



National Network Research Report

Deliverable D4.1

(No. 1 of 4)

September 2008

Document Description	
Identifier	D4.1 (1 of 4)
Title	National network research report
Authors	QMUL: Naeem Ramzan TUD: Alan Hanjalic, Maarten Clements, Bart Kroon TUB: Kai Clüver, Sebastian Schmiedeke, Pascal Kelm, Nicolas Neubauer, Matthias Zappe, Engin Kurutepe EPFL: Jong-Seok Lee U. Geneva: Stéphane Marchand-Maillet
WP Code	JPRA.1
Status	final version
Date	15 September 2008

Table of Contents

Table of Contents.....	2
1 Introduction.....	2
2 Fundamental Research of Four Special Interest Groups.....	2
2.1 Content Distribution.....	2
2.1.1 NIRICT.....	2
2.1.2 IM2.....	2
2.1.3 MMKM.....	2
2.1.4 HC3.....	2
2.1.5 Cross-evaluation.....	2
2.2 Processing.....	2
2.2.1 NIRICT.....	2
2.2.2 IM2.....	2
2.2.3 MMKM.....	2
2.2.4 HC3.....	2
2.2.5 Cross-evaluation.....	2
2.3 Indexing.....	2
2.3.1 NIRICT.....	2
2.3.2 IM2.....	2
2.3.3 MMKM.....	2
2.3.4 HC3.....	2
2.3.5 Cross-evaluation.....	2
2.4 Social Content Retrieval.....	2
2.4.1 NIRICT.....	2
2.4.2 IM2.....	2
2.4.3 MMKM.....	2
2.4.4 HC3.....	2
2.4.5 Cross-evaluation.....	2
3 Conclusion.....	2
References.....	2

1 Introduction

PetaMedia is built on a large volume of activities which encompasses the existing fundamental research activities within the national research efforts for social/peer-to-peer (SP2P) networking, multimedia content analysis and other related research fields. These activities are organized and conducted within the existing and future (inter)national projects involving the research groups from NIRICT (Netherlands), MMKM (UK), IM2 (Switzerland) and HC3 (Germany) networks.

This report aims at describing the fundamental research in these four national research clusters and performing cross-evaluation between them by classifying the efforts of different national clusters, identifying possible redundancies and needs for NoE internal competition and benchmarking. This document will serve as the main input into the yearly technology integration meetings.

In the following section, we describe the researches in the four national networks for four specific special interest groups (SIGs): content distribution, processing, indexing, and social content retrieval. Finally, concluding remarks are given in Section 3.

2 Fundamental Research of Four Special Interest Groups

2.1 Content Distribution

2.1.1 NIRICT

2.1.1.1 Peer-to-peer network

Tribler¹, the TU Delft's peer-to-peer (P2P) network client, is an application that enables its users to find, and share content such as video, audio, pictures, and other types of content. A P2P network is different from a centralized service in the sense that all peers contribute their bandwidth to the network. Therefore, no central computer is required.

The most recent developments on Tribler are the real-time and video-on-demand (VOD) extensions on the P2P protocol, which make it possible to have a click-and-play VOD experience, as opposed to a start-download-and-wait experience that other P2P clients' offer. Also, Live Streaming is now supported. Several trials have been performed recently, in which we tested these extensions in a live network with thousands of users. The extensions have proved to be working. We have proposed Give-to-Get, a P2P VOD algorithm which discourages free-riding by letting peers favour uploading to other peers who have proven to be good uploaders. As a consequence, free-riders are only tolerated as long as there is spare capacity in the system. Our simulations show that even if 20% of the peers are free-riders, Give-to-Get continues to provide good performance to the well-behaving peers. In particular, they show that Give-to-Get performs very well for short videos, which dominate the current VOD traffic on the Internet.

Furthermore, we have investigated the use of Multiple Description Coding (MDC) as a means of adding error resilience and resilience to bandwidth variability. Video that is MDC encoded and transmitted over a P2P network via different peers is robust against packet loss, link outages and to peers leaving and joining the network.

Other ongoing work is the recommendation feature of Tribler which presents the user a list of recommendations based on a profile of his tasted matched to profiles of others.

2.1.2 IM2

2.1.2.1 Multimodal quality metrics for multimedia content abstraction

¹ <http://www.tribler.org>

Measurement of perceived quality plays a fundamental role in the context of multimedia services and applications. For example, quality assessment of multimedia data is critical for the performance evaluation and optimization of many signal processing algorithms (i.e. coding, enhancement) as well as for the in-service monitoring of multimedia applications like videoconference and content streaming. The alternatives to assess the quality of multimedia data are two: on one side, subjective tests can be performed in order to collect feedbacks from the end user; on the other side, objective metrics can be used to automatically predict the subjective assessment. The subjective test activity has the obvious drawback of being expensive and time-consuming. Furthermore, it cannot be applied for the real time monitoring of multimedia applications (e.g. videoconferencing scenario) and there is a lack of extensive standard guidelines tuned for the multimedia scenario. On the other hand, the objective quality assessment is a very open challenge. In this research, starting from the investigation of the existing approaches for the objective quality assessment of audio and visual contents, we focus on the “no-reference” scenario, where the original signal (reference) is not available and the quality assessment is based only on the analysis of the test signal. We aim at developing a new approach of no-reference objective audio-visual quality assessment which integrates high-level features of the human perception. An important part of this research concentrates on the understanding and the modelling of the multimodal perception of quality, in order to design a metric for the assessment of the more complex concept of Quality of Experience (QoE) in a multimedia context.

The algorithms of different state-of-the-art image quality metrics have been collected in the “full-reference” scenario and implemented (De Simone et al, 2008). In the scenario of full-reference quality evaluation, we also developed a new approach for the design of an objective quality metric for the assessment of color pictures. Our goal is to build a multi-channel metric based on the perceptual weighting of single-channel metrics. A psycho-visual experiment has thus been designed in order to define the values of the weighting factors for each color channel.

We also implemented a module for the no-reference quality assessment of video sequences in the videoconference scenario, based on frame-based and shot-based artifacts detection and measurement. Current work includes the pooling step of the measures of the different artifacts (blur, blocking, jerkiness and freezing artifacts).

We target the investigation of audiovisual quality assessment methodologies both in the subjective and the objective scenarios. In particular, we focus on the analysis of the interactions between perceived audio quality and perceived visual quality when considering multimedia contexts starting from the videoconference scenario and going up to more complex high quality multimedia experiences.

2.1.2.2 Multimedia collection modelling and exploration

In this research we create a base-level approach for the extraction of global structures on multimedia document collections. A key issue in this task is to highlight inherent collection structures as a preparation for search and browsing. From our base theoretical modelling of a data collection, we have derived two algorithms for dimension reduction preserving structures (e.g. clusters) within the collection. The first technique is based on conditioning classical dimension reduction techniques (e.g.

IsoMap) by density estimation in high-dimensional spaces. The second technique looks at characterising clusters by nearest-neighbour operations in the high dimensional space so as to map the data into a new representation (called the cluster space) where dimension reduction is shown to be structure-preserving.

2.1.3 MMKM

Current UK research in multimedia content distribution spans a variety of problems including multimedia data management, secure distribution of multimedia content, content adaptation, and scalability of multimedia contents.

2.1.3.1 Scalable video transmission

The research on scalable video transmission focuses on the following topics: scalable generation and coding of motion vectors for scalable video coding (Mrak et al, 2005), fast error protection schemes for transmission of embedded coded images over unreliable channels using the rate-distortion characteristics of the source bitstream and dynamic programming (Sprljan et al, 2005), joint source-channel coding for scalable video bitstream (Ramzan et al, 2007), and perceptually adaptive joint deranging-deblocking filtering for scalable video multicast over wireless networks (Wan et al, 2007).

2.1.3.2 Secure distribution of multimedia contents

This research deals with the secure distribution of contents with privacy and confidentiality by ensuring that the digital content is accessible only to authorised users. It includes: watermarking protection for multimedia contents (Damnjanovic et al, 2006); system architecture for privacy preserving video-based applications which uses image and video segmentation (Cavallaro, 2004); handling of duplicates and document overlaps for meta-search engines (Wu et al, 2004); spatial quality evaluation for reproduced sound (Rumsey, 2002); development of visual attention algorithms for region of interest coding that produce an “interest ordered” progressive bit-stream in JPEG2000 so that the regions highlighted by the algorithm are presented first in bit-stream (Bradley and Stentiford, 2003); supporting searching on small devices through hierarchical-query biased summaries (Sweeney and Crestani, 2004); representing and retrieving images from a collection of scientific articles formatted in XML using the text encapsulated in the logical structure of the articles (Kong and Lalmas, 2007).

Some further examples for digital preservation include the Digital Curation Centre, UK, PrestoSpace² (Preservation towards storage and access) in the area of digital preservation of media as well as other European projects such as DigitalPreservationEurope³, PLANETS⁴ and CASPAR⁵. It is notable that these projects all are engaged to further the state of the art in one or more of the dimensions of standards, automation, intelligent handling of data and creativity.

² <http://prestospace.org>

³ <http://www.digitalpreservationeurope.eu>

⁴ <http://www.planets-project.eu>

⁵ <http://www.casparpreserves.eu>

2.1.3.3 Resource management of multimedia contents

There has been a move to add self-management to multimedia servers. Also known as Autonomic systems (Huebscher and McCann, 2008), essentially the system is able to measure user context and system trends to decide how best to serve the multimedia artefact to the user. It does this through self-configuration, self-optimisation and self-healing (given that many of these systems use unreliable networks and servers that may disconnect). Early work by McCann et al (2000) on the EPSRC-funded Kendra project measured current network capacity and adapted the multimedia data stream encoding at runtime to avoid delivery failure or delay. Here a single stream may be dynamically encoded into many codecs in a given download to adapt to the change in bandwidth observed and predicted near future trends. More recently the Kendra work has been used in self-adaptive (world wide) CDN server sites and P2P multimedia serving which again monitors user trends and general environmental conditions to adapt the actual content as it is delivered within QoS bounds which may also involve world wide distributed cache hopping, whilst also aiming to cope with Flash Crowd events (Jawaheer and McCann, 2005). Both projects focus on news-like sites and the latter work specifically also aims to further use the system's self-knowledge to lower the distributed server's carbon footprint when serving the query.

This wide spread of research activity is also mirrored and supported by formal projects with UK participation and leadership such as the nationally funded GATE⁶ (General Architecture for Text Engineering) project, a framework architecture for text engineering; FP6 AXMEDIS⁷ (Automating Production of Cross Media Content for Multichannel Distribution) that aims to speed up and optimise content production and distribution for production-on-demand; FP6 NM2⁸ (New Millennium, New Media) led by BT that develops new media forms that take advantage of the unique characteristics of broadband networks; FP6 3DTV⁹ (Integrated Three-Dimensional Television — Capture, Transmission, and Display), a project to capture, transmit and display 3D television; and an EU e-content project LIRICS¹⁰ (Linguistic Infrastructure for Interoperable Resources and Systems) that provides standards for language technology to facilitate the exchange and reuse of multilingual language resources.

2.1.4 HC3

2.1.4.1 Sprite coding for video

In spite of recent progress in the development of hybrid block-based video codecs, it has been shown that for low-bitrate scenarios there is still a coding gain applying object-based techniques. We have developed a sprite-based codec, based on latest H.264. Moreover, we generate multiple sprites based on physical camera parameter estimation that overcome the main drawbacks of sprite coding techniques.

⁶ <http://gate.ac.uk>

⁷ <http://www.axmedis.org>

⁸ <http://www.ist-nm2.org>

⁹ <http://www.3dtv-research.org>

¹⁰ <http://lirics.loria.fr>

Experimental results show that this coding approach significantly outperforms latest H.264 extensions applying hierarchical B pictures (Kunter et al, 2007). This approach holds promise for scalable coding techniques, evaluation of which poses an interesting research question.

2.1.4.2 Multiple-description coding of speech

Transmission of speech or audio signals using two or more descriptions can considerably increase the robustness to channel failures. Combining a layered speech codec with forward error correction codes has been shown to yield a flexible framework for more than two descriptions which allows robust encoding with graceful quality degradation under transmission failures (Clüver et al, 2007). Further work will concentrate on optimisation of the encoder for different channel conditions and on the development of suitable layered speech and audio codecs.

2.1.5 Cross-evaluation

The four national networks have carried out researches for distinct aspects of the content distribution issue: In the NIRICT network, a platform for the SP2P network, Tribler, has been developed and being improved. The research of the MMKM network deals with scalable video transmission, secure content distribution and resource management. In the HC3 network, work on coding of video and audio has been done. The research of the IM2 network deals with rather higher level of issues such as modelling and exploration of multimedia collection and measurement of quality metrics of multimedia materials.

Working on these different dimensions of content distribution will be a valuable contribution to the goal of PetaMedia if successful convergence of the researches is made. Thus, further consideration of compatibility, interoperability and appropriateness of the developed technologies within SP2P network environments would be required.

2.2 Processing

2.2.1 NIRICT

The activities in this direction concentrated in the past years on the development of unsupervised mechanisms for discovering semantically relevant information in audiovisual signals. The rationale for pursuing this research direction is the need for alternatives to traditional semantic inference approaches relying on statistical machine learning performed offline and using large training data sets. While these traditional approaches proved effective in many domain-specific application contexts, other important applications emerge where these approaches either do not provide satisfactory results or are not applicable at all. A good example is a scenario involving a large-scale network of smart surveillance cameras, each of which covers a different scene and is expected to automatically react to any suspicious objects appearance or behavior related to that particular scene. Another representative scenario relates to automated content-based management and retrieval in large and dynamic audiovisual collections characterized by content diversity that is impossible to be represented by a training set being good enough for producing meaningful results of classification-based indexing. The paragraphs below highlight the main activities related to content processing.

2.2.1.1 Unsupervised detection of dominant objects

In the development of our unsupervised approach to smart surveillance, we focused on the problem of detecting relevant objects in the observed scene. A common approach to building object detectors consists of annotating large data sets and using them to train a detector. However, due to inevitable limitations of a typical training data set, this supervised approach is unsuitable for building a generic surveillance system applicable to a wide variety of scenes and camera setups. To make a step towards a more generic object detection solution, we developed an unsupervised method capable of learning and detecting the dominant object class in a general dynamic scene observed by a static camera (Celik et al, 2008).

2.2.1.2 Unsupervised content discovery from composite audio

In many applications and scenarios dealing with the audiovisual content of sports, broadcasts, movies, news, and radio programs, audio signals appearing therein contain not only speech and music, but also various audio effects, such as cheering and applause. For instance, in a radio program, speech may be frequently interrupted by music or sound effects, while in an action movie a much more complex sound track containing speech, music, and various sounds of explosion, gun-shots, car-chasing, and screaming can be found. These sounds are typically not only temporally interleaved (*temporally composite*), but often also spectrally mixed (*spectrally composite*) when occurring simultaneously. Therefore, to be able to support multimedia information retrieval in a general case, and to make an audio indexing system less sensitive to unpredicted mixtures of different audio categories, we worked with a general composite audio signal as input, and concentrated on discovering meaningful, semantically coherent structure elements of an input audio signal that we

will further refer to as *semantic segments* or *audio scenes*. While the classical approach to audio segmentation infers audio scenes based on a direct analysis of low-level features, we considered an alternative approach that builds on the analogy to the text document analysis. This approach requires an intermediate analysis step resulting in mid-level semantic descriptors (Lu and Hanjalic, 2008). Our two-step segmentation approach was proved to be able to lead to a significant increase in segmentation robustness compared to the traditional approach. Once audio scenes are detected, we investigated the possibilities to automatically group them together into meaningful clusters to facilitate further steps in audio and multimedia content management. In the development of our clustering approach, we again rely on the same mid-level semantic descriptors that were applied in the segmentation step. This opens the possibility to deploy alternative clustering concepts, such as *co-clustering*, which also proved to result in a considerable increase in performance compared to the classical clustering methods (Cai et al, 2008).

2.2.2 IM2

2.2.2.1 Tagged media-aware multimodal content annotation

The content analysis alone cannot answer to the requirements for automatic and semantic handling of multimedia data. The extreme success of social networking inspired new strategies in solving multimedia content annotation problem. The social networks environment allows collecting information about multimedia content from users in a non-intrusive way. When users in social networks annotate and rate multimedia content, they provide objective cognitive information and subjective opinions at a level that multimedia content analysis cannot reach. Starting from the investigation of existing approaches for multimedia content access we focus on finding new models of interaction between automatic multimedia content analysis and social tagging.

Initial research on the state-of-the-art of multimedia annotation techniques from semantic point of view has been pursued. We have looked at approaches for the fusion of the results obtained by image analysis with textual tags. We started developing a novel system based on the bottom-up attention model, following (Walther and Koch, 2006). We have also studied approaches to introduce top-down information in the attention model. With the top-down information it is possible to adapt results of the attention system to particular task such as face detection.

2.2.2.2 Automated shallow dialogue analysis

In this research we develop and adapt to the multimodal dialogue domain, language processing tools for the detection of dialogue-specific semantic features. The tools remain at the shallow dialogue analysis level in order to preserve their robustness. The task includes the integration of several feature extractors into a system where each extractor can use the annotations previously produced by the other ones.

The work on dialogue act tagging has focussed on the direction of normalizing dialogue act tagsets. This work precedes the design of an automatic dialogue act

tagger. The work on the automatic identification of discourse markers in multiparty dialogues was completed (Popescu-Belis and Zufferey, 2007). Also, significant performance results were obtained for automatic thematic segmentation by using latent models and discriminative support vector machines, and were synthesized in a common framework for finding patterns in vocabulary use for thematic segmentation.

2.2.2.3 Combined use of vision and text for recognition

In this work we propose to combine the visual analysis with additional resources on the Web, using text search and in case location plays a role, location tags. We developed a novel system which automatically mines objects such as landmarks from community photo collections such as Flickr¹¹. We start by downloading images from that source for certain geographic areas. The downloaded photos are clustered into potentially interesting entities through a processing pipeline of several modalities, including visual, textual and spatial proximity. The resulting clusters are analyzed and are automatically classified into objects and events. Using mining techniques, we then find text labels for these clusters, which are used to again assign each cluster to a corresponding Wikipedia article in a fully unsupervised manner. We have tested this approach on several urban areas, mining over 200,000 photos.

We extended our work on mining entities from Flickr. The system was partly redesigned to allow better parallelisation of tasks and batch-processing using Condor batch queuing. This way we gathered images and meta-data in the order of about 1 million items and are currently processing those (Quack et al, 2008). Furthermore, we extended the system for annotation of images: the goal is to annotate detected objects in images with a bounding box and to make recognition and tagging of novel images faster. To that end we cross-match features between images and keep for each image only those, which receive a substantial amount of votes. This will be used for both bounding box annotation and more efficient indexing. For the indexing we started a novel visual vocabulary-based method.

2.2.2.4 Social network analysis for multimedia indexing

This research aims at mapping of the social interactions between individuals involved in the data into high-level information. We apply the social network analysis (SNA) and lexical analysis for recognition of roles in meetings. Meeting participants have been represented through the pattern of their relationships with others (extracted using Social Affiliation Networks) and their lexical choices (extracted from speech transcriptions through N-gram statistics). The two representations are mapped separately into roles using two different classifiers, and the classification outputs are then combined resulting into an improvement of the best individual classifier. The experiments have been performed over a collection of 120 meetings (around 45 hours) and 80% of the data time is correctly classified in terms of role.

SNA and social signal processing (SSP) techniques have been combined to detect agreement and disagreement in political debates. SNA has been used to discriminate between the moderator and the actual debate participants, while SSP has been used to group the participants according to their opinions. The experiments have been

¹¹ <http://flickr.com>

performed over a collection of 45 debates (around 30 hours of material). The turn-taking pattern (who talks with whom) has been automatically extracted and modelled with a Markov chain to detect pairs of individuals in mutual disagreement. The results show that in 65% of the debates, the participants are grouped correctly on the basis of their opinion (same group, same opinion).

2.2.2.5 Large-scale visual mining and retrieval

In this work, we use in full the context of mobile phones where location (GPS or cell tower ID) is a relevant feature and combine it with other visual and textual features for indexing and collection mining. We have extended our method for mining frequent feature configurations as representatives for object class. Here, our method was able to select the most relevant features for challenging object classes in common benchmark data (such as bikes, cars or giraffes). We have extended our system for object recognition from mobile phones. We extended the range of applications from meeting room slides to city guides, where we analyzed in the multi-modal spirit of the IM2 project combination of visual data with location information. The location information was obtained in form of either cell-tower IDs or GPS location from mobile phone clients, the search space in the database was restricted accordingly. For our recognition we have analyzed several methods to scale retrieval in local image features. Our results showed that the metric trees offered the best speed-up over linear search. With these results we decided to continue with metric trees for large datasets. We scaled metric trees by building them from random subsets of the data and combining multiple such trees to forests. With these measures, we could achieve results competitive with other methods on datasets with up to 1 million images and 300 million 64 dimensional SURF features at retrieval times of about 1 second.

2.2.2.6 Sociometric analysis of media

The goal of this research is to investigate the effectiveness of SNA, i.e. the domain studying the interaction between people sharing a common environment, as a tool for multimedia content analysis. We focus on the application of SNA in content abstraction problems. Two major subjects have been addressed so far: the first is role recognition and the second is content based segmentation of audio recordings.

The goal of role recognition is to assign each speaker in a multiparty recording a role belonging to a predefined set. In our experiments we applied role recognition to radio broadcast news. The role recognition approach applied in this research extracts a social network from the audio using a speaker clustering system, then uses the features of the social network to assign each speaker a role. The experiments have been performed over a corpus of around 45 hours of material and the results show that around 80 percent of the data time can be labelled correctly in terms of role.

The goal of content based segmentation is to split recordings into segments coherent from a semantic point of view. The experiments performed so far focused on the segmentation of broadcast news into stories, i.e. into segments where only a single and specific issue is presented. The approach applied in the experiments is based on the detection of social groups, i.e. of groups of speakers characterized by a high level of mutual interaction. The hypothesis is that social groups correspond to stories because people interacting with each other are more likely to talk about the same

topics than people that do not interact with each other. The experiments have been performed over two corpora: the first is a collection of talk-shows from Radio Suisse Romande (27 hours), the second is TRECVID 2003 (120 hours). The results show that respectively 80% and 68% of the times that the system identifies a topic boundary at a given time t , there is an actual topic boundary in a one minute long window centered in t . Considered that the single items of the two corpora are one hour and half a hour long, respectively, the performance can be considered satisfactory.

2.2.3 MMKM

Multimedia content processing is the constant area of research in the MMKM network. The principal research in MMKM focus on to obtain a better knowledge of the relevance of processing result obtained by: feature extraction, multimedia content segmentation, and content clustering.

2.2.3.1 Content processing

In this research, some of the work that has particular relevance and applicability to multimedia include the design of methods for extracting conceptual hierarchies from arbitrary domain-specific collections of text (Gillam et al, 2005); a critical survey by Calic et al (2005) of the methods for video representation targeting semantic analysis, outlining the importance of multimodal approach to multimedia knowledge management; the development of a system to formally annotate medical images captured to aid the diagnosis and management of breast cancer (Hu et al, 2003); ontology mapping by concept similarity using Web search engines to locate relevant material, and mapping the ontology of interest to other ontologies that exist on the semantic Web (Villa et al, 2004); the analysis and treatment of semantic relations between terms (Cai and van Rijsbergen, 2005) evaluations and comparison of knowledge representation models based on attributes (e.g., colour) with models based on values (e.g., red) using lexical clustering (Almuhareb and Poesio, 2004); semantic-based image classification and learning concepts for adding knowledge to the image descriptions (Djordjevic et al, 2007); and the creation of eminently browsable graphical networks between multimedia objects to represent semantic and content-based facets of their respective similarity (Heesch and Ruger, 2005).

The most salient UK national funded project in the area of general Knowledge Management Technologies is the EPSRC funded Advanced Knowledge Technologies (AKT)¹² project. Throughout the last decade there have been several major EU/EPSC projects that have correlated the efforts in individual ontology projects: On-To-Knowledge¹³ (Content-driven Knowledge-Management through Evolving Ontologies); OntoWeb¹⁴ (Ontology-based information exchange for knowledge management and electronic commerce); Knowledge Web¹⁵; WonderWeb¹⁶ (Ontology Infrastructure for the Semantic Web); ASPIC¹⁷ (Argumentation Service

¹² <http://www.aktors.org/akt/>

¹³ <http://www.ontoknowledge.org>

¹⁴ <http://www.ontoweb.org>

¹⁵ <http://wonderweb.semanticweb.org>

¹⁶ <http://knowledgeweb.semanticweb.org>

¹⁷ <http://www.argumentation.org>

Platform with Integrated Components); DIP (Data, Information, and Process Integration with Semantic Web Services); KB20 (KnowledgeBoard 2.0 — The European Knowledge and Capabilities Management Working Space); REVERSE¹⁸ (Reasoning on the Web with Rules and Semantics); X-Media¹⁹ (Large scale knowledge management across Media); NeOn²⁰ (Lifecycle Support for Networked Ontologies); SUPER²¹ (Semantics Utilised for Process Management within and between Enterprises); TAO²² (Transitioning Applications to Ontologies).

2.2.3.2 Annotation processing

There is a wide spread of knowledge-based methods involved in this area, some only focusing on improving the combination and selection of visual features (i.e., low-level knowledge entities such as colour and texture usage) for retrieval (Howarth and Ruger, 2005; Hilaire and Jose, 2007). On the other end of this spectrum Potter et al (2007) developed and analysed an integrated system that applies recent ideas and technologies from the fields of artificial intelligence and semantic web research to support sense- and decision-making at the tactical response level, and demonstrate it with reference to a hypothetical large-scale emergency scenario.

Others, e.g., Chakravarthy et al (2006), develop strategies and interfaces for cross-media knowledge creation and sharing that will make references between text and images in multimedia documents explicit with a view to increase the value of the document itself. In a similar spirit Gardoni et al (2005) developed a groupware tool that supports asynchronous work on (industrial) drawings and sketches, where the meaning of symbols and elements in the sketches is more often than not defined by synchronous agreement and as such ephemeral in nature.

In the area of e-learning, UK research has sparked a new direction of dynamic server-oriented approach that supports learning objectives with an automated allocation of services (as opposed to a manual composition of learning data) using current semantic web service technology and mappings between different learning metadata standards as well as ontological concepts for e-learning (Dietze et al, 2007).

2.2.4 HC3

2.2.4.1 Summarization

We are working on video summarization as well as audio summarization. On the audio side we are working on extracting the structure of the audio data. In the case of broadcast news, the content is pretty much structured around speakers, the anchor person introducing new stories, journalists being usually associated to specific themes, etc. Therefore, we are developing a speaker diarization framework that aims at finding “who spoke and when” within an audio data. This work will be further integrated in an audio summarization framework. On the video side, we have

¹⁸ <http://reverse.net>

¹⁹ <http://www.x-media-project.org>

²⁰ <http://www.neon-project.org>

²¹ <http://www.ip-super.org>

²² <http://www.tao-project.eu>

conducted extensive work on summarizing raw video data using various techniques (Dumont et al, 2008).

2.2.4.2 Camera motion characterization

We have developed a camera motion characterization system using motion vector information contained in the encoded video and using frame-by-frame optical flow computation (Haller et al, 2007). Various machine learning approaches are employed to learn features from the ground truth camera movement data and to later classify video shots into different camera motion classes such as pan, zoom or tilt. This shot based classification can be useful for shot boundary detection, scene classification and video summarization

2.2.4.3 Automated segmentation

We are working on a video object segmentation system, which generates temporally local background sprites for a video sequence and by blending of those sprites removes the foreground objects. The initial results of this approach show potential (Krutz et al, 2008). In addition to video object segmentation techniques, we are also working on automated objective evaluation of segmentation algorithms (Goldmann et al, 2008).

2.2.4.4 2D to 3D conversion

We have developed an automated 2D to 3D conversion system, which computes a virtual stereo image for a given video frame using information contained in neighboring time instances. However, this method requires a moving camera and static scenes for best performance (Knorr and Sikora, 2007). We are currently working on video analysis techniques to intelligently detect scenes where the automated conversion performs well and to assist manual conversion in other types of scenes.

2.2.4.5 Text detection and recognition on video

There is a wealth of amateur video recordings available now. Most of these recordings contain text information in some form or the other. Therefore, it is very important to be able to extract text information from average quality, uncontrolled video recordings. The obtained and recognized text can be very helpful for automated tagging applications. We are developing a method to first detect and then reliably track a text region. Through fusion over multiple time instances the recognition rate is improved significantly, analogous with super-resolution techniques.

2.2.4.6 Video scene classification

Video scene classification and segmentation are fundamental steps for multimedia retrieval, indexing and browsing. Our research topic contains shot boundary detection and video summarization as a first step for genre recognition.

Our shot boundary detection includes the two types of shot changes – abrupt cuts and gradual changes. Representative key frames per shot are also extracted for video

summarization. The focus of our research is the classification of video sequences in genres by analysing audiovisual features of consecutive frames in real time. This is part of the well-known video-genre-classification problem, where popular TV-broadcast genres and user-generated categories are studied (Glasberg et al, 2008a).

A large set of new audiovisual descriptors is being implemented, e.g. specific concept detectors (sport field, scoreboard, news ticker, logo, indoor/outdoor, building, text, etc.). Multiple approaches to achieve the aim of multi-genre classification are applied, e.g. decision fusion of single-genre classifiers or design of one multi-genre classifier (Glasberg et al, 2008b). Caused of high computational complexity a dimensional reduction is applied by using automatic feature selection algorithms and principle component analysis (PCA).

Following classification methods are deployed: support vector machines, multilayer perceptron, ID3 decision tree, hidden Markov model and Bayesian classifier. The actual results demonstrate a high identification rate based on a large representative collection of 100 video sequences (20 sequences per genre) gathered from free digital TV-broadcasting in Europe.

In our further work, there will be an attention to the extraction of features in compressed domain of H.264.

2.2.5 Cross-evaluation

From the description given in this section, one can see that there is the most ongoing research by the four national networks in this processing research field among the four SIGs, which may be because the problem of multimedia processing has been quite a classical issue dealt with by many researchers in comparison to the other three fields. However, it should be noted that the work being done in the national networks includes not only automatic processing and understanding of multimedia content, but also processing of additional information such as annotation, location and user-provided tags, and merging/integration of such information with automatically processed information. Especially, the latter is an important issue to bridge the ‘semantic gap’ between automatically extracted low-level information and humans’ high-level information in the context of SP2P networks.

The research in the NIRICT network emphasizes development of unsupervised mechanisms for discovering semantic information in audiovisual signals. Especially, the research on content discovery from composite audio is distinct from those in the other national networks. In the IM2 network, much work is being done for using information of annotation, location and tagging, along with automatic multimedia content processing. Also, the research for semantic analysis of multimedia based on the social interactions appearing in the data is being performed in this network. In the MMKM network, there exist a wide range of research work and projects on both content processing for various applications and annotation processing for drawings/sketches and e-learning. The research in the HC3 network emphasizes techniques for content analysis of audiovisual materials.

While there is much work on multimedia content processing in the four national networks, there is not much overlap between their research activities. Nevertheless, there are two desirable directions for further researches in the networks: First, it will be necessary to create cooperative work of much effort in each network in the context of PetaMedia, so that a lot of synergy can be produced for the relevant research field. Second, more work on bridging the result of automatic content processing and the user-created information would be desirable in the future research of PetaMedia.

2.3 Indexing

2.3.1 NIRICT

The work in the indexing field concentrated in the past years on two main topics: detection of faces and person-related face clusters, and affective content analysis and indexing.

2.3.1.1 Face-related research

Aiming at a generic, robust face detector, we started our face-related research with the development of an omnidirectional face detector. Our approach was based on combining multiple Viola-Jones detectors (Viola and Jones, 2004) with different classifiers. For this purpose, we also developed a face detector/pose estimator-combination that not only detects faces but also detects the pose in terms of a 3-D rotation.

The ability to match faces in video is a crucial component for many multimedia applications such as searching and recognizing people in semantic video browsing, surveillance and home video management systems. Unfortunately, most face matching methods were designed for and tested on frontal face images only, which does not comply with the professional and home video scenarios. In video, faces appear at different poses and scales, and the image quality may vary as well. We analyzed the quality of face matching approaches and specifically the relation with face pose and eye position quality. For the first research we found that quality of all face matching methods go down with increasing deviation of the pose from the frontal pose, but contrary to popular belief PCA appears to be more robust than elastic bunch graph matching (EBGM). For the second research (which uses frontal images) we found criteria that when imposed on an eye localizer provide similar face matching results as when using annotated eye positions. In this case we also show that PCA and EBGM provide similar results and thus the less complex PCA algorithm can be used without sacrificing performance. Besides the detailed analysis of face matching aspects we also demonstrated a face search system that comprises a face detector, face tracker and face matcher (Kroon et al, 2007).

The face matching research made us realize the importance of accurate eye localization. We addressed the problem of eye localization in low and standard definition content, such as webcam-generated and TV images. We developed a probabilistic eye localization method based on well-known multiscale local binary patterns (LBPs), which provide a simple and powerful spatial description of texture, and are robust to the noise typical to low and standard definition content. Our primary contribution is in the proposed method of combining the LBPs that is targeted towards achieving spatial accuracy under mentioned conditions. Evaluation performed on a standard dataset of webcam-quality images showed that our approach has superior performance with respect to the state of the art, while having a reasonable complexity and a low memory footprint. We have also shown that our eye localizer meets the requirements for efficient face matching that we have formulated (Kroon et al, 2008).

2.3.1.2 Affective content analysis and indexing

The theory and algorithms of multimedia content analysis (MCA) aim at enabling automation of tedious and time-consuming processes inherent to audiovisual content indexing, management and retrieval. Past research in the MCA field has mainly considered the extraction of the cognitive content information from audiovisual signals. This information includes facts about the temporal content structure (shots, scenes, plot points) and spatiotemporal content elements (objects, persons, events, topics). Although the cognitive MCA solutions are not yet mature enough, the need has emerged for a concurrent research effort aiming at the extraction of the mood, feeling or emotion (jointly referred to as *affective content* or *affect*) from audiovisual signals. This need stems from the critical role affect plays in indexing, management and retrieval/delivery of audiovisual content, like in the case of personalized content recommendation (“I am in the mood for this type of movie/music!”), highlights extraction (“Find 10 most exciting minutes of a sport event!”) and surveillance (“Detect surveillance video segments that do not feel right!”).

In this research direction, we developed a methodology for affect extraction from audiovisual signals, which is based on the *dimensional approach to affect* known from psychophysiology. Using this approach, the difficult problem of affect extraction can be reduced to the problem of extracting two underlying affect dimensions: *arousal* (affect intensity) and *valence* (degree of (un)pleasantness). The advantage of the proposed approach, compared to other reported affect extraction attempts, is the possibility to employ arousal and valence broadly, beyond the controlled situations, and without the need to deal with abstract and ambiguous affect categories (e.g. “happy”, “sad”) (Hanjalic, 2006).

While we aim at developing conceptually generic arousal and valence models, we are currently investigating how these models can be optimized for a given context/user/application iteratively, using interactive learning. For evaluating the proposed methodology we consider two representative application scenarios. The first scenario involves easy, non-linear access to interesting pieces of concert videos available in a large Internet music collection. Next to the methods that we developed for this purpose and that involved mining and discovery of content structure of an arbitrary audiovisual signal (Naci and Hanjalic, 2008), the core of the approach consist of the construction of an arousal-based highlights time-curve that can be personalized and enable easy selection of highlighting concert segments in the desired length (Naci and Hanjalic, 2007). We will then expand the application scope of the developed concept to soccer videos with the objective of bringing soccer highlights to the user via a mobile device in an effortless, intuitive and personalized fashion (Hanjalic, 2005).

2.3.2 IM2

There is not much work related to indexing in the IM2 network.

2.3.3 MMKM

In order to develop truly useful information retrieval systems, the challenge is to associate the low-level features extracted from the audio and visual signals representing the content with higher-level concepts that come naturally to humans, known as the semantic gap.

2.3.3.1 Multimedia feature extraction

An important technological objective of the UK network is to investigate how to structure the retrieval task given an index of the content in terms of semantic concepts associated with low-level features. An associated challenge is to perform retrieval of irrespective of the nature of the content. In order to achieve this research challenge, efficient multimedia (e.g. image, video, audio, speech, etc) content representation and indexing process are critical.

Some of the fundamental tasks for generating metadata automatically are rooted in machine learning tasks of classification and pattern analysis. One notable current EU-funded Network of Excellence in this area, PASCAL2²³, is UK coordinated and for some work in this area (Rousu et al, 2006).

The Humanities Advanced Technology and Information Institute (HATII), University of Glasgow, has been examining methods of achieving automated semantic metadata extraction from digital documents as a crucial step in realising automated ingest, selection and management of digital material. As a step in direction, Kim and Ross have been looking at automated genre classification as an essential area of study that will bind efforts in genre-specific automated metadata and provide structural classification of unstructured text for mining, both from a digital library perspective (2007) and from a language processing perspective (2008).

2.3.3.2 Clustering

There is a host of UK research that supports the bridging of the semantic gap via automated annotation: Hare and Lewis (2004) use salient interest points and the concept of scale to the selection of salient regions in an image to describe the image characteristics in that region; they then extended this work (2005) to model visual terms from a training set that can then be used to annotate unseen images; Yavlinsky et al (2005) developed non-parametric density estimation for labels in image feature space with novel kernels that utilise the Earth mover's distance; Magalhaes and Ruger (2006) developed a clustering method that is more computationally efficient than the currently most effective method of non-parametric density estimation, which they later (2007) integrated into a unique multimedia indexing model for heterogeneous data; Baillie and Jose (2004) use audio analysis of the crowd response in a football game to detect important events in the match, which was later extended for both segmentation and then classification of football video by Baillie and Jose (2004); Cavallaro and Ebrahimi (2004) proposed an interaction mechanism between the semantic and the region partitions which allows to detect multiple simultaneous objects in videos; Dorado et al (2004) use fuzzy logic and rule mining techniques to learn from human expert annotations for the task of automatically assigning keywords

²³ <http://www.pascal-network.org>

from a lexicon to non-annotated video clips; Burghardt and Calic (2006), focusing on the semantic annotation in the wildlife video domain, have utilised robust feature tracking and a specific interest model in order to classify animal locomotive processes from a predefined taxonomy; on a higher level Salway and Graham (2003) developed a method to extract information about characters emotions in films and suggested that this information can help describe higher levels of multimedia semantics relating to narrative structures; Salway et al (2005) contribute to the analysis and description of semantic video content by investigating what actions are important in films; Durand et al (2005) introduced as part of the EU-funded SAVANT²⁴ project a novel metadata model for describing scalable and interactive TV services that can be enriched with supplemental multimedia information.

AKTiveMedia (Chakravarthy et al, 2006) is a user-centred ontology based cross-media annotation (images and text) tool developed as part of the AKT project. It supports users by suggesting potential annotations by using information extracted automatically from either the text and or the images (and across). The system actively works in the background, interacting with web services and queries the central annotational store to look for context specific knowledge. The recently funded EPSRC project ANAWIKI (4.5.5) addresses the issue of creating metadata on a large scale using the community.

2.3.3.3 Tagging

One of the limiting factors of the uptake of digital contents for multimedia is the scarcity or expense of metadata for the digitised (or digitally born) media. Flickr, a popular photo sharing site, lets the users upload, organise and annotate their own photographs with tags. In order to search images in Flickr little more than the user's tags are available with the effect that many photos are difficult to find — or not at all. The same is true for the video sharing site YouTube²⁵, but to a lesser degree for commercial digital download multimedia stores such as iTunes²⁶ that sells music, movies, TV shows, audiobooks, podcasts, and games: the commercial nature makes it viable to supply metadata to the required level of granularity.

One way to generate useful tags and metadata for a multimedia object is to involve a community of people who do the tagging collaboratively (this process is also called folksonomy, social indexing or social tagging). Del.icio.us²⁷ is a social bookmarking system and a good example for folksonomies, and similarly the ability of Flickr to annotate images of other people falls also into this category. In a similar spirit von Ahn and Dabbish (2004) have invented a computer game that provides an incentive for people to label randomly selected images. All these approaches tap into “human computing power” for a good cause, the structuring and labelling of multimedia objects. Research in this area is still in the beginning, and it is by no way clear how to best harness the social power of collaborative tagging to improve metadata and access to digital museums and libraries alike. POLYMNIA²⁸ is a recently finished EU

²⁴ <http://www.elec.qmul.ac.uk/mmv/savant.html>

²⁵ <http://youtube.com>

²⁶ <http://www.itunes.com>

²⁷ <http://del.icio.us>

²⁸ <http://polymnia.pc.unicatt.it>

project that allows museum visits to be shared with friends and family via live video stream and recorded on a DVD.

There are two EU-funded Networks of Excellence that are concerned with the bridging of the semantic gap: MUSCLE²⁹ (Multimedia Understanding through Semantics, Computation and Learning) and K-Space³⁰ (Knowledge Space of Semantic Inference for automated annotation and retrieval of Multimedia Content). Another EU integrated project, aceMedia³¹ has the specific aim to generate automated annotation from content to make it easier to find and re-use content. One nationally funded project with the title “Bridging the semantic gap in visual information retrieval” has looked at the very same topic from a user point of view.

2.3.4 HC3

2.3.4.1 Usability test of tagging tools

Currently, a team of researchers from the field of machine learning, sociological technology studies, and IT management from TU Berlin prepare, with the support of HC3, a project examining usability features and emerging practices concerning a tagging tool introduced in (Ochab et al, 2008).

2.3.5 Cross-evaluation

In the field of the indexing research, some networks show many activities while some others do not. The research in the NIRICT network includes face detection/matching for implicit tagging and affective content analysis/indexing. In the MMKM network there exist researches on multimedia semantic feature and metadata extraction, annotation clustering, and automated/manual tagging. The HC3 network prepares a project for usability test of tagging tools.

One important issue which needs more researches in the future is implicit tagging. Data collected by monitoring people involved in a SP2P network play an important role for implicit tagging for the multimedia content which they are experiencing. Implicit tagging can be an important tool to create, propagate and evaluate annotations. Related research topics include human affect sensing and affect-sensitive adaptation of tags, gaze and facial expression analysis, speech and multimodal human behaviour analysis, brain-computer interfaces, perceptual and multimodal user interfaces for multimedia indexing, etc.

²⁹ <http://www.muscle-noe.org>

³⁰ <http://kspace.qmul.net>

³¹ <http://www.acemedia.org>

2.4 Social Content Retrieval

2.4.1 NIRICT

In the following paragraphs we highlight the main aspects of our recent and current research on social aspects of multimedia retrieval (Wang et al, 2006; Pouwelse et al, 2007; Clements et al, 2008).

2.4.1.1 Indexing

Since most existing data collections are collaboratively created and simultaneously accessed by thousands of individuals, social aspects play an evident role in multimedia distribution. The social openness of recently emerging internet applications stimulates users to behave in a socially accepted manner. Today's internet user is fully aware of his online identity and actively manages his reputation by accurately annotating content and leaving valuable comments in web logs. Because content-based retrieval systems are still far from bridging the semantic gap, the exploitation of the information contributed by the network users is inevitable in order to effectively access these data collections. The contributed information ranges from preference indications like ratings or play-counts to textual annotations (comments, tags) and social relations (friendships, interest groups). Besides these explicitly created annotations, information can be drawn from the user's context. Monitoring the user while he gains access to the provided data can result in a valuable affective preference description. The collected information can be exploited by facilitating directed search, but also enables the system to establish a preference profile for its users and predict autonomous recommendations. We approach integration of these information sources by integrating the ideas and methods from the fields of Information Retrieval, Collaborative Filtering and Network Analysis.

2.4.1.2 Retrieval

Most operating retrieval systems on multimedia databases rely on simple frequency based algorithms that have proven effective in text retrieval. As user contributed tags are often the only available textual annotations in collaborative media, retrieval purely based on the observed word frequencies is very limited. The graphical structure in social networks however allows us to derive latent semantic relations that can be exploited for more accurate content retrieval. Also, it allows us to derive interest groups or conceptual content clusters that can aid the user while browsing the database for relevant content or people.

2.4.1.3 Recommendation

Traditional recommender systems rely on a one-dimensional representation of user preference to derive user or content similarities. Although much work has focused on the integration of metadata into so called *hybrid recommender systems*, not much attention has gone to a more thorough representation and exploitation of preference. We are currently searching for ways to improve user preference determination. For instance, integration of more affective user information acquired through smart sensors in the user's environment can result in more accurate recommendations that

are adapted to the current context and mood of the user. Furthermore, a personalized retrieval system, based on the social surroundings of the network user, should be able to serve the user with relevant information at any time. Whether the user is explicitly querying the database or passively watching a media channel, the system should be aware of the user's context.

2.4.2 IM2

2.4.2.1 Multimodal relevance feedback in retrieval systems for images accompanied by texts

In this research we define a relevance feedback mechanism that suggests keywords that can refine the search results obtained by submitting an image as query. The performance is measured through the difference of the “precision at position N” after the relevance feedback mechanism has been applied.

2.4.2.2 Multimodal collection search and indexing

The goal of the research is to define both multimodal indexing and retrieval strategies on rich multimedia collections based on the modern concept of interactive learning. This implies the definition of robust and accessible structures underlying the original data. Starting from a comparison between the performance of the known algorithms AdaBoost and RankBoost, we have explored the direction of having a retrieval model based on rank order only rather than absolute distance measurements. This way, normalization problems inherent to the fusion of different modalities are alleviated. We have started the development of a multimodal retrieval engine (called the Cross Modal Search Engine). This package comprises a versatile feature extraction library, including most features of OpenCV (including face detection). RankBoost retrieval is supported via the FastMap indexing strategy. This C++ package may be deployed in any multimodal retrieval context and is currently tested using IM2 data in parallel with the development of an EU Project (MultiMatch) on Cultural Heritage.

2.4.2.3 Incremental multimedia content description

Incremental multimedia content description consists in exploiting the underlying data organization created by indexing and retrieval advances to organize and optimize multimodal content description. A key issue is the use of search and retrieval tools to achieve this task and take advantage of inferred relationships.

Work on two tracks is ongoing. We have developed a data annotation tool including WordNet for the disambiguation and contextualisation of (essentially manual) annotations. We aim at an incremental annotation process whereby annotation is seeded manually and propagated over a network of inter-item relationship created from external information (e.g. relevance feedback gathered with sessions of a retrieval system (Morrison et al, 2007)). Since 2008, emphasis has been placed on semantic propagation from long-term user interaction capture. We are progressively constructing a propagation model to exploit Query By Example (QBE) interaction logs. Difficulties have appeared in capturing such data. We have therefore developed

a model to simulate this data from given user interaction models and are also looking at community-based data (e.g. Flickr tagging) as a base for further data simulation.

2.4.3 MMKM

A lot of research on multimedia information retrieval system made this field established enough to make new development.

2.4.3.1 Multimedia information retrieval

A number of studies, in reporting on queries addressed by different types of users to a variety of image and video collections, have contributed to our better understanding of clients' needs. Smeulders et al (2000) provide a useful framework in which to consider these user studies. Three types of search are identified, labelled 'target search', aiming at a specific image identified by title or other unique identifier; 'category search', where the client has no specific image in mind but can describe the wanted features by means of a search statement; and 'search by association' where the client is content to browse in order to retrieve images by serendipity.

The incidence of queries that fall into these categories reflects the nature of image collections and their clientele. In the case of archival collections many requests are likely to be for target and category searches — a reflection both of the role of the archive in curating images as sources of information and the specificity enhancement to be expected from the query mediation performed by expert picture library/archive staff, see (Hollink et al, 2004). One of the earliest studies that confirmed this expectation analysed some 2,700 requests addressed by a variety of client types to the Hulton Deutsch collection — a major picture archive and now part of Getty Images (Enser and McGregor, 1992; Enser, 1993). A number of other studies have confirmed the relatively high incidence of requests for specific, named features and in the particular context of journalists' requests to newspaper picture archives. In the case of moving images, likewise, a sample of 1,270 requests for film footage addressed to eleven screen archives in the course of the VIRAMI³² (Visual Information Retrieval for Archival Moving Imagery) project, also indicated that requests for specific subjects were more evident than for generic subjects, and requests for abstract subjects were unusual, where the focus of the collection lay with archival material. The further one moves away from the specialist archival collection, expert mediation and experienced user scenario towards less constrained environments the more pronounced becomes the emphasis on search by association and category searches that seek generic objects or scenes.

Evidence available thus far about Web-based searching of image collections points to the added significance of browsing (Goodrum and Spink, 2001). In an analysis of the transaction logs of over 33,000 image requests submitted to the Excite search engine by 10,000 searchers whose characteristics were unknown: Goodrum and Spink (2001) found a high rate of search modification, with sexual or adult content terms dominating the hundred most frequently occurring terms. Jorgensen (2005) studied the search behaviour of image professionals involved in advertising, marketing and graphic design, by analysing search logs from a commercial subscription service

³² <http://www.brighton.ac.uk/cmris/research/groups/vir>

image provider. For this user group a very low proportion of the requests were for specifically named features; instead, there was heavy recourse to browsing and the use of descriptive, thematic terms. Other areas of UK research with users in the centre of multimedia knowledge management include: the measurement of the salience of targets and distractors through competitive novelty (Stentiford, 2003); Berg and Rumsey (2003) look at the evaluation of perceived spatial quality of audio signals; (Neher et al, 2003) conducted a study into the perceptual construct of ‘ensemble width’ (i.e., the lateral spacing of the outer sources contained within an auditory scene); (Tombros et al, 2005) investigated the criteria used by online searchers when assessing the relevance of Web pages for information-seeking tasks; in another study Petrelli et al (2004) examined the user needs and tasks for cross-lingual retrieval and designed and tested interface components; before finally conducting usability tests.

Search is one of the most fundamental tasks in Multimedia Knowledge Management and has been the focus of much early research in the field – thus search and retrieval systems are already highly developed and represent some of the most mature research areas. Amongst all media types TV video streams arguably have the biggest scope for automatically extracting text strings in a number of ways: directly from closed-captions, teletext or subtitles; automated speech recognition on the audio and optical character recognition for text embedded in the frames of a video. Full text search in these strings is the way in which nowadays most video retrieval systems operate, including Google’s TV search engine³³ or Blinkx-TV³⁴. This technology existed in UK research labs much earlier: for example, a final-year student project at Imperial College London that indexed videos through teletext received a national prize in the year 2000. The ANSES³⁵ project was a follow-up project that built on this student-project, while Rich News describes a similar project at another UK university, Sheffield, that was later used in the aforementioned PrestoSpace project with a focus on preservation rather than searching.

Academic research on ways to search, mine and retrieve multimedia by content (as opposed to metadata and associated text) happens in many research labs in the world — one example is the UK uBase³⁶ project, where a lot of this cutting-edge research is explored and presented. Yet, specifically for content-based multimedia search (exemplified by the “search by example paradigm”, where the user drags and drops an image into a search box or hums a tune with the expectation of the computer finding the song in the database) the uptake in industry has been very low.

A notable exception is Virage³⁷, a project suite owned by UK based company Autonomy. Virage brings together complementary technologies from multimedia, security and infrastructure specialists. Virage make the claim to offer a product set capable of television, video, audio and CCTV challenges of any kind, i.e., from making television content fully searchable and accessible via IPTV to supplying and managing complex security systems. Automation is the key for this kind of multimedia indexing and retrieval and it has been very much the focus in academic research.

³³ <http://video.google.com>

³⁴ <http://www.blinkx.tv>

³⁵ <http://kmi.open.ac.uk/technologies/anses>

³⁶ <http://ubase.open.ac.uk>

³⁷ <http://www.virage.com>

2.4.4 HC3

2.4.4.1 Image search engine

There is an ongoing development of an Image Search Engine at the TU-Berlin which is based on the low level content of available images after a first assortment of images with the help of textual information. Momentarily a number of pre-sorted images for a certain query are received from the Yahoo Image Search Engine which does not explicitly incorporate low level information in its search process. The following refining is done by using an explicit relevance feedback and several content based descriptors whose use is encouraged by the MPEG-7 multimedia content description standard. Relevance feedback means a co-operation with the user and therefore its amount needs to be minimized. As such, active learning algorithms are used to minimize the necessary number of labelled images for good retrieval results.

2.4.4.2 Spam detection in social bookmarking systems

Detecting spam entries in social bookmarking systems is a new research field emerging with the increasing success of such systems. Researchers from HC3 examined network-based approaches to spam detection (Neubauer and Obermayer, 2008). Furthermore, the extension of tagging systems to new domains was explored with a system allowing users to tag location in virtual worlds (Ochab et al, 2008).

2.4.5 Cross-evaluation

The research of the NIRICT network covers various aspects of social content retrieval, such as indexing based on user-provided information, retrieval using the graphical structure in social networks, and personalized recommendation. In IM2, issues addressing multimodality in indexing/retrieval are investigated. Also, there exists work on incremental annotation processing. The research in the MMKM network considerably focuses on relevant issues in multimedia information search/indexing/retrieval. In the HC3 network, work on image search engines and quality control in social bookmarking is ongoing.

When we compare these activities, it is observed that there exists an overlap in the research on content search and retrieval. However, since the approach and focus of each network's work on this topic are not exactly the same, effective combination and cooperation of the researches of the different networks will result in synergetic success. It should be also noted that this SIG will address several entirely new challenges which have not been dealt with in previous researches and, thus, several creative issues will be addressed in the context of PetaMedia.

3 Conclusion

We have described the research activities of the four SIGs in the four national networks and identified redundancies and needs via cross-evaluation. It has been observed that there exist not so many overlaps of the networks' work and each network covers quite different research issues regarding the whole objective of PetaMedia. However, it is still necessary to have cooperative work by the four networks and to address new research issues.

In the content distribution and processing fields, wide ranges of research issues are addressed and investigated among the four networks. Therefore, cooperative work of the networks will produce successful synergy for the research fields. Here, how to make convergent efforts from each network's work would be a crucial issue. On the other hand, in the researches on indexing and social content retrieval, there is a necessity that new challenges should be raised and explored for successfully achieving the final goals of PetaMedia.

Acknowledgement

We would like to thank the MMKM network that helps us to produce the input of MMKM research. The full snapshot of MMKM research can be found at <http://mmis.doc.ic.ac.uk/tmp/mmkm-roadmap-sota/roadmap1-sota.pdf>.

References

- M. Baillie and J. Jose (2004). An audio-based sports video segmentation and event detection algorithm. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition 2004 (Event Mining 2004 Workshop).
- J. Berg and F. Rumsey (2003). Systematic evaluation of perceived spatial quality. In Proc. Int. Conf. Audio Engineering Society.
- A. Bradley and F. Stentiford (2003). Visual attention for region of interest coding in JPEG2000. *Journal of Visual Communication and Image Representation* 14, 232–250.
- T. Burghardt and J. Calic (2006). Analysing animal behaviour in wildlife videos using face detection and tracking. *IEE Proc. on Vision, Image and Signal Processing* 153(3), 305–312.
- D. Cai and C. van Rijsbergen (2005). Semantic relations and information discovery. In Proc. Intelligent Data Mining — Techniques and Applications.
- R. Cai, L. Lu, and A. Hanjalic (2008). Co-clustering for auditory scene categorization, *IEEE Trans. on Multimedia* 10(4), 596-606.
- J. Calic, N. Campbell, S. Dasiopoulou and Y. Kompatsiaris (2005). An overview of multimodal video representation for semantic analysis. In Proc. of the Europ. Workshop on Integration of Knowledge, Semantics and Digital Media Technology, pp. 39–45.
- A. Cavallaro (2004). Adding privacy constraints to video-based applications. In Proc. European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology.
- A. Cavallaro and T. Ebrahimi (2004). Interaction between high-level and low-level image analysis for semantic video object extraction. *Journal on Applied Signal Processing, Special Issue on: Object-based and semantic image and video analysis* 2004(6), 786–797.
- H. Celik, A. Hanjalic, E.A. Hendriks, and S. Boughorbel (2008). Online training of object detectors from unlabeled surveillance video. In Proc. IEEE Online Learning for Classification Workshop, CVPR.
- A. Chakravarthy, F. Ciravegna and V. Lanfranchi (2006). Aktivemedia: Cross-media document annotation and enrichment. In Proc. Int. Semantic Web Conf.
- M. Clements, A.P. De Vries, and M.J.T. Reinders (2008). Optimizing single term queries using a personalized Markov random walk over the social graph. In Proc. ECIR Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR'08).

- K. Clüver, J. Weil, and T. Sikora (2007). Multiple-description coding of speech using forward error correction codes. In Proc. 15th European Signal Processing Conference (EUSIPCO 2007), Poznań, Poland.
- I. Damnjanovic, N. Ramzan and E. Izquierdo (2006). MPEG-2 watermarking channel protection using duo-binary turbo codes. In Proc. 31st IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Toulouse, France.
- F. De Simone, D. Ticca, F. Dufaux, M. Ansorge and T. Ebrahimi (2008). A comparative study of color image compression standards using perceptually driven quality metrics. In Proc. SPIE Optics and Photonics, Applications of Digital Image Processing XXXI, San Diego, CA, USA.
- S. Dietze, A. Gugliotta and J. Domingue (2007). Towards adaptive elearning applications based on semantic web services. In TENCompetence Open Workshop on Service Oriented Approaches and Lifelong Competence Development Infrastructures, Manchester, UK.
- D. Djordjevic, A. Dorado, W. Pedrycz and E. Izquierdo (2005). Conceptoriented sample images selection. In Proc. European workshop on Image Analysis for Multimedia Interactive Services.
- A. Dorado, J. Calic and E. Izquierdo (2004). A rule-based video annotation system. IEEE Transactions on Circuits and Systems for Video Technology 14(5), 622–633.
- E. Dumont, B. Merialdo, S. Essid, W. Bailer, H. Rehatschek, D. Byrne, H. Bredin, N. E. O'Connor, G. J.F. Jones, A. F. Smeaton, M. Haller, A. Krutz, T. Sikora, and T. Piatrik (2008). Rushes video summarization using a collaborative approach. TRECVID BBC Rushes Summarization Workshop (TVS 2008) at ACM Multimedia, Vancouver, BC, Canada.
- G. Durand, G. Kazai, M. Lalmas, U. Rauschenbach and P. Wolf (2005). A metadata model supporting scalable interactive TV services. In Int. Conf. on Multi Media Modeling, pp 386–391.
- P. Enser (1993). Query analysis in a visual information retrieval context. Journal of Document and Text Management 1(1), 25-52.
- P. Enser and C. McGregor (1992). Analysis of visual information retrieval queries. Report on Project G16412 to the British Library R&D Department, London, British Library.
- L. Gillam, M. Tariq and K. Ahmad (2005). Terminology and the construction of ontology. Terminology 11(1), 55–81.
- R. Glasberg, S. Schmiedeke, M. Mocigemba, and T. Sikora (2008a). Real-time approaches for video-genre-classification using new high-level descriptors and a set of classifiers. ICSC2008.

R. Glasberg, S. Schmiedeke, P. Kelm and T. Sikora (2008b). An automatic system for real-time video-genres detection using high-level-descriptors and a set of classifiers. ISCE2008.

L. Goldmann, T. Adamek, P. Vajda, M. Karaman, R. Mörzinger, E. Galmar, T. Sikora, N. O'Connor, T. Ha-Minh, T. Ebrahimi, P. Schallauer, and B. Huet (2008). Towards fully automatic image segmentation evaluation. Advanced Concepts for Intelligent Vision Systems (ACIVS).

A. Goodrum and A. Spink (2001). Image searching on the excite web search engine. *Inf. Process. Manage.* 37(2), 295–311.

M. Haller, A. Krutz, and T. Sikora (2007). A generic approach for motion-based video parsing”, 15th European Signal Processing Conference (EUSIPCO 2007), Poznań, Poland.

A. Hanjalic (2005). Adaptive extraction of highlights from a sport video based on excitement modeling. *IEEE Trans. on Multimedia* 7(6), 1114-1122.

A. Hanjalic (2006). Extracting moods from pictures and sounds: towards truly personalized TV. *IEEE Signal Processing Magazine*.

J. Hare and P. Lewis (2004). Salient regions for query by image content. In *Proc. Int. Conf. on Image and Video Retrieval*.

D. Heesch and S. Ruger (2003). Performance boosting with three mouse clicks — Relevance feedback for CBIR. In *Proc. European Conf. on Information Retrieval Research*.

X. Hilaire and J. Jose (2007). Enhancing CBIR through feature optimization, combination and selection. In *Proc. Int. Workshop on Content-Based Multimedia Indexing*.

L. Hollink, A. Schreiber, B. Wielinga and M. Worring (2004). Classification of user image descriptions. *Int. J. Hum-Comput. Stud.* 61(5), 601–626.

P. Howarth and S. Ruger (2005). Fractional distance measures for content based image retrieval. In *Proc. European Conf. on Information Retrieval*.

B. Hu, S. Dasmahapatra, P. Lewis and N. Shadbolt (2003). Ontology-based medical image annotation with description logics. In *Proc. IEEE Int. Conf. on Tools with Artificial Intelligence*.

M. Huebscher and J. McCann (2008). A survey of autonomic computing. *Computing Surveys*, to be published.

G. Jawaheer and J. McCann (2005). Adaptation of dynamic web pages on the edge. In *Proc. Int. WWW/Internet Conf.*

- Y. Kim and S. Ross (2007). The naming of cats: automated genre classification. *International Journal of Digital Curation* 2(1), 49-61.
- Y. Kim and S. Ross (2008). Examining variations of prominent features in genre classification. In *Proc. Hawaiian Int. Conf. System Sciences*, Waikoloa, Hawaii, USA.
- S. Knorr and T. Sikora (2007). An image-based rendering (IBR) approach for realistic stereo view synthesis of TV broadcast based on structure from motion. In *Proc. IEEE Int. Conf. on Image Processing (ICIP)*, San Antonio, Texas, USA.
- Z. Kong and M. Lalmas (2007). Using XML logical structure to retrieve (multimedia) objects. In *Proc. European Conf. on Research and Advanced Technology for Digital Libraries*, pp 100–111.
- B. Kroon, A. Hanjalic, and S. Boughorbel (2007). Comparison of face matching techniques under pose variation. In *Proc. 6th ACM Int. Conf. on Image and video retrieval*, pp. 272-279.
- B. Kroon, S. Maas, and A. Hanjalic (2008). Eye Localization for face matching: is it always useful and under what conditions? In *Proc. ACM CIVR 2008 Conference*, Niagara Falls.
- A. Krutz, A. Glantz, T. Sikora, P. Nunes, and F. Pereira (2008). Automatic object segmentation algorithms for sprite coding using MPEG-4. *50th International Symposium ELMAR-2008*
- M. Kunter, A. Krutz, M. Dröse, M. Frater, and T. Sikora (2007). Object-based multiple sprite coding of unsegmented videos using H.264/AVC. *IEEE Int. Conf. on Image Processing (ICIP 2007)*, San Antonio, Texas, USA.
- L. Lu and A. Hanjalic (2008). Audio keywords discovery for text-like audio content analysis and retrieval. *IEEE Trans. on Multimedia*.
- J. Magalhaes and S. Ruger (2006). Logistic regression of semantic codebooks for semantic image retrieval. In *Proc. Int. Conf. on Image and Video Retrieval*.
- J. Magalhaes and S. Ruger (2007). Information theoretic semantic multimedia indexing. In *Proc. Int. Conf. on Image and Video Retrieval*.
- J. McCann, P. Howlett and J. Crane (2000). Kendra: Adaptive internet system. *Journal of Systems and Software* 55(1), 3–17.
- D. Morrison, S. Marchand-Maillet, and E. Bruno (2007). Automatic image annotation with relevance feedback and latent semantic analysis. In *Proc. Int. Workshop on Adaptive Multimedia Retrieval*, Paris, France.
- M. Mrak, N. Sprljan, and E. Izquierdo (2005). Motion estimation in temporal subbands for quality scalable motion coding. *Electronics Letters* 41(19).

U. Naci and A. Hanjalic (2007). Intelligent browsing of concert videos. In Proc. 15th Int. Conf. on Multimedia (ACM MM '07), ACM Press.

U.Naci and A. Hanjalic (2008). Content-based indexing of music concert recordings based on crossing-rate features. In Proc. 6th Int. Workshop on Content-Based Multimedia Indexing (CBMI 2008), London, UK.

T. Neher, F. Rumsey, T. Brookes and P. Craven (2003). Unidimensional simulation of the spatial attribute ‘ensemble width’ for training purposes. In Proc. 114th Audio Engineering Society Convention.

N. Neubauer and K. Obermayer (2008). Predicting tag spam examining cooccurrences, network structures and URL components. Wikis, Blogs, Bookmarking Tools - Mining the Web 2.0 Workshop at ECML/PKDD 2008.

B. Ochab, N. Neubauer, and K. Obermayer (2008): Personalized recommendations for the Web 3D. In: J. Kay, P. Pu, W. Nejdl and E. Herder (eds.): In Proc. Int. Conf. Adaptive Hypermedia and Adaptive Web-Based Systems.

D. Petrelli, P. Hansen, M. Beaulieu, M. Sanderson, G. Demetriou, and P. Herring (2004). Observing users — designing clarity: A case study on the user-centred design of a cross-language retrieval system. *J. Am. Soc. Inf. Sci. Technol.* 55(10), 923–934.

A. Popescu-Belis and S. Zufferey (2007). Contrasting the automatic identification of two discourse markers in multiparty dialogues. In Proc. SIGdial Workshop on Discourse and Dialogue, Antwerp, Belgium, pp. 10-17.

S. Potter, Y. Kalfoglou, H. Alani, M. Bachler, S. Buckingham Shum, R. Carvalho, A. Chakravarthy, S. Chalmers, S. Chapman, B. Hu, A. Preece, N. Shadbolt, A. Tate, and M. Tuffield (2007). The application of advanced knowledge technologies for emergency response. In Proc. Int. Conf. Information Systems for Crisis Response and Management, Delft, The Netherlands.

J. Pouwelse, P. Garbacki, J. Wang, A. Bakker, J. Yang, A. Iosup, D. H. J. Epema, M. Reinders, M. van Steen, and H. Sips (2007). Tribler: A social-based peer-to-peer system. *Concurrency and Computation: Practice and Experience* 19, 1–11.

T. Quack, B. Leibe, and L. van Gool (2008). World-scale mining of objects and events from community photo collections. In Proc. ACM Int. Conf. Image and Video Retrieval, Niagara Falls, Canada.

N. Ramzan, S. Wan, and E. Izquierdo (2007). Joint source-channel coding for wavelet-based scalable video transmission using an adaptive turbo code. *EURASIP Journal on Image and Video Processing* 1.

J. Rousu, C. Saunders, S. Szedmak, and J. Shawe-Taylor (2006). Kernel-based learning of hierarchical multilabel classification models. *Journal of Machine Learning Research* 7, 1601–1626.

- A. Salway and M. Graham (2003). Extracting information about emotions in films. In Proc. ACM Conf. on Multimedia.
- A. Salway, A. Vassiliou, and K. Ahmad (2005). What happens in films? In Proc. IEEE Int. Conf. on Multimedia and Expo.
- A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain (2000). Content based image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Mach. Intell. 22(12), 1349–1380.
- N. Sprljan, M. Mrak and E. Izquierdo (2005). A fast error protection scheme for transmission of embedded coded images over unreliable channels and fixed packet size. In Proc. IEEE Int. Conf on Acoustics, Speech, and Signal Processing.
- F. Stentiford (2003). The measurement of the salience of targets and distractors through competitive novelty. In Proc. European Conf. on Visual Perception.
- S. Sweeney and F. Crestani (2004). Supporting searching on small screen devices using summarisation. In Proc. Mobile and Ubiquitous Information Access.
- A. Tombros, I. Ruthven, and J. Jose (2005). How users assess web pages for information seeking. J. Am. Soc. Inf. Sci. Technol. 56(4), 327–344.
- R. Villa, R. Wilson, and F. Crestani (2004). Ontology mapping by concept similarity. In Proc. Int. Conf on Digital Libraries.
- P. Viola and M. J. Jones (2004). Robust real-time face detection. International Journal of Computer Vision 52(2), 137-154.
- L. von Ahn and L. Dabbish (2004). Labeling images with a computer game. In Proc. Int. Conf. on Human Factors in Computing Systems.
- D. Walther and C. Koch (2006). Modeling attention to salient protoobjects. Neural Networks 19, 1395-1407.
- S. Wan, M. Mark, N. Ramzan, and E. Izquierdo (2007). Perceptually adaptive joint deranging-deblocking filtering for scalable video multicast over wireless networks. ELSEVIER Journal of Signal Processing: Image Communication 22(3), 235-346.
- J. Wang, A.P. de Vries, and M.J.T. Reinders (2006). Unifying user-based and item-based collaborative filtering approaches by similarity fusion. In SIGIR '06: Proc. 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, New York, NY, 501–508
- S. Wu, F. Crestani, and F. Gibb (2004). New methods of results merging for distributed information retrieval. In Proc. Recent Research in Multimedia Distributed Information Retrieval.

A. Yavlinsky, E. Schofield, and S. Ruger (2005). Automated image annotation using global features and robust nonparametric density estimation. In Proc. Int. Conf. on Image and Video Retrieval.