] L. Si, R. Jin, J. Callan, and P. Ogilvie. Language modeling framework for resource selection and results merging. In *Proc. of the Eleventh International Conference on Information and Knowledge Management (CIKM02)*. 2002.

[SK87] M. Slava and Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. In *IEEE Transactions on Acoustics, Speech and Signal Processing*, pages ASSP–35(3):400–401. March 1987.

[SO00] M. Siu and M. Ostendorf. Variable n-grams and extensions for conversational speech language modeling. *IEEE Transactions on Speech and Audio Processing*, 8:63–75, 2000.

[XWN01] J. Xu, R. Weischedel, and C. Nguyen. Evaluating a probabilistic model for cross-lingual retrieval. In *Proc. of the 24th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 105–110. 2001.

[YS99] H. Yamamoto and Y. Sagisaka. Multi-class composite ngram based on connection direction. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*. Phoenix, Arizona, 1999.

Chapter 12

Monolingual Information Retrieval

12.1 Introduction

Information Retrieval (IR) is the field of study that examines how people find information and how tools (such as search engines and catalogues) can be constructed to help people find access information. Studies examine how the organization of information affects its retrieval, the types of searches people do, the kinds of search queries people can make effectively, and what determines the relevance of retrieved information. When information is available in enormous quantities and not clearly structured, people have difficulty finding relevant information and understanding important principles embedded in the information. The World Wide Web (WWW) is one example of *Information Overload* and its expansion has generated requirements for more effective access to global and corporate information repositories. These repositories are traditionally text based but increasingly include multimedia content such as audio (e.g. spoken language or music), graphics, imagery, and video. In text IR the user's requirements are expressed as text keywords and the query results is textual data in the form of word documents. In monolingual IR query and information to be looked for are encoded in the same language. The main question is how to retrieve relevant information from large text or hypertext collections automatically and intelligently.

12.2 Approaches

Information retrieval remains an active field for research for decades and the rapid and global growth of the Internet has further increased the scientific interest in it. Consequently, the non-English content has also increased, making multilingual information access a necessity. But while the Internet is no longer monolingual, multilingual IR implies a good understanding of the issues involved in monolingual retrieval. IR, in its most simple form, is the process of gathering information on a particular subject. In its most basic terms, it is the process of matching a need to available knowledge. IR is a broad interdisciplinary and dynamic field that draws on many other disciplines. It stands at the junction of many established fields, and draws upon cognitive psychology, information architecture, information design, human information behaviour, linguistics, semiotics, information science, computer science and librarianship. Studies have typically approached IR from two major perspectives: from a rational approach which views IR as a mathematical model, as well as from a cognitive approach which views IR as an analysis of the process of information gathering done by people. In this sense, IR systems not only include search engines, but also human constructed hierarchies, annotated bibliographies, and other specialized methods of presenting materials. Nevertheless, Search Engine technology and Automatic Text IR have been fast-growing fields mainly due to the explosion of textual data available through the Web that renders inefficient the laborious task of human indexing. Therefore, statistical methods have seen significant advances in recent years and have been the dominant approaches for Text IR.

Recently, methods that try to capture more information about each document and achieve better performance have been researched and established in IR systems. These methods form three classes:

1. methods using parsing, syntactic information and Natural Language Processing (NLP) in general such as grammatical and morphological analysis (stoplists, stemming)

2. algebraic methods based on dimensionality reduction techniques that extent the VSM, such as *Generalized Vector Space Model* (GVSM) [SZRW86], *Latent Semantic Indexing* (LSI) [DDL+90] etc and

3. methods using Bayesian and neural networks and specifically spreading activation models.

Considerable advances have been made in recent years in syntactic modelling of natural language and development of efficient parsers with a broad domain. The task is to achieve automatic syntactic analysis and develop IR systems based on NLP. Progress is being made with syntax-directed semantic techniques such as lexical compositional semantics and with Artificial Intelligence techniques such as case frame analysis. But deeper semantic interpretation requires extensive knowledge engineering limiting

the breadth of systems that depend on NLP.

Full text IR is known to focus on the text itself, with semantics being handled in a rudimentary way. In traditional text retrieval the most straightforward way of locating the documents that contain a certain search term is to search all documents for the specified string (Full text scanning). Another well-known technique is the signature file approach. A fast text retrieval technique that is followed by many commercial systems is the inversion of the list keywords that represent the document content. A more sophisticated model than classical Boolean and Probabilistic models is the Vector Space Model (VSM) where a page is represented as a bag of keywords instead of a set of keywords as in the Boolean model. VSM takes frequency information into account. The language independent 'bag-of-words' representations of documents have proved surprisingly effective for text classification. Common questions regarding term and document weighting schemes, normalisation, term stemming and common word elimination have been explored in depth in current bibliography. But the optimal representation of a text document remains an open research question. Some engines break documents and queries in phrases or even *n-grams* instead of words.

Latent Semantic Indexing on the other hand has demonstrated improved performance over the traditional vector space techniques and has been successfully applied in many test IR systems. LSI, an optimal special case of multidimensional scaling, is a concept-based automatic indexing method that tries to overcome the two fundamental problems which plague traditional lexical-matching indexing schemes: synonymy and polysemy. It models the semantics of the domain in order to suggest additional relevant keywords and to reveal the "hidden" concepts of a given corpus while eliminating high order noise. The attractive point of this method is that it captures the higher order "latent" structure of word usage across the documents rather than just surface level word choice. This is done by modelling the association between terms and documents based on how terms co-occur across documents.

Recently, Latent Semantic Analysis (LSA) has come under criticism, because its probabilistic model does not match the observed data. LSA assumes that words and documents form a joint Gaussian model. However, Gaussian models can generate negative values, and it is impossible to have a negative number of words in a document. Thus, a newer alternative is Probabilistic Latent Semantic Analysis (PLSA), based on a multinomial model, and is reported to give better results than standard LSA [Hof99]. PLSA is based on a statistical method which has been called aspect model [HP98]. The aspect model is a latent variable model for co-occurrence data which associates an unobserved class variable $z \in \mathcal{Z} = \{z_1, \ldots, z_K\}$ with each observation. So, for the text we assume that these two variables, are the occurrence of a word w in a document d (observed), and the topic(unobserved). PLSA defines a generative model for word/document co-occurrences. The assumption is that each word w_j in a given document d_δ is generated from a latent topic t, i.e. a word is conditionally independent from its original document given the latent topic it was generated from. The data generation process can be described as follows:

1. Select a document index δ with probability P(δ)
2. Pick a latent topic $t = k$ with probability $P(t = k|d_\delta)$

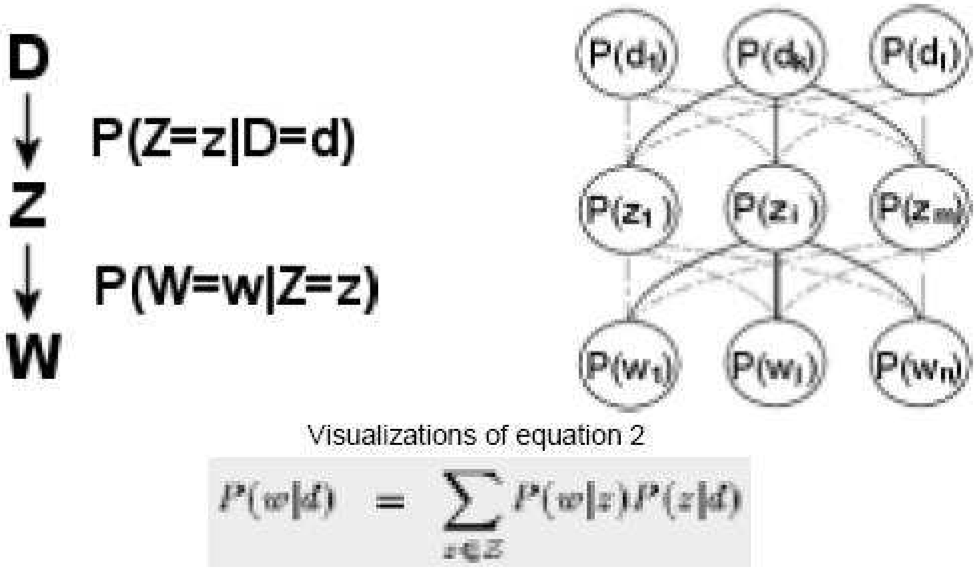


Figure 12.1: The data generation process.

3. Generate a word w_j with probability $P(w_jjt = k)$

This generative process is summarized by the joint distribution of a word w_j, a latent topic $t = k$, and a document d_δ:

$$P(w_j, t = k, d_\delta) = P(\delta)P(t = k|d_\delta)P(w_j|t = k) \tag{12.1}$$

and the joint distribution of the observed data:

$$P(d_\delta, w_j,) = P(\delta)\sum_{k=1}^{K} P(t = k|d_\delta)P(w_j|t = k) \tag{12.2}$$

So each word in a document is seen as a sample from a mixture model where mixture components are the multinomial $P(w_j|t = k)$, and the mixing proportions are $P(t = k|d_\delta)$. The PLSA algorithm maximizes the log-likelihood of the model, by using the EM algorithm[DLR77].

The PLSA model can be used to replace the original document representation by a representation in a low-dimensional ”latent” space, to perform a TC or IR task. In [Hof01], the components of the document in the low-dimensional space are $P(t = k|d)$ $\forall k$, and for each unseen document or query they are computed by maximizing the log-likelihood with $P(w_j|t = k)$ fixed. This representation scheme is referred to as PLSI, for Probabilistic Latent Semantic Indexing. It is obvious that PLSA is not a well-defined generative model of documents, since there is no direct way to assign probability to an unseen document. However, some experiments in [Hof01] report a comparison between LSI and PLSI, on several corpora. They conclude to a better performance of PLSI in all cases. In particular PLSI performs well even in the cases where LSI fails completely.

Bayesian Networks can also been applied to Text IR [TC90]. These present flexible ways of combining term weights and they can generalize previous approaches such as the Boolean model, the Binary independence model and the Probabilistic models with weaker assumptions. They also have efficient large-scale implementation. A disadvantage of these methods is that they need approximations to avoid intractable inference and the have to estimate all the probabilities by some means (whether more or less ad hoc).

12.3 Information Retrieval in the Web

The main open issues in IR have to do with the Information access methods and the Information properties. Information access includes concepts such as information transmission and visualisation, categorisation and clustering, topic detection and tracking, summarisation, query formulation, information acquisition and extraction algorithms and their performance. The Information properties refer to the type of media data (text or multimedia), its structure (unstructured, semi-structured - XML, fully structured, hyperlinked - Web, mixture of types) and its heterogeneity (mono/multi-lingual, heterogeneous structures and services). Open issues regarding the heterogeneity of Information include the standardization of non-trivial structures (e.g. Dublin Core) and services (e.g. XQuery text retrieval) and integration approaches based on uncertainty and vagueness.

The data of today are electronically distributed and are represented in diverse formats and structures. Nowadays much emphasis is given in IR systems that have to deal with an excessive amount of unstructured or semi-structured data where no explicitly well-defined syntax for the documents in the archive exists. Because of the decentralized nature of its growth, the Web has been widely believed to lack of structure and organization as a whole. Even if web documents do share a syntax, there is no well-defined semantics associated with each syntactic component. An open issue here is the size and coherence of the text repository from where we seek knowledge. At early years most of the research on information retrieval systems is on small well-controlled homogeneous collections such as collections of scientific papers or news stories on a related topic. Recently, the demand to find relevant information from large, noisy and non-homogenous corpora has become stronger. World Wide Web can be viewed as a graph, in which each node represents the page and edges connecting the nodes are the hyperlinks. The topology of this graph determines Web's connectivity and consequently how effectively can we locate information on it. The main goal of web IR is the automatic acquisition, indexing and ranking of documents in the Web. But, its enormous size, decentralized and dynamic nature and rapid growth pose a big challenge to search related pages for specific topic. Large-scale search engines struggle to cover the vast amounts of information that has been accumulated in the Web and maintain the freshness of their index. Furthermore, a very large portion of the web data is inaccessible through common web browsing or automatic crawling (hidden web). Recent studies [KL01] indicate that the Web contains a large, strongly connected core in which every page can reach every other by a path of hyperlinks. This core contains most of the prominent sites on the Web. The remaining pages can be characterised by their relation to the core. Due to good amount of resources for research in Web, many researchers are attracted into web IR area. There are many issues like extraction of the features from the pages, organizational structure of web, identifying community of pages, crawling the web, large-scale search engine and its architecture, web structure, personalized web search, page ranking methods, optimising web structure, web indexing etc. which are required for better web mining.

Information agents are programs that automatically perform customised information processing actions to deal with information overload problems. Examples of agents are the Web Crawlers which programs that traverse the hypertext structure of the Web automatically, starting from an initial hyper-document or a set of starting points (seeds) and recursively retrieving all documents referenced by that document. The recent trends in research of this field is the implementation of a focused crawler that intelligently avoids irrelevant portions of the web while visiting most relevant or promising pages early in the crawl process. This can help developing Vertical Search Engines that offer targeted and domain specific information to users. The open research problem is to efficiently reorder its crawl frontier (the queue of unvisited pages) when no content of the unvisited portion of the web graph is on hand.

What really differentiates hypertext from static text documents is the fact that the former, besides the text content, contain additional semantics, such as a document markup structure (Document Object Model - DOM), linking information that associates documents, citations and structured header (metadata) that precedes the relatively unstructured body. Link and social network analysis have been successfully applied both to academic citation data to identify influential papers and, more recently, to web hyperlink data to identify authoritative information sources. Recent techniques in web IR try to properly extract, exploit and integrate all these features in order to efficiently process and acquire information so that distributed, portable, high-performance information processing engines can be developed. Clearly, outlinking information is available and can be used to implement well known relevance metrics and ranking algorithms such as HITS [Kle98] and PageRank [BP98], two of the most prominent algorithms in web IR. The heuristic underlying both of these approaches is that pages with many inlinks are more likely to be of high quality than pages with few inlinks, given that the author of a page will presumably include in it links to pages that s/he believes are of high quality. Lately, it has been shown that the ranking of the crawl frontier can be further improved by using the textual content from links that have been already visited. Numerous methods that try to combine textual and linking information for efficient URL ordering exist in the bibliography. Many of these are modifications, improvements or extensions of either PageRank [RM00], [RD02], [Hav03] or HITS [CH01]. Chakrabarti et al. [CBD02] and Bharat & Henzinger [BH98] also propose heuristic methods for differentially weighting links. Other algorithms such as SALSA [LM00], Spectral Filtering [CDG$^+$99], HyCon [DM03] and Probabilistic HITS (PHITS) [CC00] are known to improve web search performance and provide quality pages.

Lack of domain knowledge means that user queries will inevitably have less satisfactory results. There are limitations to the amount of control an IR system has over their users' knowledge. Moreover success of query-oriented IR depends on the size of the query; short queries do not provide sufficient information to the IR system to distinguish relevant documents from irrelevant ones. But studies have shown that most of the queries consist of only a few keywords. On large scale libraries, especially over the Internet, user training is not an option to tackle with this problem. Thus, a system is needed that supports the user with additional domain knowledge. The approach taken is to refine, expand and reweight the query automatically based on the documents retrieved by the original query. The common technique for automatic query expansion is to use pseudo-relevance feedback with top-K retrieved documents per query. There is need for distinguishing important terms and applying a proper weighting scheme. An alternative method is to expand each term in the original query with synonyms or related terms drawn from a generic on-line thesaurus. A third method to query expansion is based on interactive relevance feedback from the user. The system first returns a small number of matching documents; the user scans these, marking each document as "relevant" or "irrelevant". The system then uses this feedback from the user to formulate and launch a new query that better matches what the user is seeking.

Probably the most substantial evidence for automatic indexing has come out of the SMART Project [SL65]. The SMART system, developed at Cornell, is the one of the earliest IR systems that (1) use fully automatic term indexing, (2) perform automatic hierarchical clustering of documents and calculation of cluster centroids, (3) perform query/document similarity calculations and rank documents by degree of similarity to the query, (4) represent documents and queries as weighted term vectors in a term-based vector space, (5) support automatic procedures for query enhancement based on relevance feedback. SMART has been widely used as a testbed for research into, e.g., improved methods of weighting and relevance feedback, and as a baseline for comparison with other IR methods.

The Text Retrieval Conference (TREC) [TRE] began in 1992 and serves as a major technology-transfer mechanism in the area of text retrieval. It attracts international participation from more than 100 research groups in retrieval technology, both from industry and academia. Its main goal is to accelerate the transfer of better text search and retrieval technology into commercial systems. Participating groups work with large, diverse test collections, submit their results for a common evaluation, and compare techniques and results.

12.4 Links

1. Information Retrieval: A Survey,
   http://www.csee.umbc.edu/cadip/readings/IR.report.120600.book.pdf

2. Text Retrieval Conference (TREC),
   http://trec.nist.gov

3. Special Interest Group on Information Retrieval (SIGIR),
   http://www.acm.org/sigir

4. Foreign Language Resource Center,
   http://flrc.mitre.org

5. CLEVER Project, IBM Almaden Research Center,
   http://www.almaden.ibm.com/cs/k53/clever.html

6. Linguistic Information Retrieval (Lirix) - Xerox,
   http://www.xrce.xerox.com/programs/lirix

7. Berkeley Digital Library SunSITE,
   http://sunsite.berkeley.edu

8. Latent Semantic Indexing Web Site,
   http://www.cs.utk.edu/ lsi

9. Reuters Corpus,
   http://about.reuters.com/researchandstandards/corpus

10. CMU World Wide Knowledge Base (WebKB) project,
    http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/theo-11/www/wwkb

11. CMU Text Learning Group,
    http://www-2.cs.cmu.edu/afs/cs/project/theo-4/text-learning/www

12. Center for Intelligent Information Retrieval,
    http://ciir.cs.umass.edu

13. Apache Jakarta Lucene search engine,
    http://jakarta.apache.org/lucene/docs/index.html

Bibliography

[BH98] K. Bharat and M. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *Proc. 21st ACM International Conference on Research and Development in Information Retrieval SIGIR-98*, pages 104–111. Melbourne, AU, 1998.

[BP98] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proc. WWW7 / Computer Networks*, volume 30, pages 107–117. 1998.

[CBD02] S. Chakrabarti, M. Berg, and B. Dom. Focused crawling: a new approach to topic-specific web resource discovery. *Computer Networks*, 31:1623–1640, 2002.

[CC00] D. Cohn and H. Chang. Learning to probabilistically identify authoritative documents. In P. Langley, editor, *Proc. 17th International Conference on Machine Learning (ICML 2000)*, pages 167–174. Morgan Kaufmann, jun/jul 2000. ISBN 1-55860-707-2.

[CDG$^+$99] S. Chakrabarti, B. Dom, D. Gibson, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Topic distillation and spectral filtering. *Artificial Intelligence Review*, 13(5-6):409–435, December 1999.

[CH01] D. Cohn and T. Hofmann. The missing link a probabilistic model of document content and hypertext connectivity. *Advances in Neural Information Processing Systems*, 13:430–436, June 2001.

[DDL$^+$90] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for information Science*, 41:391–407, 1990.

[DLR77] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[DM03] S. S. D. Mukhopadhyay. A hyperlink and content based topic search technique. Technical report, Haldia Institute of Technology, Department of Computer Science and Engineering, June 2003.

[Hav03] T. Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):784–796, jul/aug 2003.

[Hof99] T. Hofmann. Probabilistic latent semantic analysis. In *Proc. 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI-99)*, pages 289–296. Morgan Kaufmann Publishers, San Francisco, CA, 1999.

[Hof01] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1-2):177–196, 2001. ISSN 0885-6125.

[HP98] T. Hofmann and J. Puzicha. Unsupervised learning from dyadic data. Technical Report TR-98-042, International Computer Science Insitute, Berkeley, CA, 1998.

[KL01] J. Kleinberg and S. Lawrence. The structure of the web. *Science*, 294(5548):1849–1850, November 2001.

[Kle98] J. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proc. 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 668–677. January 1998.

[LM00] R. Lempel and S. Moran. The stochastic approach for link-structure analysis (salsa) and the tkc effect. *Computer Networks: The International Journal of Computer and Telecommunications Networking*, 33(1-6):387–401, 2000. ISSN 1389-1286.

[RD02] M. Richardson and P. Domingos. The intelligent surfer: Probabilistic combination of link and content information in pagerank. *Advances in Neural Information Processing Systems*, 14:1441–1448, 2002.

[RM00] D. Rafiei and A. Mendelzon. What is this page known for? computing web page reputations. *Computer Networks*, 33(1-6):823–835, 2000.

[SL65] G. Salton and M. Lesk. The smart automatic document retrieval system an illustration. *Communication of the ACM*, 8(6):391–398, June 1965.

[SZRW86] S.Wong, W. Ziarko, V. Raghavan, and P. Wong. On extending the vector space model for boolean query processing. In *Proc. 9th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 175–185. Pisa, Italy, September 1986.

[TC90] H. Turtle and W. Croft. Inference networks for document retrieval. In J.-L. Vidick, editor, *Proc. 13th Annual International ACM SIGIR*

Conference on Research and Development in Information Retrieval, pages 1–24. 1990. ISBN 0-89791-408-2.

[TRE] Text retrieval conference - trec.

Chapter 13

Natural Language Processing for Content Based Image Retrieval

13.1 Introduction

This chapter presents the state-of-the-art of employing natural language processing (NLP) for content-based image retrieval. After a presentation of the problem of content-based image retrieval (CBIR), we examine the current uses of natural language processing techniques in attacking the text part of this problem. We do not cover the image processing tasks which are discussed in other parts of this state-of-the-art. But just as the image processing techniques are concerned with extracting meaningful signatures from image data, the natural language processing steps discussed here are concerned with extracting normalized signatures from text. We cover the information needed to understand the textual processing tasks used within the indexing and retrieval of the recent ImageCLEF and TRECVID campaigns.

13.2 Content-based image retrieval

Content-Based Image Retrieval (CBIR) refers to the problem of retrieving images from a repository, based solely on the image content. The notion of CBIR is usually used in a wider sense, embracing the closely related problems of automatic image annotation, indexing and browsing.

CBIR emerged in the early 90s as a response to the emergence of large-scale multimedia collections and the difficulty this posed to the usual manual annotation approach [SWS$^+$00]. Manual image and video annotation not only requires a vast amount of labour but is also affected by the rich content present in images and the subjectivity of human perception. Early approaches suggested low-level features such as colour, colour layout, shape, texture and segmentation as valid cues for describing visual content in images. These descriptors have limited performance (in terms of precision and recall) in making explicit the semantics of an image, a problem known as the *semantic gap*. It soon became evident that it is far more difficult for machines to extract meaningful semantics from multimedia than it is to extract semantics from natural language text.

A wide variety of solutions were proposed in order to bridge this semantic gap. A first approach acknowledged the difficulty inherent to translating keyword or speech-based queries to visual based searches, and proposed the use of visual queries, i.e. using images as query inputs [FSN$^+$95]. Examples of this are the possibility of searching by example or by sketch or directly by performing random browsing of the repository. An intermediate approach attempts to extract semantic information from the user by using a directed visual-based search. This approach is known as relevance feedback. Most state-of-the-art image retrieval systems support one or more of the options mentioned in this paragraph [JMHA04]. Text is used to bridge the semantic gap in CBIR in three ways: manual annotation of entire images, exploiting text found attached to an image, automatically assigning keywords to images by training keyword-to-region functions.

Manual annotation of images is the approach taken by all the large commercial image repositories (e.g. GraphicObsession, Getty, Corbis)[1]. State-of-the-art research explores ways to automatically process text found near images to match the text found in user questions. There is also research in assigning words to images through image processing techniques[BDF$^+$03] that is not covered in this chapter we we concentrate on text processing aspects of conten-based image retrieval.

13.3 CBIR campaigns involving NLP

There are two major international campaigns involving content-based image and video retrieval. Both campaigns involve text and natural language processing components.

ImageCLEF is the cross-language image retrieval track which is run as part of the Cross Language Evaluation Forum (CLEF) campaign[2]. The ImageCLEF retrieval benchmark was established in 2003 with the aim of evaluating image retrieval from multilingual document collections, containing images accompanied by texts semantically related to the image (e.g. textual captions or metadata). In the ImageCLEF campaigns images can then be retrieved using low-level features based on pixels which form the contents of an image (e.g. using an image as a query) or using the associated text or a combination of both. Though the language attached to the images used in ImageCLEF is often in English, the

[1] www.graphicobsession.com, pro.corbis.com/default.aspx, www.gettyimages.com

[2] See http://clef.iei.pi.cnr.it/

retrieval task involves queries that written in many different languages in addition to English, and cross language retrieval of the images is essential to the campaign.

ImageCLEF provides tasks for both system-centered and user-centered retrieval evaluation within two main areas: retrieval of images from photographic collections and retrieval of images from medical collections. These domains offer realistic scenarios in which to test the performance of image retrieval systems, offering different challenges and problems to participating research groups. A major goal of ImageCLEF is to investigate the effectiveness of combining text and image for retrieval and promote the exchange of ideas which may help improve the performance of future image retrieval systems.

ImageCLEF has already seen participation from both academic and commercial research groups worldwide from communities including: Cross Language Information Retrieval (CLIR), Content-Based Image Retrieval (CBIR) and user interaction. More information on past ImageCLEF campaigns can be found on the University of Sheffield website[3].

For further information consult this website and these publications [CMS05, MGMMC04].

Since 2001, the yearly Text REtrieval Conference (TREC[4]) sponsored by the National Institute of Standards and Technology (NIST) has sponsored a video retrieval track called TRECVID. The main goal of TRECVID is to promote progress in content-based retrieval from digital video via open, metrics-based evaluation. Different international research and commercial groups apply to their video retrieval systems to the same data for the same tasks, and their results are independently analyzed by the NIST[5].

The data used in TRECVID comprises video (often newscasts), the audio signal, a transcription of the audio or a text produced by automatic speech recognition (donated by Jean-Luc Gauvain of the Spoken Language

Processing Group at LIMSI[GLA02]), as well as some high-level semantic features, concepts such as "Indoor/Outdoor", "People", "Speech" that is sometimes attached to the video stream. The search task is as follows: given the search test collection, a multimedia statement of information need (query), and the common shot boundary reference for the search test collection, return a ranked list of at most 1000 common reference shots from the test collection, which best satisfy the expressed query.

Another CBIR image campaign is being planned for the years 2005-2007 called ImagEval[6]. One of the tasks evaluated in ImagEval will involve improving image search using text captions found around the image.

In all three of these competitions, TRECVID, ImageCLEF and IMAGEVAL,

[3] http://ir.shef.ac.uk/ImageCLEF2005/

[4] http://trec.nist.gov

[5] Online proceeding of the campaign can be found at http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html

[6] See http://www.imageval.org

though the main goal is to find relevant queries, there is a natural language processing component involving analyzing a textual query and/or text found associated with the images to be indexed.

13.4 NLP Mechanisms used in CBIR

In this section we will look at the NLP techniques used to normalize and index text associate with images. This text can come from an image caption, from text surrounding an image in a document such as a web page, or from text that points to an image. In this last case, the text may appear in another document (for example, another web page) which is hyperlinked to the image.

13.4.1 Finding text in Document structure

When it has been decided to index text that has been associated with an image, a choice remains as to what text to use. The following types of text in electronic documents and web documents can be exploited to find a textual representation of an image:

- the name of the image file, for example carrot.gif
- the HTML title of the page in which the image is found, “Spicy Carrot Bread”
- a caption associated with the image, for example “Flames shoot off the front landing gear as the aircraft lands on the runway”
- the HTML ALT field meant to provide alternative or substitute text, primarily for use when the image is not being displayed, for example <IMG src=”toto.gif” alt=”Fox Terrier puppy”>
- the paragraph in which the image is embedded
- the visual layout block [CHL$^+$04] containing the image
- the entire document in which the image is embedded
- the text found in the anchor of another HTML page pointing to the image, for example, <A HREF=”foo.jpeg”> giant blue heron</A>.

Valid HTML pages (in which all tags are explicitly balanced) and XML pages posses a logical structure that can be exploited by the application modules such as the Document Object Model (DOM) of the W3C: n the DOM specification, the term ”document” is used in the broad sense - increasingly, XML is being used as a way of representing many different kinds of information that may be stored

Danish	Dutch	English	French	German	Italian	Norweg.	Portug.	Spanish
i	*de*	*the*	*de*	*der*	*di*	*og*	*de*	*de*
af	*van*	*and*	*la*	*die*	*e*	*det*	*a*	*la*
og	*het*	*to*	*le*	*und*	*il*	*han*	*que*	*que*
at	*een*	*of*	*à*	*den*	*che*	*i*	*o*	*el*
til	*en*	*a*	*et*	*in*	*la*	*er*	*e*	*en*
for	*in*	*in*	*des*	*von*	*a*	*på*	*do*	*y*

Figure 13.1: Most frequent short tokens per language derived from the ECI Multilingual Corpus.

in diverse systems, and much of this would traditionally be seen as data rather than as documents[7].

Such applications such as DOM and SAX[8] permit the structured extraction of logical divisions of pages that contain text [Cha01]. Images found within the logical segment can be associated with any other text up the DOM tree. Another approach uses the visual layout of the web page [CHL$^+$04] to associate the image with text that is found close to the image on the web, even though they may be separated by the logical structure of the page, for example by belonging to different HTML tables that appear one under the other when the page is displayed.

Once some combination of these elements of the document structure is chosen, we can extract a text that can then be used to index the image, using the NLP steps described in the following sections.

13.4.2 Language Identification

Before any processing of the text associated with an image is performed, it is useful to perform language identification on the text. The common approaches to language identification involve building up statistics from documents whose languages are known, for example from the ECI Multilingual corpus (see

www.elsnet.org/ecilisting.html) The statistics can involve common words or common sequences of letters from the training documents [Gre95]. For example, the word ending -*ing* is common in English, and the word ending -*que* is common in French. Some common words and trigrams for some European languages are shown in figures 13.1 and 13.2 without their frequency statistics. These word and character sequence statistics are then used to classify new documents into one of the recognized languages. The task is considered easy to implement with a high degree of accuracy [McN05].

In the CLEF'2004 medical image retrieval task[9], each image to be retrieved

[7] See www.w3.org/TR/DOM-Level-2-Core/introduction.html

[8] Found at saxproject.org

[9] See ir.shef.ac.uk/ImageCLEF2004

Danish	Dutch	English	French	German	Italian	Norweg.	Portug.	Spanish
*er_*	*en_*	*_th*	*_de*	*en_*	*_di*	*et_*	*_de*	*_de*
*en_*	*de_*	*he_*	*es_*	*er_*	*to_*	*en_*	*de_*	*de_*
for	*_de*	*the*	*de_*	*_de*	*_de*	*er_*	*os_*	*os_*
*et_*	*et_*	*nd_*	*ent*	*der*	*di_*	*_de*	*do_*	*_la*
ing	*an_*	*ed_*	*nt_*	*ie_*	*_co*	*_ha*	*que*	*el_*
_fo	*n_d*	*_an*	*_le*	*ich*	*la_*	*an_*	*_qu*	*la_*

Figure 13.2: Most frequent trigrams per language derived from the ECI Multilingual Corpus.

was accompanied by textual medical case descriptions in either English or French. Though a language tag was also connected to the descriptions, some cases had multilingual description with both languages. In order to process the text correctly these descriptions had to be isolated using language identification techniques [RS05]

Once the language of the text associated with an image is identified, automatically or manually, the following NLP steps use methods and resources specific to this language.

13.4.3 Sentence Recognition

Most NLP analysis and parsing algorithms are designed to work on single sentences. For this reason sentence segmentation needs to be performed when the image text may be more than one sentence long, as is often the case when a paragraph, long caption, or entire document is associated with the image. The main problem involved in sentence segmentation is distinguishing sentence-ending periods from abbreviations-ending periods. Since abbreviations are often domain dependant, they can be considered as resources that must be updated when new domains are treated [ZDJ03] though domain-independent methods have been proposed [Cho00]. For general language, the more resources that are used (lexicons, list of abbreviations) the better the results for recognizing sentence endings [GT94], with precisions of 95-99% attainable for English text.

13.4.4 Stemming and Stopwords

A commonly used approach to word normalization, used in the absence of more complete morphological analyzers are stemming routines [Lov68]. The Porter stemmer [Por80] is a popular algorithm that is used for removing common morphological and inflectional endings from words in English. Version of this stemmer in different programming languages can be found on Martin Porter's site[10]. Porter has more recently developed Snowball, a small string processing language

[10] www.tartarus.org/ martin/PorterStemmer

designed for creating stemming algorithms. And this has been applied to English, French, Spanish, Portuguese, German, Dutch, Swedish, Norwegian, Danish, Russian, and Finnish[11].

Stemming usually involves applying lists of suffixes to words iteratively a fixed number of times, or until a minimum numbers of letters remain as the root. Examples of endings removed for English by the Porter stemmer are: *-ies, -tion, -ence, -ance, -able, -ic,* etc. This algorithm also applies constraints on the remaining root (such as ending in consonant and containing a vowel), and deals with some morphological changes (such as double letters, *getting->get*) and exceptions (treated in the code). Some stemmers are overly aggressive producing similar roots for words such as *business, busy* or *organization,organ.* But they are widely used because they are easy to implement and control (i.e., adding exceptions to the rules can be done by any programmer).

The advantages of stemming are that it reduces a lot of morphological variation from text to be indexed, and it is rapid to implement and to execute without requiring extra linguistic resources such as lexicons and grammars.

Disadvantages of stemming are that it cannot handle internal vowel alterations, agglutinating and compounding languages, and that it provides no information about the structure of text (e.g., the parts-of-speech that a word can play) which are used in more elaborate NLP such as entity recognition and phrasal recognition.

At the same level of complexity lies the use of stopword list to further reduce the text to be indexed. Generally, stopwords are considered to be those words which bear no or little information content out of context. In English these lists contain words such as *a, an, is, that, which, of, are, with, be,* etc. and, maybe surprisingly, *not* and *no.* These lists are generally handmade containing from a few dozen to a few hundred words. For domain specific applications, there are techniques for calculating a word strength related to the retrieval effectiveness of words, and for using words with low strengths as stopwords [WS92]. Stopword lists have been made freely available[12] for English, French, German, Italian, Spanish, Portuguese, Finnish, Swedish, Arabic, Russian, Hungarian, Bulgarian, Romanian, Czech and Polish.

An example of text before and after stemming and stop word removal is shown in figure 13.3.

The text comes from one of the topics of the TREC text retrieval competition[13] and the results are from the stemming and stopword removal found in a version of the information retrieval software SMART [SM83].

Such stemming and stopword processing was used for recent entries in the ImageCLEF and TRECVID competition, such as [AOMN05, CFG$^+$04, DKN05,

[11] Online versions for these languages can be found at snowball.tartarus.org

[12] See www.unine.ch/info/clef

[13] See trec.nist.gov

TREC QUERY:

> *To be relevant, a document will discuss a pending antitrust case and will identify the alleged violation as well as the government entity investigating the case. Identification of the industry and the companies involved is optional. The antitrust investigation must be a result of a complaint, NOT as part of a routine review.*

STEMMED VERSION:

> RELEV DOCU DISCUSS PEND ANTITRUST CASE IDENTIF ALLEG VIOLAT GOVERN ENTIT INVESTIG CASE IDENTIF INDUSTR COMPAN INVOLUT OPTION ANTITRUST INVESTIG RESULT COMPLAIN PART ROUTIN REVIEW

Figure 13.3: A text reduced by first stemming tokens using the Lovins/SMART stemmer and then removing stop words.

JBJ$^+$05, MGR05], among others, in order to stem the image captions or the transcription of the video broadcast. The images were then indexed with these stemmed words.

13.4.5 Part-of-Speech Tagging and Morphological Analysis

A more complicated NLP treatment is normalizing word forms by performing part-of-speech tagging and morphological analysis depending on this part of speech. Morphological analysis is concerned with the inflectional, derivational, and compounding processes in word formation. It corresponds to the segmentation of a given word into the various smallest meaning units (morphemes) which form it, e.g. its stem and affixes. A full morphological analysis can also give morphosyntactic information about the word-form and its stem, e.g. possible part-of-speech (PoS) and/or inflectional properties (gender, number, case, person, tense, etc.), etc.

A simple approach to morphological analysis consists in using available electronic lexical databases that associate word-forms and lemmas, together with inflectional and/or derivational information. This last information is often provided as a code or model to which operations to produce all possible inflected forms are attached. For example, the MULTEXT project [IV94] has provided lexical lists of lemmas and inflected word-forms for four languages of the European Community: French, Italian, Spanish and German[14] This approach has however

[14] See http://www.lpl.univ-aix.fr/projects/multext/ The word-form dictionary for French lists 300,000 forms including proper nouns and compounds. Each character of the linguistic description specifies a value of an attribute. For example, for a verb, there are 7 attributes: PoS, type,

difficulties to deal with the quasi infinite possibilities of the derivational process, and offers no (efficient) way to analyze words not present in the database.

Though morphological analysis for a language such as English can be approximated by stemming methods as mentioned above, this is not the case for highly inflectional languages which require more sophisticated techniques. The most common approach to performing morphological analysis over other languages is to describe the morphological formation process using two-level morphology and then compiling theses rules into efficient finite-state transducers [BK03]. The primary interest of morphological analysis is to find the correct normalized form for the words and phrases indexed with images and video. For example, the word *thought* can be normalized to different lemmas according to its grammatical context:

- They thought he would come. → *think*
- That is a thought to remember. → *thought*

Proper and efficient morphological analysis is still an unsolved problem for less studied languages[15], such as Arabic, Hungarian, Turkish, Finnish, and Eastern European languages. Some languages, such as Chinese, pose the additional problem segmenting running text into word tokens.

A part-of-speech tagger will take a tokenized sentence [Gre99a] as input, assign one or more possible parts-of-speech to each token, and then usually select one possible reading of the sentence by assigning the appropriate part of speech (for example, verb, noun, adjective, adverb) to each word. It references a lexicon or morphological analyzer to find the parts of speech for known words, and unknown words are assigned possible parts of speech depending on their composition (for example, a token made up only of numbers will be assigned a numerical part-of-speech tag).

For example, given the sentence *An experimental study of the wing in a propeller slipstream was made*, a number of possible parts-of-speech are first assigned to each token by a morphological analyzer, or lexicon lookup) and then the most likely tag is chosen for each word as shown in figure 13.4.

Numerous studies in computational linguistics have dealt with how grammatical tags should be defined, and with accurate and efficient algorithms for choosing these tags for the words in a text. Problems arise on the linguistic level when we try to define a set of tags for a language, and then the subsequent rules for choosing among the tags. Computationally, we must resolve problems of exponential

mood or verbal form, tense, person, number, gender. Compounds receive the same linguistic description as simple words. CELEX http://www.ru.nl/celex/ is another large multilingual database that includes extensive lexicons of English, Dutch, and German. For each language, several types of lexicons are available: lemma, word-form, abbreviation and corpus type.

[15] See isl.ntf.uni-lj.si/SALTMIL/

BEFORE PART-OF-SPEECH TAGGING

An	*experimental*	*study*	*of*	*a*	*wing*	*was*	*made*
det	*adj*	*sn*	*prep*	*d*	*sn*	*auxb*	*vt-past*
		vt			*vi*		*vt-pastptr*

AFTER PART-OF-SPEECH TAGGING

An	*experimental*	*study*	*of*	*a*	*wing*	*was*	*made*
det	*adj*	*sn*	*prep*	*d*	*sn*	*auxb*	*vt-pastptr*

Figure 13.4: A part of speech tagged text, after morphological analysis and then after tag selection in which Here, det=determiner, sn=singular-noun, adj=adjective, vt=transitive-verb, auxb=auxiliary(be), vt-past=transitive-verb(past tense), and vt-pastptr= transitive-verb(past participle)

time and space. The main intuitions for solving the computational problems are found in most current approaches:

- Implement rules using limited context to define sequences of permissible parts-of-speech,
- use training techniques to build these succession rules,
- use frequency as a basis for deciding among competing rules,
- use ad hoc rules to treat remaining ambiguity

In order to choose the most likely tag, part-of-speech tagging systems train language models from large hand-tagged corpora, such as the 100-million word British National Corpus[16] [Lee92], and the Penn TreeBank [MSM94]. Research in part-of-speech tagging was very active in the 1990s with methods attaining performance up to 97-99% of accurately tagged text. These methods include probabilistic approaches such as HMMs [WMS$^+$93], maximum entropy [SB98], and rule-based techniques such as transformation learning [Bri95]. These results coincide with human inter-annotator agreement performing the same task [Bra00].

Both part-of-speech tagging and stopword removal (now using normalized stopwords or using part-of-speech tags other than nouns, adjectives, verbs) can be combined in the NLP of text. If we look at the example given above, after part-of-speech, lemmatisation and stopword removal, we get the following output:

In this more readable version, we see some shortcomings pf this more powerful treatment, *investigate* and *investigation* are two separate indexers for the text, and an additional step (for example, using a derivational lexicon, or synonym list, see the section on Semantic Space reduction below) is needed to conflate the two descriptors. On the other hand, a word like *optional* would not longer be conflated with a normalized form of the word *options*. The principal advantage of

[16] See www.natcorp.ox.ac.uk

TREC QUERY:

> *To be relevant, a document will discuss a pending antitrust case and will identify the alleged violation as well as the government entity investigating the case. Identification of the industry and the companies involved is optional. The antitrust investigation must be a result of a complaint, NOT as part of a routine review.*

LEMMATIZED VERSION

> RELEVANT DOCUMENT DISCUSS PENDING ANTITRUST CASE IDENTIFY ALLEGE VIOLATION GOVERNMENT ENTITY INVESTIGATE CASE IDENTIFICATION INDUSTRY COMPANY INVOLVE OPTIONAL ANTITRUST INVESTIGATION RESULT COMPLAINT PART ROUTINE REVIEW

Figure 13.5: A text reduced by first performing morphological analysis and part-of-speech tagging and then removing stop words.

part-of-speech tagging, in addition to recognizing variant word forms (including words with internal variations such as forms of the word *tenir* in French), is that the part-of-speech tag permit the recognition of larger more precise structures such as noun phrases and named entities.

Eric Brill's tagger for English can be found online [17]. The TreeTagger is a another part-of-speech tagger developed at the Institute for Computational Linguistics of the University of Stuttgart. Executable code for Sun workstations, Linux and Windows PCs and Macs as well as parameter files for English, German, Italian, Spanish, French and old French can be downloaded online[18]. Part-of-speech taggers can be tested online[19]

Part-of-speech tagging of the text associated with images was part of the approach that teams such as [PLOG05, BHMF04, SNGI05, LCC04] used in ImageCLEF. These teams used noun phrases as indexes, and also used the noun phrases structure to perform more accurate translation for cross lingual retrieval, as we discuss below.

13.4.6 Entity Recognition

Entity recognition, also called named entity recognition, means identifying proper names (e.g. names of People, Organizations, Places), numerals, and abbrevia-

[17] www.cs.jhu.edu/~brill/RBT1_14.tar.Z

[18] At www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger

[19] http://www.comp.lancs.ac.uk/computing/research/ucrel/claws/trial.html, http://www.lingsoft.fi/cgi-bin/engcg, http://www.xrce.xerox.com/competencies/content-analysis/demos/english, http://ilk.kub.nl/ zavrel/tagtest.html

tions from texts. Until 1998, a series of competitions called the Message Understanding Conferences (MUC)[20] included tasks involving finding named entities. The MUC competition used the following XML tags to mark up named entities. ENAMEX for a named person or organization, NUMEX for a numerical quantity, and TIMEX for a time. Each tag could also contain arguments specifying the type of entity further. For example, a ENAMEX could be a PERSON or an ORGANIZATION.

Here is an example of marking up a simple sentence with a named entity recognizer, in which we see two different types of ENAMEX, a PERSON and an ORGANIZATION:

Mr. Smith bought 1000 shares of ABC Corp. in December, 2005.

<ENAMEX TYPE="PERSON">Mr. Smith</ENAMEX> bought <NUMEX TYPE="QUANTITY">1000</NUMEX> shares of <ENAMEX TYPE="ORGANIZATION">ABC Corp.</ENAMEX> in <TIMEX TYPE="DATE">December, 2005</TIMEX>.

The general approach to named entity recognition has been to write a set of regular expression patterns to capture the different types of entities[Gri97]. Many systems use a combination of dictionaries and these pattern based rules. For example, finding the token *Mr.* followed by a uppercase-initial word followed by a lower case word can provoke the insertion of the <ENAMEX TYPE="PERSON"> tags around the sequence. The first systems were built using static patterns and lists of known entities, and these approached finally evolved to more open and semi-automatic approaches [CV01] For example, some systems use known entities to examine the context in which these known items are found in order to extract new identifying patterns from text [CM02].

Named entity recognition is now commonly used in the task of question answering. For content based image retrieval, the following systems have specifically identified named entities in order to favour the retrieval of images whose captions contain those entities, or video sequences whose transcriptions contain references to them [PLOG05, BHMF04, QMS$^+$04].

13.4.7 Phrase Recognition

Once text has been tagged with parts of speech tags, structures beyond simple words can be recognized. In the last section, we mentioned named entities. In addition to these, one can extract concepts which are expressed as multiword expressions. In information retrieval, noun phrases are often used as index terms, considered more precise than individual words[Fag87, ELG$^+$93]. Noun phrases are often recognized using simple regular expressions [Chu88, KCGS96] or grammars

[20] see www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings

over part of speech tagged text. For example, Pereira and Wright [PW97] provide the following simple grammar that recognizes most English noun phrases. In this grammar, Det=determiner, Art=article, Adj=adjective, N=common-noun, P=preposition, PP=prepositional-phrase, Nom=nominal-phrase, NP=noun-phrase, PN=proper-noun and P=preposition.

NP ⇒ Det Nom | PN
Det ⇒ Art | NP 's
Nom ⇒ N | Nom PP | Adj Nom
PP ⇒ P NP

This grammar will recognize noun phrases such as:

The recent Security Council meeting
the civil rights of African Americans
Ford's new offer
the three people on the boat
foreign policy

For information retrieval, simple noun phrase (not including article or prepositional phrases) are used. Simple phrases are often called "chunks." Information retrieval systems [ELG$^+$93, BdCF$^+$03] will often break up a longer noun

phrase into its components and index these components along with the simple words. These systems often give greater weights to these phrases than to simple words. Here some examples of the index terms that will be retained from the noun phrases listed above (the words in the phrases will be either stemmed or lemmatized as shown in figures 13.3 and 13.5 above):

recent_Security_Council_meeting,
recent_Security_Council, Security_Council_meeting,
recent_Security, Security_Council, Council_meeting,
civil_rights
African_Americans
Ford_new_offer, Ford_new, new_offer
tthree_ people, boat
foreign_ policy

A noun phrase chunker, based on [RM95] which attempts to insert brackets marking noun phrases in text (which have been marked with part-of-speech tags in the same format as the output of Eric Brill's transformational tagger), can be downloaded from Mark Greenwood's site[21].

Another noun phrase extractor[22] not only extracts phrases but ranks them according to their discriminating power in a corpus of text. Noun phrases have

[21] http://www.dcs.shef.ac.uk/ mark/phd/software/chunker.html

[22] Found at http://www.nzdl.org/Kea/

been used as index terms by [MFSV$^+$04, PLOG05, BHMF04] for extracting more precise index terms in ImageCLEF'2004.

Noun phrase recognition is part of shallow parsing[23]. Of course, deeper parsing methods exist, and this research forms the central concern of computational linguistics as a field, but no such methods are currently used within the CBIR community. Interested readers may find a description of deeper natural language processing at the following sites[24]

13.4.8 Semantic Space Reduction and Structuring

Information retrieval systems will take the terms extracted by any of the above means (stemmed words, lemmatized words, named entities, or noun phrases) and use them to index the text associated with an image. These treatments then allow a text-based access to images. If a user poses a query using one or more of the words that index the image, then the image can be found and returned. But this is not often the case. Furnas *et al.* [FLGD87] show, consistently across a broad range of domains, that people will use the same term to describe the same object with a probability of less than 20%. Subjects were given examples of common things and asked to give a name for that thing. For example, when shown images of common objects, a number of subjects responded *fruit* when shown pictures of nectarines, of pears, and of raisins.

There are a few ways of circumventing this problem of language variability. One way is to use a hierarchically structured lexicon, which entails more general (hypernyms) and more specific relations (hyponyms) between words. Wordnet (see wordnet.princeton.edu and [MBF$^+$90]) is the most popular such resource in the research community. See figure 13.6 for snippet from WordNet.

Terminological ontology definition goes from simple lexicons or controlled vocabulary to thesauri, taxonomies with hierarchical relations between terms, or ontologies with named concepts. Some editing tools or environments have been developed in order to ease ontology construction (Kaon, OntoEdit, Protg, WebOde, etc.). However there are still based on a large part of manual work. Automation of ontology construction can be reached by a combined use of NLP and machine learning techniques applied to texts of the concerned domain.

The most commonly used lexical ontologyWordNet was exploited in [CC04] to resolve the following problem in the TRECVID campaign. As mentioned above, the task involves finding video sequences that correspond to a user information need expressed in natural language. The video contains both an automated speech-to-text transcription and high-level features that have been added to the video stream. The problem arises in that transcribed speech (usually a reporter

[23] http://jmlr.csail.mit.edu/papers/special/shallow_parsing02.html contains recent papers on shallow parsing.

[24] http://www.essex.ac.uk/linguistics/LFG/, http://www.cs.ru.nl/agfl/, http://cslu.cse.ogi.edu/HLTsurvey/ch3node5.html♯SECTION33

```
Institution: (hyponyms)
institution, establishment
 => charity
 => religion, faith, church
              => vicariate, vicarship
              => school, educational institution
              => academy, honorary society
              => foundation
              => bank, commercial bank
institution
              => orphanage, orphans' asylum
              => penal institution
constitution, establishment, formation, initiation, founding,
foundation, institution, origination, setting up, creation,
instauration
              => colonization, settlement
```

Figure 13.6: WordNet entries under the word *institution*

talking) often precedes the actual scene of interest. Cheng and Chen [CC04] map the words found in the transcribed voice stream to the high-level descriptors in a window of shots around the stream by calculating the shortest path between each transcribed word and the WordNet node corresponding to the high-level descriptor. For example, if a transcribed word was *bank* and the high-level descriptor was *institution* the distance between the two would be 2 according to the part of WordNet shown in figure 13.6. When the distance is less than a threshold, the transcribed words are "moved" to sequence labelled with the high level descriptor.

Aslangodan *et al.* [ATY$^+$97] used Wordnet to expand the words in a user query and the words in the metadata associated with images, in a similar way to calculate a distance between a user's query words and the words used to index images in their image retrieval platform called CANVAS.

Since WordNet represented the meaning of each word as a list of other words (called *synsets* in WordNet terminology), it can also be used simply as a synonym list. Martinez *et al.* [MFSV$^+$04] uses it in this way, optionally expanding a user keyword by all of its synonyms from its synset.

In addition to using a hierarchical graph to map words onto a semantic space as in WordNet, there are other techniques for reducing lexical space, some of which are described in the Language Modelling chapter (q.v.) of this report.

Latent semantic indexing (see lsa.colorado.edu) which maps words into a small number of dimension by calculating their co-occurrence with each other [DDF+90]. In this technique, each word becomes the index of row in a matrix. the columns of the matrix correspond to documents in which the word is found. The value of the matrix is the frequency with which the word was found in the document. This very large, and very sparse matrix is then reduced to three matrices: two triangular matrices and one diagonal matrix, using singular value decomposition [Kam98]. When the matrices are reordered so that the diagonal matrix (a matrix of eigenvalues) is in descending order (from largest singular value to smallest), and then this diagonal matrix is truncated after 200 to 300 values, the original words and documents are forced into a smaller dimension spaces (of 200 to 300 dimensions). This reduction method (which preserves a maximum amount of signal) “forces” words closer to each other, and all words within a certain radius in this reduced space can be considered as synonyms for retrieval purposes. For example, using a large corpus of general text up to first year college level reading, after all the words in 37,651 general language texts, and then reducing the dimensions of the space from this size (37,651 dimensions) to 300 dimensions, the words that appear closest to *truck*, for example, are

distance	*term*
1.00	truck
0.66	trailer
0.65	parked
0.64	pickup
0.59	drove
0.57	trucks
0.56	driver
0.55	highway
0.53	cab
0.52	drive
0.52	driving
0.52	driveway
0.51	seat
0.51	garage
0.49	stopped

One advantage of this space reduction is that each word (and each document) is now represented by a 300-dimension vector, and that a collection of words (in a query) can be represented as a sum of vectors. With the documents closest to this sum being the most similar (and hopefully most relevant).

This latent semantic indexing was used by [AGC+04, KBZ04] over training documents in order to map stemmed words found in the audio stream into a smaller space with the visual characteristics and the high level semantic descriptors attached to video frames in TRECVID.

All these techniques (using synonyms lists, distance within a graph structure such as WordNet, or latent semantic indexing) are used to reduce the space of words. Instead of considering each word as a separate dimension and relying on the user to use query words which exactly match the terms that index the images, these reductions techniques reduce the dimensions to be considered.

13.5 Cross-Lingual Mechanisms using in CBIR

In the ImageCLEF competitions, and more generally in the real-world collections, text queries for CBIR can come in many different languages, even though the images themselves are usually indexed using only one language. In this section, we will discuss some of the natural language processing approaches to dealing with this problem of multilinguality. Retrieving a document that has been indexed in one language via a query in a second language is called Cross Language

Information Retrieval [Gre98a]. Cross language retrieval is more important in image retrieval than in text retrieval because such systems have a wider applicability, since anyone can understand whether an image responds to their initial query. When text is retrieved, however, the user needs at least reading knowledge of the retrieved text in order to judge its relevance to their initial query.

13.5.1 Dictionary Lookup

The simplest, and widest used, approach to cross language retrieval for CBIR is to take the users' queries and perform dictionary lookup for each word of the query. [AOMN05, CMS05] tested such an approach in ImageCLEF. The problems with such an approach

Machine readable bilingual dictionaries exist for many languages. But, despite their name which indicates that they are readily exploitable by computers, machine readable dictionaries pose many problems. Their content is geared toward human exploitation and not readily exploitable by a computer. Much of the information about translations is implicitly included in dictionary entries and making this information explicit for use by a computer program is no small task.

Finding translations useful for Cross Language Information Retrieval in machine readable dictionary raises a host of problems. Sample problems are (a) missing word forms: for example, an entry for *electrostatic* may be included in the dictionary, but the word *electrostatically* may be missing since a human reader can readily reconstruct one form from the other. Stemming headwords can mitigate this problem at the expense of increased noise, such as seeing *marine* producing translations related to *marinated*; (b) spelling norms: usually only one national form appears as a headword in a given dictionary. For example, in a bilingual dictionary concerning English, the dictionary would have a heading for only one of the spellings, *colour* or *color*; (c) spelling conventions: the use of hyphenation varies from dictionary to dictionary as it does from text to text. One can see *fallout*, *fall out* and *fall-out* in texts, but all variants may not appear in the dictionary; (d) coverage [Gre98b]: general language dictionaries contain the most common words in a language, but rarer technical words are often missing. For example, the 1-million Brown corpus[FK82] contains the word *radiopasteurization* nine times, but this word would rarely appear in translation dictionaries; (e) proper names: country names and personal names often need to be translated. For example, the Russian president's name is written *Yeltsin* in English and *Elstine* in French.

Even when the headword is present, finding the translation within the dictionary entry can be difficult. The translation may be buried in a sample use. For example, the translation of a French word like *entamer* might be contained in a phrase *enter into a discussion with someone*, in which the extra words *discussion* and *someone* appear. *Someone* may be considered part of the meta-language of the dictionary and thus eliminated but the word *discussion* is part of a sample use and must be identified as extraneous to the translations of the headword. The specific word that translates the headword may not be identifiable by any automatic means. Added to this problem, of finding which words correspond to translations and which are extra information, are other nitty little problems, such as one we stumbled across: common structural inconsistencies in the SGML markup of machine-readable dictionaries, which may or may not appear in the printed version of the dictionary, but which cause automatic processing of definitions to break down or to produce erroneous entries.

13.6 Corpus-based translation

Another option to using translation dictionaries for finding translations is using a parallel corpus, i.e., the same text written in different languages. If the corpus is large enough, then simple statistical techniques[HdK97, ON03] can be used to produce bilingual term equivalents by comparing which strings co-occur in the same sentences over the whole corpus.

The most popular tool for determining translation equivalents from parallel corpus is Giza++[25] which has been used to build translation dictionaries from scratch, for example in the John Hopkins surprise language competitions[26]. This tool, based on pioneering work in vocabulary alignment by researchers at IBM [BCP$^+$90], takes sentence aligned bilingual texts as input and outputs vocabulary alignment files in which each alignment possesses a probability measures. It can also be used to align noun phrases if there are recognizable in the input and output languages (see section above).

Another approach for using corpora to filter relations is to first find noun phrases in the source language, then generate the possible translations of noun phrases in the target language by replacing each source word by all possible

[25] Available for download at www.fjoch.com/GIZA++.html

[26] See www.clef-campaign.org/workshop2003/presentations/tides-sl03.ppt

translations, and generate candidate target language noun phrases. For example, suppose our source language is Spanish and our target language is English. Now, suppose that we have the Spanish noun phrase *agua corriente* in our query to translate. Now, also suppose that our Spanish-to-English dictionary gives the following translations:

agua	⇒ water
corriente	⇒ common
	⇒ draft
	⇒ draught
	⇒ flowing
	⇒ going
	⇒ ordinary
	⇒ power
	⇒ running
	⇒ stream
	⇒ usual

Then we can the following translation and only retain the English noun phrases that have been found in the English side of the database. It has been found that taking the most frequently attested phrase in a large corpus (or on the Web) gives the correct translation of the phrase is the great majority of cases [Gre99b].

agua-corriente	common-water
agua-corriente	current-water
agua-corriente	draft-water
agua-corriente	draught-water
agua-corriente	flowing-water
agua-corriente	going-water
agua-corriente	ordinary-water
agua-corriente	power-water
agua-corriente	running-water
agua-corriente	stream-water
agua-corriente	usual-water

This simple generation and filtering of translation has proven to be very successful [QGE02] in cross language retrieval, eliminating much of the noise that comes from adding in all the translations available in a bilingual dictionary. It was used in ImageCLEF by the LIC2M [BHMF04].

13.6.1 Machine Translation

The most common solution to the cross language retrieval problem is to use online machine translations systems [27] to translate the input query into the language of the image index.

English	Dogs rounding up sheep
Italian	Dogs that assemble sheep
German	Dogs with sheep hats
Dutch	Dogs which sheep
French	Dogs gathering of the preois
Spanish	Dogs urging on ewes
Chinese	Catches up with the sheep the dog

This solution is not optimal [Gre98a] since only one lexical item is chosen for each input term, so the correct translation may be missing, whereas using the dictionary lookup or parallel corpus techniques, described above, more than one alternative translation may be kept. In addition, technical terms are often missing from general translation lexicons, Nonetheless, this technique is easy to implement and has been used to address the cross language retrieval problem in ImageCLEF by many CBIR researchers[SNGI05, PLOG05, MFSV$^+$04, Clo05, VTG03].

Some examples of the use of machine translations on non-English versions of the ImageCLEF query "dogs rounding up sheep" back into English were the following:

13.6.2 Conclusion

In this section, we have discussed methods for associating text with images, and shown then how this text is processed by researchers involved in content based image retrieval. We have talked about where the text can be found in relation to the image, and then we examined the natural language processing steps that have been employed in order to treat this text by content-based image retrieval researchers: language identification, stemming, morphological analysis, part-of-speech tagging, entity recognition, noun phrases extraction, semantic space reduction, and finally cross language retrieval using various methods. The more in-depth processing that natural language processing can provide: deep parsing, anaphora resolution, word sense disambiguation, etc. have not been employed by CBIR researchers mainly due to the lack of availability of robust versions of these tools.

[27] babelfish.altavista.com

Bibliography

[AGC$^+$04] J. Adcock, A. Girgensohn, M. Cooper, T. Liu, L. Wilcox, and E. Rieffel. Fxpal experiments for trecvid 2004. In *Proceedings of TRECVID 2004*. NIST, USA, 2004.

[AOMN05] C. Alvarez, A. I. Oumohmed, M. Mignotte, and J.-Y. Nie. Toward cross-language and cross-media image retrieval. In *CLEF*, pages 676–687. 2005.

[ATY$^+$97] Y. A. Aslandogan, C. Thier, C. T. Yu, J. Zou, and N. Rishe. Using semantic contents and wordnet in image retrieval. In *SIGIR '97: Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 27-31, 1997, Philadelphia, PA, USA*, pages 286–295. ACM, 1997.

[BCP$^+$90] P. F. Brown, J. Cocke, S. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, 1990.

[BdCF$^+$03] R. Besançon, G. de Chalendar, O. Ferret, C. Fluhr, O. Mesnard, and H. Naets. Concept-based searching and merging for multilingual information retrieval: First experiments at clef 2003. In *CLEF*, pages 174–184. 2003.

[BDF$^+$03] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *J. Mach. Learn. Res.*, 3:1107–1135, 2003. ISSN 1533-7928.

[BHMF04] R. Besancon, P. Hede, P. Moellic, and C. Fluhr. Lic2m experiments at imageclef 2004. In *CLEF Workshop*. 2004.

[BK03] K. R. Beesley and L. Karttunen. *Finite State Morphology*. CSLI, Palo Alto, 2003.

[Bra00] T. Brants. Inter-annotator agreement for a German newspaper corpus. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000)*. Athens, Greece, 2000.

[Bri95] E. Brill. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21(4):543–565, 1995.

[CC04] Y. Cheng and H. Chen. Aligning words from speech recognition and shots for video information retrieval. In *Proceedings of TRECVID 2004*. NIST, USA, 2004.

[CFG$^+$04] E. Cooke, P. Ferguson, G. Gaughan, C. Gurrin, G. Jones, H. L. Borgue, H. Lee, S. Marlow, K. McDonald, M. McHugh, N. Murphy, N. O'Connor, N. O'Hare, S. Rothwell, A. Smeaton, and P. Wilkins. Trecvid 2004 experiments in dublin city university. In *Proceedings of TREC Video Retrieval Evaluation*. 2004.

[Cha01] S. Chakrabarti. Integrating the document object model with hyperlinks for enhanced topic distillation and information extraction. In *WWW '01: Proceedings of the 10th international conference on World Wide Web*, pages 211–220. ACM Press, New York, NY, USA, 2001. ISBN 1-58113-348-0.

[CHL$^+$04] D. Cai, X. He, Z. Li, W.-Y. Ma, and J.-R. Wen. Hierarchical clustering of www image search results using visual, textual and link information. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 952–959. ACM Press, New York, NY, USA, 2004. ISBN 1-58113-893-8.

[Cho00] F. Y. Y. Choi. Advances in domain independent linear text segmentation. In *Proceedings of the first conference on North American chapter of the Association for Computational Linguistics*, pages 26–33. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2000.

[Chu88] K. Church. A stochastic parts program and noun phrase parser for unrestricted text. *Proceedings of the 2nd Conference on Applied Natural Language Processing*, pages 136–143, 1988.

[Clo05] P. Clough. Caption and query translation for cross-language image retrieval. In *CLEF*, pages 614–625. 2005.

[CM02] J. Callan and T. Mitamura. Knowledge-based extraction of named entities. In *CIKM '02: Proceedings of the eleventh international* *conference on Information and knowledge management*, pages 532–537. 2002. ISBN 1-58113-492-4.

[CMS05] P. Clough, H. Mueller, and M. Sanderson. The clef cross language image retrieval track imageclef 2004. In C. Peters, P. Clough, J. Gonzalo, G. Jones, M. Kluck, and B. Magnini, editors, *Fifth Workshop of the Cross-Language Evaluation Forum (CLEF 2004)*. Springer, Heidelberg, Germany, 2005.

[CV01] A. Cucchiarelli and P. Velardi. Unsupervised named entity recognition using syntactic and semantic contextual evidence. *Computational Linguistics*, 27(1):123–131, 2001.

[DDF+90] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic indexing. *Journal of the American Society for Information Science*, 41(6):391–407, October 1990.

[DKN05] T. Deselaers, D. Keysers, and H. Ney. Fire - flexible image retrieval engine: Imageclef 2004 evaluation. In *CLEF*, pages 688–698. 2005.

[ELG+93] D. A. Evans, R. G. Lefferts, G. Grefenstette, S. Handerson, A. Archbold, and W. R. Hersh. CLARIT TREC design, experiments, and results. In D. Harman, editor, *The First Text REtrieval Conference (TREC-1)*. U.S. Government Printing Office, Washington, 1993. NIST Special Publication 500–207.

[Fag87] J. L. Fagan. *Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Non-Syntactic Methods.* Ph.D. thesis, Cornell University, September 1987.

[FK82] W. N. Francis and H. Kucera. *Frequency Analysis of English.* Houghton Mifflin Company, Boston, 1982.

[FLGD87] G. W. Furnas, T. K. Landauer, L. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971, November 1987.

[FSN+95] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by image and video content: The qbic system. In *IEEE Computer*, volume 28, pages 23–32. 1995.

[GLA02] J. Gauvain, L. Lamel, and G. Adda. The limsi broadcast news transcription system. *Speech Communication*, 37(1-2):89–108, 2002.

[Gre95] G. Grefenstette. Comparing two language identification schemes. In *Proceedings of the 3rd International Conference on the Statistical Analysis of Textual Data, JADT'95*. Rome, Dec 11–13 1995.

[Gre98a] G. Grefenstette, editor. *Cross-Language Information Retrieval*. Kluwer Academic Publishers, Boston, 1998.

[Gre98b] G. Grefenstette. Evaluating the adequacy of a multilingual transfer dictionary for the cross language information retrieval. In A. Rubio, N. Gallardo, R. Castro, and A. Tejada, editors, *First International Conference on Language Resource and Evaluation*, pages 755–758. Granada, Spain, May 1998.

[Gre99a] G. Grefenstette. Tokenization. In H. van Halteren, editor, *Syntactic Wordclass Tagging*, 0-7923-5896-1. Kluwer Academic Publishers, Dordrecht, 1999.

[Gre99b] G. Grefenstette. The www as a resource for example-based machine translation tasks. In *Translating and the Computer*. ASLIB, London, October 1999.

[Gri97] R. Grishman. Information extraction: Techniques and challenges. In *SCIE*, pages 10–27. 1997.

[GT94] G. Grefenstette and P. Tapanainen. What is a word, what is a sentence? Problems of tokenization. In *3rd Conference on Computational Lexicography and Text Research*. COMPLEX'94, Budapest, Hungary, 7–10 July 1994. Http://www.xrce.xerox.com/publis/mltt/mltt-004.ps.

[HdK97] D. Hiemstra, F. de Jong, and W. Kraaij. A domain specific lexicon acquisition tool for cross-language information retrieval. In L. Deroye and C. Chrisment, editors, *RIAO'97, Computer-Assisted Information Searching on the Internet*, pages 217–232. Montreal, Canada, 1997.

[IV94] N. Ide and J. Véronis. MULTEXT: Multilingual text tools and corpora. In *Proceedings of the 15th. International Conference on Computational Linguistics (*COLING 94*)*, volume I, pages 588–592. Kyoto, Japan, 1994.

[JBJ$^+$05] G. J. F. Jones, M. Burke, J. Judge, A. Khasin, A. M. Lam-Adesina, and J. Wagner. Dublin city university at clef 2004: Experiments in monolingual, bilingual and multilingual retrieval. In *CLEF*, pages 207–220. 2005.

[JMHA04] M. Joint, P. Moellic, P. Hde, and P. Adam. Piria: a general tool for indexing, search, and retrieval of multimedia content. In E. Dougherty, J. Astola, and K. Egiazarian, editors, *Image Processing: Algorithms and Systems III*, volume 5298, pages 116–125. 2004.

[Kam98] J. Kamm. Singular value decomposition-based methods for signal and image restoration, 1998.

[KBZ04] M. Kherfi, D. Brahmi, and D. Ziou. Combining visual features with semantics for a more efficient image retrieval. In *Proceedings of IEEE/IAPR International Conference on Pattern Recognition (ICPR)*. August, 2004.

[KCGS96] L. Karttunen, J. Chanod, G. Grefenstette, and A. Schiller. Regular expression for language engineering. *Natural Language Engineering*, 2(4), December 1996.

[LCC04] W.-C. Lin, Y.-C. Chang, and H.-H. Chen. From text to image: Generating visual query for image retrieval. In *CLEF Workshop*. 2004.

[Lee92] G. Leech. 100 million words of english: the british national corpus. In *Language Research*, volume 28, pages 1–13. 1992.

[Lov68] J. B. Lovins. Development of a stemming alogrithm. *Mechanical Translation and Computational Linguistics*, 11:22–31, 1968.

[MBF$^+$90] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. Introduction to WordNet: An on-line lexical database. *Journal of Lexicography*, 3(4):235–244, 1990.

[McN05] P. McNamee. Language identification: a solved problem suitable for undergraduate instruction. *J. Comput. Small Coll.*, 20(3):94–101, 2005.

[MFSV$^+$04] J. Martinez-Fernandez, A. G. Serrano, J. Villena, V. D. M. Saenz, S. G. Tortosa, M. Castagnone, and J. Alonso. Miracle at imageclef 2004. In *CLEF Workshop*. 2004.

[MGMMC04] H. Mueller, A. Geissbuhler, S. Marchand-Maillet, and P. Clough. Benchmarking image retrieval applications. In *proceedings of the Seventh International Conference on Visual Information Systems*. San Francisco, September 8-10 2004.

[MGR05] H. Müller, A. Geissbühler, and P. Ruch. Imageclef 2004: Combining image and multi-lingual search for medical image retrieval. In *CLEF*, pages 718–727. 2005.

[MSM94] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330, 1994.

[ON03] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[PLOG05] V. Peinado, F. Lopez-Ostenero, and J. Gonzalo. Uned at imageclef 2005: Automatically structured queries with named entities over metadata. In *Cross Language Evaluation Forum, Working Notes for the CLEF 2005 Workshop*. 2005.

[Por80] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[PW97] F. C. N. Pereira and R. N. Wright. Finite-state approximation of phrase-structure grammars. In E. Roche and Y. Schabes, editors, *Finite-State Language Processing*, pages 149–173. MIT Press, Cambridge, 1997.

[QGE02] Y. Qu, G. Grefenstette, and D. A. Evans. Resolving translation ambiguity using monolingual corpora. In *CLEF*, pages 223–241. 2002.

[QMS$^+$04] G. M. Quenot, D. Mararu, S.Ayache, M. Charhad, L. Besacier, M. Guironnet, D. Pellerin, J. Gensel, and L. Carminati. Clips-lis-lsr-labri experiments in trecvid 2004. In *Proceedings of TREC Video Retrieval Evaluation*. 2004.

[RM95] L. Ramshaw and M. Marcus. Text chunking using transformation-based learning. In D. Yarovsky and K. Church, editors, *Proceedings of the Third Workshop on Very Large Corpora*, pages 82–94. Association for Computational Linguistics, Somerset, New Jersey, 1995.

[RS05] M. E. Ruiz and M. Srikanth. Ub at clef2004: Cross language information retrieval using statistical language models. In *CLEF*, pages 180–187. 2005.

[SB98] W. Skut and T. Brants. A maximum entropy partial parser for unrestricted text. In *Proceedings of the 6th ACL Workshop on*

Very Large Corpora (WVLC), pages 143–151. Montréal, Canada, 1998.

[SM83] G. Salton and M. McGill. *Introduction to Modern Information Retrieval.* McGraw–Hill, New York, 1983.

[SNGI05] M. Saiz-Noeda, J. L. V. González, and R. Izquierdo. Pattern-based image retrieval with constraints and preferences on imageclef 2004. In *CLEF*, pages 626–632. 2005.

[SWS$^+$00] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content based image retrieval at the end of the early years. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 22, pages 1349–1380. 2000.

[VTG03] B. Vrusias, M. Tariq, and L. Gillam. Scene of crime information system: Playing at st. andrews. In *CLEF*, pages 631–645. 2003.

[WMS$^+$93] R. Weischedel, M. Meteer, R. Schwartz, L. Ramshaw, and J. Palmucci. Coping with ambiguity and unknown words through probabilistic models. *CL*, 19(2):359–382, 1993.

[WS92] J. Wilbur and K. Sirotkin. The automatic identification of stopwords. *Journal of Information Science*, 18:45–55, 1992.

[ZDJ03] T. Zhang, F. Damerau, and D. Johnson. Updating an nlp system to fit new domains: an empirical study on the sentence segmentation problem. In W. Daelemans and M. Osborne, editors, *Proceedings of CoNLL-2003*, pages 56–62. Edmonton, Canada, 2003.

Chapter 14

Recommendations

There are three main problems that still need to be resolved in natural language processing for multimedia understanding. One problem is easy to state but requires much work: extending natural language tools to all the European languages. Most tools have been made and tuned for English, and to a lesser extent for Western European languages such as French, Spanish, German and Italian. Despite efforts such as MULTEXT-East[1] for developing at least some resources for some Eastern European languages, most other languages are "poor cousins" with respect to the tools and resources available to process them. These tools are necessary in order to recover and normalize the features that can be used for computer understanding of text.

A second problem lies in improving the features that can be used by machine learning programs over text. These features are currently normalized (or stemmed) words as seen above. Such features can be improved in two directions. First, using more elaborate and precise text structures. We saw above that some of these structures are currently being used, namely named entities and simple noun phrases, but beyond these simple to recognize structures there are syntactic structures (agent-action-object) structures that can provide better clues for classifying and retrieving information from text. As mentioned in the last paragraph, tools for recognizing these elaborate structures need also to be extended to languages other than English. The second research axis involving features text concerns creating reduced and manipulable feature spaces that are appropriate to specific domains (just as WordNet is a structured version of English[2] for general applications). This involves understanding the way words and phrases and combinations of words (for example, the agent-action-object structures) are related to each other: which are synonymous for the domain, which are more specific, which are more general. Work on automatically structuring this lexical space, for example into ontologies, still needs to be pursued. The interdependence between

[1] http://nl.ijs.si/ME

[2] See also www.illc.uva.nl/EuroWordNet/ for some work on other language version of WordNet.

text features has to be taken into account in machine learning applications which up until now tend to use a “bag of words” approach to text.

The third problem in natural language processing is largely unexplored but nonetheless at the heart of the MUSCLE consortium concern. What is lacking from the current research approaches is a rethinking of the relation between the lexicon and what is visible in an image. We propose that research be performed for extracting visual aspects of items, from text. For example, there are no lexical resources available that tell us what part of the lexicons corresponds to objects in the image and what part corresponds to abstract objects. Though WordNet has made a first attempt on marking what are objects (one of their nodes is labelled “object” with the additional label of “can cast a shadow”) but underneath this node, we find many things which are not physical objects (e.g.“tree of knowledge”) and outside this branch we find many things which can be pictured in an image (e.g. “snow” and “sea”). We need to rethink WordNet for CBIR purposes. There are other physical aspects of objects which might be useful for CBIR and which could be found using NLP. For example, it might be possible to find the common colour of things through text processing. For example, if you search for associations of colours and objects using the Web, you find that the most common colours associated (i.e. found in the same noun phrase) with *apple* are *green, red* and *golden.* It might be possible to associate colours to all objects using only text processing and statistics. Other visual aspects such as texture or shape might also be found through NLP. In addition, for pure image processing, it might be useful to know what objects appear in what settings. If image processing is able to recognize a background as being a sky, text processing should be able to provide some clues as to what appears in a sky, thus reducing the number of possibilities to be considered during object recognition. In other words, NLP might be useful not only in dealing with textual queries on image collections as described here in this chapter, but it might also be useful for bringing information to pure image processing tasks.

Conclusion

One of the major identified lock by the multimedia content retrieval community is the semantic gap when differences arise between the targeted content and the retrieval result. To fulfil the overall objectives of Muscle we need to address this problem for the future steps in WP5. This will assume that in the WP5, we will spend less effort on basic image analysis techniques that could not have direct exploitation for information retrieval.

We will focus on content description methods for each single modality. Our primer goal will be to provide, from low-level content description methods, the appropriate input to reduce the semantic gap. This assumes to start from the effort on fidelity of content descriptors to object, or more generally *"single modality event"*, detection and recognition methods. This includes face/human class objects since this class provides highly semantic information when available. Important efforts will be spent specially in the near future on **saliency** detection, description and retrieval for each modality. As noticed all over this report, machine learning techniques and content-description methods are continously embedded together. Therefore, collaboration with WP8 (as well as WP6) is one of the strong future issues.