MUSCLE

Network of Excellence

Multimedia Understanding through Semantics, Computation and Learning

Project no. FP6-507752

Deliverable D.6.1

Cross-Modal Integration for Performance Improving in Multimedia

Report on the State-of-the-Art

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):

- Petros Maragos (maragos@cs.ntua.gr)

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:

MUSCLE Network of Excellence

Work Package 6: Cross-Modal Integration for Performance Improving in Multimedia

Deliverable 1: Report on the State-of-the-Art

Edited by Petros Maragos
National Technical University of Athens, Greece

September 2004

Contents

Chapter 1

Introduction: Objectives

Author: P. Maragos, ICCS-NTUA

In multimedia analysis, most of the tools are usually devoted to a single modality, the other ones being treated as illustrations or complementary components. For example, web search engines do not use images, image retrieval systems barely mix textual and visual descriptions, video processing is usually done separately on sound and on images. One main reason for this is that the different media concern different and sometimes very separate scientific fields. However, even without learning, performance of multimedia analysis and understanding systems (especially in terms of robustness) can be greatly enhanced by combining different modalities. Thus one of the goals of this NoE is to develop algorithms and systems processing several different media present in the same multimedia. This requires a strong collaboration between many research groups. Examples of modalities to integrate include all possible combinations of: (i) vision and speech/audio; (ii) vision (or speech) and tactile; (iii) image/video (or speech/audio) and text; (iv) multiple-cue versions of vision and/or speech; (v) other semantic information or metadata. These combinations of modalities can be either of the cross-interaction type or of the cross-integration type. Interaction implies an information reaction-diffusion among modalities with feedback control of one modality by others. Integration involves exploiting heterogeneous information cumulatively from various modalities and a data feature fusion toward improved performance. Work Package 6 of the MUSCLE NoE addresses research on the theory and applications of multimedia analysis approaches that improve robustness and performance through cross-modal interaction and/or integration.

In this report, the contributing partners try to review and evaluate the current state-of-the-art in the areas spanned by the scientific and technological objectives of Work Package 6. These could be generally classified in the following categories: 1) Cross-Modal Interaction in Multimedia Problems, and 2) Cross-Modal Integration for Multimedia Analysis and Recognition.

Chapter 2

Cross-Modal Interaction in Multimedia Problems

Interaction of multiple modalities, e.g. vision, speech-audio, and text is explored in cases where one modality is directly affected by the others and when the goal is to improve performance over single modality processing. All possible combinations of modalities and their interactions are of research interest to this NoE. However, for brevity of exposition, we outline next only two such specific problem areas: the first involves interaction between the speech and the vision modality; the second deals with many interacting modalities, i.e., vision, text, speech and, possibly, tactile information.

2.1 Audio-Visual Interaction for Speech Recognition - Part 1: Overview

Authors: C. Kotropoulos and G. Patsis, AUTH

The bimodality of human speech perception and the need for robust automatic speech recognition in noisy environments has motivated significant research effort towards audio-visual automatic speech recognition (AVASR) [104, 76, 256]. In addition to the acoustic input, AVASR utilizes speech information present in the speaker's mouth region, and has been successfully demonstrated to improve the accuracy and noise robustness of automatic speech recognition (ASR) systems for both small- and large-vocabulary tasks. For example, bimodal ASR of a small-vocabulary task under speech babble noise at 0 dB signal-to-noise ratio (SNR) is reported in [256] to achieve the performance of an acoustic-only recognizer at 10 dB, i.e., to provide an "effective" SNR gain of 10 dB in the ASR "usable" range. Significant gains are also reported on the same task even in clean acoustic conditions. Furthermore, an 8 dB "effective" SNR gain is demonstrated for large-vocabulary continuous speech recognition (LVCSR) [256].

A frame-level phonetic classification problem is considered using two single-stream Gaussian Mixture Models (GMMs) in [121]. Audio and video streams are adaptively weighted using a cumulative mean of the sample confidence values over past frames in addition to the present sample confidence value. The confidence values for audio and video decisions are computed using L-statistics (linear combination of order-statistics) of log-likelihoods against phone models. It is shown through various experiments, on a database of about 15,000 sentences from large vocabulary continuous speech, that the proposed approach results in better classification accuracy as compared to phonetic classification results to audio only or video only. Two different groups of visual features that can be used in addition to audio to improve ASR, high- and low-level visual features are compared in [32]. Facial Animation Parameters (FAPs), supported by the MPEG-4 standard for the visual representation of speech are used as high-level visual features. Principal component analysis (PCA) based projection weights of the intensity images of the mouth area were used as low-level visual features. PCA was also applied on the FAPs. An AVASR system was developed and its performance was compared for two different visual feature groups, following two approaches. The first approach assumes the same dimensionality for both high- and low-level visual features, while in the second approach the percentage of statistical variance described by the visual features used was the same. Multi-stream HMMs and a late integration approach were used to integrate audio and visual information and perform continuous AVASR experiments. Experiments were performed at various SNRs (0- 30dB) with additive white Gaussian noise on a relatively large vocabulary (approximately 1000 words) database. Conclusions were drawn on the trade off between the dimensionality of the visual features and the amount of speechreading information contained in them and its influence on the AVASR performance.

One of the most exciting research topics in joint audio-visual speech processing is the integration of the two speech informative inputs. A hybrid Hidden Markov Model (HMM)/Artificial Neural Network (ANN) architecture for audio-speech recognition is used in [143]. The system integrates the audio and visual classifiers at a likelihood, decision level, in a "state-synchronous" fashion. Various schemes to weigh the posterior likelihoods of the audio and visual-only ANNs with appropriate stream exponents are first considered. Their multiplicative combination which respects their class-conditional independence is demonstrated to be superior. An adaptive estimation of the combination weights, based on reliability of each stream of information, is also studied. Results are reported on a single-speaker, connected digits database. A different approach that concentrates on the "state-asynchronous" architectures for audio-visual fusion by means of HMMs is followed in [230]. The integration both at the feature level as well as at the likelihood level using a multitude of Bayesian network models, such as the product HMM, the factorial HMM (FHMM) and the coupled HMM (CHMM) is considered. Iterative algorithms for obtaining maximum likelihood estimates of the model parameters as well as their initial estimates are presented. The performance of CHMM and FHMM for audio-visual integration is compared with the existing models in speaker dependent audio-visual isolated word recognition. The statistical properties of both the CHMM and FHMM allow to model the state asynchrony of the audio and visual observation sequences while preserving their natural correlation over time. In the reported experiments, the CHMM performs best overall, outperforming all the existing models and the FHMM. A model based on Dynamic Bayesian Networks (DBNs) to integrate information from multiple audio and visual streams is proposed in [126]. The DBN based system is compared with a classical HMMs for both the single and two stream integration problems A new model is proposed to integrate information from three or more streams derived from different modalities and the new model's performance is compared with that of a synchronous integration scheme. A new technique to estimate stream confidence measures for the integration of three or more streams is also developed and implemented. Results from the developed implementation using the Clemson University Audio Visual Experiments (CUAVE) database indicate an absolute improvement of about 4% in word accuracy in the -4 to 10 db average case when making use of two audio and one video streams for the mixed integration models over the synchronous models.

In spite of such impressive benefits however, AVASR systems have yet to be deployed in real-life applications. This is mainly due to issues related to the extraction of visual speech information, most notably robustness and computational complexity of visual front end processing. Regarding the former, most research work has concentrated and reported on databases recorded under controlled, visually "clean" conditions. Such sets contain high-resolution video of the subjects' full frontal face, with very limited variation in head pose and subject-camera distance, rather uniform lighting, and, in most cases, constant background. In contrast, little is known about AVASR performance in realistic, non-ideal environments, where the visual channel quality is poor, thus presenting challenges to speech-informative visual feature extraction. Preliminary experiments reported in [255] show significant degradation of the visual modality ASR benefit in such visually challenging domains, for example videos recorded in moving automobiles, or by low quality web-cams. A review on the efforts in IBM T. Watson Research center to overcome the lack of robust and fast visual feature extraction procedures based on an appearance-based visual feature representation of the speaker's mouth region is presented in [257]. AVASR in realistic, visually challenging domains, where lighting, background, and head-pose vary significantly is discussed. To enhance visual-front-end robustness in such environments, an improved statistical-based face detection algorithm is employed that significantly outperforms the baseline scheme. However, visual-only recognition remains inferior to visually “clean” (studio-like) data, thus demonstrating the importance of accurate mouth region extraction. A wearable audio-visual sensor to directly capture the mouth region is proposed, thus eliminating face detection. Its use improves visual-only recognition, even over full-face videos recorded in the studio-like environment. The speed issue in visual feature extraction is addressed by discussing a real-time AVASR prototype implementation. Visual speech front-end processing is further studied in [163]. An improved appearance-based face and feature detection algorithm is proposed that utilizes Gaussian mixture model classifiers. This method is shown to improve the accuracy of face and feature detection, and thus visual speech recognition, over the baseline system. An improved audio-visual ASR resulting in a 10% relative reduction of the word-error-rate in noisy speech is reported.

A novel method based on the use of automatic lipreading in order to extract an acoustic speech signal from other acoustic signals by exploiting its coherence with speaker’s lip movements is proposed in [286]. In contrast to the classical blind source separation techniques, the authors explore the case of an additive stationary mixture of decorrelated sources with no further assumptions on independence or non-Gaussian character. It is shown that it is possible to separate a source when some of its spectral characteristics are provided to the system. If a statistical model of the joint probability of visual and spectral audio inputs is learnt to quantify the audio-visual coherence, separation can be achieved by maximizing the probability. A number of preliminary separation results on a corpus of vowel-plosive-vowel sequences uttered by a single speaker, embedded in a mixture of other voices is presented. The separation can be quite good for mixtures of 2, 3, and 5 sources.

2.2 Audio-Visual Interaction for Speech Recognition - Part 2: Tutorial Review

Authors: G. Papandreou, P. Maragos and A. Katsamanis, ICCS-NTUA

2.2.1 Introduction

Commercial *Automatic Speech Recognition* (ASR) systems are uni-modal, i.e., only use features extracted from the audio signal to perform recognition. Although audio-only speech recognition is a mature technology with a long record of significant research and development achievements [260], current uni-modal ASR systems can work reliably only under rather constrained conditions, where restrictive assumptions regarding the size of the vocabulary, the amount of noise etc can be made. These shortcomings have seriously undermined the role of ASR as a pervasive *Human-Computer Interaction* (HCI) technology and have limited the applicability of speech recognition systems to well-defined applications like dictation and low-to-medium vocabulary transaction processing systems.

On the other hand, speech recognition by humans is fundamentally multimodal. Although audio is the most important source of information for speech recognition, people also use visual cues as a complimentary aid in order to successfully perceive speech. The key role of the visual modality is apparent in situations where the audio signal is either unavailable or severely degraded, as is the case with hearing-impaired listeners or very noisy environments, where seeing the speaker's face is indispensable in recognizing what has been spoken. Human perception weighs the visual information more when two articulated sounds are not easily discernible acoustically, but can be discriminated visually due to a different place of articulation, as is the case with the phonemes /n/ and /m/(,/t/ and /p/ or /b/ and /v/), which sound very similar but look quite different [253]. This phenomenon is lucidly manifested in a well known psychological illusion, the so-called *McGurk effect* [215, 211]. In their experiments, McGurk and MacDonald found out that when somebody experiences contradictory audio and visual speech cues, he/she tends to perceive whatever is most consistent with both sensory information. For example, if in a videotape the audio syllable ”ba” is dubbed onto a visual ”ga”, then fusion of the auditory stimulus of the front consonant in ”ba” (vocal tract closed at the lips) with the visual stimulus of the back consonant in ”ga” (closure at the back of the throat) will yield in most people the perception of the middle consonant ”da”. The McGurk effect shows that human speech understanding is multi-modal, resulting from sensory integration of audio and visual stimuli.

These findings provide strong motivation for the Speech Recognition community to do research in exploiting visual information for speech recognition, thus enhancing ASR systems with speechreading capabilities [289]. Research in this relatively new area has shown that multimodal ASR systems can perform better than their audio-only or visual-only counterparts. The first such results where reported back in the early 80's by Petajan [247]. The performance gain becomes more substantial in scenarios where the quality of the audio signal is degraded, as is the case with particularly noisy environments such as a vehicle's cabin [250]. The potential of significant performance improvement of audiovisual ASR systems, combined with the fact that image capturing devices are getting cheaper, has increased the commercial interest in them.

However, the design of robust audio-visual ASR systems, which perform better than their audio-only analogues in all scenarios, poses new research challenges. Two new major issues arise in the design of audio-visual ASR systems, namely:

- *Selection and robust extraction of visual speech features*. From the extremely high data rate of the raw video stream, one has to choose a small number of salient features which have good discriminatory power for speech recognition and can be extracted automatically, robustly and with low computational cost.

- *Optimal fusion of the audio and visual features*. Inference should be based on the heterogenous pool of audio and visual features in a way that ensures that the combined audiovisual system outperforms its audio-only counterpart in practically all scenarios. This is definitely non-trivial, given that the relative quality of the audio and visual features can vary dramatically during a typical session.

A block diagram of an audiovisual ASR system depicting its main components is shown in Fig. 2.1.

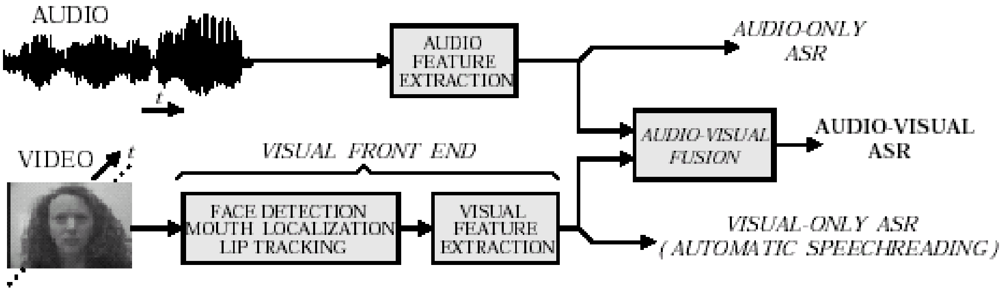
AUDIO
t
AUDIO FEATURE EXTRACTION
AUDIO-ONLY ASR
VIDEO
t
VISUAL FRONT END
AUDIO-VISUAL FUSION
AUDIO-VISUAL ASR
FACE DETECTION MOUTH LOCALIZATION LIP TRACKING
VISUAL FEATURE EXTRACTION
VISUAL-ONLY ASR (AUTOMATIC SPEECHREADING)

Figure 2.1: Block diagram of a typical audiovisual ASR system. Figure from [258].

In this report we will attempt to review the main trends in the literature for addressing the two aforementioned major research challenges in *Audiovisual ASR* (AV-ASR), namely the design of the visual front-end and the fusion of audio-visual features. These two issues are by no means trivial to solve and are the subject of current intensive research. We will also refer to the main audiovisual databases serving as benchmarks in comparing the performance of different techniques. Other reviews of the subject are [258, 289].

2.2.2 The Visual Front End of Audio-Visual Automatic Speech Recognition Systems

As emphasized in Section 2.2.1, the design of the visual front end is one of the main challenges in building an audiovisual ASR system. One can generally identify the following steps in the processing of the visual modality input [144, 258]:

- *Active speaker's face detection and tracking*. The speaker's face contains significant information for visual speechreading and thus the system must reliably locate and track it.

- *Facial model fitting*. In the case an exemplar is used to model the lips' borders (e.g. an active contour model [168]) or even the whole face (e.g. an active appearance model [88]), the exemplar is roughly initialized near the identified *Region Of Interest* (ROI) and evolves towards its final best-fit configuration.

- *Visual features extraction*. After the ROI has been identified, a number of visual features are extracted from it. These features can be appearance-based, shape-based, or a combination of them.

- *Visual features post-processing*. The stream of visual features might need some further processing steps, like upsampling to match with the stream rate of auditory features, before it can be used as input to the audiovisual fusion inference machine.

We proceed by examining the main methodologies to address these problems.

Active speaker's face detection and tracking

In order for the visual modality to be useful for the task of speech recognition, an AV-ASR system must be able to reliably detect and track the speaker's face, which contains important visual cues for speech recognition. This task usually involves detecting human faces in the first video frame which are used in the sequel to initialize a face tracker. The detector is typically triggered again periodically to accommodate for a possible tracking failure and detect any new face entering the scene in the meantime. In the case of multiple persons being present in the scene, an additional requirement for the visual front-end is to determine who of them is the active speaker. Therefore the main tasks that the visual front-end of an AV-ASR system needs to do in order to detect and track the speaker's face are *face detection, face tracking* and *active speaker detection*.

Human **face detection** in still images is one of the main problems in Computer Vision, arising in important applications, like face recognition, area surveillance and Human–Computer Interaction. Although the problem is well studied, it remains unsolved in its full generality, when many faces in various scales and orientations are allowed in the scene, the lighting conditions can vary widely and real-time performance is required. Among the face detection methods, the ones based on the statistical learning paradigm, requiring a training phase on an appropriate training set, are in the focus of current research and demonstrate the best performance. We will review here the main such methods. For another recent survey consult [323].

Although most of the latest approaches to human face detection can deal with faces at arbitrary scales, the majority of them can only be utilized to detect forward-facing faces. A notable exception is the method of Schneiderman and Kanade, which can handle faces in profile pose, after training the system with examples of humans facing at three different orientations [272]. Most methods under consideration can account for reasonable changes in the lighting conditions and camera variations through an appropriate image pre-processing step, which can involve illumination correction by subtraction of a linear best-fit function and histogram equalization, respectively [238, 265, 294].

Schneiderman and Kanade apply statistical likelihood tests, using feature output histograms to create their detector scheme in [272]. Sung and Poggio model faces and non-faces as mixtures of anisotropic Gaussians in a high-dimensional linear space [294]. Rowley and Kanade use neural network-based filters in [265], obtaining good early results in what has apparently become a benchmark of sorts for face detection schemes. In another early work, Papageorgiou et al. propose a general object detection scheme which uses a wavelet representation and statistical learning techniques [244]. Osuna et al. apply Vapnik's support vector machine technique to face detection in [238], and Romdhani et al. improve on that work by creating reduced training vector sets for their classifier in [264]. In perhaps the most impressive paper, Viola and Jones use the concept of an integral image, along with a rectangular feature representation and a boosting algorithm as its learning method, to detect faces at 15 frames per second [311]. This represents an improvement in computation time of an order of magnitude over previous implementations of face detection algorithms.

After a face has been detected in a frame, a **face tracking** module is needed to track it for the subsequent frames until the face detector is triggered again. This processing step is only needed when the face detector is not activated every single frame. The tracker might be designed to track either the speaker's lips only or his/her whole face, depending on the modeling approach. Since this is tightly connected with the model used to describe the speaker's face, we defer discussing the issue until 2.2.2.

The **detection of the active speaker's face** (in contrast to non-speakers' faces) in the case that many people are present in the visual scene is the final step to be done in order to successfully locate and track the speaker's face. This is especially important for the deployment of AV-ASR systems in realistic environments, like meeting rooms with many participants, where a number of attendants speak one after each other and the system needs to discern who is the active speaker at each particular moment. The same requirement is also posed by other applications, such as tele-conference systems, where the camera needs to zoom on the active speaker [92].

For that purpose a number of techniques have been devised. While early attempts were based on audio-only sound source localization techniques, most of the recent approaches to the problem utilize both audio and visual cues to successfully identify the speaker among the different persons present in the scene. The resulting fused system can be more robust to both vision and audio background clutter than corresponding single-modal systems. We will next describe two such techniques, presented in [43] and [78]. Other relevant references are [310], [335] and [145].

The authors in [43] present an algorithm which can track a sound-producing moving object in a cluttered, noisy scene using an array of two microphones and a single camera, without requiring any manual calibration. Their algorithm is based on graphical models that combine audio and video variables. The model they propose uses unobserved variables to describe the data in terms of the process that generates them. It is therefore able to capture and exploit the statistical structure of the audio and video data separately, as well as their mutual dependencies. Model parameters are learned from data via an Expectation-Maximization algorithm, and automatic calibration is performed as part of this procedure. Tracking is done by Bayesian inference of the object location from data.

In [78], a particle filter-based fusion framework that combines both bottom-up and top-down approaches to probabilistically fuse multi-modal cues is proposed. Three trackers (two visual-based and one audio-based) are designed to generate effective bottom-up proposals for the fuser. The fuser performs reliable tracking by verifying hypotheses over multiple likelihood models from multiple cues in a top-down fashion. The proposed framework is a closed-loop system where the fuser and trackers coordinate their tracking information. Since the reliability of each tracker varies throughout a session, they propose a method to dynamically evaluate the performance of the individual trackers and update their weights. The resulting speaker tracking system can reliably fuse object contour, color and sound source location cues in real-time.

Facial model fitting

After the speaker's face has been located, speech related information must be extracted from it. There are generally two approaches for achieving this goal. The first, model-free approach is to find a rectangular ROI around the mouth area and subsequently use a transformation of the raw pixel values in this ROI as a feature set. This approach will be discussed further in 2.2.2. The second, model-based approach is to try and match a facial shape or appearance model to the observed face. The parameters of this model can be used in the sequel for creating the feature set. In this subsection we will describe some representative approaches used to model the face or parts of it.

Active contours (also called snakes) [168, 156] have been used to model and track the speaker's lips [52, 169, 80]. In the snakes framework, a contour is defined as a concatenation of spline segments, with the coordinates of the control points and the spline coefficients being the parameters of the model. The contour is driven towards the lip borders, minimizing an energy criterion which favors both good matching of the contour with the feature of interest (external, image dependent force) and curve smoothness (internal, elastic force) [168]. The snake can be attracted towards either grey-scale edges [52] or towards edges in a transformed color space, designed specifically to maximize the contrast between the color of the lips and the color of the skin [169]. The shape of the snake is constrained to resemble the shape of the lips by making the shape parameter vector lie in a subspace learnt by means of a PCA analysis on a training set, which gives the main modes of lip shape variation (eigenlips) [52]. Accurate temporal tracking is achieved by means of a Kalman filter or a particle filter, after learning the lip motion dynamics [155, 169]. Tracking both the outer and inner lip contours rather than the outer lip contour only can increase lipreading performance [169]. It would be interesting to test in this problem the performance of non-parametric, geometric active contour models, such as geodesic active contours with shape priors [61, 188].

A related method to model parts of the face is through **deformable templates** [328, 135], utilized in the AV-ASR context in [144, 64]. Using deformable templates, the shape of the lips is modeled by a small number of curves, capturing the shape of the lips with very few parameters, as in Fig. 2.2. The template is allowed to deform by minimizing an associated cost functional via gradient descent. The local minimum achieved corresponds to a shape that matches the given image closely, provided that the initial condition is good. A heuristic is to use the final position of the template in the previous frame as initial condition in the new frame.

Another powerful approach to human face modeling is through *Active Shape Models* (ASMs) and *Active Appearance Models* (AAMs). In **ASMs**, which were first proposed in [87], an object's shape is modeled by a set of landmark points. The shape's main modes of variation are learned by means of a PCA analysis, using a training set of images where the landmarks are manually annotated. A local appearance profile in the neighborhood of each landmark is also learned during the training phase. In order to localize a novel shape, an ASM is initialized

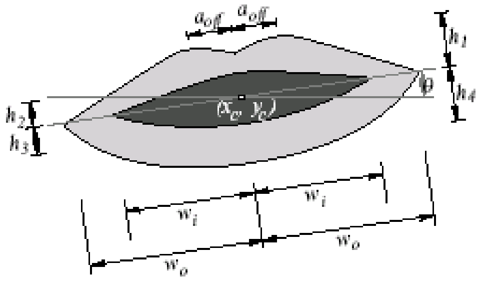

Figure 2.2: A deformable template of the lips shape, consisting of two parabolas for the inside edges and three additional curves for the outside edges. The shape is described by 12 parameters. Figure from [144].

near it and is updated iteratively by alternating between two steps: 1) the position of each landmark is updated independently in small steps by searching for the position in the landmark's neighborhood which has the best fitting local profile and 2) the resulting shape is projected back to the learned shape manifold. The algorithm converges fast, but the model can be stuck to local minima due to the local nature of computations, which means that a good initialization is crucial. ASMs were first applied to lipreading by [201].

AAMs [88] and the closely related methods of Morphable Models [166] and Active Blobs [276] are an extension of ASMs, in the sense that the main modes of variation of both the object's shape and appearance (after warped to the mean shape) are learned from the training data by applying PCA twice. A third PCA is sometimes applied to capture the correlations between shape and appearance parameters. The model fits a novel image by minimizing the appearance reconstruction error. Algorithms for this nonlinear optimization task can be found in [88, 212], with recent extensions presented in [58] and the references therein. By taking the appearance information into account, AAMs are more robust to the initial condition than ASMs, but they are more sensitive to illumination variations and their sufficient training requires more examples, which can be a problem due to the limited availability of big training sets. AAMs were first used for AV-ASR in [213], where it was shown that they perform better than ASMs in this task.

Visual Features

In order the visual information in a video stream depicting the speaker to be useful for speech recognition, a compact set of about 10-100 informative features must be extracted from each frame and be used later for statistical classification. These features should be as robust as possible when different people talk and their poses or lighting conditions vary. Unlike audio-only speech recognition, where the properties of various sets of features are well understood, research on the relative merits of alternative visual features for visual speech recognition is far less mature. One can classify the various visual features proposed in the literature into three broad categories [258], namely:

- *appearance features*, which directly use (a transform of) the pixel values in the mouth ROI.
- *shape features*, where the parameters of a shape model are used to derive the features.
- *combined shape and appearance features*, where information from both the shape and the appearance of the ROI are used to form the features.

In the sequel we further discuss these approaches. One can refer to [258] for more details.

Appearance Features This approach to feature extraction doesn't need a shape model to be fit on the speaker's face, as described in 2.2.2. It only requires that a ROI around the speaker's mouth has been identified at each video frame. This ROI, usually scale and rotation normalized, can be a rectangle containing the mouth or a larger part of the face, such as the jaw and cheeks [254]. It can also be a disk-like region around the lips, where an exponential window has been applied to accommodate for the edge effects [230]. The preliminary feature vector consists then of the concatenated greyscale or color pixel values, having length $d = N$ or $d = 3N$, respectively, where N is the number of pixels in the ROI.

However the length d of the preliminary feature vector is still too large to allow successful statistical modeling of the visual speech by means of *Hidden Markov Models* (HMMs) [260]. A dimensionality reduction technique is therefore usually applied, before these features are used for speech recognition.

One usual unsupervised learning approach to dimensionality reduction is through **Principal Component Analysis** (PCA) [206]. This method uses a training set of ROI images to learn an affine space of reduced dimensionality $k < d$ capturing the main modes of variation (eigenimages) in the class of ROI images. A novel ROI can then be described faithfully by its projection on this affine space, as a mean ROI plus a linear combination of eigenimages with a feature vector of length k. PCA-based dimensionality reduction has been used extensively in the context of speechreading (see the references in [258]).

Another popular approach for coding visual features for visual speech recognition is to use image compression techniques, based on standard **image transforms**, such as the *Discrete Cosine Transform* (DCT) and the *Discrete Wavelet Transform* (DWT) [252].

The aforementioned methodologies attack the problem of dimensionality reduction in a data compression framework. However the ultimate goal in our application is classification of visual speech samples, which means that the dimensionality reduction technique should attempt to find a reduced set of optimal discriminating features, using supervised learning methods. The simplest such method is **Linear Discriminant Analysis** (LDA). LDA finds a projection matrix such that the between-class variance of the projected data is maximized relative to their within-class variance and can be shown to be the optimal decision rule in the case that the classes are Gaussian and have a common covariance matrix [103]. For more details and enhancements to LDA, see the references in [258].

Shape Features This approach to feature extraction assumes that the shape of the ROI contains enough information for visual speech recognition. Shape based features are extracted utilizing shape models, which were described in 2.2.2. The features of interest are usually extracted from the shape of the speaker's lips area, although in some cases larger parts of the face are used [214]. The feature vector can either describe some geometric properties of the lip's shape or, alternatively, just consist of the parameters of the specific shape model used for shape fitting.

After the lip contours have been identified in the current frame with one of the methods discussed in Sec. 2.2.2, a vector of **lip geometric features** can be extracted from them to subsequently be used for visual speech recognition. Examples of such features, which have been used extensively by various AV-ASR systems, include the height, width, perimeter, as well as the area contained within the contour. Another approach to describe the shape of the lips succinctly is through lip shape moments or Fourier descriptors. See [258] for more details.

If a parametric shape model has been fitted to the speaker's face, as described in 2.2.2, it is natural to use the **shape model parameters** as shape features. The parameters of Active Contours tracking the speaker's lips have been used for speech recognition in many speechreading systems, including the ones presented in [52, 80]. The variables controlling the shape of deformable templates fitted on the lips have been utilized for the same task in [144, 64]. Finally, the parameters of ASMs have been used for AV-ASR in [201, 213], among others.

Combined Shape and Appearance Features Since appearance or shape only features are useful for visual speechreading, it seems plausible that features encoding both shape and appearance information can be most efficient in capturing visual information for speech recognition. Some researchers have therefore tried to integrate joint shape and appearance features in their AV-ASR systems. In most early attempts to achieve this goal, features from each category are just concatenated. In [201], for example, shape parameters from a fitted ASM model were combined with the intensity profiles around each landmark of the ASM (see 2.2.2) in order to enhance the ASM with appearance information. The advent of statistical tools such as the AAMs, described in 2.2.2, which can model both the shape and the appearance of the face in a unified framework, has resulted in a more principled way to describe visual speech information with combined shape and appearance features. The first AV-ASR system utilizing AAMs in its visual front-end is reported in [213].

Visual Feature Comparison An efficiency comparison between different sets of visual features is complicated, because the various researchers usually test their methods on different AV-ASR corpora and on different tasks, ranging from connected digit recognition to *Large Vocabulary Continuous Speech Recognition* (LVCSR). The lack of adequately large databases also makes it sometimes difficult to sufficiently train both visual models (such as AAMs) and inference engines for the recognition task, which means that reduced reported performance of some models might in some cases be not due to their inadequacy but due to undertraining.

Despite these difficulties, a couple of comparisons among different visual features are worth mentioning. First of all, a number of studies have shown appearance information, in the form of either appearance features or joint shape and appearance features, is indispensable in visual speech recognition. For example, AAMs outperform ASMs in the work reported in [213] and simple DCT appearance features give better results than lip contour geometric features in [252]. The experiment on speaker-independent LVCSR task documented in [214] shows that simple, image transform, appearance-based features, which require no particular training, perform better than AAMs, whose efficiency critically depends on careful training. For more references, one can consult [258].

The advent and wide availability of big AV-ASR corpora will certainly ease comparisons between different visual features and boost research and development on audiovisual speech recognition, the same way that audio-only speech recognition benefited from the availability of big audio databases to the wider research community.

Visual Feature Postprocessing After one has decided on a proper set of visual features, some final processing steps need to be done, before these features enter the audiovisual speech inference engine. These postprocessing steps are necessary in order to make the stream of visual features compatible with the stream of audio features. The main action to be taken refers to the synchronization of the two streams. Since the auditory features are extracted at a high rate of about 100 Hz and the video rate is typically only about 25 Hz, the visual features stream is usually upsampled to the audio features stream rate. A couple of additional post-processing techniques, inspired by the traditional, audio-only speech recognition technology, have been proposed in order to normalize the visual features by mean subtraction and capture the dynamics of visual speech. For details about these techniques, one can consult [258] and the references therein.

A block diagram summarizing a typical visual front-end of an AV-ASR system, depicting its main components, is shown in Fig. 2.3.

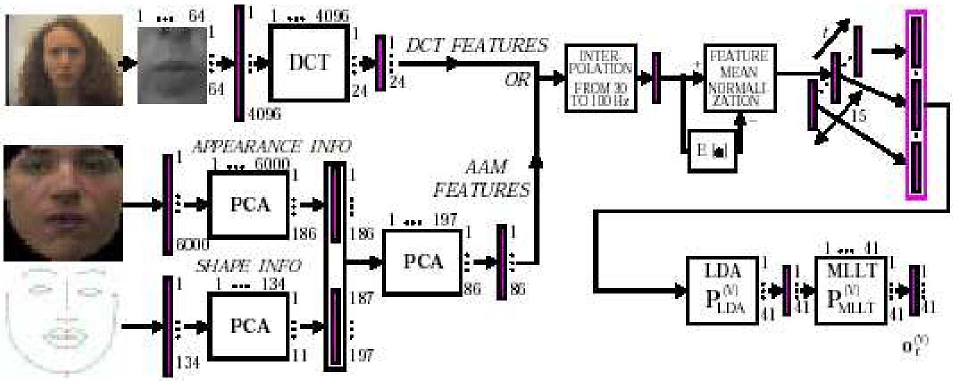

Figure 2.3: Block diagram of a typical audiovisual ASR system. Figure from [233].

2.2.3 Audio Visual Integration for Speech Recognition

To successfully address the problem of audiovisual speech recognition, it certainly does not suffice to complement the set of robust audio features with an informative set of visual features from the video stream, following the methods described in the previous section. The main task that needs to be addressed next is the fusion of the heterogenous pool of audio and visual features in a way that ensures that the combined audiovisual system outperforms its audio-only counterpart in all practical scenarios [27]. This task is complicated due to a couple of issues, the main of them being:

- Audio and visual speech asynchrony. Although the audio and visual observation sequences are certainly correlated over time, they exhibit state asynchrony, with visual activity preceding auditory activity by as much as 120 ms [52], close to the average duration of a phoneme. As we will see, this asynchrony renders modeling audiovisual speech with conventional HMMs [260] problematic.

- The relative speech discriminative power of the audio and visual streams can vary dramatically during a typical session in unconstrained environments, making their optimal fusion a challenging task.

Therefore successful audio and visual feature integration requires utilization of advanced techniques and models for cross-modal information fusion. This research area is currently very active and many different paradigms have been proposed for addressing the general problem. In the sequel we will confine ourselves to reviewing the main research trends for feature fusion in the context of audiovisual feature integration. The general problem of cross-modal information fusion is reviewed separately in Section 3.2.

One can generally classify the various approaches to audio and visual feature integration into three main categories [144], depending on the stage that the audio and visual streams are fused, namely early, intermediate and late integration techniques. The early integration paradigm tries to deal with the speech recognition problem utilizing a single classifier (usually a conventional HMM), which acts on the concatenation of the audiovisual features, often after they have undergone an appropriate transformation. The intermediate integration class comprises classification methods that explicitly model the two different modalities and their interaction. The overall class conditional likelihood used in recognition with these models can then be computed by combining the class conditional likelihood of each modality. The inference engines used for these models are usually various HMM extensions, belonging to the general class of *Dynamic Bayesian Networks* (DBNs) [96]. We note here that research is conducted on DBN-based classifier architectures for audio-only speech recognition as well [336]. Finally, late integration models utilize different, independent classifiers for the audio and visual features and the final classification decision is reached by combining the partial outputs of the uni-modal classifiers. In the following, after we discuss general issues related to audiovisual speech modeling for ASR, we will examine representative instances of the different fusion techniques and comment on their relative strengths and weaknesses.

Speech Units for Audio-Visual Automatic Speech Recognition

A central issue in modeling speech for ASR purposes is deciding on the basic acoustic units to be used to form the speech models for recognition.

It is well known that the basic phonetic/linguistic speech units are the *phonemes*, each of them corresponding to a different excitement or configuration of the vocal tract. The categorization of auditory distinct speech sounds into phonemes is often subjective and differs for various languages. For the American English their number is estimated to be around 45 [260]. The visual analogue of phonemes in speechreading are the *visemes* [289], which correspond to the visually distinct configurations of the speaker's visible articulators, such as the lips, the teeth and the tongue. Since only a few of the articulators are visible, visemes are fewer than phonemes. In the experiments reported in [233], 13 visemes have been used.

However, it has been shown in the audio-only ASR technology that using the phonemes per se as the basic acoustic unit for speech recognition with HMMs leads to rather poor performance [260]. The acoustic variability of the phonemes due to context is large and makes them inappropriate for this role. Instead, whole-word models have been used as the basic speech unit, both for isolated or connected word recognition tasks, because their acoustic representation varies less. In the case of LVCSR tasks, where having a different model for each word becomes impractical, context-dependent subword units are successfully used as the basic speech building blocks. These units are usually learned from a training set in a top-down clustering fashion [327]: One starts with a complete set of context-dependent elements, typically all possible tri-phones, and then sequentially merges the most similar of them until a desired number of units is reached. This number of distinct units to retain is subject to a tradeoff: more such units can capture the context-dependent variability of speech better, but their sufficient training requires bigger corpora. A few thousands (typically 1000–3000) context-dependent subword units are used in practice.

In the case of audio-visual speech recognition, the problem of selecting the basic audiovisual unit for speechreading is generally more difficult. For small vocabulary, isolated or connected word recognition experiments, which constitute the majority of AV-ASR experiments till now, whole words have been most often utilized as elementary speech units [213, 230]. The situation is more complicated in the case of LVCSR experiments, which have only recently been reported [233, 214, 230]. Obtaining context-dependent audiovisual subword speech units is more difficult than in the audio-only case due to the unavailability of sufficiently large audiovisual training sets. Therefore the researchers either use context-dependent speech classes obtained after audio-only training [233, 214] or just use context-independent audio-visual speech units [230]. A more principled treatment of the problem that will retain the benefits of context-dependent modeling in the audiovisual case is still an open problem [233].

Early Integration Techniques for Audio-Visual ASR

The simplest approach to audio-visual feature integration is through early integration methods. This class of techniques utilize a single classifier, avoiding the explicit modeling of the two different speech modalities.

In early integration approaches to audiovisual integration one simply concatenates the audio and visual feature vectors to obtain a single combined audiovisual vector [27]. In order to reduce the length of the resulting feature vector, dimensionality reduction techniques like LDA are usually applied before the feature vector finally feeds the recognition engine [253]. Additional details on AV-ASR techniques based on audiovisual feature fusion can be found in [258].

The classifier utilized by most early integration audiovisual ASR systems is a conventional HMM, which is trained using the mixed audiovisual feature vector. The structure of one such left-to-right HMM, with conditional probabilities modeled as mixtures of Gaussians, is depicted in Fig 2.4. In that figure, following the convention used in the Bayesian Networks literature [89], different symbols for hidden and observed nodes have been used.

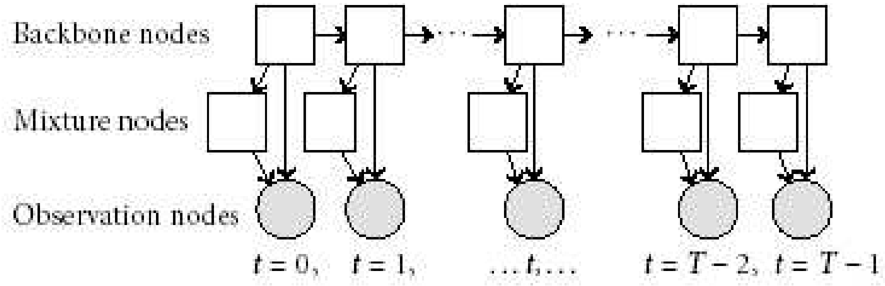

Figure 2.4: Recognition in the early integration case by means of an HMM. Figure from [230].

The main advantage of early integration approaches to audio-visual integration is their conceptual simplicity and the utilization of conventional HMMs for recognition. This makes them particularly easy to implement using readily available HMM tools which are very popular with the speech recognition community. However, since these techniques avoid the explicit modeling of the multimodal nature of speech, they fail to model both the fluctuations in the relative reliability and the asynchrony problems between the two distinct audio and visual streams. These issues are addressed by the intermediate integration techniques discussed next.

Intermediate Integration Techniques for Audio-Visual ASR

In order to address the shortcomings of early integration techniques, the multimodality of audiovisual speech needs to be modeled more faithfully than conventional HMMs allow. A number of HMM extensions, belonging to the class of DBNs, have been proposed in the literature in an attempt to address this goal. These models have two aspects in common, namely:

- They attempt to explicitly capture the reliability of each modality by letting the class conditional observation likelihood to be the product of the observation likelihoods of the single-stream components, raised to appropriate stream exponents that vary depending on the confidence of each stream.

- They allow modeling the state asynchrony of the audio and visual streams while preserving their natural correlation over time.

We will next examine the most popular HMM extensions which are useful for audiovisual speech recognition, following mainly the comprehensive exposition of [230], where the interested reader can resort for more details.

A first representative of the class of models under discussion is the *multistream HMM* (MS-HMM). The MS-HMM preserves the state synchrony of HMMs but allows weighting the observation likelihoods of the different streams according to their reliability, as shown in the left diagram of Fig. 2.5. It has also been used for audio-only speech recognition, where each stream is devoted to an heterogenous class of auditory features. For its training simple extensions of standard HMM training techniques suffice [327]. MS-HMMs have been often utilized for audiovisual speech recognition [258].

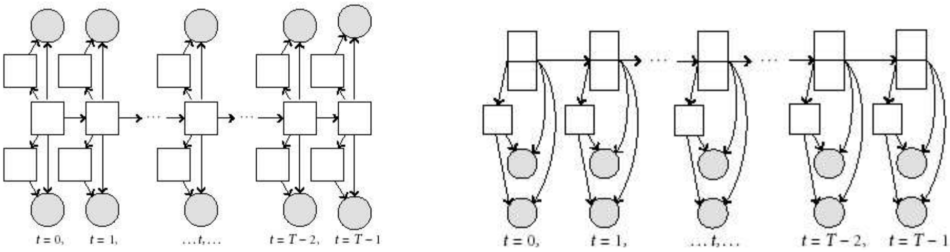

Figure 2.5: Recognition using a Multistream (left) and a Product (right) HMM. Figures from [230].

The *product HMM* (P-HMM) [127] relaxes the state synchrony constraint of the MS-HMM in order to account for the observed state asynchrony between the audio and the visual streams. This is achieved by having two backbone nodes per state, one for each stream, as depicted in the right diagram of Fig. 2.5. In that diagram one can also notice that both the observation likelihood and the transition probabilities depend on both hidden variables, which preserves the natural correlation over time of the audio and visual features.

However, this joint modeling of the transition probabilities and observation likelihoods in the P-HMM can be superfluous, leading to an excessive number of parameters that need to be estimated. Two successful models that take a more moderate approach to the tradeoff between independent and joint modeling of the transition probabilities and observation likelihoods are the factorial HMM and the coupled HMM.

The topology of the *factorial HMM* (F-HMM) [120] can be seen in the left diagram of Fig. 2.6. One can observe in this diagram that the F-HMM uses two backbone nodes, which jointly influence both observation likelihoods. However, unlike P-HMM, the transition probabilities of the backbone nodes of the F-HMM are assumed to be independent of each other.

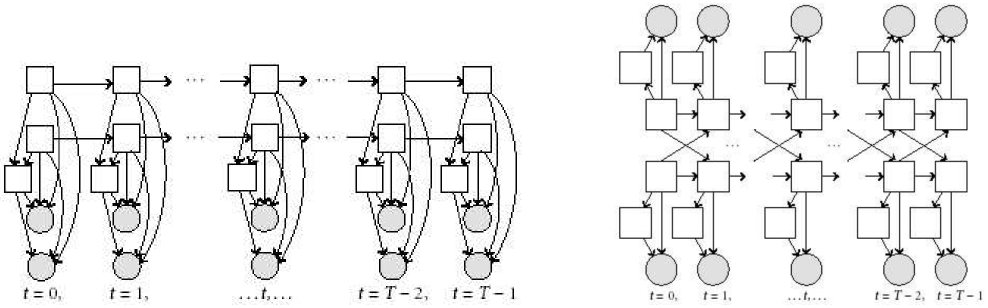

Figure 2.6: Recognition by means of a Factorial (left) and a Coupled (right) HMM. Figures from [230].

The *coupled HMM* (C-HMM) [51] is a model often used for various applications requiring multimodal fusion. Unlike the F-HMM, the two backbone nodes of the C-HMM are allowed to interact directly with each other, letting their transition probabilities depend on both of them. At the same time, the observation likelihoods of each stream depend on its corresponding backbone only. These properties can be seen in the right diagram of Fig. 2.6. Application of the C-HMM for audiovisual speech recognition has been reported, among others, in [83, 230, 290]. A comparative study of the relative performance of different DBN-based architectures (including the MS-HMM, the P-HMM, the F-HMM and the C-HMM) for the isolated word recognition task has shown that the C-HMM outperforms the other models in almost all cases.

As far as model training is concerned, Expectation-Maximization algorithms for parameter estimation of both F-HMMs and C-HMMs are presented in [230]. It should be noted here that a good initial estimate of the model parameters is needed in order a good local optimum to be attained. Viterbi-type algorithms can be used for that purpose.

Late Integration Techniques for Audio-Visual ASR

Late integration models utilize two independent HMMs, one for the audio and one for the visual features stream, which can be trained separately. The final classification decision is reached by combining the partial outputs of the uni-modal classifiers. The correlations between the visual and acoustic channels are not captured by these models.

In more detail, for small-vocabulary, isolated word speech recognition, late integration can be easily implemented by combining the audio- and visual-only log-likelihood scores for each word model in the vocabulary, given the acoustic and visual observations [27]. However, this approach is intractable in the case of connected word recognition or LVCSR, where the number of alternative paths explodes. A good heuristic alternative in that case is through lattice rescoring [327]. The n most promising hypotheses are extracted from the audio-only recognizer and they are rescored after taking the visual evidence into account. The hypothesis that has the highest combined score is then selected. More details about this approach can be found in [233].

Stream Reliability Modeling

A final issue that needs to be discussed here is how to choose the relative weight of audio and visual observations when computing the combined observation score. This problem is also known as *stream exponent* estimation when one linearly combines log-likelihoods, as is usually the case. This is an important point for both intermediate and late integration techniques.

A first approach to this problem is to try out different values of the exponents for a given audio-channel SNR level and find the weighting factors that minimize the the recognition error on a held-out validation set. Then, in the working phase of the AV-ASR system, an estimate of the SNR level can be used to select the appropriate pre-computed values of the stream exponents [27, 91, 230]. The values of these exponents can be updated at the utterance or even state level to account for abrupt change in the environmental conditions.

Some further refinements are also plausible, although they pose more demands during the training phase of the system. One such improvement is to allow the stream exponent to depend on the speech class, too. This is reasonable because different speech classes manifest themselves more in the audio domain and some others more on the visual domain [233]. Another enhancement is to also take the reliability of the visual stream into account when determining the stream exponents [27]. For example, when a face tracker has lost track of the speaker's face, the visual features are not reliable and their exponent must be reduced. For further approaches to this problem, one is advised to consult [258].

2.3 Various Cross-media Interaction Scenarios

2.3.1 Multimodal Human-Computer Interfaces

Authors: M. Perakakis, M. Toutoudakis and A. Potamianos, TUC

As defined in [241] multimodal interfaces process two or more combined user input modalities such as speech, pen, touch, manual gestures, gaze and head and body movements in a coordinated manner with multimedia system output. This is a paradigm shift away from conventional WIMP interfaces towards more flexible, efficient and powerfully expressive means of human computer interaction. Multimodal interfaces are expected to be easier to learn and use, more robust and more adaptable to the user, tasks and usage environment.

In this section we give some motivations for multimodal interfaces and briefly review some examples of multimodal applications. Then we examine topics such as multimodal interaction (mainly fusion and integration techniques) and dialogue management handling. Finally, we examine issues regarding to system architectures and conclude with various efforts for standardizing multimodal interaction specially on the web.

Motivation

As shown in [85], multimodal interfaces may have many advantages: they prevent errors, bring robustness to the interface, help the user to correct errors or recover from them easier, bring more bandwidth to the communication and add alternative communication methods to different situations and environments. Disambiguation of error-prone modalities using multimodal interfaces is the main motivation for the use of multiple modalities in many systems. As shown in [239] error-prone technologies can compensate each other, rather than bring redundancy to the interface and reduce the need for error correction.

It should be noted, however, that multiple modalities alone do not bring these benefits to the interface: currently there is too much hype in multimodal systems, and the use of multiple modalities may be ineffective or even disadvantageous. Oviatt [240] has presented common misconceptions (myths) of multimodal interfaces most of them related to the use of speech as an input modality.

Example applications

¿From the historical perspective, multimodality offers promising opportunities, as presented Bolt's “ Put-That-There” system [49]. Combined pointing and speech inputs offered a natural way to communicate, and later authors added gaze direction tracking to disambiguate other modalities [49]. Other early systems used speech input along with keyboard and mouse in an effort to support greater expressive power for complex visual manipulation. Speech technology advances in late 1980s allowed speech to become an alternative to keyboard leading to map and tourist information systems like CUBRICON [228] and Georal [282].

Bimodal systems that combine speech and pen-input or speech and lip-movements emerged in 1990s leading to work on integration and synchronization issues and the development of new architectures to support them. Speech and pen-input (2D or 3D gestures) involving hundreds of different interpretations beyond pointing have advanced rapidly leading to mature research (e.g. Quickset [85]) and commercial systems. Speech and lip movement systems exploit the detailed classification of human lip movements (visemes) and viseme-phoneme mappings that occur during articulatory speech. Lip movement offers speech recognition robustness in noisy environments and animated character systems with coordinated text-to-speech output and lip movement.

Examples of such systems include (*talking heads* or *speaking agents*) include the Rea system [62], KTH's August, Adapt and Pixie systems ([130, 129] and [128]). These systems use audiovisual speech synthesis and anthropomorphic figures to convey facial expressions and head or body movements. Systems with animated interactive characters have also been constructed such systems built at DFKI (see [34]). These systems mainly focus on multimedia presentation techniques and agent technologies. Information kiosks (*intelligent kiosks*) such as SmartKom ([312]) project use speech and haptics to provide interface for users in public places (e.g. museums).

Recently, systems combining 3 or more modalities such as person identification and verification systems which use both physiological(retina, fingerprints) and behavioral (voice, handwriting) modalities have been developed. Also there is an increased interest in *passive input modes* which refer to naturally occurring user behavior that is unobtrusively monitored by a computer (e.g., facial expressions). *Ambient intelligence* and blending of active and passive modes is a promising direction to this end.

Multimodal interaction

When the results of multiple modalities are combined, fusion techniques are needed for their integration. Early multimodal interfaces were based on a specific control structure for multimodal fusion. For example Bolt's demo, searches for a synchronized gestural act that designates the spoken referent. To support more broadly functional multimodal systems though, general processing architectures have been developed which handle a variety of multimodal integration patterns and support joint processing of modalities.

Multimodality levels According to [235], we can differentiate four different uses of multimodal inputs and outputs depending on fusion type and use of modalities: *exclusive* (independent fusion, sequential modalities), *concurrent* (independent fusion, parallel modalities), *alternate* (combined fusion, sequential modalities) and *synergistic* (combined fusion, parallel modalities).

	use of modalities	
fusion type	Sequential	Parallel
Independent	**Exclusive**	**Concurrent**
Combined	**Alternate**	**Synergistic**

Table 2.1: Levels of multimodality

The *exclusive* use of modalities is the most straightforward, since independent modalities can be used at different times. This mode imposes the least requirements for a multimodal system. With *concurrent* fusion, modalities are used concurrently but their results are not combined in any way (they can for example be used for different tasks). Conversely, modalities are *alternative* when they are used at different times but their results are combined in some way. Finally, with synergistic use (the most sophisticated use of multimodality), modalities are combined at the same time. This puts heavy demands on the system and is seldom used.

Fusion techniques Multimodal systems usually integrate signals at the feature level (*early fusion*) or at a higher semantic level (*late fusion*). In an early fusion architecture, the signal-level recognition process in one mode influences the course of recognition in the other and so, is considered more appropriate for closely temporally synchronized input modalities, such as speech and lip movements. Systems using the late fusion approach have been applied to processing multimodal speech and pen input or manual gesturing, for which the input modes are less coupled temporally and provide different but complementary information. Late semantic integration systems use individual recognizers that can be trained using unimodal data and can be scaled up easier in number of input modes or vocabulary size.

Alternatively, one can consider fusion at *lexical*, *syntactic* or *semantic* levels. Lexical fusion is used when hardware primitives are mapped to application events, syntactic fusion synchronizes different modalities and forms a complete representation of these and semantic fusion represents functional aspects of the interface by defining how interaction tasks are represented using different modalities.

Integration techniques Multimodal systems based on late (semantic) fusion integrate common meaning representations derived from different modalities into a combined final interpretation. This requires: a common meaning representation framework for all modalities used and a well-defined operation for integrating the *partial meanings*.

Meaning representation uses data structures such as *frames* [221] and *feature structures* [170] or *typed feature structures* [59]. Frames represent objects and relations as consisting of nested sets of attribute/value pairs while feature structures goes further to use shared variables to indicate common substructures. Typed feature structures are pervasive in natural language processing, and their primary operation is unification, which determines the consistency of two representational structures and, if they are consistent, combines them.

Various integration techniques have been derived so far: *frame-based integration* techniques use a strategy of recursively matching and merging attribute/value data structures (e.g., [277]) while *unification-based integration* techniques use *logic-based* methods for integrating the *partial meaning fragments*. Unification-based architectures have only recently been applied to multimodal system design [165, 164]. Some important unification-based integration techniques include feature-structure and symbolic unification. *Feature-structure unification* is considered well suited to multimodal integration, because unification can combine complementary or redundant input from both modes, but it rules out contradictory input. *Symbolic unification* which combined with statistical processing techniques results *hybrid symbolic/statistical* architectures which represent a new direction for multimodal system development and achieve very robust functioning, compared with either an early or late fusion approach alone.

Fission techniques Fission is a process in which modalities are selected for outputs. For example, in multimodal speech systems outputs can be expressed by using synthesized speech, non-speech audio, text or graphics. Fission techniques have not gained as much attention as fusion techniques, and this is often thought of as a simple practical issue. Work has been done mainly in the context of multimedia systems, for example in the area of automated multimedia systems ([34]). The focus in these systems is often more on the rendering of the information for different medias than in the selection of the media for different elements.

Dialogue management

The dialogue manager controls the overall interaction between the system and the user by finding suitable system actions which corresponds to the user input, which can be seen as a mapping from a user action to a system action, or from one system state to an another system state. Many dialogue managers (especially those used in text based systems), tend to extend their functionality to natural language understanding and generation as well. The communication with the data source such as the database is often one of the tasks of the dialogue manager.

Although dialogue management is a fairly mature research area, and many sophisticated text-based systems have been constructed, they have not been proven to be very successful. Although speech is different from text, many of the principles found in these systems can be used in speech dialogue systems.

Dialogue initiative strategies One of the key aspects in dialogue management is how the initiative is handled. The dialogue management strategy used may be system-initiative, user-initiative or mixed-initiative. We briefly review and compare each strategy in this section.

System-initiative dialogue strategy With system-initiative dialogue, the computer asks questions from the user, and when the necessary information has been received, a solution is computed and a response is produced. It can be highly efficient since the paths which the dialogue flow can take are limited and predictable. It is most suitable for well-defined, sequential tasks where the system needs to know certain pieces of information in order to perform a task (e.g. a database queries for bus timetables or flights).

One key advantage is the predictable nature of the dialogue flow, which makes it possible to use context-sensitive recognition grammars, for every dialogue state, helping the recognizer to achieve more robust recognition results. Also, since the system asks questions, it can easier guide the user to reach his/her goal, making sure all necessary steps will be performed. This makes the user feel comfortable with the system and prevent disorientation (specially for the novice user). The main disadvantage is the clumsiness of interaction with experienced users, because only single pieces of information are exchanged in every dialogue turn, making the dialogue advance slowly. The system may let experienced users pass certain dialogue turns by using more complicated expressions, but this may make the dialogue management complicated and recognition grammars more complex.

User-initiative dialogue strategy User-initiative dialogue strategy assumes that the user knows what to do and how to interact with the system. The system waits for user inputs and reacts to these by performing corresponding operations. The main advantage is that experienced users are able to use the system freely and perform operations any way they like without the system getting in their way. This is also natural in open-ended tasks which have many independent subtasks. The main weakness is that they assume (and require) that users are familiar with the system and know how to speak. This imposes very open language models to the system and cognitive load to the user, which are both difficult to handle.

Mixed-initiative dialogue strategy There is no single dialogue management strategy which is suitable for all situations. Different users and application domains have different needs, and different dialogue handling strategies may needed even inside a single application. With mixed-initiative strategy, the initiative can be taken either by the user or the system. The user has freedom to take the initiative, but when there are problems in the communication, or the task requires it, the system takes the initiative and guides the interaction. If properly constructed, a mixed- initiative system can help the user by employing system-initiative strategy while still preserving the freedom and efficiency of user-initiative strategy.

Dialogue control models Various models have been devised for the overall structure of the dialog flow, like event and plan based, agent-based or even theorem-proving ones. In practice however models based on finite-state machines and frame-based are the ones most usually used.

Finite-state machines Most of the current commercial speech applications use finite-state machines for dialogue control, because they are well known and straightforward to use. Finite-state machine consists of a set of nodes representing dialogue states and a set of arcs between the nodes, which move the dialogue from one state to another. The resulting network represents the whole dialogue structure, and paths through the network represent all the possible dialogues which the system is able to produce. If there are numerous states, and a lot of transitions between states, the complexity of the dialogue model increases rapidly, so they are mostly suitable for small-scale and system-initiative applications

Frame-based systems Frame-based systems use collections of information (templates) as a basis for dialogue management and the purpose of the dialogue is to fill necessary information slots and then perform a query or similar operation on the basis of the frame. In contrast to the state-machine approach, they are more open, since there is no predefined dialogue flow (dialogue control is centralized and usually specified with a single algorithm), but instead the required information is fixed. Multiple slots can be filled by using a single utterance, and the order of filling the slots is usually free. It is a more natural choice for implementing mixed-initiative dialogue strategy, since the computer may take the initiative by simply asking for the required fields. VoiceXML applications for example, use the frame-based approach for standard dialogue control and the event-based one for cases like error handling.

Architectures

Software architecture in multimodal systems was considered to be mostly a practical issue and it was often not modeled explicitly. This has already been noticed [216] and as such systems become more complicated, even more focus will be needed on system architectures. This is important, since systems can be more efficient and easier to build and maintain if proper architectural models are used.

Most systems are very complex in terms of architecture and software design, so they usually mix and exploit many software architectural styles and models like the pipe-and-filter, finite-state machine, event-based model, client-server, object-oriented and agent-based ones. For example spoken dialogue systems are usually structured either in a pipeline fashion or using the client-server model with a central component, which facilitates the interaction between other components, like the Galaxy-II architecture. Multimodal systems are based on even more sophisticated architectures like [180] or agent architectures (see 2.3.1).

In this paragraph we briefly examine the differences in requirements between GUI and multimodal architectures, then we turn our attention to the most common style of multimodal architectures and conclude with the development of multimodal frameworks which facilitate easier development of such complex applications.

GUIs vs Multimodal architectures In Contrast with GUIs which assume that there is a single event stream that controls the underlying event loop, multimodal interfaces process continuous and simultaneous input from parallel incoming streams. Also GUIs assume that the basic interface actions, such as selection of an item, are atomic and unambiguous events, while multimodal systems process input modes using recognition-based technologies, which are designed to handle uncertainty and entail probabilistic methods of processing. Finally, multimodal interfaces that process two or more recognition-based input streams require time-stamping of input, and the development of temporal constraints on mode fusion operations.

Multimodal agent-based architectures The most common infrastructure that has been adopted by the multimodal research community involves *multi-agent architectures*, such as the *Open Agent Architecture* [210] and *Adaptive Agent*

Architecture [180]. Multi-agent architectures provide essential infrastructure for coordinating the many complex modules needed to implement multimodal system processing, and they permit doing so in a distributed manner. In a multi-agent architecture, the many components needed to support the multimodal system (e.g., speech recognition, gesture recognition, natural language processing, multimodal integration) may be written in different programming languages, on different machines, and with different operating systems. Agent communication languages are being developed that can handle asynchronous delivery, triggered responses, multi-casting and other concepts from distributed systems.

Using a multi-agent architecture, for example, speech and gestures can arrive in parallel or asynchronously via individual modality agents, with the results recognized and passed to a *facilitator*. These results, typically an nbest list of conjectured lexical items and related time-stamp information, then are routed to appropriate agents for further language processing. Next, sets of meaning fragments derived from the speech, etc. arrive at the multimodal integrator which decides whether and how long to wait for recognition results from other modalities, based on the system's temporal thresholds. It fuses the meaning fragments into a semantically-and temporally-compatible whole interpretation before passing the results back to the facilitator. At this point, the system' s final multimodal interpretation is confirmed by the interface, delivered as multimedia feedback to the user, and executed by any relevant applications.

Multimodal frameworks Despite the availability of high-accuracy speech recognizers and the maturing of multimodal devices such as gaze trackers, touch screens, and gesture trackers, very little applications take advantage of these technologies. One reason for this may be that the cost in time of implementing a multimodal interface is prohibitive. One desiring to equip an application with such an interface must usually start from scratch, implementing access to external sensors, developing ambiguity resolution algorithms, etc. However, when properly implemented, a large part of the code in a multimodal system can be reused. This aspect has been identified and many multimodal application frameworks have recently appeared such as VTT's *Jaspis* and *Jaspis2* frameworks [302, 301], Rutgers CAIP Center framework [115], the embassi system [109] and more.

Standards

W3C standards The number of different kinds of devices that can access the Web has grown from a small number with essentially the same core capabilities to many hundreds with a wide variety of different capabilities like mobile phones, smart phones, personal digital assistants, kiosks, automotive interfaces, etc.

Device Independence The range of capabilities for input and output and the range of markup languages and networks supported greatly complicate the task of authoring web sites and applications that can be accessed by users whatever device they choose to use. The W3C *Device Independence Working Group* encompasses the techniques required to make such support an affordable reality. In particular the activity focuses on methods by which the characteristics of the device are made available for use in the processing associated with device independence methods to assist authors in creating sites and applications that can support device independence in ways that allow it to be widely employed. The group has overtaken work by *Composite Capability/Preference Profiles Working Group* (CC/PP see [10]), and through coordination with *Web Accessibility Initiative* [17] and *MultiModal Interaction Working Group* [13] continues it's work on avoiding the fragmentation of the Web into spaces that are accessible only from subsets of devices.

Multimodal Interaction Activity Mobile profiles have emerged using a number of W3C specifications like XHTML, making mobile access more close to reality. Recently, a tremendous growth of interest in using speech as a means to interact with Web-based services over the telephone (*Voice Browser Activity*), but spoken interfaces (based upon VoiceXML), only prompt users with pre-recorded or synthetic speech and understand simple words or phrases. There is now an emerging interest in richer forms of interaction, combining speech with other modalities. Multimodal interaction will enable the user to speak, write and type, as well as hear and see using a more natural user interface than today's single mode browsers.

The *Multimodal Interaction Activity* is extending the Web user interface to allow multiple modes of interaction(aural, visual and tactile), offering users the means to provide input using their voice or their hands via a key pad, keyboard, mouse, or stylus. For output, users will be able to listen to spoken prompts and audio, and to view information on graphical displays. By allowing multiple modes of interaction (GUI, speech, vision, pen, gestures, haptic interfaces, etc), to any device it facilitates the dream of *accessibility to all*.

The Working Group was launched in 2002 following a joint workshop between the W3C and the WAP Forum with contributions from SALT [9] and XHTML+Voice (X+V) [6]. It's major contributions include: *Multimodal Interaction Use Cases*, *Multimodal Interaction Use Requirements*, the *W3C Multimodal Interaction Framework* [14]. Work has also been done on dynamic adaptation to device configurations, user preferences and environmental conditions (*System and Environment Framework*) [15], on integration of composite multimodal input and modality component interfaces such as interfaces for ink and keystrokes which will enable the use of grammars for constrained input, and the context sensitive binding of gestures to semantics (speech and DTMF modalities are developed by the *Voice Browser Working Group* [16]).

Group's work has also stimulated the creation of mark-up languages such as EMMA, and InkML. EMMA (*Extensible MultiModal Annotation Markup Language*) [11], formerly known as *Natural Language Semantics Markup Language*, is a markup language intended to represent semantic interpretations of user input (speech, keystrokes, pen input etc.), together with annotations such as confidence scores, timestamps, input medium etc. The interpretation of the user's input is expected to be generated by signal interpretation processes, such as speech and ink recognition, semantic interpreters, and other types of processors for use by components that act on the user's inputs such as i interaction managers. InkML [12], defines an XML data exchange format for ink entered with an electronic pen or stylus as part of a multimodal system, which will enable the capture and server-side processing of handwriting, gestures, drawings and other specific notations.

Salt and X+V Until W3C standards emerge and mature, other related efforts have been shown, namely SALT and XHTML + Voice. *Speech Application Language Tags* (SALT) [9], is a lightweight set of extensions to existing markup languages, allowing developers to embed speech enhancements in existing HTML, XHTML and XML pages, enabling multimodal and telephony-enabled access to information, applications, and Web services from PCs, telephones, tablet PCs, and PDAs. *XHTML+Voice* [6], by IBM, Motorola and Opera Software, is yet another effort exploiting the combined use of XHTML and parts of VoiceXML through *XML events* to support for visual and speech interaction. In contrast with SALT, X+V provides a standard visual markup language (XHTML) and an event model, has reacher voice interaction and makes development easier by allowing separation of visual and voice programming. Development tools from IBM are already available and so is a multimodal browser from Opera for Sharp's Zaurus PDA.

2.3.2 Bimodal emotion recognition

Authors: C. Kotropoulos and G. Patsis, AUTH

Emotion recognition is one of the latest challenges in intelligent human/machine communication. Most of previous work on emotion recognition focused on extracting emotions from visual or audio information separately. A review of recent approaches for bimodal emotion recognition is attempted in [243].

Chen et al. proposed a rule-based method for singular classification of input audiovisual data into one of the following emotion categories: happiness, sadness, fear, anger, surprise, and dislike [75, 74]. 36 video clips of a Spanish speaker and 36 video clips of a Sinhala speaker have been used to test the proposed method. The speakers were asked to portray each of the six emotions considered 6 times using both vocal and facial expressions. From the speech signals, pitch, intensity, and pitch contours were extracted as acoustic features. Facial features such as lowering and rising of the eyebrows, opening of the eyes, stretching of the mouth, and presence of a frown, furrow, and wrinkles were manually measured from the input images. A set of rules for the classification of the acoustic and facial features into pertinent emotion categories were defined and evaluated on the aforementioned data set. However, a clear picture on the actual performance of this method cannot be obtained. De Silva and Ng also proposed a rule-based method for singular classification of input audiovisual data into one of the six emotion categories examined in the work of Chen et al. [281]. The input data utilized were 142 2-seconds long video clips of two English speakers. Each speaker has been asked to portray 12 emotion outbursts per category by displaying the related prototypic facial expression while speaking a single English word of his choice. The audio and visual material has been processed separately. The optical flow method was used to detect the displacement and its velocity of the following facial features: the mouth corners, the top and bottom of the mouth, and the inner corners of the eyebrows. Pitch and pitch contours have been extracted from the speech signals. A nearest-neighbor method has been used to classify the extracted facial features, and an HMM has been used to classify the estimated acoustic features into one of the emotion categories. Manually derived rules have been used for emotion classification of the input audiovisual material. A correct recognition rate of 72 % for a reduced input data set has been reported.

A hybrid approach to singular classification of input audiovisual data into one of the following “basic” emotion categories: happiness, sadness, anger, surprise, and neutral has been proposed in [326]. 100 video clips of one female Japanese professional announcer have been employed in the tests. The subject was asked to pronounce a 2-syllable Japanese name while portraying each of five emotions 20 times. From the speech signals, pitch, intensity, and pitch contours have been extracted as acoustic features. The acoustic features were classified into one of the emotion categories considered by applying an HMM. Both visible light and infrared cameras have been used to obtain ordinary and thermal face images. The images correspond to the points where the intensity of the speech signal was maximal for the syllables of the word pronounced. The eyebrows and the eye region were extracted from each of the selected images separately. Each image segment has been compared to the relevant “neutral” image segment in order to generate a differential image. A discrete cosine transform has been applied to both visible and infrared images to extract a feature vector, respectively. An ANN has been used further to classify each of these feature vectors into one of the emotion categories. The classification scores obtained by the examination of the visual stimuli were added to those obtained from the speech signal to decide the final output category. A correct recognition rate of 85 % has been reported to a reduced input data set. It is not known whether and with which precision this method could be used for emotion classification of audiovisual data from an unknown subject.

A novel approach that uses both visual and audio from video clips to recognize human emotions is presented in [287]. A tripled HMM is introduced to perform the recognition which allows the state asynchrony of the audio and visual observation sequences while preserving their natural correlation over time. The experimental results demonstrate that such an approach outperforms separate use of visual or audio information.

2.3.3 Audio and Text Interaction

Author: Andreas Rauber, TUWIEN-IFS

Analysis of audio files, particularly music, has seen a significant increase in interest recently. Like video, music is intrinsically multi-modal, having both audio as well as textual aspects, i.e. the actual sound signal, but also song texts, note representations, artist biographies, etc. Furthermore, the audio-aspect itself can been seen both from a signal-processing (i.e. frequency spectra coding) as well as text-processing perspective when symbolic notations are considered, such as the popular MIDI file format.

While initially these different aspects of music were primarily dealt with individually, different disciplines are now collaborating to integrate these different points of view on music to get a more comprehensive representation of the complex domain of structuring and searching music.

The various aspects of music information retrieval, its sub-disciplines and different goals, and work in these directions, are comprehensively reviewed in a recent review of music information retrieval literature [102]. These, as well as the wealth of work addressing music retrieval from the various individual modalities shall not be considered here.

One of the few exceptions to this prevalence of single-modality analysis for audio files is [316]. The authors present an automatic style detection system that uses both the acoustical representation of audio, i.e. frequency data, as well as textual data, referred to as “cultural representation” from community metadata to analyse and organize pieces of music according to musical style and genre. The individual feature sets used for analyzing music have both been used in similar settings before individually. For the audio-based style classification, the audio signal is downsampled to 11.025kHz mono, and transformed to zero mean and unit variance, subsequently extracting the spectral density of the signal. In order to incorporate cultural information about the artists, web pages were indexed using a specific weighting scheme to determine higher term-weights for words that appear closer to the artists names within the webpage as being more descriptive. The two representations for artists, based on their musical performance as well as on the textual descriptions, were subsequently used for style-based classification using a feedforward time-delay neural network. Results show that some musical styles, such as Rhythm&Blues and Rap have a high cross-over in the textual domain, whereas others, such as IDM (Electronic - Intelligent Dance Music) were very difficult to detect properly from the audio signal features alone. By combining both the textual as well as the audio dimension, significant classification performance improvements could be reached.

Chapter 3

Cross-Modal Integration for Multimedia Analysis and Recognition

Nowadays we are witnessing a rapid explosion of multimedia data. They are produced by a variety of sources including video cameras, TV and other digital entertainment, digital audio-visual libraries, and the multimodal web. This rapid explosion of multimedia data creates an increasing difficulty of finding relevant information, e.g. in the web, which has spurred enormous efforts to develop tools for automatic semantic analysis of multimedia content. Most of these efforts, however, concentrate on using the available textual information and ignore other types of information. The multimedia explosion also poses several ambitious technical challenges; two of which are: (i) Natural access and interaction with multimedia databases, and (ii) Analyzing and Recognizing objects/events and human behavior in surveillance or sports indexing by processing combined video-audio-text data. The integration-fusion of multiple modalities is explored toward the goals of analyzing multimedia content and recognizing entities, given the information provided by several cues including visual, audio, speech, text, and tactile information.

3.1 Multimodal Video Analysis

Authors: A.D. Bagdanov, A. Smeulders, C. Snoek and M. Worring, UvA

More and more video information repositories are becoming available every day. Indexes are essential for effective browsing, searching, and manipulation of collections of video sequences. Such indexes are central to applications such as digital libraries storing multimedia information. To support effective use of video information, and to cater to ever-changing user requirements, these indexes must be as rich and complete as possible.

Until recently, video indexing was mostly carried out by human experts who manually assigned textual descriptions to video content. The specialized knowledge required to perform indexing makes this approach laborious, costly, and error-prone. To alleviate these limitations, automatic classification of video content and assignment of semantic labels to video elements is needed. This process of automatically assigning content-based labels to video content is referred to as video indexing [136].

An aspect of video indexing that sets it apart from other types of indexing tasks is the presence of multiple information channels, or *modalities*. For video sequences, we can identify three primary modalities:

- *Visual modality:* contains everything, either natural or artificially created, that can be seen the the video sequence.
- *Auditory modality:* contains the speech, music, and environmental sounds that can be heard in the video.
- *Textual modality:* contains textual resources that describe the content of the video.

Most current solutions to video indexing address only a single modality of video sequences. Good books [118, 139] and review papers [48, 54] exist describing unimodal approaches to video indexing.

Human indexers of video content are very adept at integrating all three of these modalities into meaningful semantic interpretation of video concepts. Consider the assignment of the semantic description:

Heated argument between Donald Rumsfeld and Kofi Anan

to a scene from a news report. This determination requires visual interpretation of the scene to decide that there are two people in shot, auditory interpretation to asses the “heat” of the scene, and potentially textual interpretation of visible captions. The main shortcoming of unimodal approaches is their inability to integrate information from these three modalities. Effective indexing requires a multi-modal approach in which either the most appropriate modality is selected or the different modalities are used collaboratively.

One review of multi-modal indexing is given in [314]. A recent review of multi-modal video indexing is given in [284]. This review builds on the work of [48, 54, 314], and combined they form a complete overview of the field of multi-modal video indexing. We take our general discussion of multi-modal video analysis from [284].

3.1.1 Structural Segmentation

The authors of [284] treat video analysis and indexing as the inverse of an authoring process, arguing that video is the product of the authoring actions of writers, directors, editors, etc. They even coin the phrase "video document" to emphasize the similarity of the process to document analysis. The first stage in this inversion process is the segmentation of a video document into structural layout components and their associated content. First, patterns of interest must be distinguished that can be used to support decisions about layout and content categories. According to [158] the four best approaches to this are:

- *Template matching:* patterns to be recognized are compared to a learned template.

- *Statistical classification:* patterns to be recognized are classified based on the learned distribution of patterns in a feature space.

- *Structural pattern matching:* patterns to be recognized are compared to a small set of learned primitives and grammatical rules for combining them.

- *Neural networks:* patterns to be recognized are given as inputs to a network which has a learned nonlinear input-output relationships.

Examples of all four of these approaches abound in the literature on video indexing, but statistical approaches are the most frequently encountered. Four specific techniques that appear often are:

- *Bayesian classification:* assigns a pattern to a class with maximum posterior probability [158].

- *Decision trees:* assigns patterns to a class based on a hierarchical decomposition of the feature space [158].

- *k-Nearest Neighbor:* assigns a pattern to a class according the majority class assignment of the *k* nearest training samples [158].

- *Hidden Markov Models:* assigns a pattern to a class based on a sequential model of state and transition probabilities [205, 261].

It should be noted that each modality independently gives rise to features and classification techniques that can be used in subsequent stages for segmentation, analysis, and indexing.

After the initial identification and classification of interesting patterns, the layout of the video document is reconstructed. Since layout, or the structural/temporal composition of the video sequence guides the human viewer in his experience of the video, so should it guide analysis.

Several techniques operating on the visual modality exist to segment video documents into distinct camera shots, known as *shot boundary detection*. An extensive review of cut detection algorithms is given in [54]. Transition edits are an important cue for shot boundaries, and since such transitions are gradual, comparison of successive frames is insufficient. The first approach to exploit this observation is described in [329].

In the auditory modality, detection of abrupt cuts can be accomplished by identifying silences and transition points. In [245] it is shown that average energy, E_n, is a sufficient measure for detecting silence segments. Average energy is computed for a window of size n, and if the average for all windows in a segment falls below a given threshold, a silence is recorded. Another approach based on average energy is given in [331], where E_n is combined with the zero-crossing rate (ZCR), where a zero-crossing is said to occur if successive samples have different signs. A segment is classified as silence if E_n is below a threshold or if most ZCRs are below a threshold. Li et al [191] use silence detection to separate audio segments into silence and signal segments. Besides silence detection, the approach of [191] also detects silence transition points in the signal segments using break detection and merging.

Since structural composition is very modality dependent, a multi-modal approach is not very effective. Multi-modal integration, however, can be a successful approach to improving structural segmentation results.

3.1.2 Content Segmentation

As low-level feature detection in video data streams becomes more reliable, the natural trend is toward the segmentation of high-level semantic concepts in videos. In this section we describe some approaches to identification of semantic concepts in images and video. Classification of such items is a general problem, and there is much overlap here with the fields of machine learning, image processing, and content based image retrieval.

People detection

Many approaches to detecting people in videos reduce the problem to that of detecting faces in single video frames [323]. This type of detection is a non-trivial problem due to variations in location, scale, lighting conditions and orientation. Face variations caused by differences in facial expression, facial hair, the presence of glasses, and occlusion complicate the problem additionally. The best face recognition technique to date is the proposed by Rowly [266]. It is a neural network-based approach, and is able to correctly detect approximately 90% of all upright and frontal faces [249].

More advanced techniques for locating people detect not only the face or the head, but the whole human body [223]. The algorithm applies independent detectors for head, legs and arms. After geometric configuration of the detected parts is validated a second level classifier combines the part-detector results to classify a candidate as a person or non-person.

The auditory modality can also provide indicators of the existence of people in-shot. Classification of different signal segments as speech provide the most useful clues. In [331] five features are used to identify speech signals. The features are computed through analysis of the average energy E_n and the zero-crossing rate (ZCR). A more elaborate approach is proposed in [191]. The technique identifies not only speech signals, but speech together with noise and music with an accuracy of about 99%. Once a segment has been classified as containing speech, speaker identification can be applied to identify speakers, thus providing a much richer semantic description of the scene. A generic speaker identification technique is proposed in [246].

In the textual modality, the strongest indicator of the existence of people in a scene is the presence of words in captions or transcripts that are identifiable as proper names. Natural language processing techniques are used in [271] to locate names in transcripts of news video segments. The location of names is actually a special case of the named entity recognition problem in computational linguistics [45]. This approach treats name recognition as a binary classification problem, where every word is classified as being (part of) a name or not. The authors use a variant of an HMM based on a bigram language model.

People detection is an excellent example of the necessity for multi-model integration. Identifying people in the visual modality is unreliable due to scene-accidental variance in orientation, pose, and occlusion. While speech detection and speaker identification are sensitive to environmental noise, and more research is needed to improve detection of names in the textual modality, the combination of these three types of evidence can be used to improve overall performance of person detection algorithms.

Object detection

Object detection is a generalization of the people detection problem. Specific objects can be detected by applying specialized visual, motion, sound, and appearance detectors. An example of detecting specific objects based on visual appearance is given in [273]. The authors describe a technique that detects the presence of passenger cars in video frames. They use a product of histograms, where each histogram represents the joint statistics of a subset of wavelet coefficients and their position in the object.

In the absence of prior knowledge about the classes of objects we want to search for, motion-based techniques can be applied. Since the appearance of objects can be highly variable and sensitive to sensing conditions, motion is a much more reliable feature. In [234] the authors describe a system for segmenting a video frame into independently moving objects. The method takes a bottom-up approach, beginning with a color-based decomposition of the frame. Regions are merged bases on their motion parameters.

In [113], Fergus, Perona, and Zisserman present a method to learn and recognize object class models from unlabeled and unsegmented cluttered scenes in a scale invariant manner. Their paper won the best paper award at CVPR03, and rightfully so. Objects are modeled as flexible constellations of parts. A probabilistic representation is used for all aspects of the object: shape, appearance, occlusion and relative scale. An entropy-based feature detector is used to select regions and their scale within the image. In learning the parameters of the scale-invariant object model are estimated. This is done using expectation-maximization in a maximum-likelihood setting. In recognition, this model is used in a Bayesian manner to classify images. The flexible nature of the model is demonstrated by excellent results over a range of datasets including geometrically constrained classes (e.g. faces, cars) and flexible objects (such as animals).

Specific objects can also be detected by analyzing the segmented audio signal from a video. Environmental sounds are analyzed for the presence of specific object sound patters. In [318, 331] specific sounds are detected including dog barks, ringing telephones, and musical instruments. In the presence of such sound patterns it is safe to assume the presence of a specific object.

Setting detection

Motion is not as relevant to the detection of setting in video, as settings tend to be static. The challenge in improving setting detection in video streams is the incorporation of visual, auditory, and textual information into the analysis.

In [296] frames are classified as either indoor or outdoor. Three types of visual features are used: color, texture, and frequency. Outdoor images are further classified into city and landscape images in [307]. The features used are color histograms, color coherence vectors, DCT coefficients, edge direction histograms, and edge coherence vectors. Through analysis of subblocks, the same authors report a technique for detecting the presence of sky and vegetation in images of outdoor scenes [306].

Setting detection using the auditory modality can is achieved by detecting specific environmental sound patterns. In [318] the authors summarize an audio segment using a small set of parameters such as loudness, pitch, brightness, bandwidth, and harmonicity. Statistical techniques are used to build classifiers and retrieval algorithms over a variety of specific sound patters including laughter, crowds, and water. In [331] classes of natural and synthetic sound patterns are distinguished using an HMM. The authors are able to classify scenes as containing applause, explosions, rain, flowing rivers, thunder, and windstorms.

While the visual and auditory modalities are well suited for setting recognition in videos, the textual modality from a transcript of the video can also be incorporated to support more precise information about setting [82]. Fusion of these three modalities is needed in the future to support detection of richer semantic concepts.

3.1.3 Multi-modal Analysis

After low-level reconstruction (shot-boundary detection) and medium-level reconstruction (semantic concept detection) of the layout of video segments, the next stage is the analysis of these primitive physical and logical components with the goal of extracting a rich semantic index of the video content. The rest of this chapter discusses state-of-the-art approaches that build on the type of low- and medium-level concepts discussed in this section to achieve high-level descriptions of multimedia content.

3.2 Video Analysis and Integration of Asychronous Time-evolving Modalities

Authors: P. Gros and E. Kijak, INRIA / TEXMEX

It is absolutely reasonable to think that integrating several sources of information should improve the results of content-based video analysis, as shown by several authors [314, 285]. Several domains of analysis were particularly studied: speech recognition where the visual data are lips' contours [35, 231] or facial animation parameters [33], temporal segmentation [149, 50, 161], logical units identification in news reports [157, 174, 259, 148], dialog detection [270, 25, 30], creation of video abstracts [195, 293], style classification [150, 151], and semantic concept detection like ”rocket launch” [26].

However, the integration of features issued from various media is not a trivial task. Two of the problems encountered by this process are the following:

a decision problem which is common to all systems based on information fusion: what should be the final decision when the various media or sources of information provide contradictory data?

a synchronization problem which is specific to multi-modal integration. As a matter of fact, the sampling frequency of low-level attributes depends on the media:

- the elementary unit of video signal is the image; when the sampling frequency is 25 Hz, it is possible to obtain low level features every 40 ms ;
- the elementary unit of audio signal is the frame ; when the sampling frequency is 100 Hz, it is possible to obtain low level features every 10 ms. For statistical studies, the unit used by most authors is the clip, a set of consecutive frames whose length ranges between 1 and 3 s. In this interval, the signal is assumed to be quasi-stationary.

Of course, the boundaries of a visual segmentation into shots and those of an audio segmentation in homogeneous segments do not coincide.

To solve these two problems, several ways of integration of audio and visual features can be used. A first way consists in a successive use of audio and video analysis. Another way combines the audio and visual features in a single audiovisual feature vector before any classification or decision (early integration). Finally the third way consists of two independent classifications with respect to each of the modalities and fuse their results (late integration).

Successive analysis The principle of this analysis is the following: The audio or textual signal is sued in a first stage to detect interesting segments. Image analysis (tracking, spatial segmentation, edge / line / face detection) is then used in the regions previously detected to identify a particular event [175, 69], or more simply to identify the video segment boundaries [152]. In this first case, audio, or text, are used to restrict the temporal window where video analysis will be used. An implicit assumption of such a method is that interesting segment detection is faster with these modalities (applauds in the sound track or keywords in the textual stream). This constitutes the first stage of a prediction verification method, whose second stage is a verification and localization step done on the audio or on the visual stream [69, 38].

The use order of the various media may be inverted as proposed in [190, 187]: in a first stage, visual features are used to detect interesting events. In a second stage, the state of excitement of the speaker or the public is measured to filter the most interesting shots. This process is no more a prediction / verification process, but the audio signal is used in order to order the visual segments by level of importance.

Early integration It consists on integrating the audio and video features into a single vector before the classification stage. This process assumes that data extracted from all media are synchronized. This synchronization can be done at several levels and with a priority given to one media or the other:

- at the audio fame level [225]: the image features are then over-sampled by interpolation or simple repetition;
- at the audio clip level [322, 199]: the video attributes are averaged on this duration;
- at the image level [138];
- at the video shot level [137, 173].

Late integration In this approach, each modality is classified independently. Integration is done at the decision level and is usually based on heuristic rules. For example, audio and video streams are segmented and classified by two separate Hidden Markov Models. Dialogs are identified as segments where audio signal is mainly speech while visual information is an alternation of two views. The detection of such particular scenes is done by fusion of the decisions [25].

Multimodality and HMMs Early integration is simple but very costly in computation time. The computation complexity increases with the dimension of the vector space, as does the number of data required for the learning stage. Let V be the dimension of the visual feature vector, and A be that of the audio feature vector. From the HMM point of view, the concatenation of both audio and visual vectors induces the use of a single stream HMM, with probability densities estimated in a space of dimension A+V. This approach usually assumes that the features are synchronized.

A way to remedy these problems is to use late integration. This method does not take into account the dependencies between the features of the various modalities. When the audio and video features are synchronized, multi-stream HMM can be used, where audio and video can bring separate contribution to the observation probability. When the features are not synchronized, product HMM are uses: these HMM are a generalization of multi-stream HMM whose topology authorizes asynchronism between the states. As a matter of fact, they are built by the combination of two independent HMM, one for video analysis, the other for audio analysis. The likelihood of each sequence of observation is a combination of the likelihood obtained for both modalities.

3.2.1 Combined Audio/Speech and Image Processing in Videos.

Authors: C. Kotropoulos, C. Cotsaces, M. Kyperountas, G. Patsis and N. Nikolaidis, AUTH

Introduction

The advances in digital video technology and the ever increasing availability of computing resources have resulted in the last few years in an explosion of digital video data, especially on the Internet. However, the increasing availability of digital video was not accompanied by an increase in accessibility. This is because the nature of video data is unsuitable for traditional forms of access, indexing, search and retrieval. Additionally, the amount of data contained in video is such that manually summarizing or annotating it is at best a laborious and economically undesirable process. Therefore, techniques have been sought that organize video data into more compact forms or extract information from it [100]. This is useful because it can serve as a first step for a number of different data access tasks like browsing, retrieval and fingerprinting. In the following we present a review of some of the recent developments of multimedia analysis techniques that involve integration of information from audio/speech and video modalities.

Temporal Image Sequence Segmentation and Analysis

The field of temporal image sequence (video) segmentation is not a new one, as it dates to the first days of motion pictures, well before the introduction of computers. Motion picture specialists perceptually segmented, and still segment, their works into a hierarchy of partitions. A video (or film) is completely and disjointly segmented into a sequence of scenes, which are subsequently segmented into a sequence of shots. Scenes (also called story units) are a concept that is much older than motion pictures, ultimately originating in the theater. Traditionally, a scene is a continuous sequence that is temporally and spatially cohesive *in the real* *world* — but not necessarily in the projection of the real world on film. Shots, on the other hand, originate with the invention of motion cameras, and are defined as the longest continuous sequence that originates in a single take of the camera — a take being what the camera images in an uninterrupted run. Shot detection is the most basic temporal video segmentation task, as it is intrinsically and inextricably linked to the way that video is produced. As such it is the natural choice for segmenting a video into more manageable parts and is thus very often the first step in algorithms that accomplish other tasks.

In general, the automatic segmentation of a video into scenes is a very difficult task. This is due to two reasons. First, it is a very subjective task that depends on human cultural conditioning, professional training and intuition. Second, since it focuses on real-world actions and temporal and spatial configurations of objects and people, it requires the ability to extract physical meaning from images, a task well known to be extremely difficult for computers.

On the other hand, the segmentation of a video into shots is both exactly defined and also characterized by distinctive features of the video stream itself. It should be noted however that (by its very nature) the audio track of a video is continuous across a scene's shots. As such it is mostly useless when trying to detect shots, but somewhat useful when trying to group them into scenes. As a consequence, shot detection algorithms are based solely on video information and do not try to incorporate audio information. Thus in the following we will review only scene boundary detection methods and in particular those that combine audio and visual information to achieve their goal. It should be noted however that the majority of scene boundary detection algorithms rely solely on video information and thus are not reviewed here.

In [334], the authors present a scene change detection method based on audio and visual features, which analyzes both auditory and visual sources and accounts for their inter-relation and synergy to semantically identify video scenes. First, the video input is split into audio stream and video stream. Then, the audio stream is classified into four classes: speech, music, environmental sound and silence. Speech data are further decomposed into different elements according to different speakers. Meanwhile, visual analysis partitions the video stream into shots. By combining visual and audio features, by considering certain temporal expectations, the scene extraction accuracy is enhanced, and more semantic segmentations are achieved. Specifically, when a shot break and an audio break are detected simultaneously, the boundary of the sequence of shots is marked as a potential scene boundary. When the potential scene boundary is consistent to an audio break that relates to an audio class change (e.g. music to speech or speech to environment noise), a scene change is set. When the potential scene boundary relates to a speaker change, whether the scene boundary is accepted depends on the correlation of a sequence of shots near the scene boundary.

The algorithm proposed in [111] consists of two phases. Initially, the video shot boundaries are detected using an unsupervised segmentation algorithm along with an object tracking technique. The second phase analyzes extracted audio features based on the results of the video shot detection. Nine different audio features are analyzed: volume, energy, sub-band energy, low shot-time energy ratio, zero crossing rate, frequency centroid, frequency bandwidth, spectral flux and cepstral flux. Unlike most audio feature-based segmentation algorithms, the content of the audio is not used; a scene change is indicated on the basis of the differences of the audio features from the corresponding adjacent shots. Shots that lie close to one another, in the temporal domain, are merged. The authors show that the proposed method is better than those that separately segment the audio and video data into scenes and then integrate the detection results.

In [269], audio was distinguished into four pre-selected classes (silence, speech, music and noise), and this information was later combined with the probability value for a visual cut detection that segmented the video into shot segments. For video abrupt cuts, the energy difference between two successive frames was evaluated, while for fades and dissolves the same difference but at a larger distance, in terms of video frames, was used. To find scene changes, information from both the video and audio classifiers was used in order to determine if a correlation between adjacent shots exists. Specifically, music characteristics, information regarding the identity of a speaker and silence detection is used to indicate a possible scene change. A scene change is set if the speaker or music characteristics change, or if silence is detected, near where a shot change has been detected.

The work introduced in [292] uses a finite-memory model to independently segment the audio and video data into scenes; then two ambiguity windows are used to merge the audio and video scenes. The audio segmentation algorithm determines the correlations amongst the envelopes of audio features. The video segmentation algorithm determines the correlations amongst shot key-frames. The scene boundaries in both cases are determined using local correlation minima. Then, the authors fuse the resulting segments using a nearest neighbor algorithm that is further refined using a time-alignment distribution derived from the ground truth.

In [149], scene change detection is carried out by using audio information along with image and motion information to accomplish segmentation at different levels. To detect audio breaks, a dissimilarity index function based on a set of 12 audio features that are computed over one second long clips is applied. To detect motion breaks, the phase correlation functions computed between every two frames are used. For color changes, the color histograms of each pair of adjacent frames are compared. For shot segmentation, the results from detecting both color and motion breaks are compared. Finally, to detect scene breaks, the authors seek frames for which both visual and audio breaks are detected.

The authors in [325] propose a method for scene boundary detection by exploiting both audio and video features, which takes editing pattern of movies and audio features into account. Background noise extraction and its classification is cooperatively evaluated with visual feature extraction in order to characterize a sequence of shots as a single scene. Candidate scene boundaries are extracted from video data based on the detection of visual effects that are frequently used in scene changes such as dissolve or fade in/out. In addition, they are also extracted from audio features by detecting ’audio cuts’ based on background noise/music classification. Then, the audio data, that are located close (in a temporal sense) to where a shot change occurs, are analyzed using the average power of sub-bands. In case that a large change is found between the starting and ending audio frame of each shot a scene change is set.

In [162], speech and non-speech segments were detected with the non-speech segments being further classified to music and environmental sound. This classification was based on audio periodicity and other audio features. Audio breaks were detected using one-second long audio frames. Then the position of these breaks was compared with the video shot boundaries. All the shot boundaries within a one-second interval from an audio break were set as scene candidates. Sequentially, a color correlation algorithm was used for shot clustering.

In [182] audio frames are projected to an eigenspace that aims to discover the changes in the audio track that are caused by the variations of background audio. The distance of the audio frames from a reference audio frame that is the mean of a set of audio frames that correspond to scene changes is found in the selected subspace. The scene change indications from the audio track are identified by processing this distance vector. Video information is used to align audio scene change indications with neighboring shot changes in the visual data by considering certain timing restrictions, and accordingly to reduce the false alarm rate. Moreover, video fade effects are identified and used independently in order to track scene changes.

Apart from the scene boundary detection algorithms presented above a number of other audio-visual algorithms for analyzing video sequences and extracting a variety of semantic information have been reported in the literature. Some of these algorithms are reviewed below.

Low level processing of audio and video for extracting semantics is proposed in [25]. Simple but effective methods to segment the audio stream are developed. Audio segmentation in homogeneous segments of speech and music is obtained using two different approaches: the first is based mainly on zero-crossing (ZC) rate and Bayesian classification, while the second is based on a multilayer perceptron. The average audio loudness is related to shots for example in sport events. Joint audio-visual analysis is applied. Soccer video indexing is achieved by using MPEG motion vectors and scene classification is obtained by employing Hidden

Markov Models (HMMs).

Multi-modal dialogue scene analysis can be categorized into methods of low, medium, and high complexity. Low-level complexity methods rely on silence detection in the audio track, while medium-level complexity resort to a classification of audio segments into silence, music, and speech [270]. The methods of high complexity require speaker identification. Of course the interaction with the visual track facilitates and improves the decisions taken by considering only the audio signal. Such a system is described in [29].

A content-based video parsing and indexing method is proposed in [300]. The method analyzes both auditory and visual sources of information and accounts for their inter-relations and synergy to extract high-level semantic information. Both frame-based and object-based access to the visual information is employed. Semantically meaningful video scenes are extracted and semantic labels are assigned to them. Due to the temporal nature of video, time has been accounted for in order to create time-constrained video representations and indices. The method has been used to detect the presence or absence of speakers or persons. To this end end-point detection and voiced-unvoiced discrimination has been used to detect voiced audio frames that are subsequently analyzed in order to compute the Mel Frequency Cepstrum Coefficients MFCCs. MFCCs drive a speaker classification based on Vector Quantization that concludes the analysis of audio channel. Shot boundary detection has been applied to the video signal. Shots where persons are present are further found by employing face detection and the likelihood of a person being present in a face shot is computed. A similar strategy is applied in [31] where the problem of reducing the amount of data to be analyzed when indexing audiovisual sequences is addressed. The reduction is based in selecting those parts of the video sequence where it is likely that there is a face talking. This can be useful since usually this kind of scenes contain important and reusable information such as interviews. The proposed technique is based on our a priori knowledge of the editing techniques used in news sequences. The results show that with this algorithm it is possible to discard around 76% of the news sequence with minimal processing. In another closely related work first isolated speech segments from the background are identified by applying video shot detection, audio classification, and adaptive silence detection [194]. Then a decision is made based on the calculated likelihood between the incoming speech data and pre-trained speaker/background models. Experimental results indicate that the proposed algorithm can achieve approximately 84% identification accuracy by integrating multiple media cues.

Condensed Representations of Video

An important functionality when retrieving information in general is the availability of a condensed representation of a larger unit of information. This can be the summary of a book, the theme or ground of a musical piece, the thumbnail of an image, or the abstract of this paper. This can allow us to assess the relevance or value of the information before committing time, effort and computational and communication resources to process the whole information unit. It can also allow us to extract high level information when we are not interested in the whole unit, especially during manual organization, classification and annotation tasks. While for most types of information there exists one or more standard forms of condensed representation, this is not the case for video. The results of video representation can be in the form of a piece of text, a symbolic description, a collection of images or a shorter video consisting of pieces of the original video.

In the literature the term key-frame extraction or storyboarding is used for video representation resulting in images, while skimming is sometimes used for representation resulting in a shorter video. We will use the term *representation* to describe in general the process of summarization, key-frame extraction, storyboarding and skimming.

One might object by noting that the image sequences produced by key-frame extraction, if treated as a video, the same thing as the video sequences produced by skimming. However, since video (and its accompanying audio) is only meaningful at frame rates approaching the original, the difference between skimming and key-frame extraction are clear: the former is comprised of continuous video sequences and thus can be played back, while the latter cannot.

Representation methods can be classified into manual, semi-automatic and automatic depending on the amount of user interaction necessary. Below we review some of the most interesting audio-visual algorithms for the automatic extraction of condensed video representations. It should be noted that most of the work in the literature focuses on extracting the representation solely from the visual component of video.

The work of Hanjalic at Delft University of Technology [138] estimates the "excitement" caused from each frame in the video by combining motion activity, frequency of shot changes and sound energy. It uses combination of nearby maxima of the above quantities, which are weighted to favor adjacent maxima in all three time curves, to obtain a "highlight time curve". This can be thresholded to obtain key-frames. The authors' approach works well in videos with relatively homogeneous content, such as the soccer videos which they used for testing. However in the case of videos with large variation in content, such as motion pictures the use of a static threshold would mean that the low excitement sections would not be represented in the condensed representation, even if they are very large.

Xiong et al. at Mitsubishi Electric Research Laboratories [322] propose a dual audio-video summarization method. It classifies audio information from video into a number of classes (speech, music, applause etc) by using Gaussian Mixture Models. This is then combined with the most basic (MPEG7) motion descriptor, which quantizes the motion in a frame into 5 levels. The key-frames are defined as the maxima of the relative entropy metric, which is defined as the dissimilarity of the feature distribution within a small temporal window compared with the feature distribution in a much larger one. The fact that the audio part of the algorithm depends on specific predefined types of sound, and the simplicity of the visual part mean that its application is limited to specific classes of video content.

Gong and Liu at NEC Laboratories [124] present a shot-based skimming extraction algorithm. First the video is down-sampled temporally as well as spatially. Each resulting frame is characterized by its color histogram and SVD is performed on the resulting feature matrix **A** to reduce its dimensionality. For SVD-reduced features $\psi_i = [\upsilon_{i1}\ \upsilon_{i2}\ \ldots\ \upsilon_{in}]^T$ the authors define a new metric

$$||\psi_i|| = \sqrt{\sum_{j=1}^{rank(A)} \upsilon_{ij}^2}$$

which is inversely proportional to the cardinality of the feature within the matrix, and thus to how typical the frame is within the video. Similarly, for video segments S_i the metric $\mathrm{CON}(S_i) = \sum_{\psi_i \in S_i} ||\psi_i||$ is defined as a measure of visual content. It should be noted that the author's expectations for these metrics are optimistic, since $||\psi||$ takes into consideration only equality between reduced feature vectors and not similarity, and $\mathrm{CON}(S_i)$ includes information about the uniqueness of its constituent frames with respect to the whole video and not with respect to the segment itself. The next step is clustering the reduced feature vectors. The most common frame is used to seed the first cluster, which is grown until a heuristic limit is reached. The other clusters are grown in such a way that they have $\mathrm{CON}(S_i)$ similar to that of the first cluster. Then the longest shot is extracted from each cluster, and an appropriate number of these shots is selected to form the summary, according to user preferences. Alternatively, as described in [123], the clustering can be performed by computing a Minimum Spanning Tree from all shots in the video, based on their $\mathrm{CON}(S_i)$, and then pruning its longest branches. In addition, a speech transcript is extracted from the audio track by a speech recognition module. This transcript is then analyzed by latent semantic analysis in order to find the most salient sentences and construct a condensed audio track. The two representations are then aligned by domain-specific heuristics. Specifically, when a face is detected in the video skimming the alignment is made directly, but otherwise it is performed by a graph-matching algorithm and a set of heuristics. The problem with the above audio approach is that it is designed to work only with a specific type of video (news) and that it ignores non-verbal sounds.

A thorough attention-based method for the computation of various types of condensed representations of video is presented by Ma et al. at Microsoft Research [204]. It is based on the computation and processing of attention curves, which model the degree to which a feature would attract a potential user's attention. Three visual attention curves are extracted from the magnitude, spatial and temporal variance of MPEG macro-block vectors. Another three are extracted from the spatial color, intensity and orientation contrasts of each frame. Finally a face-based attention curve is constructed using a face detector on each frame. These curves are linearly combined using heuristic weights and are additionally weighted by another attention curve which is extracted from estimated camera motion. A similar attention scheme is used for the construction of an audio attention curve by first classifying audio segments into classes (music, speech etc) and then summing the class ratios. This curve is then weighted by an audio saliency attention curve which is derived from the sum of audio energy and and audio energy peaks. The result is linearly combined with the visual attention curve. Extraction of the key-frames is performed by detecting maxima in the overall attention curve. This approach can be modified for shot-based key-frame extraction by choosing as key-frame the frame with the maximal attention value within the shot. A way to produce a skimming of the video is also shown, by choosing segments around key-frames based on the attention curve and a number of heuristics. The work is thorough in its selection of video features, theoretically well founded, and flexible in its resulting representation. It is only compromised by the weights used for the fusion of the different attention curves, which are both constant and heuristic.

Because the objective of the above methods is to extract a semantically informative representation of the video, its evaluation is a very difficult affair. In practice, it is only feasible by actual human observation. This is in fact what most authors use for experimental evaluation, depending on panels of experts that quantify whether a specific video representation contains all salient information in the original video. Of course there are other measures of the characteristics of each algorithm, such as the size of the resulting representation but these are characteristics of the algorithms rather than qualitative result measures.

3.2.2 Multimodal Integration: Application in sport video analysis

Use of audio and visual features for sports video abstraction

Author: P. Bouthemy and F. Coldefy, INRIA / VISTA

Video abstraction is motivated by the growing need of fast or selective visualization of TV broadcasts or videos. It may be an efficient tool to browse a video by picking out the main excerpts of the program or by removing all the irrelevant sequences according to the program theme. We are focusing here in sports video summarization or highlights extraction from sports video.

A first category of approaches exclusively deals with image analysis. Tekalp *et al.* in [108] present an automatic and deterministic goal detection method based on dominant color extraction and shot classification. The succession of particular shot types such as close-ups and replays is decisive to the goal detection. In [332], Zhong et al. have developed visual modules including color clustering, object segmentation and line detection to classify tennis and baseball shots. Li *et al* [189] propose an hidden Markov model (HMM) using color, motion and shape features to discriminate between play and break sequences in baseball, soccer and sumo videos. Xie *et al* [321] have designed a similar but more complete model for soccer videos.

A second category introduces a combination of audio and visual features to perform highlights extraction. Leonardi et al. [187] present an audio and visual model also exploiting a HMM to classify each pair of successive shots in soccer videos. Hanjalic [138] has proposed a deterministic excitement criterion based on the mean dominant motion magnitude per shot, the density of cuts, and the audio loudness to detect goals in soccer video. However, the audio short time energy which is the unique audio descriptor used in [187][138] is too rudimentary to deal with complex soundtracks in which the stadium noise is very loud. In [232], Nepal et al. use crowd cheers, scoreboard display and camera motion analysis to detect goals in basketball videos. Crowd cheer might not be always a reliable cue to detect highlights in sports videos because it may occurs independently of the interest of the actions. Furthermore, the features involved in the cheer detection [232] may not be robust enough in a noisy configuration.

Rui *et al.* [267] use more elaborated audio features such as pitch statistics and Mel-cepstral coefficients (MCC) to extract emphasized segments of speech in baseball TV programs. They use Support Vector Machines (SVM) to perform the highlights detection. Petkovic *et al.* [248] include pause rate in the audio features combined with word spotting and video analysis (text, motion and color analysis) to detect highlights in Formula 1 TV programs. Their method is based on a Dynamic Bayesian Network (DBN). Finally, Xie *et al.* [320] present a Hierarchical Hidden Markov Model to automatically discriminate between a pool of features and to classify play and break segments in soccer videos.

As Rui *et al* [267] and Petkovic *et al.* [248], we believe that the soundtrack content is crucial to detect highlights in most sports. More precisely, excited speech is most often simultaneous to an interesting action in the game. Combined with appropriate visual features, excited speech detection can be considered as a reliable gauge to detect highlights in sports video. Moreover, it allows for the design of an efficient, simple and robust event detection scheme.

Cowie *et al* [90] characterize excitement in speech by a high energy, a relative high pitch value and a wide pitch range. Simple statistics over a temporal window such as maximum, mean value and range of the pitch and of the energy subband (0-4400Hz or 680-4400Hz) are commonly used in emotional speech classification.

The work cited above does not involve unsupervised dedicated detectors of excited speech based on pitch and energy measures. A learning scheme (Gaussian Mixture models, HMM or SVM) is usually applied to estimate the model parameters or the classification function [267],[248]. However, model learning may be a long process since it requires precise and manual indexing of numerous videos so that enough examples are available for the training stage. Conversely, in [86] Coldefy *et al.* have recently designed an unsupervised method for soccer video summarization which especially involves an efficient detector of excited speech segments combined with visual features (related to dominant colour and camera motion). Satisfactory results on seven soccer videos (with different speakers and corresponding to almost 20 hours of TV programs) are reported.

Highlight detection in sport events

Authors: C. Kotropoulos, C. Cotsaces, M. Kyperountas, G. Patsis and N. Nikolaidis, AUTH

Algorithms and applications for detecting highlights in sport programs using multimedia cues are described in [93]. Several methods that employ keywords, audio features, or visual processing for content-based extraction of important video segments have been proposed in the literature. Another challenge is to determine the starting and end point of such segments. To detect highlights in sport programs based on audio, the energy level is measured and compared to a threshold. Usually highlights are correlated with increased crowd activity. However, clustering techniques should be applied to group of high-energy frames in order to create meaningful highlights.

The analysis of football audiovisual sequences is studied in [186]. Audio data is divided into short sequences of duration 1-1.5 sec which are classified into several classes (speaker, crowd, and referee whistle). Every sequence is further analyzed depending on the class it belongs to. Two-class and three-class segmentation algorithms are tested. Crowd can easily discriminated from speaker's voice by computing the Euclidean distance between a theoretical sinusoidal curve and the 2-D projection from cepstrograms of audio segments and comparing it with a threshold. HMMs and C-means with multidimensional HMMs (M-HMMs) are also tested. In the latter case, the C-means algorithm is used to classify every audio segment into several classes so that the variation within the class is less important. A set of 11 features is measured, namely non-silence ratio, volume standard deviation, standard deviation of ZC rate, volume dynamic range, standard deviation of pitch period, smooth pitch ratio, non-pitch ratio, frequency centroid, frequency bandwidth, 4Hz modulation energy, and energy ratio of subband 1-3. The classification is based on five features: non-silence ratio, smooth pitch ratio, non-pitch ratio, volume dynamic ratio, and 4Hz modulation energy. Within every class and for each of the 3 possible states (whistle, crowd, speaker) a M-HMM is built. The observation data for the M-HMM is composed of the remaining six features and a seventh one that represents the cepstrum. High recall and precision rates are reported.

Audiovisual Event Detection in Sports

Authors: R. Dahyot and A. Kokaram, TCD

Joint Audio-Visual Event Detection in Sports [94] By combining a visual analysis tool for detecting court views, with an audio tool for detecting short duration, loud noises, we have proposed in Trinity College a method to parse a Tennis game at a semantic level [94]. The visual cues correspond to the second moment of the Hough transform of the edges computed for each image of the sequence. This visual feature allows to detect relevant shots corresponding to large views of the court that exhibit straight edges. A rally in the tennis game is then defined as the shot of the video showing large view of the court and a characteristic audio activity measured by the racket hit sounds. A racket hit detector has been designed using a statistical learning technique (Principal Component Analysis) in the spectral domain of the audio data of the video. The analysis of the visual (for large court views) and the audio data (for racket hits) are performed independently. The audio-visual information is then combined to detect the rallies in the tennis game.

Inlier Modeling for Mixed Media analysis [95]. This work, to be published in late 2004, is the result of a collaboration between the University of Cambridge and Trinity College Dublin. A new process for modelling inliers in audio and video streams, is proposed in [95]. It is applied to object segmentation in images and silence detection in audio data.

3.3 Combining Text and Vision for Semantic Labeling of Image Data

3.3.1 Structural and Textual Information for Semantic Interpretation of Image Data

Author: P. Duygulu, Bilkent Univeristy

Huge amounts of digital multimedia content are currently available in large archives due to the recent developments in technology. There is a huge amount of information, but it is not possible to access or make use of this information unless efficient and effective organization and retrieval is provided.

Early work on image retrieval systems is based on text input, in which the images are annotated by text and retrieval is performed on text (see [65] for a survey on text based image retrieval systems). However, two major difficulties are encountered with text based approaches: First, manual annotation, which is a necessary step for these approaches, is labor-intensive and becomes impractical when the collection is large. Second, keyword annotations are subjective; the same image/video may be annotated differently with different annotators.

In order to overcome these difficulties, content-based retrieval is proposed in the early 1990's. Instead of text-based annotations, images are indexed, searched or browsed by their visual features, such as color, texture or shape. Many systems are introduced for searching image and video archives using their visual contents (see [154, 209, 268, 125, 278, 117] for recent surveys on image and video indexing and retrieval technologies).

In most of the systems, images are matched based on low-level features, like color and/or texture, extracted from the entire image or from image regions. With the exception of systems that can identify faces and cars [272], people [114], or pedestrians [236] matching is not usually directed towards object semantics. However, user studies show that the users seem to be interested in mostly the semantics [110, 237, 207]. Therefore, such systems which are based on low level features do not satisfy the user needs.

Due to the limitations of only text based and only content based systems, recently many systems are proposed to make use of multimodal data. It is shown that, performance of multimedia analysis and understanding systems can be greatly enhanced by combining different modalities such as image, video, audio and text [60, 73, 285, 141, 198].

The goal is to develop useful multimedia systems that employ annotation and search technologies based on semantic concepts that are natural to the user. However, extracting the semantics from the images is a very difficult and long-standing problem. Learning the semantics linked to image features requires carefully labeled data, which is very difficult to require in large quantities. Instead, in recent studies it is shown that, such relationships can be learned from multimodal data sets that provide a loosely labeled data in large quantities. Such data sets include Corel photographs annotated with a few keywords, museum collections with associated decriptions and news photographs on the web.

With careful use of such available data sets, it is shown that semantic labeling of images is possible. In [208], Maron et. al. use multiple-instance learning to train classiffiers for identifying particular keywords from image data using labeled bags of examples (an image is positive if it contains a tiger somewhere in the image, and negative if it doesn't). Wenyin et al. [315] propose a semi-automatic strategy for annotating images using the users feedback of the retrieval system. The query keywords which receive positive feedback are collected as possible annotation to the retrieved images. Li and Wang [192] model image concepts by 2-D multiresolution Hidden Markov Models and label an image with the concepts best fit the content.

Recently, probabilistic models are proposed to capture the joint statistics between image regions and caption terms. Mori et al. [224] proposed a model for automatic annotation of images using the co-occurrences of words with image regions created using a regular grid. Barnard and Forsyth [42] proposed a model which cluster image representations and text, to produce a representation of a joint distribution linking images and words. The model is a multi-modal extension of Hofmann's hierarchical model for text [146] and combines the assymetric clustering model which maps documents into clusters and the symmetric clustering model which models the joint distribution of documents and features (aspect model). Jeon et. al [160] attacked the annotation problem analogous to the cross-lingual retrieval problem and used a cross-media relevance model (CMRM) to perform both image annotation and ranked retrieval. Blei and Jordan [46] extended the Latent Dirichlet Allocation (LDA) Model and proposed a Correlation LDA model which relates words and images. In [105, 40], Duygulu et.al. considers the problem of learning the correspondences between image regions and words as a translation process, similar to the translation of text in two different languages. First, images are segmented into regions, then the regions are clustered in the feature space, categorizing the regions into a finite set of blobs. The correspondences between the blobs and the words are learned, using a method adapted from Statistical Machine Translation [53]. Once learned, these correspondences are used to predict words corresponding to particular image regions (region naming), or words associated with whole images (auto-annotation). Similar approach is also applied to news videos to solve the correspondence problem between visual information and speech transcripts for better annotation and retrieval [107, 106]

3.3.2 Relations Between Concepts in Ontologies and Images

Authors: M. Crucianu, M. Ferecatu, N. Boujemaa, INRIA / IMEDIA

When accessing image or video databases, a user usually wants to retrieve content corresponding to a concept. The use of keywords poses several problems: the manual annotation of images and video is very expensive and inherently incomplete, the relation between words and concepts is sometimes complex due to such phenomena as synonymy (different words denote the same concept) or homonymy (same word denotes different concepts) and some concepts can't be described by a few keywords. Alternatively, the use of the visual content also has limitations due to the "semantic gap" and to the practical difficulty of formulating visual queries. If we consider either the information provided regarding the target concept or the possibilities of interaction between the user and the system, keywords and visual content appear to be rather complementary to each other and it may be valuable to rely on both of them for the retrieval of images.

To fully exploit this potential, one must first establish a comprehensive relation between keywords and visual content. The extension of annotations from one visual entity (entire image, image region, etc.) to another is a by-product of such a relation. One should note that some keywords found in manual annotations don't refer to the visual appearance, even if for some specific database they may occur for images sharing some common visual characteristics; their association with visual content can produce spurious retrieval results.

Part of the work attempting to establish a relation between keywords and visual content consists in the modelling of the visual appearance of images or of image regions corresponding to given concepts. In [105] (following earlier work in [41]) the authors are searching for a correspondence between image *regions* and keywords that were only provided for *entire* images but refer to regions; the method is based on the development (using expectation maximization) of a joint statistical model of the occurrence of keywords and low-level visual descriptions, and can be related to *multiple-instance learning*. Hierarchical aspect models and latent Dirichlet allocation are evaluated in [40], where the authors also study the extension of annotations to other entire images. Supervised learning is used in [26] (see also [283]) for obtaining models (Markov models or support vector machines) of the "visual content" of "atomic concepts" that can be objects, scenes or events and are associated to keywords. In [217], descriptions of image regions are directly associated to user-provided rough visual descriptions—in terms of color, position, size, shape—of concepts in an ontology.

However, it may not be possible to obtain meaningful "visual models" for all the concepts that are related to the visual appearance. But the relation between keywords and visual content can also have other expressions. In some cases, it is implicit in the relations between sets of keywords and sets of images. We first mention [183], where vectorial representations are produced for the texts associated to images and *latent semantic indexing* is performed. Every image is then described both by a vector of visual features and by the latent semantic index (vector) of the text associated to the image; text-based similarity between latent semantic vectors complements the similarity defined by visual features.

By marking several images as “relevant” during a relevance feedback (RF) session, a user usually defines a similarity between these images that goes beyond what can be directly obtained from low-level visual features. Considering that this similarity is related to the presence of common keywords in the annotations of some images marked as “relevant”, in [330] (see also [200]) the authors link these keywords to the images top ranked by RF. A relation between the keywords and the images is thus gradually developed. In a rather analogous setting, the association of keywords to different images marked as “relevant” during an RF session serves in [333] to update similarities between these keywords; the similarity matrix can be initialized using the synonymy relations from an ontology. A “soft” extension of annotations is then performed: a keyword-based feature vector is defined for every image and contains not only keywords that directly annotate this image but also, to some degree, keywords that were found to be similar to these. The resulting similarities between keywords can capture (to some extent) general synonymy but also contextual or user-dependent synonymy and can help in dealing with homonymy. Again, keyword-based similarity complements the similarity defined by the visual features.

The relation between keywords and visual content can also be more explicit. Starting from a collection of images (represented as visual feature vectors) and associated texts (with “bag of words” representations), in [176] the authors search for independent components in the joint representations and rely on these components for the subsequent classification of images. The visual and textual feature vectors are kept apart in [140], but kernel canonical correlation analysis is employed for finding a “semantic image subspace” and a “semantic textual subspace” (having the same dimensions) where the representations of images are maximally correlated to the representations of the associated texts. The two semantic subspaces are then considered as one and the inner product in this subspace is used as a similarity measure between any two vectors.

Once established, the relations between keywords and visual content can be used to identify the images corresponding to a keyword-based query or, alternatively, to suggest keywords for a new image. The results returned by a keyword-based query usually need to be refined and this can be done on the visual features alone, as in [217]. However, the presence of joint representations (including both visual and textual features) makes *combined* search possible, often using some form of RF as in [183], [283], [330] or [333].

The development of models of the visual appearance of images or image regions corresponding to given concepts, put forward in [105], [40], [26] or [217], can also serve as a basis for combined search mechanisms, most useful when the target concept is complex. A development of these approaches can (and should) take more into account differences in the nature of concepts and in the relations between concepts, as defined in an ontology.

3.4 Integrated Multimedia Content Analysis

Authors: N. Simou, K. Rapantzikos, G. Stoilos, V. Spyrou, T. Athanasiadis, Y. Avrithis and G. Stamou, ICCS-NTUA

3.4.1 Introduction

Digital audiovisual information is growing rapidly and this brings out the need of its transformation into applicable knowledge or in other words into a machine-processable representation of its semantics.

The semantics of a media unit depends on the context in which it is used, where the meaning of the context here is twofold. It represents the use of the material in the current application and also represents the overall placement of the information in the domain the application is applied to.

The semantics of a media unit are divided in two categories surface structures (expression) and deep structures (content). Expression is used to represent the media itself, whereas content is the representation of the conceptual items, which are expressed through the media. Both expression and content however depend on the substance and form, where substance represents the natural material for content and expression and on the other hand form represents the abstract structure of relationships, which a particular media demands.

In order to realize the meaning of the four aspects we can consider them in a sports video example. The substance of expression is the mpeg format of the video, the form of expression are the cuts, the substance of content is represented by the ball, and the form of content is given by the structure of the event sequences, in sports usually the sequence of highlights.

The goal of the Multimedia Content Interface [20] is to provide a standardized means of describing audiovisual data content in multimedia data, so that this data can be searched for, browsed, filtered or interpreted either by search engines, filters agents or any other program.

This document is structured as follows. We first present the MPEG-7 standard and the way that it is structure saying few words for each of its descriptors.Then we discuss the techniques of multimedia content analysis and we also give a brief review of work on the multimedia knowledge integration. The following section is related with ontologies presenting the needs that bring ontologies out to Knowledge Based Systems. In he next two sections we will try to review the technologies that exist today for the creation, maintenance, administration, and inferencing with ontologies. Ending the last section discuss the problem of conflicting representations of expression-based media semantics or differently the ontology problem.

3.4.2 Multimedia Content Description

MPEG-7 Standard

MPEG-7 offers a set of audiovisual description tools in the form of descriptors (Ds) and description schemata (DS) describing the structure of the metadata elements, their relationships and the constraints a valid MPEG-7 description should adhere to. These structures form the basis for users to create application specific content descriptions i.e. a set of instantiated description schemata and their corresponding descriptors. The standard is organized in 8 parts, each of them is responsible for a particular aspect of the functionality.

Systems specifies the tools for preparing descriptions for efficient transport and storage, compressing descriptions, and allowing synchronization between content and description. [20]. The *Description Definition Language* (DDL) that is the next part specifies the language for defining the standard set of description tools (Description schemata (DS), descriptors (Ds), and datatypes) and for defining new description tools. The main parser requirements are defined here. [21]. Following *visual* consists of structures and descriptors that cover basic visual features, such as color, texture, shape and motion and appear as a powerful tool for classification, search and comparison of visual content[22]. Similarly *audio* specifies a set of low-level descriptors for audio features (e.g. spectral, parametric and temporal features of a signal) and high-level description tools that are more specific to a set of applications. [23]. The part *Multimedia Description Schemes* (MDS) specifies the generic description tools pertaining to multimedia including audio and visual content [24]. *Reference Software* provides references software to the standard [18]. Ending *Conformance* specifies the guidelines and procedures for testing conformance of implementation of the standard [18].

MPEG-7 Descriptors

Color Descriptors MPEG7 uses a few color spaces including monochrome, RGB, HSV, YCrCb and the new HMMD. Monochrome color space corresponds to the Y component of the YCrCb color space. RGB is a well known color space where color is represented via 3 primary colors Red, Green, Blue, while YCrCb is a luminancechrominance transformation of it. HSV approximates the way humans perceive color. The transformation from RGB to HSV is nonlinear but reversible and can be found in [298]. The HSV H component corresponds to Hue, S corresponds to Saturation and V corresponds to Value. Hue represents color (e.g. Green, Red), Saturation of a color can be changed by adding white, while Value corresponds to the brightness. When the HSV space is represented as a 6-sided inverted pyramid, its top corresponds to V=1. However the most common representation is cylindrical. MPEG7 supports a new color space called HMMD where H is the same as in HSV, while M,M are the maximum and minimum values among those of the RGB color space. Finally D stands for Difference and is defined as the difference between max and min [217]. All these color spaces are allowed for various visual MPEG-7 descriptors. *Dominant Color* [98] is probably the most useful MPEG 7 descriptor for applications like similarity retrieval using color. The *Scalable Color Descriptor* (SCD) is defined in the hue-saturation-value (HSV) color space with fixed color space quantization, and uses a novel Haar transform encoding. The *Color Structure Descriptor* expresses local structure in an image using an 8x8 structuring element. It counts the number of times a particular color is contained within the structuring element as the structuring element scans the image.

Colors present in an image are clustered so as the total number of the remaining colors will be small. These colors are not fixed in the color space but are computed based on the given image. Clustering of colors is followed by the calculation of their percentages and optionally their variances. The spatial coherency is a single number that represents the overall spatial homogeneity of the dominant colors in an image. The method of the dominant color extraction is described in [19]. Each image could have up to a maximum of 8 dominant colors, however experimental results show that 3-4 colors are generally sufficient to provide a good characterization of the region colors. *Color Layout Descriptor* is a compact MPEG7 visual descriptor designed to capture the spatial distribution of color in an image or an arbitrary-shaped region. It uses the YCrCb color space. The given picture is divided into blocks and the average color of each block is calculated. However the representative color of each block is only implicitly recommended to be the average color. A DCT is performed into the series of the average colors and a few low-frequency coefficients are selected using zigzag scanning. The CLD is formed after quantization of the remaining coefficients, as described in

[19]. In conclusion, the CLD is an effective descriptor in applications such as sketch-based image retrieval, content filtering using image indexing and visualization.

Texture Descriptors In addition to color descriptors, MPEG7 contains a few texture descriptors, as they seem to be very powerful for search and retrieval applications [319], image classification [79], [134]. Very shortly, the main texture descriptors are: the Texture Browsing Descriptor which characterizes a texture regularity, directionality and coarseness, the *Homogeneous Texture Descriptor* (HTD), that provides a quintative characterization of texture and the *Local Edge Histogram* Descriptor that captures the spatial distribution of edges [217]. HTD is an easy to compute and robust descriptor. At first, the image is filtered with a bank of orientation and scale sensitive filters. The texture features are extracted from the frequency space which is divided in 30 channels, as described in in [19]. Then, the energy and the energy deviation of each channel are computed.

Shape Descriptors MPEG7 Shape descriptors are a powerful tool for object recognition, as shape often carries semantic information [47]. These descriptors can be characterized as region-based and contour-shaped as both notions of similarity are concerned. The 3-D shape descriptor is based on the shape spectrum and is an extension of the shape index [19]. *Region-Based Shape Descriptor* expresses pixel distribution within a 2-D object region and can describe both simple and complex objects. The *Contour-Based Shape Descriptor* expresses properties of the contour and is probably the most useful descriptor mainly as it can distinguish between shapes that have similar region-shape properties but different contour-shape properties. It is based on the *Curvature Scale Space* (CSS) representation of the contour. A short description of CSS representation is given in [19]. The last Shape Descriptor is 2-D/3-D descriptor which combines 2-D descriptions of a certain object viewed from different angles, in order to form a 3-D view-based representation of it.

Motion Descriptors MPEG7 Motion Descriptors aim to capture essential motion characteristics from the motion field [298] in a more efficient way, to be used in certain types of applications such as similarity matching [159]. It is also possible to combine motion descriptors with other visual descriptors extracted from still scenes, such as color, texture and shape [217],[47], as shown in [68]. These descriptors are classified in two categories: these for a video segment and those for a moving region. The first category consists of: *Motion Activity*, which captures overall motion activity, Camera Motion, which describes the movement of the camera (e.g. pan, zoom, tilt etc), or the motion of the viewing point and finally Warping Parameters which capture global motions. The second category consists of *Motion Trajectory*, which captures a few successive positions of moving objects and Parametric Motion which is similar to the Warping Parameters and enables object retrieval by comparing similar motions. More details can be found in [19].

3.4.3 Multimedia Content Analysis

Descriptor Matching

MPEG-7 determines certain measures for descriptors similarity. Often these measures are explicit. There exist several measures for the descriptors presented above. The L1 norm is usually used for similarity matching with good retrieval accuracy for the SCD, CSD and EHD descriptors. A simple and flexible trajectory distance measure is a linear weighting of distances between object positions, speeds and accelerations. More details may be found in [159].

Pattern recognition methods may be used in conjunction with similarity metrics to produce recognition results. A quite new method in pattern recognition in the area of Neural Networks are the Support Vector Machines [55], [309], [142], introduced by Vapnik. SVMs are capable of solving classification problems that are non-separable by a hyperplane in the input space, by transforming into a higher-dimension feature space, where the problem may be linearly separable. One interesting application of the SVMs can be found in [71], where they are used for Histogram-Based Image classification. Neurofuzzy networks can also be used for recognition. Fuzzy systems are numerical model-free estimators. While neural networks encode sampled information in a parallel-distributed framework, fuzzy systems encode structured, empirical (heuristic) or linguistic knowledge in a similar numerical framework. Although they can describe the operation of the system in natural language with the aid of human-like if-then rules, they do not provide the highly desired characteristics of learning and adaptation. The use of neural networks in order to realize the key concepts of a fuzzy logic system enriches the system with the ability of learning and improves the subsymbolic to symbolic mapping [177], [196],[303], [288].

Knowledge Assisted Analysis

Although the use of segmentation techniques improved significantly image analysis, indexing and multimedia retrieval systems, it did not improve the extraction of semantic knowledge from low-level features. Manipulation of low-level features, such as pixel luminance,the region contour, motion activity and other, can not provide a human understandable representation of the image or video. An automatic way for transition from the low level features to semantic entities or equivalently the automatic extraction of high level characteristics is an extremely hard task [67], . The latest efforts have been focused in the extraction of medium level features, such as automatic summarization and key frames extraction [66],[304], [81], [56]. Similarly, automatic categorization of images in pre-defined classes, such as indoors/outdoors, city/landscape, faces/non faces can be achieved after a training phase [202]. The use of a priori knowledge of well structured, specific domain applications (e.g. sports and news broadcasting) can facilitate the extraction of higher level semantics [324],[28]. In [299], semantic entities, in the context of the MPEG-7 standard, are used for knowledge assisted video analysis and object detection, thus allowing for semantic-level indexing. In [313], fuzzy ontological relations and context-aware fuzzy hierarchical clustering are employed to interpret multimedia content for the purpose of automatic thematic categorization of multimedia documents. In [227] the problem of bridging the gap between low-level representation and high-level semantics is formulated as a probabilistic pattern recognition problem. Finally, in [77], [63], hybrid methods extending the query-by example strategy are developed. Ontology modelling and ontology-based metadata creation currently address mainly textual resources [274] or simple annotation of photographs [317],[203]. Ontologies facilitate inference based on rules and knowledge and are capable of creating new knowledge.

Domain Specific Analysis

Knowledge Assisted Analysis can be performed in a few domains such as football (soccer), tennis and snooker, and can also be used for tasks like face detection and human recognition. In [280], reliable skin segmentation despite wide variation in illumination during tracking is achieved. In [179], an intelligent system that locates human faces within images, using neuro-fuzzy networks is presented, while in [178], the construction of a skin pixel detector is described. 3D human figures tracking in monocular image sequences is performed in [279]. In the sports domain, tennis broadcasts are enhanced with ball tracking and some impressive virtual replays in [251], while in [173] HMMs are used for structure analysis of tennis videos using visual and audio cues. Event detection and summarization from snooker broadcasts is presented in [263]. As for soccer, a fully automatic framework for analysis and summarization is presented in [108] and does not require strictly the use of object-based features, but can be efficient using only cinematic features. HMMs are also used in [321] to analyze the structure of soccer programs, and more specifically to detect play and break and in [36]to detect soccer highlights.

3.4.4 Multimedia Knowledge Integration

Knowledge is usually defined as facts about the world and is often represented as concepts and relationships among the concepts i.e. semantic networks. Concepts are abstraction of objects, situations events or perceptual patterns in the world (e.g. a color pattern and concept Car); relationships represent interactions among concepts (e.g. color pattern one visually similar to color pattern two, and ”sedan” specialization of ”car”).

Automatic knowledge integration summarization and evaluation or in other words multimedia knowledge representation, is essential for multimedia applications because multimedia applications often deal with multimedia knowledge at different abstraction levels such as perceptual and semantic knowledge (e.g., image clusters and word senses, respectively), which can be extracted using different techniques. This diverse multimedia knowledge needs to be integrated to be used in a coherent and meaningful way by applications. Furthermore it is often necessary to reduce the multimedia knowledge, before or after the knowledge integration. Hence ways to quantify the consistency, completeness and conciseness of the multimedia knowledge are essential to evaluate and compare of these knowledge integration and summarization techniques.

Related work on multimedia knowledge integration includes generic pattern classification techniques. In particular, Bayesian Network (BNs) allows the discovery of the statistical structure of a domain but they are not optimized for multimedia. There is a lot of work in the literature on building and fine-tuning classifiers for recognition of objects and scenes in images [242, 296, 305] among other multimedia; however these are usually constrained to a specific domain and trained on skewed data sets. Prior work on multimedia knowledge summarization has been limited to efforts in network and concept reduction such as EZWordNet [218] and VISAR [84]. Ez Word net 1-2 are coarser versions of the English dictionary WordNet generated by collapsing similar word senses and by dropping rare word senses [218]. This process is governed by five rules manually designed by researchers for WordNet so they are not applicable to other knowledge bases or other kinds of knowledge such as perceptual knowledge. WordNet organizes English words into set of synonyms (e.g., ”rock stone”) and connects them with semantic relations (e.g generalization) [219] . VISAR is a hypertext system for the retrieval of textual captions [84]. One of the functionalities of the VISAR system is the representation of the relationships. Several reduction operators are used in this process (e.g replace two concepts for a common ancestor) but the reduction operators are again manually defined and lacking generality. Furthermore the methodology followed by some of the reduction operators is not clearly specified. Prior work relevant to multimedia knowledge evaluation includes manual evaluation of semantic ontologies [122] and automatic but application-oriented evaluation of multimedia knowledge.

3.4.5 Ontologies

Until recently construction of new Knowledge Based Systems required the construction of new Knowledge Bases from scratch. This meant that knowledge had to be recorded from the beginning, even if the exact same knowledge had been previously coded in another Knowledge Based System, reasoners had to be implemented and specialized for the specific domain we wanted to model as well as new inference engines and algorithms for them. As it can be easily understood this process involved a huge amount of effort, time and cost and resulted in small powerless systems. It was in 1991, in the ARPA Knowledge Sharing Effort [229], that researchers tried to deal with this big drawback in the construction of new Knowledge Based Systems. What they envisioned was the construction of new Knowledge Based Systems by assembling reusable components. In that way previously declared knowledge could be reused and the only effort that a System Developer had to make was to create specific knowledge and reasoners that did not previously existed. Building systems in such a way would entail another advantage. Systems then would be capable of communicating each other using knowledge and reasoning not possessed by them and thus accomplishing tasks that were out of their initial capabilities. These features would result in the construction of bigger and more complete Knowledge Based Systems with the minimum amount of cost and effort. The means to accomplish this vision was *ontologies*.

Ontologies have gained a lot of attention the last decade. Their promising features like, knowledge sharing, machine interoperability and intercommunication, extensibility, scalability and inferencing, has caught the attentions of many researchers from different fields. Today there is a lot of effort on using ontologies in medical science [262], word-wide web [44, 7], multimedia [153], schema matching in databases [133] and video processing [1]. But due to their delicate nature as well as the broad power of their capabilities, researchers of the AI field were forced to revise and reconsider the tools and languages that they had available for constructing knowledge based systems. The outcome was the development of a whole new family of languages.

3.4.6 Languages

Knowledge Representation Languages

Knowledge representation languages exist in our world for over 2000 years. From Aristotle's categorical propositions and Boole's Propositional Logic to Frege's

First-Order Logic. In our century new knowledge representation languages have emerged which had divided them into two categories, those that are logic-based, and those that are not. In the logic-based family, languages usually are fragments of first-order logic. They have well defined semantics and mathematical foundations. This makes them general-purpose and very powerful in representing our world. On the other hand they feature many drawbacks like undecidability, inability in representing hierarchies and difficulty in visualizing the described knowledge. These drawbacks drove researchers in creating languages that belong to the second category. Such representation languages are, Semantic Networks [132] and Frame Systems [220]. These languages model knowledge using network-shaped cognitive structures as well as concepts that resemble those of classes and hierarchies from object oriented modeling. These features made these new languages more appealing in applications where the visualization of concepts, their properties and their relations was a crucial factor. Unfortunately they were not fully satisfactory because of their usual lack of precise semantic characterization. The end result of this was that every system behaved differently from the others, despite virtually identical-looking components. As it is obvious none of these languages can be used today for representing ontologies where both decidability and well-defined semantics are required.

Some years before the ARPA effort research on a new formalism for knowledge representation was initiated. This new formalism is a descendant of semantic networks and frame systems with revised constructs to overcome their ambiguities. This formalism is today known as Description Logic [37]. The building blocks of Description Logic are atomic concepts (unary predicates), atomic roles (binary relations) and individuals (constants). The language also features a small set of epistemologically adequate constructors that always keep the language decidable. The language is capable of inferring hierarchies between concepts as well as instance relations between individuals and concepts automatically unlike IS-A relations in semantic networks, which are explicitly stated. As it will be clear by our presentation of ontology languages Description Logic plays an important role in ontology development today and is almost the de facto formalism used by ontology creation languages.

Another formalism developed the last decade and which some ontology development tools use is the F-Logic [172]. F-Logic allows for concise definitions with object oriented-like primitives (classes, attributes, OO-style relations, instances). Furthermore, it also has PL-1 like primitives (predicates and function symbols). In addition, F-Logic allows for axioms that further constrain the interpretation of the model. Axioms may either be used to describe constraints or they may define rules, e.g. in order to define a relation R by the composition of the two other relations S and Q. F-Logic rules have the expressive power of Horn-Logic with negation and may be transformed into Horn-Logic rules. The semantics of

F-Logic are well defined [308]. The semantics are close to first-order semantics. Unlike Description Logic, F-Logic does not provide means for subsumption, but it provides efficient reasoning with instances and the capability to express arbitrary powerful rules.

Ontology Representation Languages

Knowledge Interchange Format (KIF) One of the working groups created in ARPA knowledge-sharing effort was the Interlingua Working Group at Stanford. Their goal was to solve the problem of the heterogeneity of knowledge representation languages. In order to interchange knowledge between heterogeneous programs, they realized that a formal language was needed, which, like an interlingua, allowed knowledge in a given representation language to be expressed in another. The interlingua had to:

1. Be a language with declarative semantics and independent of any interpreter;
2. Be a language with sufficient expressive power to represent the declarative knowledge contained in typical applications system knowledge bases;
3. Have a structure that enabled semiautomatic translations into and out of typical representation languages;

The result was KIF [119], a prefix version of first-order predicate logic, with extensions to improve its expressiveness, such as: definition of terms, representation of knowledge about knowledge, reifying functions and relations, specifying sets and nonmonotonic reasoning. An example definition of an author ontology in KIF is the following:

```
(define-class AUTHOR (?author)
:def (and (person ?author)
(= (value-cardinality ?author AUTHOR.NAME) 1)
(value-type ?author AUTHOR.NAME biblio-name)
(¿= (value-cardinality ?author AUTHOR.DOCUMENTS) 1)
( (author.name ?author ?name)
(person.name ?author ?name))))
```

This definition states that an author must be also a person. He must have exactly one name with the additional constraint that the name is of type ”biblio-name”, he must have at least one document written and at last that the name of an author coincides with his person name.

RDF and RDF Schema for Simple Ontologies XML is already widely known, and is the basis for a rapidly growing number of software development activities. Along with XML Schema XML documents promise data interoperability and reuse. World wide web is currently heavily based on huge amounts of xml documents, which transfer data among computers and form configuration files. The incorporation of ontologies in the current web and the advent of semantic web, combined with the advantages and the need of XML has given ontology languages a new direction. The syntax of almost all today's ontology languages is completely based on xml or they at least have an xml serialization. The first outcome of this combination was RDF and RDFS. The Resource Description Framework (RDF) [8] is a recent W3C recommendation, designed to standardize the definition and use of meta-data descriptions of web-based resources. However, RDF is equally well suited to representing data, like XML. The basic building block in RDF is an object-attribute-value triple commonly written as A (O, V). That is an object O has an attribute A with values V. RDF allows values and objects to be interchanged, thus resulting in a graph structure. An example of RDF is the following :

```
<rdf:Description rdf:about = ”http://www.w3.org/employee/id1321”>
   <hasName rdf :resource= ”Jim Lerners”>
</rdf:Description>
```

which states that the object ”http://www.w3/org/employee/id1321” has the value ”Jim Lerners” for its attribute ”hasName”.

It's important to note that RDF is designed to provide a basic object-attribute-value data model for meta-data. Other than this intended semantics, described only informally in the standard, RDF makes no data-modelling commitments. RDF Schema (RDFS) takes a step further into richer representation formalism and introduces basic ontological modelling primitives into the web. With RDFS we can talk about classes, subclasses, subproperties, domain and range restrictions of properties, and so forth in a web-based context. Despite its similarity with XML Schema RDFS fulfills a different role. RDFS only provides information about the interpretation of the statement given in an RDF data model, but it does not constrain the syntactical appearance of an RDF description. RDFS lets developers define a particular vocabulary for RDF data, such as hasName, and specify the kinds of objects to which these attributes can be applied. In other words, the RDFS mechanism provides a basic type system for RDF models. For example the previous RDF declaration would be in RDFS like this:

```
<rdfs:Property rdf:about hasName>
   <rdfs:domain rdf:resource=”♯ HTMLPage”>
   <rdfs:range rdf:resource=”♯ EmployeeName>
</rdfs:Property>
```

Ontology Inference Layer (OIL) As we can easily see RDFS can be regarded as a very simple ontology creation language. Though it provides us with a mechanism for creating ontologies there are many thing that cannot be stated in RDFS. Such examples are restriction in cardinalities, lack of support for primitive data types and declaring properties to vary within a certain set of range values. These drawbacks led researches to an improvement of RDFS called OIL [112].

OIL language is designed to combine frame -like modeling primitives with the increased expressive power, formal rigor and automated reasoning services of expressive description logic. OIL comes equipped with both XML and RDFS serializations. The frame structure of OIL is based on XOL [167], an XML serialization of the OKBC-lite knowledge model [72]. In these languages, classes are described by frames, whose main components consist of a list of super-classes and a list of slot-filler pairs. OIL extends the basic frame syntax so that it can capture the full power of an expressive description logic, thus overcoming the problem of frame based systems where it was unclear if a slot filler intended the existential or the universal quantifier.

OIL semantics are based on SHIQ (D), which refers to the SHIQ description logic [147], enhanced with simple concrete data types. Within OIL we can define classes, subclasses, slots, slot constraints, which may regard the cardinality as well as the value of the slot fillers and axioms. OIL also provides primitive data types like, string and integer and constructors like conjunction, disjunction and negation for complex class expressions. An example of an OIL ontology is the following:

```
class-def defined Grandmother
      subclass-of Mother
      slot-constraint hasChild
          has-value Parent.
```

DAML+OIL DAML+OIL [2] is similar to OIL in many respects, but is more tightly integrated with RDFS, which provides the only specification of the language and its only serialization. DAML+OIL tries to build on XML and RDFS to produce a language that is well suited for building the Semantic Web. It follows the same path for representing data and information in a document as XML and provides rules and definitions similar to RDFS. Additionally, like OIL, it provides rules for describing further constraints and relationships among resources, including cardinality, domain and range restrictions, conjunction, disjunction, negation and transitive rules. The presence of these role constructors makes the DAML+OIL language theoretically undecidable. In practice, however this is not a very serious problem as it would be easy for a DAML+OIL processor to detect the occurrences of such constraints and warn the user for their consequences. An example of a DAML+OIL ontology is the following:

```
<daml:Class rdf:ID=”Mother>
   <daml:intersectionOf rdf:parseType=”daml:collection”>
      <daml:Class rdf:about=”♯Woman”/>
      <daml:Restriction>
         <daml:onProperty rdf:resource=”♯hasChild”/>
         <daml:hasClass rdf:resource=”♯Person”/>
   </daml:Restriction>
   </daml:intersection>
</daml:Class>
```

Web Ontology Language (OWL) The OWL [99] language is now considered a standard by W3C for the creation of ontologies in the semantic web. It is a revision of DAML+OIL language, which tries to overcome the difficulties that there were present. OWL comes in three flavors, OWL-Lite, OWL-DL and OWL-Full with increasing expressive power, respectively. The semantics of OWL are those of Description Logics incorporated in a frame-like syntax. With OWL we can construct class and property hierarchies, property constraints on values as well as on cardinality, declare roles to be transitive, functional, inverse and symmetric. Conjunction, disjunction and negation of classes and roles can be used as well as many other Description Logic constructors, like equivalent classes, the ”same as” constructor and the ”one of” constructor. An example of an OWL ontology is the following :

```
<owl:Class rdf:ID=”Mother”>
   <rdfs:subClassOf rdf:resource=”♯Woman”/>
   <rdfs:subClassOf>
      <owl:Restriction>
         <owl:onProperty rdf:resource=”♯hasChild”>
         <owl:minCardinality rdf:datatype=” xsd;nonNegativeInteger”>
            1
         <owl:minCardinality>
      <owl:Restriction>
   <rdfs:subClassOf>
<owl:Class¿ ¡owl:ObjectProperty rdf:ID=”hasChild”>
   <rdfs:domain rdf:resource=”♯Wonam”/>
   <rdfs:range rdf:resource=”♯Person”/>
```

```
</owl:ObjectProperty>
```

Semantic Web Rule Language (SWRL) Now that OWL is a standard recommendation of the W3C, techniques to extend the use of OWL ontologies come more into focus. The Joint US/EU ad hoc Agent Markup Language Committee has proposed an OWL rule language called SWRL [3], a Horne clause rules extension to OWL. SWRL extends OWL in a syntactically an semantically coherent manner: the basic syntax for SWRL rules is an extension of the abstract syntax for OWL-DL and OWL-Lite; SWRL rules are given formal meaning via an extension of the OWL-DL model-theoretic semantics; SWL rules are given an XML syntax based on the OWL XML presentation syntax; and a mapping from SWRL rules to RDF graphs is given based on the OWL RDF/XML exchange syntax. Although SWRL provides a fairly minimal rule extension to OWL, the consistency problem of SWRL ontologies is still undecidable, due to property compositions.

3.4.7 Ontology Inferencing

As stated previously one of the biggest advantage of ontologies, as well as to all knowledge based systems, is our ability to inference with them.Since ontologies are built on new knowledge formalism, new inference engines are required. In this section we present some of the inference engine developed for Description Logic as well as for F-Logic.

FaCT

FaCT [147] is an inference engine for the Description Logic SHIQ. FaCT was one of the first description logic systems that used such an expressive set of description logic constructors. The system showed that although the runtime behavior of expressive description logic can be exponential in the worst case, optimization techniques could be found that prevent the system from running into combinatorial explosion and the system could still stay sound and complete. FaCT currently does not support reasoning with instances. The system is implemented in CommonLisp and it is open source for research purposes. It has a CORBA interface to interact with network-aware applications.

RACER

Racer [131] is quite similar to FaCT. It also uses SHIQ and the same optimization techniques but with many improvements and additions. This gives it the ability to reason with individuals. Racer uses techniques to dynamically select appropriate optimization techniques. It is implemented in CommonLisp with a Socket based Java interface also available.

Ontobroker

Ontobroker [97] in an ontology tool created by the ontoprise [5] company. It is indented to provide support for reasoning to other tools of the ontoprise like the popular ontology editing tool OntoEdit [295]. The inference engine of Ontobroker has two key parts: the one that does the translation from the rich modeling language, F-Logic, to a restricted one, Horn-Logic, and the part that does the evaluation of expressions in the restricted language. The queries for ontologies stored in Ontobroker are formulated in F-Logic syntax.

Cerebra

Cerebra [4] is another powerful commercial inference engine used for ontologies. It supports ontologies created in the OWL language, thus the inference is performed for description logic. The ontologies are stored in any relational database, Oracle, MySQL, SQL Server, and the queries are formulated using XQuery, J2EE or .NET.

3.4.8 The ontology problem

An important problem concerning multimedia and knowledge representation is the mapping of the semantics or the ontology problem. The description of a domain provides particular semantic concepts that are hard if not impossible, to be mapped onto ontologies that cover the same or a similar semantic space. This is due to the difficulties in the use of language or conceptual structure. In this section we discuss the representational problems that are caused because the ontological commitment in MPEG-7 is implicit. Furthermore the same problem also applies to the semantics of low-level features and is made explicit in the Audi and Video parts.

Semantics are always purpose driven and because of this a large number of schemata in the standard establish ontological structures. Most of the schemata in MPEG-7 are inspired by the domain of broadcasting and audio-visual entertainment (e.g. the VideoEditingSegment, the AgentDS, the PlaceDS or the user preference schemata in the MDS). So the large number of schemata, often describing similar aspects of the same semantic problem and their interlocked nature, indicate the ontological role at least of the MDS. However the attempt of abstraction to achieve domain independence makes it impossible to use those schemata as ontology items. Furthermore the goal of abstraction is mainly responsible for the syntax and semantic problem.

However there is an attempt in MPEG-7 to address the language problem within ontologies i.e. the classification schema. The classification schema facilitates the organizational wrapper for a controlled vocabulary build out of terms ant the relations between them. Terms are organized by the relations in the form of a hierarchy, indicating if one term is broader or narrower in its meanings than another, a synonym or in the given set of relations, the one of highest relevance. Hence the classification schema in a way covers aspects of a thesaurus and also allows the incorporation of other classification schema, though no indication is given if this feature only takes account of the inclusion of other MPEG-7 classification schemata or also the insertion of or connection to other ontologies. Unfortunately there is also no information provided about how the mapping from previously unconnected terms should be achieved.

It is one achievement of MPEG-7 in particular the parts Audio and Visual that it made the problem of semantic mapping for the level of expressional explicit. The optical and audible patterns in audiovisual media communicate on the basis of precise perception however this low-level data needs interpretation.Thus even on the level o low- level feature description we have to provide collections of objective measurements for media units representing prototypical style-detail based on their own ontology as described in Nack et all [226], Schreiber et al [275] and Dorai and Venkatesh [101]. Each of these descriptors varies on the level of structural depth and the use of feature descriptors, through the approaches describe the same or similar semantics.

3.5 Integration-Fusion Methods

Authors: G. Stamou and N. Simou, ICCS-NTUA

3.5.1 Introduction

Neural networks and fuzzy logic are two bio-mimetic techniques that gained great interest the last few years and they are used to provide approximations to real-world problems. The main advantage of these techniques is that are known to be robust alternatives to conventional deterministic and programmed models.

Fuzzy logic is used to represent qualitative knowledge, and provides interpretability to system models. In other words a system model is explicit and understandable to a knowledge or systems engineer. This makes inspection of the mode easier, and hence validation and maintenance processes are simplified. Zadeh [171] has summarized fuzzy-logic as a body of concepts and techniques for dealing with imprecision, information, granulation, approximate reasoning and computing with words.

Neural networks on the other hand are used in different kinds of problems. They are applied in order to include knowledge of functional relationships from instances or sampled data. This is useful in cases that despite the model is observable the development of analytic model from first principles is not possible. The main difference here, comparing to fuzzy-logic systems, is that this knowledge is not readily understandable to the system engineer because it is encapsulated in the so called black box. Another difference between these two techniques is that while traditionally fuzzy knowledge is obtained from human experts, neural networks relationships are usually automatically learned from a training process that iterates through a sample data. Having these in mind, easily someone can understand that the combination of fuzzy and neural systems provides a synergy such that the marriage of each of their strengths overcomes some of their individual drawbacks and can lead to greatly enhanced systems. In particular, fuzzy system design does not incorporate any learning, while neural networks do not posses mechanisms for explicit knowledge representation.

Fuzzy processing is desirable in computer vision because of the uncertainties that exist in many aspects of image processing. These uncertainties include additive and non-additive noise in low-level image processing, imprecision in the assumptions underlying the algorithms, and ambiguities in interpretation during high-level image processing. As an example for computational convenience we can think of the common process of edge detection that usually models edges as intensity ridges. Nevertheless in practice this assumption only holds approximately, leading to some of the deficiencies of these algorithms. Given the complexity of visual information and the attendant difficulty of determining the fundamental underlying models despite much research in a wide range of areas such as physics, physiology and psychophysics neural networks are useful to computer vision for learning image classification image recognition and general image processing. Neural network and fuzzy logic together with genetic algorithms have in fact emerged as the basis for so-called intelligent systems [171].

Neural-fuzzy or fuzzy-neural hybrid systems exist in two forms. First, there are implementations designed to represent a fuzzy linguistic algorithm in a multilayered network [197]. In this approach neural networks are used in order to improve the performance of fuzzy systems by tuning the rules or the membership functions. The second approach, are implementations that aim to explicitly replicate the processes of fuzzy inference and reasoning through the use of connectionist structures [297]. In this approach, fuzzy concepts, such as linguistic attributes can be build into neural networks to enable knowledge-based interpretation. Ending there is also a third approach in which fuzzy and neural systems are cascaded in any order to achieve one objective or another.

3.5.2 Applications

An area that demonstrates the potential for computer vision systems based on fuzzy-logic and neural networks is robotic applications typically involving recognition and manipulation of objects. Lee and Qian [184] work presents a two-component system for picking up moving objects from a vibratory feeder. The first component that is a fuzzy system selects an object of interest and tracks it and the second component, a neural network, predicts the position at which the robot picks up the object. Problems of this kind involve non-linear dynamics that are often impractical to model accurately. Neural-fuzzy approaches behave better than computing the inverse Jacobian as in the usual feature based feedback control [291], in such problems since they can be used to approximate the nonlinear dynamics.

Another area of emerging significance in computer vision is fuzzy application in this area include tissue identification from magnetic resonance imaging (MRI) data, a system that utilizes features extracted from colour images of poultry viscera to categorise them into normal and abnormal classes [70], and neuro-fuzzy vision system developed to monitor cell populations during fermentation.

A challenge to the use of computer vision is the uncertainty associated with the environments within which systems have to operate. These occasions usually require a reasoning approach in order to cope with these uncertainties. In such cases multiple sources of information are used, from which the final solution is derived through information aggregation. Li et al [193] for example implemented a system for identification of machine-tool wear, which uses a neural network embedded with fuzzy classifiers to analyse several images obtained by laser scattering from the machine surfaces of the work piece. Another similar system is the robotic die polishing system reported by Kuo [181] that also uses multiple images of the die texture to determine the polishing direction. This system operates by analyzing the image using a neural network and integrating together the multiple decisions from different networks using an additional network. Furthermore, the learning of the networks is also controlled using fuzzy models. Mirhosseini et al. [222] describe a face recognition system in which eyes, mouth, and nose locations are detected and each facial feature provides evidence for neural classifiers with varying degrees of reliability. The decisions of individual classifiers are then combined using a fuzzy information fusion technique.

3.5.3 Recent progress

Although general neuro-fuzzy systems have been around for over a decade neurofuzzy vision systems are still in their infancy. In this section we are going to review some of the diversity of neuro-fuzzy approaches and to demonstrate their applicability to tackling computer vision tasks

One of the central issues in neural network research over the years has been the improvement of their effectiveness and efficiency by the development of better architectures and learning algorithms. Canuto et al [57]. presented a variation of the well-known fuzzy-ARTMAP by the addition of reinforcement learning. Their architecture referred to as RePART behaves better over the fuzzy multiplayer perceptron in terms of the training time required.

Another improvement was made by Hou [185] et al. that have proposed a new training algorithm. Their approach combines gradient descent and least mean squares learning. The system presents an interesting combination of image feature extraction, fuzzy C-means clustering, and RBF classification for the unusual application of automatically counting bullet holes in paper targets.

Knowledge acquisition and representation are two of the open issues in neurofuzzy systems. In the paper by Shanahan et al. [39] they describe a system for learning compact models, which are also understandable. This is made possible by a combination of evolutionary learning and Cartesian granule feature modeling offering the advantage of adapting models to changing environments. Their system is applied to object recognition in outdoor scenes.

The final two papers show how neuro-fuzzy techniques can be applied to benefit two diverse image analysis problems. Fisher and Kohlhepp perform the reconstruction of 3D models from multiple range maps and cope with uncalibrated data sets, occlusion, and noise. This is achieved by employing an evolutionary algorithm to generate correspondences between surface patches which are evaluated by neuro-fuzzy similarity measures. These are generated by taking a set of fuzzy rules provided by a human expert, and training a neural network to have the same transfer function. The advantage of this latter step is that in the advent of environmental changes the neural network can be automatically updated by simply extending training with new data without recourse to the expert.

Ending Foody's work [116] is in the area of remote sensing, in particular, the classification of image pixels into land cover types. Given the nature of the data a crisp classification is not always appropriate. Foody's focused on the interpretation of the neural network outputs to provide a soft classification of data. The concept of entropy is applied in combination with the maximum and summed neural network outputs to indicate the appropriateness of crisp versus fuzzy class assignments.

These papers highlight the variety of neuro-fuzzy combinations. For instance, Canuto et al. combine the fuzzy and neuro aspects within a single integrated architecture. Hou et al., on the other hand, present a fuzzy clustering component which then provides a configuration for the neuro component. Yet another variation is given by Foody who provides a neural classifier followed by a fuzzy interpretation stage. At present there appears to be no consensus on what is a good neuro-fuzzy approach and hence the architecture chosen normally only reflects the researcher's preferences and previous experience.

3.5.4 Conclusion

Ending an important question to ask is how useful neuro-fuzzy approaches are. In order to answer to this question we must think of the neuro-fuzzy features and to compare them with other techniques. Hence neuro-fuzzy systems are capable of operating in a dynamic and uncertain environment that need to adaptable to changes. This ability of neuro-fuzzy systems is presented by the work of Fisher and Kohlhepp that transcribed a fuzzy rule base into a neural network, which can be updated in response to new data. Second, ensuring understandability of a system is important on several fronts: user interaction to ensure acceptance of the output, and system development is simplified in terms of debugging, verification, and maintenance. Shenahan et al. have shown that system accuracy comparable with competing techniques can be achieved while maintaining transparency. Shenahan et al. and others also demonstrate that their neuro-fuzzy approach results in more compact models compared to decision trees neural and Bayesian networks proving that compactness of the system model is beneficial both in terms of improving computational efficiency as well as understandability. Finally, it is clear that even a simple noncrisp interpretation of neural network outputs can provide insight into the network's classification, as shown by Foody's paper. Hence, obviously the answer to the question that we addressed at the beginning of the section is that neuro-fuzzy systems are very useful to many domains.

Appendix A

Contributing Authors

AUTH-AIIA

C. Kotropoulos	costas@zeus.csd.auth.gr
N. Nikolaidis	nikolaidis@zeus.csd.auth.gr
C. Cotsaces	
M. Kyperountas	
G. Patsis	

Bilkent University, Turkey

P. Duygulu	duygulu@cs.bilkent.edu.tr

INRIA Imedia

N. Boujemaa	Nozha.Boujemaa@inria.fr
M. Crucianu	michel.crucianu@inria.fr
M. Ferecatu	

INRIA-Texmex

P. Gros	patrick.gros@inria.fr
E. Kijak	

INRIA-Vista

P. Bouthemy	patrick.bouthemy@inria.fr
F. Coldefy	

ICCS-NTUA

P. Maragos	maragos@cs.ntua.gr
G. Stamou	gstam@softlab.ntua.gr
Y. Avrithis	iavr@image.ntua.gr
G. Papandreou	gpapan@cs.ntua.gr
A. Katsamanis	nkatsam@cs.ntua.gr
N. Simou	nsimou@image.ntua.gr
K. Rapantzikos	rap@image.ntua.gr
G. Stoilos	gstoil@image.ntua.gr
T. Athanasiadis	thanos@image.ntua.gr
V. Spyrou	espyrou@image.ntua.gr

TCD

A. Kokaram	anil.kokaram@tcd.ie
R. Dahyot	dahyot@mee.tcd.ie

TUC

A. Potamianos	potam@telecom.tuc.gr
M. Perakakis	
M. Toutoudakis	

TU Vienna-IFS Vienna University of Technology, Austria

A. Rauber	rauber@ifs.tuwien.ac.at

University of Amsterdam, Holland

A.D. Bagdanov	andrew@science.uva.nl
A. Smeulders	
C. Snoek	
M. Worring	

Bibliography

[1] Acemedia: Integrating knowledge, semantics and content for user-centered intelligent media services.

[2] http://www.daml.org.

[3] http://www.daml.org/2003/11/swrl.

[4] http://www.networkinference.com.

[5] http://www.ontoprise.com.

[6] Ibm x+v site. http://www-306.ibm.com/software/pervasive/multimodal/.

[7] Knowledge web: Realizing the semantic web. semantic based knowledge systems.

[8] Resource description framework.

[9] Salt forum. http://www.saltforum.org/.

[10] W3c composite capability/preference profiles working group. http://www.w3.org/Mobile/CCPP/.

[11] W3c extensible multimodal annotation markup language (emma). http://www.w3.org/TR/emma/.

[12] W3c ink markup language. http://www.w3.org/TR/InkML/.

[13] W3c multimodal interaction working group. http://www.w3.org/2002/mmi/.

[14] W3c multimodal interaction working group : Multimodal interaction framework. http://www.w3.org/ TR/mmi-framework/.

[15] W3c multimodal interaction working group : System and environment framework. http://www. w3.org/TR/sysenv/.

[16] W3c voice browser activity. http://www.w3.org/voice/.

[17] W3c web accessibility initiative. http://www.w3.org/WAI/.

[18] Overview of the mpeg-7 standard. 2001.

[19] Text of iso/iec 15938-3 multimedia content description interface part 3: Visual. final comitee draft. *ISO/IEC/JTC1/SC29/WG11, Doc.N4062*, 2001.

[20] Text of iso/iec cd 15938-1 information technology - multimedia content description interface - part 1 systems. 2001.

[21] Text of iso/iec cd 15938-1 information technology - multimedia content description interface - part 2 description definition language. 2001.

[22] Text of iso/iec cd 15938-1 information technology - multimedia content description interface - part 3 visual. 2001.

[23] Text of iso/iec cd 15938-1 information technology - multimedia content description interface - part 4 audio. 2001.

[24] Text of iso/iec cd 15938-1 information technology - multimedia content description interface - part 5 multimedia description schemes. 2001.

[25] N. Adami, A. Buggati, R. Leonardi, and P. Migliorati. Low-level processing of audio and video information for extracting the semantics of context. In *Proc. Fourth IEEE Workshop on Multimedia Signal Processing*, Cannes, France, October 2001.

[26] W. H. Adams, G. Iyengar, C.-Y. Lin, M.R. Naphade, C. Neti, H.J. Nock, and J.R. Smith. Semantic indexing of multimedia content using visual, audio and text cues. *EURASIP Journal on Applied Signal Processing*, 3(2):170–185, 2003.

[27] A. Adjoudani and C. Benoît. On the integration of auditory and visual parameters in an HMM-based ASR. In D.G. Stork and M.E. Hennecke, editors, *Speechreading by Humans and Machines*, pages 461–471. Springer, Berlin, Germany, 1996.

[28] W. Al-Khatib, Y.F. Day, A. Ghafoor, and P.B. Berra. Semantic modeling and knowledge representation in multimedia databases. *IEEE Transactions on Knowledge and Data Engineering*, 11(1):64–80, 1999.

[29] A. Alatan, A. N. Akansu, and W. Wolf. Comparative analysis of Hidden Markov Models for multi-modal dialogue scene analysis. In *Proc. 2000 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, volume IV, pages 2401–2403, Istanbul, Turkey, June 2000.

[30] A.A. Alatan, A.N. Akansu, and W. Wolf. Multi-modal dialog scene detection using hidden markov models for content-based multimedia indexing. *Multimedia Tools and Applications*, 14(2):137–151, 2001.

[31] A. Albiol, L. Torres, and E. J. Delp. Video preprocessing for audiovisual indexing. In *Proc. 2002 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, volume IV, pages 3636–3639, Orlando F.L., U.S.A., May 2002.

[32] P. S. Aleksic and A. Katsaggelos. Audio-visual continuous automatic speech recognition. In *Proc. 2004 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, volume V, pages 917–920, Montreal, Canada, May 2004.

[33] P.S. Aleksic and A.K. Katsaggelos. Product hmms for audio-visual continuous speech recognition using facial animation parameters. In *Proceedings of the IEEE International Conference on Multimedia and Expo, Baltimore, Maryland, USA*, volume 2, pages 481–484, July 2003.

[34] E. André and T. Rist. Presenting through performing: On the use of multiple lifelike characters in knowledge-based presentation systems. In Proceedings of the Second International Conference on Intelligent User Interfaces (IUI 2000): 1-8, 2000.

[35] R. André-Obrecht, B. Jacob, and N. Parlangeau. Audio-visual speech recognition and segmental master slave hmm. In *Proceedings of the Audio-Visual Speech Processing Workshop, Rhodes, Greece*, September 1997.

[36] J. Assfalg, M. Berlini, A. Del Bimbo, W. Nunziat, and P. Pala. Soccer highlights detection and recognition using hmms. In *ICME 2002*, pages 825–828, 2003.

[37] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. Patel-Schneider. *The Description Logic Handbook: Theory, Implementation and Applications*. 2003.

[38] N. Babaguchi and N. Nitta. Intermodal collaboration: A strategy fr semantic content analysis for broadcasted sports video. In *Proceedings of the 10th IEEE International Conference on Image Processing, Barcelone, Espagne*, September 2003.

[39] J.F. Baldwin, T.P. Martin, and J.G. Shanahan. System identification of fuzzy cartesian granules feature models using genetic programming. *Selected and Invited Papers from the Workshop on Fuzzy Logic in Artificial Intelligence,* pages 91–116, 1997.

[40] K. Barnard, P. Duygulu, N. de Freitas, D. A. Forsyth, D. Blei, and M. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[41] K. Barnard, P. Duygulu, and D. Forsyth. Clustering art. In *IEEE Conference in Computer Vision and Pattern Recognition,* volume II, pages 434–441, 2001.

[42] K. Barnard and D. A. Forsyth. Learning the semantics of words and pictures. In *Int. Conf. on Computer Vision,* pages 408–415, 2001.

[43] M. J. Beal, N. Jojic, and H. Attias. A graphical model for audio-visual object tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(7):828–836, July 2003.

[44] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 2001.

[45] D. Bikel, R. Schwartz, and R.M. Weischedel. An algorithm that learns what's in a name. *Machine Learning*, 34(1-3):211–231, 1999.

[46] D.M. Blei and M. I. Jordan. Modeling annotated data. In *26th Annual International ACM SIGIR Conference*, July 28-August 1, 2003, Toronto, Canada.

[47] M. Bober. Mpeg-7 visual shape descriptors. *IEEE Trans. on Circuits and Systems for Video Technology*, 11(6), 2001.

[48] R.M. Bolle, B.-L. Yeo, and M.M. Yeung. Video query: Research directions. *IBM Journal of Research and Development,* 42(2):233–252, 1998.

[49] R. Bolt. Put-that-there : Voice and gesture at the graphics interface. Computer Graphics, 14(3): 262-270, 1980.

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

[51] M. Brand, N. Oliver, and A. Pentland. Coupled hidden markov models for complex action recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 994–999, 1997.

[52] C. Bregler and Y. Konig. Eigenlips for robust speech recognition. In *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing*, pages 669–672, 1994.

[53] P.F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[54] R. Brunelli, O. Mich, and C.M. Modena. A survey on the automatic indexing of video data. *Journal of Visual Communication and Image Representation*, 10(2):78–112, 1999.

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

[56] J. Calic and E. Izquierdo. A multiresolution technique for video indexing and retrieval. In *IEEE Int. Conf. On Image Processing*, 2002.

[57] A.M.d.P. Canuto, G. Howwels, and M. Fairhurst. A comparative performance evaluation of the repart neuro-fuzzy network. *Sixth workshop of Fuzzy systems*, pages 225–229, 1999.

[58] Carnegie Mellon University, The Robotics Institute. Real-time AAM fitting algorithms. `http://www.ri.cmu.edu/projects/project_448.html`, August 2004.

[59] R. Carpenter. The logic of typed feature structures. Cambridge University Press, 1992.

[60] M. La Cascia, S. Sethi, and S. Sclaroff. Combining textual and visual cues for content-based image retrieval on the world wide web. In *Proc. IEEE Workshop on Content-Based Access of Image and Video Libraries*, 1998.

[61] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. *Int'l Journal of Comp. Vision*, 22(1):61–79, February 1997.

[62] J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjlmsson, and H. Yan. Embodiment in conversational interfaces: Rea. Proceedings of the Association for Computing Machinery (ACM) Special Interest Group on Computer Human Interaction (SIGCHI), Pittsburgh, PA., May 1999, pp. 520-527, 1999.

[63] S.S.M. Chan, L. Qing, Y.Wu, and Y. Zhuang. Accommodating hybrid retrieval in a comprehensive video database management system. *IEEE Trans. on Multimedia*, 4(2):146–159, 2002.

[64] D. Chandramohan and P. L. Silsbee. A multiple deformable template approach for visual speech recognition. In *Proc. Int'l Conf. on Spoken Language Processing*, pages 50–53, 1996.

[65] S. Chang and A. Hsu. Image information systems: Where do we go from here? *IEEE Trans. on Knowledge and Data Enginnering*, 4(5):431–442, October 1992.

[66] S. Chang and H. Sundaram. Structural and semantic analysis of video. In *IEEE International Conference on Multimedia and Expo (II)*, 2000.

[67] S.-F. Chang. The holy grail of content-based media analysis. 9(2):6–10, 2002.

[68] S.F. Chang and W.Chen. A fully automated content-based video search engine supporting multi-objects spatio-temporal queries. *IEEE Trans. on Circuits and Systems for Video Technology*, 8:602–615, 1998.

[69] Y.L. Chang, W. Zheng, I. Kamel, and R. Alonso. Integrated image and speech analysis for content-based video indexing. In *Proceedings of the IEEE International Conference on Multimedia Computing and Systems, Hiroshima, Japan*, pages 306–313, 1996.

[70] K. Chao, Y. R. Chen, H. Early, and B Park. Roles of soft computing and fuzzy logic in the conception, design and deployment of information/intelligent systems. *Color image classification systems for poultry viscera inspection*, 15:363–369, 1999.

[71] O. Chapelle, P. Haffner, and V. Vapnik. Support vector machines for histogram-based image classification. *IEEE Trans. on Neural Networks*, 10(5), 1999.

[72] V. Chaudhri, A. Farquhar, R. Fikes, P. Karp, and J. Rice. Okbc: A programmatic foundation for knowledge base interoperability. In *In Proc. of the 15 National Conderence on Artificial Intelligence*, 1998.

[73] F. Chen, U. Gargi, L. Niles, and H. Schutze. Multi-modal browsing of images in web documents. In *SPIE Document Recognition and Retrieval*, 1999.

[74] L. S. Chen and T. S. Huang. Emotional expressions in audiovisual human computer interaction. In *Proc. 2000 IEEE Int. Conf. Multimedia and Expo*, pages 423–426, 2000.

[75] L. S. Chen, T. S. Huang, T. Miyasato, and R. Nakatsu. Multimodal human emotion/expression recognition. In *Proc. IEEE Face and Gesture Recognition Workshop*, pages 396–401, Nara, Japan, 1998.

[76] T. Chen. Audiovisual speech processing. lip reading and lip synchronization. *IEEE Signal Processing Magazine*, 10(1):9–21, January 2001.

[77] W. Chen and S.-F. Chang. Vismap: an interactive image/video retrieval system using visualization and concept maps. In *IEEE Int. Conf. on Image Processing*, pages 588–591, 2001.

[78] Y. Chen and Y. Rui. Real-time speaker tracking using particle filter sensor fusion. *Proceedings of the IEEE*, 92(3):485–494, March 2004.

[79] Y.C. Cheng and S.Y. Chen. Image classification using color, texture and regions. *Image and Video Computing*, 21:759–776, 2003.

[80] G. Chiou and J.-N. Hwang. Lipreading from color video. *IEEE Trans. on Image Processing*, 6(8):1192–1195, August 1997.

[81] M. Christel, A. Hauptmann, H. Wactlar, and T. Ng. Collages as dynamic summaries for news video. In *ACM Multimedia*, 2002.

[82] M. Christel, A. Olligschlaeger, and C. Huang. Interactive maps for a digital video library. *IEEE Multimedia*, 7(1):60–67, 2000.

[83] S. Chu and T. Huang. Audio-visual speech modeling using coupled hidden markov models. In *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing*, pages 2009–2012, 2002.

[84] P. Clitherow, D. Riecken, and M. Muller. Visar: A system for inference and navigation in hypertext. *ACM Conference on Hypertext*, pages 5–8, 1989.

[85] P. Cohen and S. Oviatt. The role of voice in human-machine communication. In Voice Communication Between Humans and Machines. Roe, D., Wilpon, J. (editors). National Academy Press, Washington D.C.: 34-75, 1994.

[86] F. Coldefy and P. Bouthemy. Unsupervised soccer video abstraction based on pitch, dominant color and camera motion analysis. In *ACM Multimedia*, 2004.

[87] T. F. Cootes, Taylor C.J., Cooper D. H., and J. Graham. Active shape models - their training and application. *Comput. Vis. Image Underst.*, 61(1):38–59, 1995.

[88] T.F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *Proc. Europ. Conf. on Comp. Vision*, volume II, pages 484–498. Springer-Verlag, 1998.

[89] R.G. Cowell, A.P. Dawid, S.L. Lauritzen, and D.J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer-Verlag, 1999.

[90] R. Cowie, E.Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 2001.

[91] S. Cox, I. Matthews, and A. Bangham. Combining noise compensation with visual information in speech recognition. In *European Tutorial Workshop on Audio-Visual Speech Processing*, pages 53–56, 1997.

[92] R. Cutler, Y. Rui, A. Gupta, J. Cadiz, I. Tashev, L.-W. He, A. Colburn, Z. Zhang, Z. Liu, and S. Silverberg. Distributed meetings: A meeting capture and broadcasting system. In *ACM Conf.Multimedia*, pages 123–132, 2002.

[93] S. Dagtas and M. Abdel-Mottaleb. Extraction of TV highlights using multimedia features. In *Proc. Fourth IEEE Workshop on Multimedia Signal Processing*, Cannes, France, October 2001.

[94] R. Dahyot, A. C. Kokaram, N. Rea, and H. Denman. Joint audio visual retrieval for tennis broadcasts. In *IEEE proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Hong Kong, April 2003.

[95] R. Dahyot, N. Rea, A. Kokaram, and N. Kingsbury. Inlier modeling for multimedia data analysis. In *IEEE International Workshop on MultiMedia Signal Processing*, Siena, Italy, September 2004.

[96] T. Dean and K. Kanazawa. A model for reasoning about persistence and causation. *Artificial Intelligence*, 93(1-2):1–27, 1989.

[97] S. Decker, M. Erdmann, D. Fensel, and R. Studer. Ontobroker: Ontology based access to distributed and semi-structured information. *Database Semantics: Semantic Issues in Multimedia Systems*, 1999.

[98] Y. Deng, B.S. Manjunath, C. Kenney, and M.S. Moore. An efficient color representation for image retrieval. *IEEE Trans. on Image Processing*, 10(1), 2001.

[99] R. Dieng and S. Hug. Owl web ontology language. *Working Draft W3C*, 2003.

[100] N. Dimitrova, H.-J. Zhang, B. Shahraray, I. Sezan, T. Huang, and A. Zakhor. Applications of video-content analysis and retrieval. *IEEE Multimedia Magazine*, 12(3), Jul 2002.

[101] C. Dorai and S. Venkatesh. Bridging the semantic gab in content managment systems: Computational media aesthetics. In *Proceedings of the First Conference on Computational Semiotics for Games and New Media-COSIGN*, pages 94–99, 2001.

[102] 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`.

[103] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley, 2nd edition, 2000.

[104] S. Dupont and J. Luettin. Audio-visual speech modeling for continuous speech recognition. *IEEE Trans. Multimedia*, 2:141–151, 2000.

[105] P. Duygulu, K. Barnard, N.d. Freitas, and D. A. Forsyth. Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. In *Seventh European Conference on Computer Vision (ECCV)*, volume 4, pages 97–112, Copenhagen, Denmark, May 27 - June 2 2002.

[106] P. Duygulu and Alex Hauptmann. Whats news, whats not? associating news videos with words. In *The 3rd International Conference on Image and Video Retrieval (CIVR 2004) Ireland*, July 21-23, 2004.

[107] P. Duygulu and H. Wactlar. Associating video frames with text. In *Multimedia Information Retrieval Workshop, in conjuction with the 26th annual ACM SIGIR conference on Information Retrieval, August 1, 2003, Toronto, Canada*.

[108] A. Ekin, A. M. Tekalp, and R. Mehrotra. Automatic soccer video analysis and summarization. *IEEE Trans. on Image Processing*, 2003.

[109] C. Elting, S. Rapp, G. Mohler, and M. Strube. Architecture and implementation of multimodal plug and play. ICMI 03, November 5 7, 2003, Vancouver, British Columbia, Canada, 2003.

[110] P.G.B. Enser. Query analysis in a visual information retrieval context. *Journal of Document and Text Management*, 1(1):25–39, 1993.

[111] S-C. Chen et. al. Scene change detection by audio and video clues. In *International Conference on Multimedia and Expo*, volume 2, pages 365 – 368, 2002.

[112] D. Fensel, F. Harmelen, I. Horrocks, S. Decker, M. Erdmann, and M. Klein. Oil in a nutshell. In *Proc. Of the 12th European Workshop on Knowledge Acquisition*, 2000.

[113] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc. IEEE Conf. on Comp. Vision and Pattern Recogn.*, pages II: 264–271, 2003.

[114] M.M. Fleck, D.A. Forsyth, and C.Bregler. Finding naked people. In *4th European Conference on Computer vision*, volume 2, pages 591–602, 1996.

[115] F. Flippo, A. Krebs, and I. Marsic. A framework for rapid development of multimodal interfaces. ICMI 03, November 5 7, 2003, Vancouver, British Columbia, Canada, 2003.

[116] G. M. Foody. Land cover mapping from remotely sensed data with a neural network: accommodating fuzziness. In *Proc. 1st COMPARES Workshop*, 1996.

[117] D. A. Forsyth and J. Ponce. *Computer Vision: a modern approach*. Prentice-Hall, 2002.

[118] B. Furht, S.W. Smoliar, and H.-J. Zhang. *Video and Image Processing in Multimedia Systems*. Kluwer Academic Publishers, Norwell, USA, 2th edition, 1996.

[119] M. Genesereth and R. Fikes. Knowledge interchange format. 1992.

[120] Z. Ghahramani and M.I. Jordan. Factorial hidden markov models. In *Proc. Advances in Neural Information Processing Systems*, volume 8, pages 472–478, 1995.

[121] A. Ghosh, A. Verma, and A. Sarkar. Using likelihood L-statistics to measure confidence in audio-visual speech recognition. In *Proc. Fourth IEEE Workshop on Multimedia Signal Processing*, Cannes, France, October 2001.

[122] A. Gomez-Perez. Evaluation of taxonomic knowledge in ontologies and knowledge bases. *Workshop on Knowledge Acquisition*, pages 16–21, 1999.

[123] Y.-H. Gong. Summarizing audio-visual contents of a video program. *EURASIP Journal on Applied Signal Processing*, Feb 2003.

[124] Y.-H. Gong and Xin Liu. Video summarization and retrieval using singular value decomposition. *ACM Multimedia Systems Journal*, 11(5):0 – 0, Sep 2003.

[125] A. A. Goodrum. Image information retrieval: An overview of current research. *Informing Science*, 3(2):63–66, 2000.

[126] J. Gowdy, A. Subramanya, C. Bartels, and J. Bilmes. DBN-based multi-stream models for audio-visual speech recognition. In *Proc. 2004 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, volume I, pages 993–997, Montreal, Canada, May 2004.

[127] G. Gravier, G. Potamianos, and C. Neti. Asynchrony modeling for audio-visual speech recognition. In *Human Language Technology Conference*, 2002.

[128] J. Gustafson. *Developing Multimodal Spoken Dialogue Systems. Empirical Studies of Human-Computer Interaction*. PhD thesis, Departmnet of Speech, Music and Hearing, KTH, 2002.

[129] J. Gustafson, L. Bell, J. Beskow, J. Boye, R. Carlson, J. Edlund, B. Granstrm, D. House, and M. Wirn. Adapt - a multimodal onversational dialogue system in an apartment domain. In Proceedings of 6th International Conference on Spoken Language Processing (ICSLP 2000): 134-137, 2000.

[130] J. Gustafson, N. Lindberg, and M. Lundeberg. Experiences from the development of august - a multi-modal spoken dialogue system. In Proceedings of the ESCA tutorial and research workshop on Interactive Dialogue in Multi-Modal Systems, IDS 99, 1999.

[131] V. Haasrlev and R. Moller. Description of the racer system and its applications. In *In Proceedings of the 2001 Description Logic Workshop*, 2001.

[132] F. Hakimpour and A. Geppert. Word concepts: A theory and simulation of sime basic capabilities. *Behavioral Science*, 1967.

[133] F. Hakimpour and A. Geppert. Resolving semantic heterogeneity in schema integration: An ontology based approach. *Formal Ontology in Information Systems*, 2001.

[134] G.M. Haley and B.S. Manjunath. Rotation-invariant texture classification using a complete space-frequency model. *IEEE Trans. on Image Processing*, 8(2):759–776, 1999.

[135] P.W. Hallinan, G.G. Gordon, A.L. Yuille, P. Giblin, and D. Mumford. *Two– and Three–dimensional Patterns of the Face*. A. K. Peters, Ltd., 1999.

[136] A. Hampapur, R. Jain, and T. Weymouth. Feature based digital video indexing. In *IFIP 2.6 Third Working Conference on Visual Database Systems*, Lausanne, Switzerland, 1995.

[137] M. Han, W. Hua, and Y. Gong. An integrated baseball digest system using maximum entropy method. In *Proceedings of the ACM International Conference on Multimedia, Juan-les-Pins, France*, pages 347–350, December 2002.

[138] A. Hanjalic. Generic approach to highlights extraction from a sport video. In *Proceedings of the 10th IEEE International Conference on Image Processing, Barcelone, Espagne*, September 2003.

[139] A. Hanjalic, G.C. Langelaar, P.M.B. van Roosmalen, J. Biemond, and R.L. Lagendijk. *Image and Video Databases: Restoration, Watermarking and Retrieval*. Elsevier Science, Amsterdam, The Netherlands, 2000.

[140] D.R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis; an overview with application to learning methods. Technical Report CSD-TR-03-02, Royal Holloway University of London, 2003.

[141] A. Hauptmann. Informedia at trecvid 2003: Analyzing and searching broadcast news video. In *Proceedings of (VIDEO) TREC 2003 (Twelfth Text Retrieval Conference), Gaithersburg, MD*, November 17-21, 2003.

[142] S. Haykin. *Neural Networks: A comprehensive foundation, Second Edition, Prentice Hall*. 1999.

[143] M. Heckmann, F. Berthommier, and K. Kroschel. Noise adaptive stream weighting in audio-visual speech recognition. *EURASIP Journal of Applied Signal Processing*, (11):1260–1273, November 2002.

[144] M.E. Hennecke, D.G. Stork, and K.V. Prasad. Visionary speech: Looking ahead to practical speechreading systems. In D.G. Stork and M.E. Hennecke, editors, *Speechreading by Humans and Machines*, pages 331–349. Springer, Berlin, Germany, 1996.

[145] J. Hershey and M. Case. Audio-visual speech separation using hidden markov models. In *Proc. Advances in Neural Information Processing Systems*, volume 14, 2002.

[146] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning Journal*, 42(1):177–196, 2001.

[147] I. Horrocks. Fact and ifact. In *In Proc. of the 1999 Description Logic Workshop*, 1999.

[148] W.H.M. Hsu and S.F. Chang. A statistical framework for fusing mid-level perceptual features in news story segmentation. In *Proceedings of the IEEE International Conference on Multimedia and Expo, Baltimore, Maryland, USA*, volume 2, pages 413–416, July 2003.

[149] J. Huang, Z. Liu, and Y. Wang. Integration of audio and visual information for content-based video segmentation. In *International Conference on Image Processing*, volume 3, pages 526 – 530, 1998.

[150] J. Huang, Z. Liu, and Y. Wang. Integration of multimodal features for video scene classification based on hmm. In *Proceedings of the IEEE Workshop on Multimedia Signal Processing, Copenhagen, Denmark*, pages 53–58, September 1999.

[151] J. Huang, Z. Liu, and Y. Wang. Joint video scene segmentation and classification based on hidden markov models. In *Proceedings of the IEEE International Conference on Multimedia and Expo, New York, New York, USA*, volume 3, pages 1551–1554, August 2000.

[152] Q. Huang, Z. Liu, A. Rosenberg, D. Gibbon, and B. Shahraray. Automated generation of news content hierarchy bt integrating audio, video, and text information. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Los Alamitos, California, USA*, volume 6, pages 3025–3028, March 1999.

[153] J. Hunter. Adding multimedia to the semantic web-building an mpeg-7 ontology.

[154] F. Idris and S. Panchanathan. Review of image and video indexing techniques. *Journal of Visual Communication and Image Representation*, 8(2):146–166, 1997.

[155] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In *Proc. Europ. Conf. on Comp. Vision*, volume I, pages 343–356, 1996.

[156] M. Isard and A. Blake. *Snakes: Active Contour Models*. Springer, 1998.

[157] U. Iurgel, R. Meermeier, S. Eickeler, and G. Rigoll. New approaches to audio-visual segmentation of tv news for automatic topic retrieval. In *Proceedings of the IEEE International Conference on Acoustics, Speech, ans Signal Processing, Salt Lake City, Utah, USA*, May 2001.

[158] A.K. Jain, R.P.W. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000.

[159] S. Jeannin and A. Divakaran. Mpeg-7 visual motion descriptors. *IEEE Trans. on Circuits and Systems for Video Technology*, 11(6), 2001.

[160] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *26th Annual International ACM SIGIR Conference*, July 28-August 1, 2003, Toronto, Canada.

[161] H. Jiang, T. lin, and H. Zhang. Video segmentation with the support of audio segmentation and classification. In *Proceedings of the IEEE International Conference on Multimedia and Expo, New York, New York, USA*, volume 3, pages 1551–1554, August 2000.

[162] H. Jiang, T. Lin, and H.J. Zhang. Video segmentation with the assistance of audio content analysis. In *International Conference on Multimedia and Expo*, pages 1507–1510, 2000.

[163] J. Jiang, G. Potamianos, H. Nock, G. Iyengar, and C. Neti. Improved face and feature finding for audio-visual speech recognition in visually challenging environments. In *Proc. 2004 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, volume V, pages 873–876, Montreal, Canada, May 2004.

[164] M. Johnston. Unification-based multimodal parsing. Proceedings of COLING-ACL, 1998.

[165] M. Johnston, P. R. Cohen, D. McGee, S. L. Oviatt, J. A. Pittman, and I. Smith. Unification-based multimodal integration. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and of the 8th Conference of the European Chapter of the Association for Computational Linguistics, Madrid, Spain, 7 12 July 1997, pages 281 288, 1997.

[166] M. J. Jones and T. Poggio. Multidimensional morphable models: A framework for representing and matching object classes. *Int'l Journal of Comp. Vision*, pages 107–131, 1998.

[167] P. Karp, V. Chaudhri, and J Thomere. Xol: An xml-based ontology exchange language. 1999.

[168] M. Kass, A.P. Witkin, and D. Terzopoulos. Snakes: Active contour models. *Int'l Journal of Comp. Vision*, 1(4):321–331, January 1988.

[169] R. Kaucic and A. Blake. Accurate, real-time, unadorned lip tracking. In *Proc. Int'l Conf. on Comp. Vision*, 1998.

[170] M. Kay. Functional grammar. Proceedings of the Fifth Annual Meeting of the Berkeley Linguistics Society, 142-158, 1979.

[171] O. Kaynak, L. A. Zadeh, B. Turksen, and I. J. Rudas. Roles of soft computing and fuzzy logic in the conception, design and deployment of information/intelligent systems. *Computational Intelligence: Soft Computing and Fuzzy-Neuro Integration with Applications*, 162:1–9, 1998.

[172] M. Kifer, G. Lausen, and J. Wu. Logical foundations of object-oriented and frame-based languages. *Journal of the ACM*, 1995.

[173] E. Kijak, G. Gravier, P. Gros, L. Oisel, and F. Bimbot. Hmm based structuring of tennis videos using visual and audio cues. In *Proceedings of the IEEE International Conference on Multimedia and Expo, Baltimore, Maryland, USA*, volume 3, pages 309–312, July 2003.

[174] J.G. Kim, H.S. Chang, Y.T. Kim, K. Kang, M. Kim, J. Kim, and H.M. Kim. Multimodal approach for summarizing and indexing news video. ETRI *Journal*, 24(1):1–11, February 2002.

[175] K. Kim, J. Choi, N. Kim, and P. Kim. Extracting semantic information from basket-ball video based on audio-visual features. In *Proceedings of the International Conference on Image and Video Retrieval, London, England*, volume 2383 of *Lecture Notes in Computer Science*, pages 278–288. Springer-Verlag, July 2002.

[176] T. Kolenda, L.K. Hansen, J. Larsen, and O. Winther. Independent component analysis for understanding multimedia content. In H. Bourlard, T. Adali, S. Bengio, J. Larsen, and S. Douglas, editors, *Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing*, pages 757–766, Piscataway, New Jersey, 2002. IEEE Press. Martigny, Valais, Switzerland, Sept. 4-6, 2002.

[177] B. Kosko. *Neural networks and fuzzy systems: a dynamical approach to machine intelligence, Prentice Hall, Englewood Cliffs*. 1992.

[178] A.Z. Kouzani. Statistical color models with application to skin detection. *Cambridge Research Laboratory Technical Report Series, CRL 98/11*, 1998.

[179] A.Z. Kouzani. Locating human faces within images. *Computer Vision and Image understanding*, 91:247–279, 2003.

[180] S. Kumar and P.R. Cohen. Towards a fault-tolerant multi-agent system architecture. Fourth International Conference on Autonomous Agents 2000, 459-466. Barcelona, Spain: ACM Press, 2000.

[181] R. J. Kuo. A robotic die polishing system through fuzzy neural networks. *Comput. Industry*, 32:273–280, 1997.

[182] M. Kyperountas, Z. Cernekova, C. Kotropoulos, M. Gavrielides, and I. Pitas. Scene change detection using audiovisual clues. In *Norwegian Conference on Image Processing and Pattern Recognition*, 5 2004.

[183] M. La Cascia, S. Sethi, and S. Sclaroff. Combining textual and visual cues for content-based image retrieval on the world wide web. In *IEEE Workshop on Content-Based Access of Image and Video Libraries*, pages 24–28, 1998.

[184] K. M. Lee and Y. F. Qian. Intelligent vision-based part-feeding on dynamic pursuit of moving objects. *J. Manufacturing Sci. Engrg.Trans*, 120:640–647, 1998.

[185] S.-J. Lee and C.-L. Hou. An art-based construction of rbf networks. *IEEE Transactions on Neural Networks*, 13:1308–1321, 2002.

[186] S. Lefévre, B. Maillard, and N. Vincent. 3 classes segmentation for analysis of football audio. In *in Proc. 14th IEEE Int. Conf. Digital Signal Processing*, volume II, pages 975–978, Santorini, Greece, July 2002.

[187] R. Leonardi, P. Migliorati, and M. Prandini. Semantic indexing of sport program sequences by audio-visual analysis. In *Proceedings of the 10th IEEE International Conference on Image Processing, Barcelone, Espagne*, September 2003.

[188] M.E. Leventon, W.E.L. Grimson, and O. Faugeras. Statistical shape influence in geodesic active contours. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume I, pages 316–323, 2000.

[189] B. Li and I. Sezan. Event detection and summarization in sports video. In *CVPR*, 2001.

[190] B. Li and M.I. Sezan. Event detection and summarization in american football broadcast video. In *Proceddings of the —sc is& t/spie Conference on STorage and Retrieval for Media Databases*, volume 4676, pages 202–213, 2002.

[191] D. Li, I.K. Sethi, N. Dimitrova, and T. McGee. Classification of general audio data for content-based retrieval. *Pattern Recognition Letters*, 22(5):533–544, 2001.

[192] J. Li and J.Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(10):14, 2003.

[193] X. Q. Li, Y. S. Wong, and A. Y. C Nee. Intelligent tool wear identification based on optical scattering image and hybrid artificial intelligence techniques. In *Proc. Inst. of Mechanical Engineers Part B*, volume 213, pages 191–196, 1999.

[194] Y. Li, S. Narayanan, and C. -C. Jay Kuo. Identification of speakers in movie dialogs using audiovisual cues. In *Proc. 2002 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, volume II, pages 2093–2096, Orlando F.L., U.S.A., May 2002.

[195] R. Lienhart, S. Peiffer, and W. Effelsberg. Video abstracting. *Communications of the ACM*, 40(12):54–62, 1997.

[196] C.-T. Lin and C.S. Lee. *Neural fuzzy Systems: A neuro-fuzzy synergism to intelligent systems, Prentice Hall, Englewood Cliffs, NJ*. 1995.

[197] C. T. Lin and S. G. Lee. Reinforcement structure/parameter learning for neural network based fuzzy logic systems. *IEEE Trans. Fuzzy Systems*, 2:46–63, 1994.

[198] C.-Y. Lin. Ibm research trecvid-2003 video retrieval system. In *Proceedings of (VIDEO) TREC 2003 (Twelfth Text Retrieval Conference), Gaithersburg, MD*, November 17-21, 2003.

[199] Z. Liu and Q. Huang. etecting news reporting using audio/visual information. In *Proceedings of the 6th IEEE International Conference on Image Processing, Kobe, Japan*, volume 1, pages 324–328, October 1999.

[200] Y. Lu, C. Hu, X. Zhu, H.-J. Zhang, and Qiang Yang. A unified framework for semantics and feature based relevance feedback in image retrieval systems. In *Proceedings of the 8th ACM International Conference on Multimedia*, pages 31–37. ACM Press, 2000.

[201] J. Luettin. *Visual Speech and Speaker Recognition*. PhD thesis, Univ. of Sheffield, May 1997.

[202] J. Luo and A. Savakis. Indoor vs outdoor classification of consumer photographs using low-level and semantic features. In *IEEE Int. Conf. On Image Processing*, 2001.

[203] J. Luo, A. Singhal, S.P. Etz, and R.T. Gray. A computational approach to determination of main subject regions in photographic images. *Image and Vision Computing*, 22(3):227–241, 2004.

[204] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li. A user attention model for video summarization. In *ACM Multimedia*, Dec 2002.

[205] C.D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, USA, 1999.

[206] K. Mardia, J. Kent, and J. Bibby. *Multivariate Analysis*. Academic Press, 1979.

[207] M. Markkula and E. Sormunen. End-user searching challenges indexing practices in the digital newspaper photo archive. *Information retrieval*, 1:259–285, 2000.

[208] O. Maron and A. L. Ratan. Multiple-instance learning for natural scene classification. In *The Fifteenth International Conference on Machine Learning*, 1998.

[209] M. De Marsicoi, L. Cinque, and S. Levialdi. Indexing pictorial documents by their content: A survey of current techniques. *Image and Vision Computing*, 15(2):119–141, 1997.

[210] D. L. Martin, A. J. Cheyer, and D. B. Moran. The open agent architecture: A framework for building distributed software systems. Applied Artificial Intelligence, 13, 91-128, 1999.

[211] D. Massaro and D. Stork. Speech recognition and sensory integration. *American Scientist*, 86(3):236–244.

[212] I. Matthews and S. Baker. Active appearance models revisited. *Int'l Journal of Comp. Vision*, 60(2):135–164, 2004.

[213] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey. Extraction of visual features for lipreading. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(2):198–213, February 2002.

[214] I. Matthews, G. Potamianos, C. Neti, and J. Luettin. A comparison of model and transform-based visual features for audio-visual LVCSR. In *Proc. Int'l Conf. on Mult. and Expo*, 2001.

[215] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976.

[216] M. McTear. Spoken dialogue technology: Enabling the conversational interface. ACM Computing Surveys, 34, 1, March 2002: 90-169, 2002.

[217] V. Mezaris, I. Kompatsiaris, and M. G. Strintzis. Still image segmentation tools for object-based multimedia applications. *International Journal of Pattern Recognition and Artificial Intelligence*, 18(4):701–725, 2004.

[218] R. Mihalcea and D. Moldovan. Automatic generation of a coarse grained wordnet. *Workshop on WordNet and Other Lexical Resources, in the North American Chapter of the Association for Computational Linguistics*, 2001.

[219] G.A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41, 1995.

[220] M. Minski and J. Haugeland. A framework for representing knowledge. *Mind Design*, 1981.

[221] M. Minsky. A framework for representing knowledge. In P. Winston (Ed.), The Psychology of Computer Vision (pp 211-277). New York: McGraw-Hill, 1975.

[222] A. R. Mirhosseini, H. Yan, K. M. Lam, and T. Pham. Human face image recognition: An evidence aggregation approach. *Computer Vision Image Understanding*, 71:213–230, 1998.

[223] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(4):349–361, 2001.

[224] Y. Mori, H. Takahashi, and R. Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *First International Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.

[225] M.Petkovic, V. Mihajlovic, W. Jonker, and S. Djordjevic-Kajan. Multimodal extraction of highlights fro tv formula 1 programs. In *Proceedings of the IEEE International Conference on Multimedia and Expo, Lausanne, Switzerland*, volume 1, pages 817–820, August 2002.

[226] F. Nack and W. Putz. Designing annotation before its needed. In *In Proceedings of the 9th ACM International Conference on Multimedia*, pages 251–260, 2001.

[227] M.R. Naphade, I.V. Kozintsev, and T.S. Huang. A factor graph framework for semantic video indexing. *IEEE Trans. on Circuits and Systems for Video Technology*, 12(1):40–52, 2002.

[228] J. Neal and S. Shapiro. Intelligent multi-media interface technology. In J. Sullivan and S, Tyler (Eds.) Intelligent User Interfaces . Addison-Wesley. 11-43, 1991.

[229] R. Neches, R. Fikes, T. Finin, T. Gruber, R. Patil, T. Senatir, and W.R. Swartout. Enabling technology for knowledge sharing. *AI Magazine*, 1991.

[230] A.V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy. Dynamic bayesian networks for audio-visual speech recognition. *EURASIP Journal on Applied Signal Processing*, (11):1–15, 2002.

[231] A.V. Nefian, L. Liang, X. Pi, L. Xioaxiang, C. Mao, and K. Murphy. A coupled hmm for audio-visual speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, ans Signal Processing, Orlando, Florida, USA*, May 2002.

[232] S. Nepal, U. Srinivasan, and G. Reynolds. Automatic detection of goal segments in basketball videos. In *ACM Multimedia*, 2001.

[233] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou. Audio-visual speech recognition: Final workshop 2000 report. Technical report, Center for Language and Speech Processing, The Johns Hopkins University, 2000.

[234] H.T. Nguyen, M. Worring, and A. Dev. Detection of moving objects in video using a robust motion similarity measure. *IEEE Transactions on Image Processing*, 9(1):137–141, 2000.

[235] L. Nigay and J. Coutaz. A design space for multimodal systems: concurrent processing and data fusion. Human Factors in Computing Systems. In Proceedings of INTERCHI'93, ACM Press: 172-178, 1993.

[236] M. Oren, C. Papageorgiou, P. Sinha, and E. Osuna. Pedestrian detection using wavelet templates. In *Computer vision and pattern recognition*, 1997.

[237] S. Ornager. View a picture, theoretical image analysis and empirical user studies on indexing and retrieval. *Swedis Library Research*, 2-3:31–41, 1996.

[238] E. Osuna, R. Freund, and F. Girosi. Training support vector machines: An application to face detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 130–136, 1997.

[239] S. Oviatt. Mutual disambiguation of recognition errors in a multimodal architecture. In Proceedings of Conference on Human Factors in Computing Systems: CHI '99, New York, N.Y., ACM Press: 576-583, 1999.

[240] S. Oviatt. Ten myths of multimodal interaction. Communications of the ACM, 1999.

[241] S. Oviatt. Multimodal interfaces. Chapther to appear in Handbook of Human-Computer Interaction, (ed. by J. Jacko & A. Sears), Lawrence Erlbaum: New Jersey, 2002.

[242] S. Paek and S.-F. Chang. The case for image classification systems based on probabilistic reasoning. *IEEE Interational Conference of Multimedia and Expo*, 2000.

[243] M. Pantic and L. J. M. Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390, September 2003.

[244] C. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection. In *Proc. Int'l Conf. on Comp. Vision*, pages 555–562, 1998.

[245] N.V. Patel and I.K. Sethi. Audio characterization for video indexing. In *Proceedings SPIE on Storage and Retrieval for Still Image and Video Databases*, volume 2670, pages 373–384, San Jose, USA, 1996.

[246] N.V. Patel and I.K. Sethi. Video classification using speaker identification. In *IS&T SPIE, Proceedings: Storage and Retrieval for Image and Video Databases IV*, San Jose, USA, 1997.

[247] E.D. Petajan. *Automatic Lipreading to Enhance Speech Recognition*. PhD thesis, Univ. of Illinois, Urbana-Campaign, 1984.

[248] M. Petkovica, V. Mihajlovic, W. Jonker, and S. Djordjevic-Kajan. Multi-modal extraction of highlights from TV formula 1 programs. In *ICME*, 2002.

[249] T.V. Pham and M. Worring. Face detection methods: A critical evaluation. Technical Report 2000-11, Intelligent Sensory Information Systems, University of Amsterdam, 2000.

[250] R. Pieraccini, K. Dayanidhi, J. Bloom, J.-G. Dahan, M. Phillips, B. Goodman, and K.V. Prasad. Multimodal conversational systems for automobiles. *Communications of the ACM*, 47(1):47–49, January 2004.

[251] G. Pingali, A. Opalach, and Y. Jean. Ball tracking and virtual replays for innovative tennis broadcasts. In *ICPR00*, 2000.

[252] G. Potamianos, H.P. Graf, and E. Cosatto. An image transform approach for HMM based automatic lipreading. In *Proc. Int'l Conf. on Image Proc.*, volume III, pages 173–77, 1998.

[253] G. Potamianos and C. Neti. Automatic speechreading of impaired speech. In *Int'l Conf. on Auditory-Visual Speech Processing*, pages 177–182, 2001.

[254] G. Potamianos and C. Neti. Improved ROI and within frame discriminant features for lipreading. In *Proc. Int'l Conf. on Image Proc.*, volume III, pages 250–253, 2001.

[255] G. Potamianos and C. Neti. Audio-visual speech recognition in challenging environments. In *Proc. European Conf. Speech Technology*, pages 1293–1296, 2003.

[256] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior. Recent advances in the automatic recognition of audio-visual speech. *Proceedings of the IEEE*, 91(9):1306–1326, September 2003.

[257] G. Potamianos, C. Neti, J. Huang, J. H. Connell, S. Chu, V. Libal, E. Marcheret, N. Haas, and J. Jiang. Towards practical deployment of audio-visual speech recognition. In *Proc. 2004 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, volume III, pages 777–780, Montreal, Canada, May 2004.

[258] G. Potamianos, C. Neti, J. Luettin, and I. Matthews. Audio-visual automatic speech recognition: An overview. In G. Bailly, E. Vatikiotis-Bateson, and P. Perrier, editors, *Issues in Visual and Audio-Visual Speech Processing*, chapter 10. MIT Press, 2004.

[259] W. Qi, L. Gu, H. Jiang, X.R. Chen, and H.J. Zhang. Integrating visual, audio, and text analysis for news video. In *Proceedings of the 7th IEEE International Conference on Image Processing, Vancouver, Canada*, volume 3, pages 520–523, September 2000.

[260] L. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, NJ, USA, 1993.

[261] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[262] A. Rector, P. Zanstra, W. Solomon, J. Rogers, R. Baud, W. Ceusters, W. Classen, J. Kirby, J. Ridrigues, A. Mori, E. Haring, and J. Wagner. Reconciling users needs and formal requirements: Issues in developing a re-usable ontology for medicine. *IEEE Transactions on Information Technology in BioMedicine*, 1999.

[263] H. Renman, N.Rea, and A. Kokaram. Content-based analysis for video from snooker broadcasts. *Computer Vision and Image understanding*, 92, 2003.

[264] S. Romdhani, P. Torr, B. Schölkopf, and A. Blake. Computationally efficient face detection. In *Proc. Int'l Conf. on Comp. Vision*, volume II, pages 695–700, 2001.

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

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

[267] Y. Rui, A. Gupta, and A. Acero. Automatically extracting highligths for TV baseball programs. In *ACM Multimedia*, 2003.

[268] Y. Rui, T. S. Huang, and S. Chang. Image retrieval: Current techniques, promising directions, and open issues. *Journal of Visual Communication and Image Representation*, 10(4):39–62, 1999.

[269] C. Saraceno and R. Leonardi. Audio as support to scene change detection and characterization of video sequences. In *International Conference on Acoustics Speech and Signal Processing*, volume 4, pages 2597–2600, 1997.

[270] C. Saraceno and R. Leonardi. Identification of story units in audio-visual sequences by joint audio and video processing. In *Proceedings of the 5th IEEE International Conference on Image Processing, Chicago, Illinois, USA*, volume 1, pages 363–367, October 1998.

[271] S. Satoh, Y. Nakamura, and T. Kanade. Name-It: Naming and detecting faces in news videos. *IEEE Multimedia*, 6(1):22–35, 1999.

[272] H. Schneiderman and T. Kanade. A statistical method for 3d object detection applied to faces and cars. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume I, pages 746–751, 2000.

[273] H. Schneiderman and T. Kanade. A statistical method for 3D object detection applied to faces and cars. In *IEEE Computer Vision and Pattern Recognition*, Hilton Head, USA, 2000.

[274] H.-P. Schnurr, M. Erdmann, A. Maedche, and S. Staab. From manual to semi-automatic semantic annotation: About ontology-based text annotation tools. In *COLING 2000 Workshop on Semantic Annotation and Intelligent Content*, 2000.

[275] A.T.G Schreiber, B. Dubbeldam, J. Wielemaker, and B. Wielinga. Ontology-based photo annotation. *IEEE Intelligent Systems*, 213:66–74, 2001.

[276] S. Sclaroff and J. Isidoro. Active blobs. In *Proc. Int'l Conf. on Comp. Vision*, pages 1146–1153, 1998.

[277] A. Shaikh, S. Juth, A. Medl, I. Marsic, C. Kulikowski, and J. Flanagan. An architecture for multimodal information fusion. Proceedings of the Workshop on Perceptual User Interfaces (PUI 97), 91-93. Banff, Canada, 1997.

[278] L. Shapiro and G. Stockman. *A new Computer Vision Textbook*. Prentice-Hall, 2001.

[279] H. Sidenbladh, M.J. Black, and D.J. Fleet. Stochastic tracking of 3d human figures using 2d image motion. In *European Conf. on Computer Vision*, 2000.

[280] L. Sigal, S. Sclaroff, and V. Athitsos. Estimation and prediction of evolving color distributions for skin segmentation under varying illumination. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2000.

[281] L. C. De Silva and P. C. Ng. Bimodal emotion recognition. In *Proc. IEEE Face and Gesture Recognition Workshop*, pages 332–335, 2000.

[282] J. Siroux, M. Guyomard, F. Multon, and C. Remondeau. Oral and gestural activities of the users in the georal system. In Proc. of the Intl. Conf. on Cooperative Multimodal Communication, vol. 2, 1995, 287–298, 1995.

[283] J.R. Smith, S. Basu, C.-Y. Lin, M.R. Naphade, and B. Tseng. Integrating features, models and semantics for content-based retrieval. In *Proceedings of the international workshop on MultiMedia Content-Based Indexing and Retrieval (MMCBIR'01)*, pages 95–98, September 2001.

[284] C.G.M. Snoek and M. Worring. Multimedia event based video indexing using time intervals. *IEEE Transactions on Multimedia*, 2004. Accepted for publication.

[285] C.G.M. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 2005 (in press).

[286] D. Sodoyer, J. -L. Schwartz, L. Girin, J. Klinkisch, and C. Jutten. Separation of audio-vsual speech sources: A new approach exploiting the audio-visual coherence of speech stimuli. *EURASIP Journal of Applied Signal Processing*, (11):1165–1173, November 2002.

[287] M. Song, C. Chen, and M. You. Audio-visual based emotion recognition using tripled hidden markov model. In *Proc. 2004 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, volume V, pages 877–880, Montreal, Canada, May 2004.

[288] G.B. Stamou and S.G. Tzafestas. Resolution of composite fuzzy relational equations based on archimedean triangular norms. 10(5):395–407, 2001.

[289] D.G. Stork and M.E. Hennecke, editors. *Speechreading by Humans and Machines*. Springer, Berlin, Germany, 1996.

[290] A. Subramanya, S. Gurbuz, E. Patterson, and J.N. Gowdy. Audiovisual speech integration using coupled hidden markov models for continuous speech recognition. In *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing*, 2003.

[291] I. H. Suh and T.W. Kim. Fuzzy membership function-based neural networks with applications to the visual servoing of robot manipulators. *IEEE Trans. Fuzzy Systems*, 2:203–220, 1994.

[292] H. Sundaram and S.-F. Chang. Video scene segmentation using video and audio features. In *International Conference on Multimedia and Expo*, pages 1145 – 1148, 2000.

[293] H. Sundaram and S.F. Chang. Determining computable scenes in films and their structures unsing audio-visual memory models. In *Proceedings of the ACM International Conference on Multimedia, Los Angeles, California, USA*, pages 95–104, November 2000.

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

[295] Y. Sure, S. Staab, J. Angele, D. Wenke, and A. Maedche. Ontroedit: Guiding ontology development by methodology and inferencing. 2002.

[296] M. Szummer and R.W. Picard. Indoor-outdoor image classification. In *IEEE International Workshop on Content-based Access of Image and Video Databases, in conjunction with ICCV'98*, Bombay, India, 1998.

[297] H. Takagi, N. Suzuki, T. Koda, and Y. Kojima. Neural networks designed on approximate reasoning architecture and their applications. *IEEE Trans. Neural Networks*, 3:752–760, 1992.

[298] M. Tekalp. *Digital Video Processing, Prentice Hall*. 1995.

[299] G. Tsechpenakis, G. Akrivas, G. Andreou, G. Stamou, and S.D. Kollias. Knowledge-assisted video analysis and object detection. In *European Symposium on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems*, 2002.

[300] S. Tsekeridou and I. Pitas. Content-based video parsing and indexing based on audio-visual interaction. *IEEE Trans. Circuits and Systems for Video Technology*, 11(4):522–535, 2001.

[301] M. Turunen. *Jaspis - A Spoken Dialogue Architecture and its Applications*. PhD thesis, University of Tampere, Department of Information Studies, 2004.

[302] M. Turunen and J. Hakulinen. Jaspis2 - an architecture for supporting distributed spoken dialogues. In Proceedings of Eurospeech 2003: 1913-1916, 2003.

[303] V. Tzouvaras, G. Stamou, and S. Kollias. Knowledge refinement using fuzzy compositional neural networks. In *Proceedings of International Conference of Artificial Neural Networks,* 2003.

[304] S. Uchihashi, J. Foote, A. Girgensohn, and J. Boreczky. Video manga: Generating semantically meaningful video summaries. In *ACM Multimedia*, 1999.

[305] A. Vailaya, A. Jain, and H.J. Zhang. On image classification: City vs. landscape. *IEEE Workshop on Content-Based Access of Image and Video Libraries*, 1998.

[306] A. Vailaya and A.K. Jain. Detecting sky and vegetation in outdoor images. In *Proceedings of SPIE: Storage and Retrieval for Image and Video Databases VIII*, volume 3972, San Jose, USA, 2000.

[307] A. Vailaya, A.K. Jain, and H.-J. Zhang. On image classification: City images vs. landscapes. *Pattern Recognition*, 31:1921–1936, 1998.

[308] A. Van Gelder, K. Ross, and Schlipf. The well-founded semantics for general logic programs. *Journal of the ACM*, 1991.

[309] V. Vapnik. *Statistical Learning Theory, Second Edition, Prentice Hall*. 1999.

[310] J. Vermaak, A. Blake, M. Gangnet, and P. Perez. Sequential monte carlo fusion of sound and vision for speaker tracking. In *Proc. Int'l Conf. on Comp. Vision*, pages 741–746, 2001.

[311] P. Viola and M.J. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume I, pages 511–518, 2001.

[312] W. Wahlster, Reithinger N., and A. Blocher. Smartkom: Multimodal communication with a life-like character. In Proceedings of Eurospeech 2001: 1547-1550, 2001.

[313] M. Wallace, G. Akrivas, P. Mylonas, Y. Avrithis, and S. Kollias. Using context and fuzzy relations to interpret multimedia content. In *CBMI03*, 2003.

[314] Y. Wang, Z.Liu, and J.C. Huang. Multimedia content analysis using both audio and visual cues. *IEEE Signal Processing Magazine*, pages 12–36, November 2000.

[315] L. Wenyin, S. Dumais, Y. Sun, H. Zhang, M. Czerwinski, and B. Field. Semi-automatic image annotation. In *INTERACT2001, 8th IFIP TC.13 Conference on Human-Computer Interaction*, Tokyo, Japan July 9-13, 2001.

[316] B. Whitman and P. Smaragdis. Combining musical and cultural features for intelligent style detection. In *Proceedings of the 3rd International Conference on Music Information Retrieval*, Paris, France, October 13-17 2002.

[317] J. Wielemaker, A.Th. Schreiber, B. Dubbeldam, and B.J. Wielinga. Ontology-based photo annotation. *IEEE Intelligent Systems*, 2001.

[318] E. Wold, T. Blum, D. Keislar, and J. Wheaton. Content-based classification, search, and retrieval of audio. *IEEE Multimedia*, 3(3):27–36, 1996.

[319] P. Wu, B.S. Manjunath, S.D. Newsam, and H.D. Shin. A texture descriptor for image retrieval and browsing. In *IEEE Workshop on Content-Based Access of Image and Video Libraries*, 1999.

[320] L. Xie, S. Chang, A. Divakaran, and H. Sun. Learning hierarchical hidden Markov models for video structure discovery - tech. rep. 2002-006. Technical report, ADVENT Group, Columbia Univ., 2002.

[321] L. Xie, P. Xu, S.F. Chang, A. Divakaran, and H. Sun. Structure analysis of soccer video with domain knowledge and hmm. *Pattern Recognition Letters*, 25:767–775, 2004.

[322] Z. Xiong, R. Radhakrishnan, and A. Divakaran. Generation of sports highlights using motion activity in combination with a common audio feature extraction framework. In *Proceedings of the 10th IEEE International Conference on Image Processing, Barcelone, Espagne*, volume 1, pages 1–5, September 2003.

[323] M. H. Yang, D. J. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(1):34–58, January 2002.

[324] A. Yoshitaka and T. Ichikawa. A survey on content-based retrieval for multimedia databases. *IEEE Transactions on Knowledge and Data Engineering*, 11(1):81–93, 1999.

[325] A. Yoshitaka and M. Miyake. Scene detection by audio-visual features. In *International Conference on Multimedia and Expo*, pages 49–52, 2001.

[326] Y. Yoshitomi, N. Miyawaki, S. Tomita, and T. Kitazoe. Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face. In *Proc. ROMAN*, pages 178–183, 2000.

[327] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. The HTK book (for HTK version 3.2). Technical report, Cambridge University Engineering Department, December 2002.

[328] A.L. Yuille, P.W. Hallinan, and D.S. Cohen. Feature extraction from faces using deformable templates. *Int'l Journal of Comp. Vision*, 8(2):99–111, 1992.

[329] H.-J. Zhang, A. Kankanhalli, and S.W. Smoliar. Automatic partitioning of full-motion video. *Multimedia Systems*, 1(1):10–28, 1993.

[330] H.-J. Zhang and Z. Su. Improving CBIR by semantic propagation and cross-mode query expansion. In *Proceedings of the international workshop on MultiMedia Content-Based Indexing and Retrieval (MMCBIR'01)*, pages 83–86, September 2001.

[331] T. Zhang and C.-C. Jay Kuo. Hierarchical classification of audio data for archiving and retrieving. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 6, pages 3001–3004, Phoenix, USA, 1999.

[332] D. Zhong and S. Chang. Structure analysis of sports video using domain models. In *ICME*, 2001.

[333] X.S. Zhou and T.S. Huang. Unifying keywords and visual contents in image retrieval. *IEEE Multimedia*, 9(2):23–33, 2002.

[334] Y. Zhu and D. Zhou. Scene change detection based on audio and video content analysis. In *International Conference on Computational Intelligence and Multimedia Applications*, pages 229–234, Sep 2003.

[335] D. N. Zotkin, R. Duraiswami, and L. S. Davis. Joint audio-visual tracking using particle filters. *EURASIP J. Applied Signal Processing*, 11:1154–1164, 2002.

[336] G. Zweig, J. Bilmes, T. Richardson, K. Filali, K. Livescu, P. Xu, K. Jackson, Y. Brandman, E. Sandness, E. Holtz, J. Torres, and B. Byrne. Structurally discriminative graphical models for automatic speech recognition: Final workshop 2001 report. Technical report, Center for Language and Speech Processing, The Johns Hopkins University, 2001.