MUSCLE Research Roadmaps

Version 1.2 — Edited by Eric Pauwels

eric.pauwels@cwi.nl

First draft January 2, 2006, – Current draft: April 12, 2006

Internal document: for Network members only

Contents

1 Introduction

The scientific work in the Network is typically urged on by two different driving forces: On the one hand there is the bottom-up, grass roots momentum created by individual researchers pushing their interests and trying to set up collaboration on a one-to-one basis. In addition however, there is also room for a second and complementary top-down approach: one in which ambitious high-level goals are set and research groups are challenged to contribute to the grand vision. Clearly, the Grand Challenges in the Integration WP fit squarely within the latter category. However, to further emphasize this top-down pulling-force, the Steering Committee has decided to endorse a number of Intermediate Challenges and to invite researchers (both within and outside the Network) to contribute to the corresponding Roadmaps. The difference between the Roadmaps and the more traditional Workplan (with its tasks and deliverables as detailed in the Description of Work) is twofold:

- Whereas the WP's Workplan specifies a number of contractual obligations (in terms of tasks and deliverables), the Roadmaps should be thought of as an invitation to collectively define and explore an ambitious vision. As a consequence, there is more room for curiosity-driven blue-sky research that is able to respond with appropriate alertness to emerging opportunities.
- In contradistinction to WPs, Roadmaps are not as strictly constrained by a very tight (12 or 18 month) time-schedule. Indeed, whereas for WP-deliverables, timeliness is of the essence, the priority in the case of Roadmaps is to indicate research directions that through their "pulling-power" will enthuse researchers to pool resources and to collectively make a difference. They can therefore span longer periods, even extending beyond the nominal end-date of the Network (thus ensuring lasting collaboration).

This document (which is meant to be frequently updated to reflect the evolving interests of Consortium) is an attempt to identify major research trends and directions within the Network and outline their embedding in the wider context of European ICT research and development. To that end, we have consulted the **Strategic Research Agenda** issued by the *European Technology Platform on Networked and Electronic Media* (NEM cf. www.nem-initiative.org). The scenarios discussed in this extensive report, reinforce our conviction that audio-visual content analysis and the subsequent extraction of searchable semantic meta-data will remain high on the priority list for the next 10 to 15 years. The following quotes (from version 3.0, 31 Jan 2006) are particularly relevant

- (p.6) The users will be faced with an enormous amount of information accessible in various places. Without assistance it will become impossible to retrieve the relevant content from media libraries, which stresses the importance of **metadata** and **semantic search engine technologies.**
- (p.7) Technologies that will enable the NEM vision to happen include . . . technologies aimed at retrieving information (search engines, metadata, semantic web).
- (p. 17) Topics to be addressed: semantic search algorithms should be integrated in audio-visual search applications.

- (p.27) State of the Art in Content Indexation: the lack of proper indexing and retrieval systems is rendering useless most of the huge collections of digital multimedia content that is available. Only a few indexing techniques can be considered as mature and effectively deployed: audio and video segmentation, image and music identification, detection of recurrent shots and speech-to-text transcription in very favourable conditions.

- (p.54) Enabling Technologies: Media libraries must be searchable. Even in its most simple form (searching text-based meta-data) this puts huge demands on RTD. Indexing digital content in an efficient and standardized manner manually (e.g. speech communication), semi-automatically (e.g. by detecting scene cuts or highlights and then require manual input) or even automatically (e.g. by content analysis and synchronization with electronic content guides) is not solved today, thus NEM should target solutions in these technological areas.

The roadmaps outlined below try and outline a MUSCLE response to the challenges identified in the NEM-document, taking into account the Network's resources and time-frame.

2 Creating a Showcase for Content Analysis and Audio-Visual Search

2.1 Objectives and Motivation

The goal of this is to bring together the wide range of semantic analysis and annotation capabilities that are present within MUSCLE to show on a single show-case demo. We would compile a video sequence, based on contributions from each partner, by combining TV short recordings of different kinds (say news reports, music clips, etc.) and home videos. These would be integrated and shared by all team members, who are then invited to perform whatever semantic extraction and analysis (single- or multi-modal) they can apply to this video, such as all kind of low-level feature extraction, face detection, moving objects, fire-and-smoke detection, logo detection, music genre analysis, speech recognition, text detection and recognition, etc.

Notice that the aim in this initiative is quite different from the benchmarking initiative pursued in the Benchmarking WP. Whereas in a typical benchmarking scenario different groups are pitted against each other to determine a relative ranking with respect to some predefined criterion, the ambition in this initiative is to act as an *inspiration pump*, facilitating the emergence of new and serendipitous collaborations and breakthroughs. The hope is that the sheer act of applying a varied set of techniques to a common corpus will suggest new methodologies that will further the field.

To achieve this vision, partners are allowed to use all kind of algorithms, additional external information, as well as additional data they may have and use within their own labs to enhance the information extracted from the video.

The results will be:

- A video to be used for analysis, serving as a common, joint showcase on what the consortium can achieve by applying all our techniques from different perspectives;

- A set of different features for audio, video, images, and text extracted from this video that in a later stage can be used by other partners within MUSCLE;

- A comprehensive set of annotations performed on this video using the variety of Machine Learning techniques employed by MUSCLE members;

The result will therefore provide a kind of practical state-of-the-art showcase of the capabilities and competences on feature extraction and semantic annotation within MUSCLE.

2.2 Tentative Workplan and Timeframe

- **Months 1-2, Data acquisition:** Discussion of some restrictions concerning the video to be created to ensure a certain heterogeneity of content, duration, etc.. Subsequently, all partners will record a short video sequence, which will then be joined to form a single video for analysis. This video will be made available to eTeam partners in several forms (eg. video formats, audio only, set of images).
- **Months 3-7, Analysis by individual partners:** Partners will analyze the data using their standard techniques, extracting the set of features they have implemented, and sharing these with the other partners. Results will be presented and discussed during a meeting at the end of this period, and a joint showcase wil be developed.
- **Months 8-12, Creating synergy between contributing partners:** The various feature sets extracted will be combined and used to extract higher-level and more robust semantic descriptors. In order to ensure know-how transfer as well as foster collaboration, a set of pairwise exchanges (1-2 weeks) is foreseen in this period. Results stemming from the combination of different feature sets will be published at international conferences, as well as demonstrated as a protoype showcase.

2.3 Envisaged Contributions by Partners

All partners: will contribute in recording and building a joint video from heterogenous video sources (preferably TV recordings from news broadcasts, music video clips, commercials) in order to have a wide range of different characteristics, objects and events present in the video. Subsequently, the video will be segmented in the audio and video stream, as well as a set of keyframes extracted. These will be used separately or combined to allow partners to apply their respective image, audio and video indexing tools to extract features.

Specifically, we are looking for approx. 10 minutes of each of the following:

- Sports TV recordings (single sport as well as sport news)
- News
- Music video clips
- Commercials
- Tele-Shopping
- Soap operas
- Movies (Color, B/W, different genres,...)
- Home Movies (birthdays, vacations, ...)

Recordings should be taken, if possible, from national TV stations rather than international ones to get different flavours of broadcastings, different languages, etc. Recordings should be MPEG-1 encoded. The only special item in the list above are News, where it would be interesting to record them on the same day, assuming that this would reveal some very interesting similarities in terms of names, places, faces and figures shown - so the task would be to agree on a potential day for recording the 10 minutes of news.

TUWIEN-IFS will focus on the analysis of the audio features from the resulting video stream. Specifically, we will extract a range of features comprising

- Rhythm Patterns
- Statistical Spectrum Descriptors
- Roughness Features
- Set of standard features like energy, zero-crossings, etc.

These will then be used to analyze the audio stream both statistically as well as with machine learning techniques (clustering, classification) to identify

- speech/audio/speech+audio separation, classification into male/female speech
- classification into different music genres
- cluster analysis of the different audio characteristics present
- integration and comparison with other audio sources, standard reference sets from e.g. the MIREX competition

University of Amsterdam will focus on the automatic detection of 30-50 high-level semantic concepts in the video stream, such as table, car, crowd, and mountain. Specifically we will analyse the visual modality using our wiccest features, which combine color invariance with natural image statistics. In case, other participants make a speech transcript available, we will perform a multimodal analysis that will improve the performance of the detected concepts further. In addition, we can also provide basic video analysis techniques such as:

- camera motion results
- shot-detection results

Cambridge University We propose to perform the unified extraction of different types of low-level multiscale low-level features which cover the major visual types of visual saliencies: blobs, corners, edges, ridges. Detected features can be visualised as an overlay on the video. These features can serve as a strong basis in the future for various higher-level tasks such as object recognition, scene classification, etc. Tradeoff between detection accuracy and speed will be considered depending on the real-time constraint of the application.

TAU-Speech will contribute video recordings from Israeli TV, as well as participate in analysing audio.

2.4 Relevant E-teams

- E-team on Content Analysis Showcase (Andi Rauber, Nicu Sebe)

3 Visual Saliency, Focus of Attention and Habituation

3.1 Objectives and Motivation

In spite of the massively parallel character of computations performed by the brain, it is clear that biological systems employ a selection processing strategy for managing the enormous amount of information in the visual inflow. This selection of relevant information is known as **visual attention.** Two related concepts are **saliency** and **habituation**, the former refers to the ability of a visual element to capture the attention, while the latter alludes to suppression of a demand for attention once the originator of this demand has become familiar or repetitive.

It is clear that accurate computational models for saliency and attention would be enormously valuable in the following scenario's (non-exhaustive list!):

- **Foreground-background segmentation** as salient items tend to belong to the foreground;
- **Compression:** Making compression rates dependent on the saliency score of each image region would allow for high compression rates without noticeable loss of quality. Similar considerations hold for progressive transmission.
- **Similarity Scoring and Retrieval:** Similarity scoring between images (and subsequent retrieval from databases) depends to a large extent on the matching of salient items in the images. Being able to accurately predict what will be perceived as salient will vastly improve the retrieval accuracy.
- **Visual inspection** (for both industrial or military applications) hinges on the capability of the automated inspection system to spot unusual or unexpected occurrences;
- **Perceptual organisation** of images crucially depends on our (as yet unsurpassed) ability to extract non-accidental spatial or temporal alignments (e.g. continuation, symmetry, proximity, size, etc.) which impart additional significance to specific image elements.

3.2 Planned Activities and Contributions

For all of the above reasons, MUSCLE researchers are convinced that the problem of saliency is sufficiently generic and important to merit special attention (. . . pun actually intended). A dedicated E-team has been created and number of different approaches will be explored.

- **Probabilistic Probing:** A powerful, purely data-driven (i.e. bottom-up) mechanism to extract salient features is furnished by *probabilistic probing* (cf. Fred Stentiford, UCL). In this set-up an image is "probed" with randomly generated local filters (so-called forks) and the statistics thus generated can be used to automatically single out various saliencies (such as edges, symmetries, shape-similarities and the like). The goal is to further explore further applications of the this appealingly simple, yet powerful, paradigm.

- **Geometrical configuration of local interest points:** Whereas a single interest point usually has little saliency value, a local configuration of such points (and its invariant description) might be very powerful. Applications hereof will be studied by INRIA Imedia.
- **Spatiotemporal Visual Attention:** In video-streams visual attention is also influenced by the temporal aspect of events. This requires researchers to extend saliency filters to operate on spatiotemporal datacubes (contributions by ICCS-NTUA).
- **Habituation:** One should not lose track of the fact that the robust performance of attention focusing is only partially due to saliency detection. Active suppression of irrelevant information is also very much part of the equation. *Habituation* refers to capacity of biological organisms to stop paying attention to repetitive or familiar stimuli. he characterisation of dynamic textures (Sztaki, Bilkent, CWI) is an attempt to accurately detect (and then – whenever appropriate – suppress) background activity that would generate spurious saliency alarms (e.g. when trying to detect walking pedestrians, one does not want attention to be diverted by the erratic waving of a nearby tree).

3.3 Relevant E-teams

- E-team on Saliency (Nozha Boujemaa)
- E-team on Dynamic Texture Analysis and Detection (Enis Cetin)

4 Detecting and Recognizing Humans and Human Behaviour

4.1 Objectives and Motivation

A large percentage of the material in audio-visual databases revolves around humans and their various activities (holiday and home videos are a case in point). Therefore, automatic recognition of humans and human behaviour would go a long way towards taming the information overload that confronts us when accessing personal image and video collections. Different MUSCLE groups will try and set up a coordinated action. Notice that this also ties in closely with the Content Analysis Showcase.

4.2 Planned Activities and Contributions

- **Body detection:** The use of background learning techniques for object detection in single and multi-camera environments. This problem will be addressed both for situations in which there is context knowledge (e.g. smart rooms) or unconstrained environments (e.g. street or outdoor scenes). (UPC, Tszaki)
- **Body analysis:** The task of analyzing basic actions and gestures from a group of people inside a multiple camera environment has been studied by first generating a 3D virtual reconstruction and then processing this data in a pattern analysis framework. Foreground regions (moving objects) are extracted from images obtained by the N cameras. Two approaches are being studied towards human action detection, the first one is to develop tools for detection and tracking of body parts in every image and then extracting 3D information from them. The second approach consists in extracting a 3D voxel reconstruction of the space and detecting actions over it. The current research is devoted to the fitting of a

hierarchical human body model to the data in 3D (voxels) and to perform an accurate tracking over it. From this data we expect to train a system in order the classify actions from the data obtained by the model fit to the data.

- **Person Tracking in Crowded Scenes**
- **Adaptive detection:** TU Graz's expertise with respect to body detection is people detection based on an On-line Adaboost method, which is embedded in a learning framework that can train a Person detector without hand labelling. Appearance based tracking of people based on an on-line classifier. Both methods are based on integral orientation histogram features and are able to run in real-time on a standard PC. We can contribute to the team our expertise, various sequences we use for testing our methods (some of them with ground truth). We are interested in combining our methods with other techniques to improve the robustness and applicability.
- **Emotion recognition** New modalities for human-computer interaction such as facial expressions, voice, and gestures are emerging. A very important aspect previously ignored is the emotional aspect. In addition to recognizing what is said and who is saying it, it would also be very helpful to figure out how things are said, i.e. the emotional/affective channel of information. The goal is to develop visual and multimodal (e.g. combining vision and audio) features that allow recognition of primal emotional states such as laughter, surprise, fear and the like, in audio-visual material.

4.3 Relevant E-teams

- E-team on Body detection, tracking and analysis