TOSCA-MP

State of the art on semantic retrieval of AV content beyond text resources

Deliverable D3.1

TOSCA-MP identifier:	TOSCAMP-D3.1-v1.0.docx
Deliverable number:	D3.1
Author(s) and company:	Ozelín López, Guillermo Álvaro, Sinuhé Arroyo, Carlos Romero (PLY) Marie-Francine Moens, Gert-Jan Poulisse (K.U.Leuven) Mike Matton (VRT)
Internal reviewers:	Antje Linnemann (HHI)
Work package / task:	WP3
Document status:	Final
Confidentiality:	Public

Version	Date	Reason of change
0.1	2012-01-26	Document created (initial input…)
0.2	2012-02-17	First raw contents from partners
0.3	2012-03-16	Completed most sections
0.4	2012-03-21	Completed section 4.2
0.5	2012-03-23	Completed section 3
0.6	2012-03-26	Updates on section 4, merged PLY/KUL references
0.7	2012-03-28	Updates on section 5, merged PLY/KUL/VRT references (Version for internal review)
0.8	2012-04-20	Refined version after internal review
1.0	2012-04-26	Final version, submitted
1.1	2012-09-28	Document amended to include section 4.1.6

Acknowledgement: The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 287532.

Table of Contents

List of Figures

List of Tables

1 Executive Summary

This deliverable describes the state of the art in the area of semantic retrieval of multimedia (audiovisual) content beyond text resources, i.e., considering the nature of the content to be retrieved. In this line, the characteristics of images, video and audio, are exploited for improving accuracy of retrieval results.

One important issue regarding the retrieval of content (and multimedia is not an exception) is that it has to be considered by taking into account the characteristics of the data itself as well as of the associated metadata. Semantic technologies are able to provide enhanced retrieval results and are to be addressed from the point of view of broadcasting content. Therefore, the whole spectrum of the semantic approach, ranging from the multimedia data itself to semantic annotations and search/retrieval aspects, is considered in this deliverable.

From the point of view of data, different aspects are tackled, including the definitions of data, metadata, content, essence, material or asset, as well as the characteristics of important metadata formats. Semantic technologies and ontologies in the area of broadcasting are covered, and in particular the Linked (Open) Data approach is analysed.

The annotation of audiovisual content is described from two different yet complementary perspectives: automatic and manual annotation techniques. On the one hand, automatic annotation techniques include visual analysis, concept detection, speech-to-text methods, etc. On the other hand, manual techniques range from repurposing existing metadata and manual annotation tools to different crowdsourcing approaches. Hybrid methods that combine the best results from the large spectrum of techniques are able to provide better annotations and thus facilitate the retrieval process.

Finally, the semantic retrieval of audiovisual content is considered from the perspective of different techniques, from speech oriented solutions to multimodal approaches that combine exploration and navigation features.

2 Introduction

2.1 Purpose of this Document

The purpose of TOSCA-MP Deliverable 3.1 is to describe the state of the art on semantic retrieval of audiovisual content beyond text resources.

2.2 Scope of this Document

The present deliverable addresses i) the characteristics of data and metadata from the broadcasting domain, with a special emphasis on their semantic aspects, ii) different automatic and manual annotation techniques for such content, and iii) semantic retrieval of audiovisual content that relates to the models and annotation techniques described.

2.3 Status of this Document

This is the final version of D3.1.

2.4 Related Documents

N/A

3 Multimedia Data, Metadata and Semantics

Advanced solutions for the retrieval of content need the characteristics of the data they are dealing with in order to provide relevant search results. Audiovisual content is not an exception, and the characteristics of multimedia data have to be understood and considered for enhancing retrieval aspects of asset-management solutions.

In this section, we address the main characteristics of data and metadata, covering from the definition of terms frequently used in broadcasting such as data, metadata, essence, material, content or asset, to metadata standards which are relevant for the area (subsection 3.1). Semantic technologies and ontologies in particular are able to improve retrieval results, and thus we also cover the perspective of data in the broadcasting domain from the point of view of semantics (subsection 3.2).

3.1 Data, Metadata Extraction

3.1.1 Introduction: essence, material, metadata and content

In the literature, many concepts recur frequently: data, essence, material, assets, metadata… As these concepts are not always used in the correct manner, we will start this section with a few definitions, before diving into the concept of metadata itself.

The main source of information is **data**. Data is just any form of information that is translated into a form that is convenient to move or process. Important to notice is that data is abstracted from its physical carrier. E.g. a collection of bytes representing for instance a text document (the data) can be stored in a file on a computer, on a digital tape, printed on a piece of paper, or in many other forms. In fact, data is the basic concept necessary for defining all the other related concepts.

Cox et al. have define these concepts applied to the broadcasting domain (audiovisual data) as follows [Cox et al., 2006]:

- **Essense** is any data or signal necessary to represent any single type of visual, aural, or other sensory experience (independent of the method of coding).
- **Material** is any one or more combination of video, sound and other data essences.
- **Metadata** is data which conveys information about the material.
- **Content** is material in combination with associated metadata.

Next to these concepts, another concept is frequently used in broadcasting: **asset**. An asset is content that is associated with a special type of extra metadata: intellectual property rights.

In the next section, we will dig a little bit deeper into the concept of metadata.

3.1.2 Metadata

As explained, metadata is just another type of data. Important to notice is that metadata is always associated with essence. However, it is possible that metadata already exists before the actual essence exists (e.g. the title of a video is already chosen before the video is actually produced).

Several types of metadata can be identified:

- **semantic metadata** provides a description of the contents of the data.
- **technical metadata** provides technical information about the essence. This technical information is usually required in order to be able to read, decode or process the essence.
- **administrative metadata** is metadata that includes business and legal aspects for the essence.

We will now apply these different metadata types to the broadcasting domain.

Semantic metadata

Descriptive metadata, as the name suggests, is metadata that describes the essence. It provides substantive information about the material.

Three main types of semantic metadata for videos can be identified. The first type provides information which is applicable to the whole video. Examples of such metadata are title, actors performing in the video, genre of the video, key topics of the video, a textual description of the video, etc.

A second type of semantic metadata is time-coded metadata, or segmented metadata. This type of metadata only applies to a specific time segment in the video. Examples are a speech transcript (which provides time codes for every word), scene locations (which provides location information of a specific scene in the video), a penalty in a soccer match, etc.

The third type of semantic metadata is regional metadata. This kind of metadata usually applies to a specific area of the video, often at a specific time. Examples are faces and objects appearing in the video at some point in time.

Technical metadata

A second type of metadata is technical metadata. Technical metadata provides information on technical aspects of the material and the material carrier. Examples are the number of tracks (audio, video, other data) in the material, the type of carrier (file on disk, tape…), the codecs and corresponding parameters that are used for encoding the material, the resolution of the video...

The technical metadata is usually generated when the material is recorded and/or transcoded.

Administrative metadata

A third type of metadata frequently occurring in broadcasting is administrative metadata. This administrative metadata in turn can again be subdivided in two main types. The first type is business-oriented metadata. This type of metadata is associated with the creation of the material. Examples are the date of production, the names of the people in the production crew, information on the type of camera's that are used, etc.

The second kind of administrative metadata is intellectual property (IP) metadata. This kind of metadata described the IP rights holder(s) of the material. It also describes specific particularities of the material (such as parts of a video that may never be reused, or an authorization that is required before reusing some material…).

3.1.3 Metadata standards

In this section, we provide an overview of some important standards for structuring metadata.

Dublin Core

Dublin Core[1] is a metadata element set that is intended to be a common set of elements that can be used across many different media types. It has been approved as a U.S. National Standard (ANSI/NISO Z39.85) in 1995.

The standard contains 15 basic descriptive elements, which can be tagged with a qualifier and which can occur many different times. The elements are: contributor, coverage, creator, date, description, format, identifier, language, publisher, relation, rights, source, subject, title and type. Note that this list contains descriptive, technical and administrative metadata fields.

While Dublin core is widely agreed upon as a common standard, the expressiveness of it for broadcast video content is usually too limited. Moreover, Dublin Core cannot cope with segmented and regional descriptive metadata.

The specification of Dublin Core can be found in the Dublin Core Metadata Element Set, version 1.1[2].

EBUCore

The existence of the EBUCore[3] metadata standard dates back to 2000. The standard was originally devised as a refinement of Dublin Core for audio archives, but it has been extended several times.

The current scope of EBUCore is now identified as the minimum information needed to describe radio and television in broadcasting. It addresses creation, management and preservation of audiovisual content. Next to a formal description, it is also available as an RDF (Resource Description Framework, [Klyne & Carroll, 2004]) ontology, and it is consistent with the W3C Media Annotation Working Group ontology. For more information we refer to the EBU Technical document 3293[4].

As with Dublin Core, also EBUCore has difficulties to cope with segmented and regional descriptive metadata.

NewsML-G2

NewsML-G2[5] is a XML-based metadata standard that targets news content exchange. It has been standardised by the International Press Telecommunications Council (IPTC). The standard can be used as en container for news items, or can be a structured set of links to news items. It also contains metadata that describe the news items and the relations between them. It includes descriptive information about the news content, as well as administrative and technical information.

[1] http://dublincore.org/

[2] http://dublincore.org/documents/2010/10/11/dces/

[3] http://tech.ebu.ch/lang/en/MetadataEbuCore

[4] available at http://tech.ebu.ch/docs/tech/tech3293v1_3.pdf

[5] http://www.iptc.org/site/News_Exchange_Formats/NewsML-G2/

MPEG-7

Unlike its predecessors, MPEG-1, MPEG-2 and MPEG-4, the MPEG-7 standard[6] is not an AV coding standard. Instead it is a multimedia content description standard, standardized in ISO/IEC 15938. The standard identifies three kinds of elements. The first one is a set of description schemes (DS) and descriptors (D) that can be used to describe technical, administrative of descriptive metadata. The second element of the standard is a language for specifying the D and DS, called Description Definition Language (DDL). A third element in the standard is a scheme for coding the description in order to provide a standard to store it, or to MUX it with the content.

MPEG-7 AVDP

Recently, a special profile MPEG-7 has been defined, specially devised for integrating results of automatic audiovisual feature extraction tools. The profile is called Audio Visual Description Profile (AVDP). Currently it is in the final stage of becoming a standard.

The AVDP profile has been created because the standard MPEG-7 is conceived as too generic and too complex. Therefore it is not easily adopted on the industrial site.

The AVDP specification and AVDP schema are currently in final ballot before becoming an official standard. When they are finally standardised they will be part of part 9 and part 11 of the MPEG-7 standard specification.

3.1.4 Metadata extraction

Metadata extraction is defined as the creation of metadata based on the content. The task can be performed manually or automatically. An extensive overview of manual and automatic extraction techniques is presented in chapter 4 of this document.

3.2 Linked Data and Multimedia Ontologies

In this subsection, we address the data perspective from a semantic point of view, with the implications it has for exposing, organising and interlinking content at a Web scale, and in particular with respect to the area of multimedia content.

Semantic Technologies

Semantic Web technologies and solutions, which focus on formally representing semantically structured knowledge, have enabled the possibility of data to be "understood" and processed directly and indirectly by machines. The "Semantic Web" (a term coined by inventor of World Wide Web and W3C director Tim Berners-Lee) extends the network of hyperlinked human-readable web pages by inserting machine-readable metadata about pages and how they are related to each other, enabling automated agents to access the Web more intelligently and perform tasks on behalf of users.

A fundamental concept in the area of semantic technologies is that of "ontology". According to [Gruber, 1993], an ontology is a "formal, explicit specification of a shared conceptualisation", and thus ontologies are structural frameworks for organizing

[6] http://mpeg.chiariglione.org/standards/mpeg-7/mpeg-7.htm

information in a formal and reusable way. Ontology languages are formal languages used to encode ontologies. Particularly relevant for ontologies on the Web, the Resource Description Framework (RDF, [Klyne & Carroll, 2004]) is a family of World Wide Web Consortium (W3C) specifications designed as a metadata data model, currently used as a general method for conceptual description or modelling of information implemented in Web resources.

Linked Data

The most successful trend within the Semantic Web community is arguably Linked Data, a publishing paradigm in which not only documents but also structured data can be interlinked and become more useful, enabling a global data space based on open standards, namely the so-called Web of Data [Heath & Bizer, 2011]. Linked Data builds upon standard Web technologies such as HTTP and URIs (Uniform Resource Identifiers), but rather than using those to serve unstructured documents (Web pages) to humans, information is shared in a way that can be accessed automatically by computers, enabling data from different sources to be connected and queried.

The term Linked Data was also introduced by Tim Berners-Lee, in his Web architecture note Linked Data [Berners-Lee, 2006], where he formulated the now known as four “Linked Data principles”:

1. Use URIs as names for things
2. Use HTTP URIs, so that those names can be looked up
3. When someone looks up a URI, provide useful information, using standards (RDF, SPARQL)
4. Include links to other URIs, so that more things can be discovered

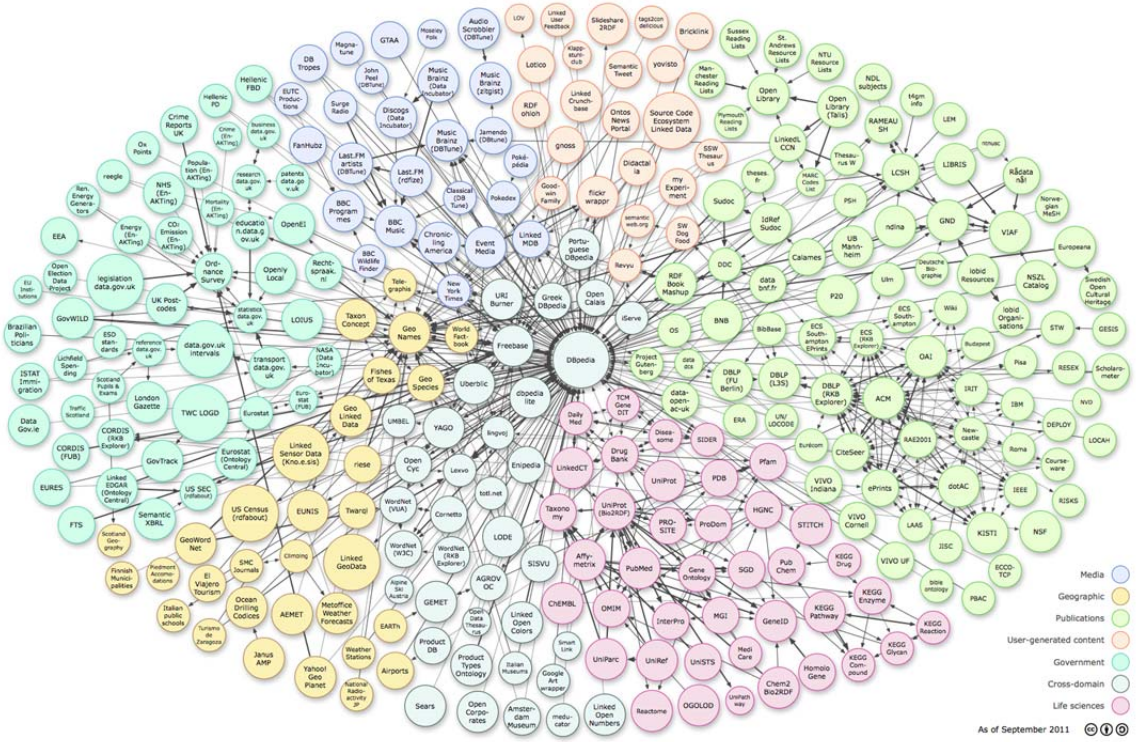

Figure 1 - Linking Open Data cloud diagram[7]

While the Linked Data paradigm has gained a huge momentum by providing means to interlink datasets (cf. Figure **1**), thus contributing to a rich user experience on the Web, methods to interlink data do not cover multimedia content in a sufficient way. [Bürger & Hausenblas, 2011] argue that interlinking multimedia requires more than just putting resources globally in relation to each other, and propose a set of principles and requirements to bridge the gap and successfully interlink multimedia content on the Web.

Moreover, multimedia content and interlinking at a Web scale comes as a particularly relevant topic given that, in words of Tim Berners-Lee again, *"the next generation Web should not be based on the false assumption that text is predominant and keyword-based search will be adequate for all reasonable purposes. (…) the issues relating to navigation through multimedia repositories such as video archives and through the Web are not unrelated (…) The Web is a multimedia environment, which makes for complex semantics."* [Berners-Lee et al., 2006]

MPEG-7

The multimedia content description standard MPEG-7, standardized in ISO/IEC 15938 [MPEG-7, 2001], is intended to provide complementary functionality to the previous MPEG standards, representing information about the content (metadata), not the content itself. The descriptions are associated with the content in order to allow fast and efficient searching for material that is relevant for the user. Therefore, the standard does not deal with the actual encoding of moving images and audio, like previous standards MPEG-1, MPEG-2 and MPEG-4 do, but it rather uses XML to

[7] LOD cloud diagram by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/

structure and store metadata, which can be attached to time instants in order to associate particular events along the duration of the multimedia asset.

The functionality of MPEG-7 is the standardization of multimedia content descriptions by using the Description Definition Language ("DDL"), a scheme for coding the description, a Description “D” consists of a Description Scheme “DS” (structure) and the set of Descriptor Values (instantiations) that describe the Data. It is worth noting that functionalities like feature extraction algorithms are not inside the scope of the standard.

COMM Core Ontology for Multimedia

The Core Ontology for Multimedia [Franz et al., 2011], based on the MPEG-7 standard and the DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering) foundational ontology [Masolo et al., 2001], enables semantic descriptions of media available on the Web to be used to facilitate retrieval and presentation of media assets and documents containing them, providing formal description of a high quality multimedia ontology that is compatible with existing (semantic) Web technologies.

COMM also considers issues like fragment identification for annotating particular subparts of the multimedia asset, e.g., regions of the image, sequences of the video.

W3C Ontology for Media Resources

The Ontology for Media Resources 1.0 (W3C Recommendation 09 February 2012) [Lee et al., 2012] is a vocabulary that aims at bridging the different descriptions of media resources, thus providing a core set of descriptive properties. It defines a core set of metadata properties for media resources, along with their mappings to elements from a set of existing metadata formats, and it is mostly targeted towards media resources available on the Web, as opposed to media resources that are only accessible in local repositories. An implementation of the abstract ontology suitable for the semantic Web using RDF/OWL is also available.

HTML5

Though not an ontology itself, it is important to consider the new syntactical features in HTML5, the latest work-in-progress revision of the HTML standard, where audio and video have become first-class citizens on the Web the same way that other media types like images did in the past. New markup elements like `<video>` and `<audio>` are semantic replacements for previous generic tags like `<object>`.

The new APIs provided by HTML5 make it easy to include and handle multimedia content, giving developers access and control to timeline data and network states of multimedia assets, like reading and writing raw data to audio files (Audio Data API) or manipulating captions in videos (Timed Track API). Additionally, Audio and Video elements can be combined with other technologies of the Web stack, like Canvas, SVG, CSS or WebGL.

Media Fragments URI

Media Fragments URI 1.0 (currently W3C Candidate Recommendation [Troncy et al., 2011]) specifies the syntax for constructing media fragment URIs and explains how to handle them when used over the HTTP protocol. The syntax is based on the

specification of particular name-value pairs that can be used in URI fragment and URI query requests to restrict a media resource to a certain fragment.

The aim of the specification is to enhance the Web infrastructure for supporting the addressing and retrieval of subparts of time-based Web resources, as well as the automated processing of such subparts for reuse. It provides media-format independent, standard means of addressing media fragments on the Web using URIs, by considering media fragments along four different dimensions: temporal, spatial, and tracks. Temporal fragments can be marked with a name and then addressed through a URI using that name, using the id dimension. While the specified addressing schemes apply mainly to audio and video resources, the spatial fragment addressing may also be used on images.

4 Semantic Annotation and Indexing of AV Content

In this section, we address the semantic annotation and indexing of audiovisual content from two different perspectives, namely via automatic annotation techniques (subsection 4.1) and through manual annotation ones (subsection 4.2).

4.1 Automatic Annotation Techniques

The techniques regard the automatic segmentation of video and assignment of semantic descriptors.

4.1.1 Segmentation of News Video

The Informedia Digital Library Project was one of the earliest projects which aimed at the indexing and retrieval of full length news broadcast video [Hauptmann & Witbrock, 1998]. The success of the project was defined as depended on successfully transcribing broadcast audio with ASR, and on broadcast video segmentation into stories useful for information retrieval. In order to maintain domain independence, researchers refrained from using broadcaster specific features such as logos, recognizing anchor faces, jingles or known timings of stories. As broadcasts were full length, part of the task also involved detecting commercials and separating them from the resultant stories. Closed caption text was aligned with ASR output to obtain an accurate transcript with timing information. The purpose was twofold: to provide timing information to closed caption text, which tended to be more accurate; and to fill in the text missing from the closed caption text. The features used for segmentation, however, were primarily visual and acoustic. On the visual side shots were identified, as well as motion activity in the shots themselves with the idea that high motion activity shots typically did not occur near story boundaries. News readers typically opened and concluded a story, and as such the identification of the news reader was also performed through a face detection and the identification of the most recurring faces, as well as clustering the color histograms of all shot key-frames, with the most frequently occurring shot representing the studio background. Finally, black frame detection was useful in identifying commercials as black frames tended to precede commercials. On the acoustic side the following features were identified: silences, with long silences indicative of story boundaries; and changes in acoustic environment, perceived as changes in background noise, recording channel, or speaker changes. Acoustic changes were clustered into a number of acoustic classes where the change in class was indicative feature. Prior to story segmentation, commercials were marked by a heuristic that identified a succession of rapidly occurring shots accompanied by black frames. Story boundaries were placed in non-commercial portions of the video where there were long silences, and by cues where indicated in the closed caption transcript.

While successful, the Informedia Digital Library Project outputs provided scope for improvement in terms of feature utilization in combination with better story boundary placement through more sophisticated machine learning approaches [Hauptmann & Witbrock, 1998]. Many of the developed research features recur in systems submitted for the TRECVID 2003-2004 story segmentation task. A selection of Informedia

systems which achieved a top-ranking in the segmentation of video news broadcasts is made below.

The National University of Singapore (NUS) Trecvid entry in 2003 proved the best performing system in story segmentation when using features from all modalities [Chaisorn et al., 2003; Chaisorn et al., 2003]. The two-tiered system, which first classified shots into seventeen categories such as "sports", "anchor", "two anchors", "people", "speech/interview", "live-reporting", "introduction", "commercial" etc. Special specific classes, such as "lead-in/out shot", and "top story logo shot", were also introduced to accommodate broadcaster specific behaviour. Shot classification used a decision tree. The features included: the background audio class, whether it was speech, music, silence, noise, speech and music, speech and noise, noise and music; the motion in the shot, ranging from high to low, shot length, number of faces present in the shot, whether the shot was a close up or farther away, and the number of lines of text present on the screen, and whether these were centralized or not. A background audio class containing speech and music would indicate an "introduction" shot, while "sports" would be accompanied by speech and noise. Low motion activity was associated with "speech/interview" shots, while high motion activity with "Sports". The number of faces detected, in combination with close-up determined whether a shot was an "anchor", "two-anchor", "speech/interview", "people" or another shot class. Centered text often was an indicator of "sports" type shots, with match scores displayed centrally on screen. The second step of the system used a HMM to perform the actual story segmentation at a shot level, based on the underlying shot class, presence of cue phrases at the beginning of the shot, and whether a change in shot class had occurred.

While the NUS system performed admirably with a F metric of 0.944 in TRECVID 2003, one constraint is its broadcaster specific dependence; both at the shot classification level and the tendency to learn a program structure at the HMM level, with an intensive annotation requirement.

The IBM system [Amir et al., 2004; Hsu et al., 2005] focused on a domain independent approach incorporating all modalities. The approach extracted features around candidate story boundaries -in this case the union of shot boundaries and long pauses. Features from all modalities were extracted, and in an early fusion approach used to train a SVM classifier. The primarily lexical features were based on prior work [Franz et al., 1999]. Cue words (learned via mutual information criterion), silence duration, and a comparison of noun distributions across boundaries were used to train three decision trees, which were then combined in a weighted scoring function. The same features were also used in a maximum entropy model, with some additional features. These included additional tri-gram cue phrases, as well as a feature modelling speaker rate, based on the idea that news readers speak faster at the start of a new story. The last features included were model broadcaster specific program structure such as specific time slots for commercials. Although the maximum entropy model generally outperformed the decision tree story boundary model, a fusion of both models performed best and was used in the video story segmentation system. Speech prosody was also considered, such as word rate, pause duration, duration of voiced segments, pitch features, and pitch slope. By extracting the mean, variance, minimum and maximum of pitch features, at either side of a candidate boundary point and at various window lengths, over 70 prosody features were considered. Visual features included the output of commercial detectors, as well as optical character recognition to

recognize short duration sports segments. The primary visual features, however, were Visual Cue Clusters, intermediate features automatically induced from raw features such as color, texture, and motion of each shot according to the mutual information they have with a target class label, in this case a story boundary.

The IBM system came as a close-second in the story segmentation task in TRECVID 2004, with the best run F metric of 0.65 incorporating all modalities. Of all text only runs, the IBM system significantly outperformed other submissions, with a F metric 0.55 score. A heavy emphasis was put on the automatic induction of features given a set of training examples such that minimal human annotation intervention was required.

The top system in TRECVID 2004 [Hoashi et al., 2004], with a F metric of 0.69 is noteworthy because their feature set entirely omits the text modality. An initial trained SVM classifier associated story boundaries with features extracted at a shot level, such as the average audio RMS (root mean square) of the shot, RMS of the first frames, audio class of the shot (silence, speech, music, noise), motion-both overall and in the horizontal and vertical components, shot duration and density, and the color components of the first, middle, and last frames. A ranking type approach was adopted, in which the start of the top-N shots were taken as story boundaries. The value of N was the average number of boundaries in the training set; the training set was broadcaster specific. A specialized SVM classifier was trained for a-typical cases within a news broadcast, such as when headlines were read out. The a-typical cases constituted separate stories, but within a single shot, and often had background music playing as well. The headline specific SVM classifier was applied on segments corresponding to the headlines, based on the detection of "jingles", the music tunes used to introduce and conclude headline sections. Another SVM classifier was trained to recognize news anchors, and story boundaries were recognized where long pauses were detected in the corresponding audio signal. A final post-filtering step removed all boundaries which were not associated with long pauses, and that were not preceded or followed by an anchor shot.

Although scoring only a median F metric at the 2003 TRECVID segmentation task, the [Besacier et al., 2004; Quénot et al., 2004] efforts only used a simple Boolean combination of feature detectors to arrive at segmentation, thus giving an insight into the relative performance gain by each feature. The union of all shot and long pauses in the video gave a recall metric of 0.963 when taken as story boundaries, and as such, these boundaries served as candidate boundaries for evaluation. A pause detector alone gave a F metric of 0.44, and this score was maintained when an audio change feature was included (boolean AND). The introduction of a jingle detector (pause and audio change or jingle) raised the F metric to 0.45, and identifying whether the news anchor was speaking (via speaker diarization) at the candidate boundary raised the F metric to 0.47 (pause and audio change or jingle || audio news anchor detection). Including cue phrases from the ASR and removing boundaries (pause and audio change or jingle or audio news anchor detection or cue phrases or commercial detection) placed in commercials gave a final F metric of 0.53.

One of the aims of the TRECVID 2003-2004 story segmentation task efforts was to successfully segment video news broadcasts relying primarily on visual and audio sources, thus omitting ASR transcript text medium [Kraaij et al., 2004]. The efforts by [Hoashi et al., 2004] and [Chaisorn et al., 2003; Chaisorn et al., 2003] illustrate that this is entirely possible. However both approaches required extensive specialization in

video detectors used; approaches that are broadcaster dependent and require extensive annotation of training examples in advance. [Chaisorn et al., 2003; Chaisorn et al., 2003] performed extensive analysis of each broadcaster to arrive 17 categories of genre type shots for which specific detectors were trained. The [Hoashi et al., 2004] system, slightly more generic, still relied on broadcaster specific jingle- and anchor detection. Only [Amir et al., 2004; Hsu et al., 2005] aimed to maintain broadcaster independence in their system development. The text feature on its own gave a F metric of 0.55, and when fused with audio and visual features a final F metric of 0.65. When also considering the [Besacier et al., 2004; Quénot et al., 2004] system, TRECVID 2003-2004 output suggests that text features tend to provide a substantive core contribution in overall news video segmentation performance. Speech prosody in the audio channel clearly plays a major role in the segmentation task; silences clearly are most discriminant, but speaker intonation as captured by numerous pitch related features also can help. An open question is how much a more extensive analysis of lexical features can contribute to the news story segmentation task, as nearly all TRECVID systems restricted themselves to only utilizing cue phrases.

4.1.2 Sport Video Analysis

Sports video analysis is sometimes presented as a segmentation task, this is true insofar that a sports game is divided up into the underlying match events. These sporting events can serve as browsing indexes, and give a semantic understanding of events at the corresponding temporal position in the video. A summarization engine may select particularly exciting events, such as when a goal is scored in a football game, to provide viewers with a concise overview of the most salient events, which are also referred to as highlights. In contrast with content-based approaches for news video segmentation described in section 4.1.1, the approaches in sport video analysis often use specialized models designed to capture sporting events defined by prior knowledge of a game structure or production effects. As such, they can also be seen as a form of multi-modal pattern recognition albeit in a more specialized domain-restricted setting. Low level features such as colors and motion in images, or pitch and spectral shape in sound, are often extracted and used in an intermediate representation. For example, color may be interpreted as a particular view of the court or field of play, sound may be characterized as excited cheering or normal match commentary. In combination, these intermediate features may be used to infer match specific events.

The inter-modal collaboration strategy for semantic content analysis in broadcast sports video was designed to analyse sports video, specifically baseball and American football [Babaguchi & Nitta, 2003]. Highlights were detected by examining the text stream for domain specific keyword phrases such as "touchdown" and then finding the corresponding time interval in the video stream. Crowd cheering was determined by the short time energy feature of the audio stream. Using the idea that crowd cheering was indicative of a highlight moments, a more sophisticated detector was developed by excluding highlights without cheering. A Bayesian network was used to classify the closed caption text into segments which were either "live", "replay", "commercials", or "other". These segments were then aligned with the video stream, to identify the corresponding visual segments. "Live" segments were further annotated by player names and the type of plays occurring. In a final step, external knowledge sources were consulted to augment events missed in the closed captions. These were synchronized with the existing visual stream by means of OCR of the in-game time on

screen overlay. This system is a typical example of how specialized domain knowledge can readily provide a successful solution for game specific indexing and annotation. The extensibility of the system is however open to question.

The event detection in basketball video using multiple modalities system adopted a rule based approach to describe structural basketball events (section beginning, section ending, in play, and out of play) along with five regular game events (jump ball, foul, penalty, shot, and goal) [Liu et al., 2006]. Video shots were classified by certain viewpoints, and typically associated with specific events. In combination with audio clues, such as speaker excitement or the whistle of a referee, basketball events could be inferred.

Kijak segmented a tennis game video using a hierarchy of Hidden Markov Models that characterized basic tennis game structure and TV production rules [Kijak et al., 2003]. Audio events consisted of Gaussian mixture models that classified the audio stream into "speech", "applause", "ball hits", "noise" and "music". Visual features consisted of shots, their duration, shot dissolve detection, and similarity to a global view model. All features together trained HMMs that classified a shot as either "missed first serve", "rally", "replay", and "break". Shots were combined in an overarching HMM which modelled the structure of a tennis match in terms of points, games, and sets.

Sadlier presented a framework for analysing a common class of sport, field sports, such as soccer, rugby, hockey and Gaelic football [Sadlier & O'Connor, 2005]. The video stream was segmented into shots, and commercial segments were removed as pre-processing. From low level features, specialized detectors were created to recognize a player close up (skin tone and shirt color), crowd detection, speech activity detection, change in on screen graphic, such as when the score count changed, logo presence detection (typical during an update), motion activity detection, and field orientation. The output from these detectors formed a set of intermediate features that described a match. Feature fusion was performed at a shot level, and the combined set of features trained SVMs for various sporting events such as (goal, tries, penalties, etc.). The dataset consisted of three different field sports and demonstrated the feasibility of using feature detectors common to multiple sports within the field sport video domain.

The TIME framework adopted a multi-modal approach to address context- and time synchronization common in the news video and sports (soccer) domains [Snoek & Worring, 2005]. TIME segmentation evaluation used three classifiers, C4.5 decision trees, maximum entropy, and SVMs. The choice for statistical classifiers was made in order to provide for a robust performance in domains such as soccer, where events are sparse, context dependent, and unpredictable. Likewise the TIME framework also provided for accurate fusion on the time domain, such as in the news domain, where structure and thus time synchronization is of greater importance. Low level concept detectors operating on the video stream detected various multi-modal events, such as camera shot type, microphone shot, text shots, panning camera, speech, speech excitement, motion intensity, close-up, and goal related keywords. These low level features had additional context information and when added by temporally relations using the Allan time relations [Allen, 1984] (precedes, meets, overlaps, starts, during, finishes, equals), thus produced events. Events were assumed to always have at least a time distance due to noise. If events were separated by an interval, then events were assumed to have no temporal relationship with each other. High level semantic concepts were thus modelled as a combination of time ordered low level events within

a certain interval. This exercise in pattern recognition was performed by a classifier. In the soccer domain, high level concepts were detected for goal, yellow card, substitutions. Of the three evaluated classifiers, C4.5 decision trees gave the poorest performance on the soccer domain. Maximum Entropy (MaxEnt) and SVM algorithms detected all semantic events equally well. What differentiated the algorithms was that the SVM classifier required considerably less training time than the MaxEnt algorithm to achieve results. The SVM algorithm outperformed the C4.5 and MaxEnt algorithms, both performed similarly, in the news domain, where events such as reporting anchor, monologue, split-view interview, and weather-report were sought. In an additional experiment to test the effectiveness of the TIME framework, the SVM based classification on the news domain was performed with temporal relations enabled and disabled. For most semantic concepts, the additional information provided by the TIME framework yielded increased performance, except for the weather report, where results were comparable. The merit of this work lies in the fact that it demonstrates that it is possible to add additional contextual information, in this case a temporal ordering, to low level features. This additional information results in better performance of the classifier than when it is not provided. It also shows the effectiveness of SVM classifiers over C4.5 decision tree- and MaxEnt classifiers in two different domains.

A similar approach was taken for the development of a baseball detector [Fleischman et al., 2007; Fleischman et al., 2007]. Distinct unimodal detectors formed intermediate level features out of the low-level data streams. For example, a shot was characterized as either a pitching scene, a field scene, or other scene. Camera motion was also estimated, pan, tilt, and zoom. Cheering, music, or speech was detected in the audio stream. In [Fleischman et al., 2007] decision trees were used to learn temporal feature representation for baseball events (home run, outfield hit, infield hit, strikeout, outfield out, infield out, and walk) in a discriminative setting. In [Fleischman et al., 2007] chi-square analysis was performed to automatically learn significant baseball events based on repeated temporal-feature sequences, and then map events to words from the closed caption transcripts, thus permitting later retrieval. Bertini et al. used Finite State Machines (FSM) to model events in sport games [Bertini et al., 2005]. Camera motion, play field zone estimation, and the speed and position of athletes were extracted from low level visual features. These values were encoded as state transition conditions in a FSM, thus modelling game events. Time constraints, such as the before and during operator defined by Allen [Allen, 1984], could also be imposed on state transitions for imposing a temporal ordering.

Xu et al. proposed a system which is somewhat domain (team-sport) independent while capable of handling semantic events that do not have significant audio/video features, such as when players are given yellow/red cards in soccer when most audio/video patterns are insufficiently distinct to recognize such semantic events [Xu & Chua, 2006]. The system approach detected generic video concepts, using HMM, from the audio/video stream, such as shot category, focal distance, special view category, field zone, camera motion direction, and motion activity. Another HMM classifier was used to detect the transition between these events in the video stream. Domain dependencies were introduced in the form of external text streams detailing game rules (important for field type and match duration), player names (facilitates text analysis), and event types (linking event types with audio-visual patterns detected by the HMM), used to detect more detailed semantic concepts. The assumption was that only noteworthy events were included in, for instance, a match report. The more

detailed semantic concept events were aligned against the generic video events detected. Xu compared three fusion methods, a rule-based scheme, a probabilistic aggregation scheme, and one using Bayesian inference. The rule based scheme aligned text events with the number of matches between text events and the domain specific model events, provided externally, within a temporal window. Knowing that the text stream may contain more detail than the video stream, additional events located in the text stream, such as the example with the yellow/red cards (which does not appear as an event type in the video), might be determined in the video stream. Text and video stream events were usually misaligned by some offset, depending on the nature of the sport. For example, in soccer, the match report transcription was offset from the accurate video stream, because of the time lag in the human transcription process. The aggregation scheme models offsets as a likelihood problem. Xu's system allows for a reasonably generic system for sports video analysis, with good precision and recall metric. The systems supports extension, a caveat is that every sport needs external, domain specific parameters. Xu argued that this data can often automatically be retrieved and parsed, whereas event models are non-volatile after construction. Provided there is some operator assistance to develop these models, the system developed could support a large number of sports.

Typically in sports video analysis, unimodal semantic detectors were created which capture intermediate concepts such as crowd excitement, position in a playing area, or camera motion. These were linked temporally, either through hand-crafted rules or through machine learning, to classify the desired sporting events. The amount of supervision was substantial, both to create intermediate semantic detectors, and then to link them, and it was questionable whether most frameworks were sufficiently robust to handle variations in broadcaster production style apart from redeployment for a different sport. Fleischman [Fleischman et al., 2007] was an exception, as he learned intermediate level features through sequence mining.

4.1.3 Scene Segmentation in Video

Video segmentation has often been justified from an information retrieval- as well as a browsing point of view when video is partitioned into semantically coherent units. News video is a fairly specialized domain characterized by high information content and number of features unique to the domain (anchor detection, silences, jingles) which can be exploited for the segmentation task. There are many other video genres, for example Youtube videos, movies and TV series, sports broadcasts, and documentaries. Current research focuses on defining generic techniques of video segmentation in semantically coherent segments that work for different genres. Often the term 'scene' segmentation is used in this setting.

Scene detection using film production rules

Alatan et al. perceives scenes as the resultant of a post-editing process in a studio, and as such targets the movies and television series domain [Alatan et al., 2001]. A Hidden Markov model is used to identify scenes. Alatan models a scene as consisting of three elements, people, conversation, and a location. People are detected using face detection, while audio is classified as either music, speech or silence. Shifts in location are detected by analyzing the histograms of several consecutive shots. The results of each detector are then used as inputs of a Hidden Markov model to detect, and classify, shots as either establishing, dialogue, or transitional, the three types most

commonly used by film directors. A scene is defined as starting with an establishing shot, followed by a sequence of dialogue shots, and concluding with a transitional shot.

Also [Tavanapong & Zhou, 2004] model a scene in terms of shots produced using common continuity- editing techniques from film making. These are determined using region specific color descriptors, which characterize the type of shot. Scene boundaries are established when a transitional shot is detected.

Li et al. provides a definition for salient scenes in movies suitable for retrieval [Li et al., 2004]. These higher semantic scenes are ones which contain 2-speaker dialogue, multiple-speaker dialogue, and hybrid events. Shot segmentation via color histograms and audio classification followed by speaker diarization form the audio-visual features. Graph analysis of the sequence clusters allows for matching with the pre-established models (i.e. 2-speaker dialogue etc.).

Graph based scene detection

Yeung et al. perform segmentation using a graph based approach, referred to Scene Transition Graph (STG) segmentation [Yeung et al., 1998]. In this methodology, clustered shots form vertexes in a graph. A directed edge is drawn from one vertex to another representing the video progression, one shot transitioning to another. Edges which, if removed, divide the graph into two disconnected graphs are known as "cut-edges". After removing all cut-edges from the STG, each disconnected sub-graph represents a scene, with boundaries at the cut-edges. A scene consisting of multiple shots presents as a cycle in a STG. Rasheed et al. uses a similar graph-clustering approach [Rasheed & Shah, 2005]. Shots are linked based on a similarity function weighted by the temporal proximity. In contrast to [Yeung et al., 1998] who partitions the graph using complete links, the graph is partitioned recursively using normalized cuts. In addition, the initial shot clustering is performed using a similarity function influenced by a decaying temporal distance function, while in [Yeung et al., 1998] the similarity function is only applicable within a temporal window. A cut is the summation of weights associated with the edges being removed, and edges where this is at a minimum are candidates for removal. In order to find a global optimum solution, the association degree also has to be considered. This is defined as the total connection (e.g. weights) from all nodes in the proposed sub-graph with respect to all nodes in the parent graph. The formulation for the normalized cut presents as the cut cost as a fraction of the total edge connections (association degree). By minimizing the normalized cut values in a recursive bi-partitioning procedure, a video can be partitioned into scenes.

Sidiropoulos et al. improves on the STG approach of [Yeung et al., 1998], by including an additional set of features [Sidiropoulos et al., 2009]. Both [Yeung et al., 1998] and [Rasheed & Shah, 2005] used shot similarity only, Sidiropoulos however includes audio similarity in the form of speaker diarization and the background class. Audio and shot similarity jointly improve on the visual only approach using complete linkage of [Yeung et al., 1998].

Other methods

A Markov Chain Monte Carlo approach for finding scene boundaries is used by [Zhai & Shah, 2005]. In this approach, scene boundaries are randomly placed and then moved

until equilibrium is reached in the Markov process. Scenes can be merged or split by the shifting of boundaries, based on a likelihood function comparing visual similarities.

Chen et al. perform scene determination after first segmenting a shot into a foreground and background region, of which only the background is relevant, based on the perception that the background scenery stays consistent within a single scene, while foreground objects may move around [Chen et al., 2008]. The foreground and background segmentation is done by analysing motion vectors within a shot. An image mosaic is built up from all frames in a shot, representing all static and background objects in the shot. Similarity is computed by comparing the average of four spatial features computed on each image mosaic for a shot. Inspired by film production rules, scenes are formed by examining the shot similarity over three consecutive shots. If the first and third shots are similar, all three shots are merged together to form a scene; if only the first two are similar the third shot starts a new scene; else the next scene starts at the second shot.

Goela et al. use a supervised approach, sampling visual and audio features around scene changes in a variety of video genres [Goela et al., 2007]. The features extracted include the MFC coefficients, the type of audio class (music/speech/laughter/silence), presence of shot cuts, and motion and pixel level differences surrounding a scene boundary. These were then used to train a SVM. Cour et al. use scripts and closed captions common to movies and tv shows as a source of textual information regarding scene transitions [Cour et al., 2008]. Scene transitions are inferred using a generative framework, based on visual features, and then aligned using dynamic programming against cues in the text source. Chasanis et al. operates purely on the visual modality, using sophisticated features such as SIFT and contrast context histogram (CCH) [Chasanis et al., 2009]. The concatenation of the two forms the representative feature vector for a shot, which is then mapped onto a set of visual words (bag of words approach) to form a shot histogram. A temporal smoothing kernel is applied, so that the shot histogram is smoothed with the histogram information of neighbouring shots, thus preserving some context information. Changes in visual word content indicate potential changes in scene, and these are identified by finding the local maxima of the Euclidean distance between successive smoothed histograms.

4.1.4 Concept Detection in Video

Traditional image features such as color and texture, or text descriptors due to social media tags or file descriptor, do not adequately describe the semantic content in an image or video. This is known as the semantic gap, which can be formulated as [Smeulders et al., 2000]:

> *"...the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation."*

In order to bridge this semantic gap in video retrieval, intermediate representations are created which describe the low level multimedia features. These intermediate representations are known as semantic concepts, which provide a text annotation of the underlying content. These concepts can be related to objects, such as "airplane" or "car", scenes, such as "city scape" or "desert", people, such as "Bill Clinton" or "female human close-up", acoustic, such as "speech" or "music", genre, such as "weather" or "sports", and production "camera motion" or "blank frame" [Hauptmann et al.,

2007][Chang et al., 2005]. Ontologies which define concept categories for video collections include the LSCOM ontology LSCOM Lexicon Definitions and Annotations Version 1.0, DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia 2006 which compromises over 2600 concepts although only around 300 exist in the TRECVID 2005-2009 dataset, and the MediaMill Challenge [Snoek et al., 2006] which defines 101 concepts over the same dataset. The Trecvid 2010 Semantic indexing task defined 130 semantic concepts, growing to 346 in 2011, which include all concepts used in earlier TRECVID efforts and some from the LSCOM ontology. Relations relating concepts were also provided [Over et al., 2010]. In the TRECVID semantic indexing task, concepts are learned from low level multimedia features in a supervised setting, typically using SVMs based on annotated examples. It should be noted in the semantic indexing task, the presence of each concept is assumed to be binary, i.e., it is either present or absent in the given shot, and a concept is present in a shot if it is present in a single frame within the shot. Most TRECVID participants treat concept detection as a single frame image analysis task.

Detecting concepts in TRECVID

A typical approach for end to end concept indexing system is the top performing MediaMill system [Snoek et al., 2010] in the TRECVID semantic indexing task. Salient points which are robust against viewpoint changes are identified using a Harris-Laplace point detector. Dense sampling is also performed for concepts like scenes, which have many homogenous areas. Sampling is extended over several frames beyond the current keyframe under analysis, and spatial pyramids [Lazebnik et al., 2006] are applied over sparse and dense keypoints to aggregate the different resolutions. Opponent- Sift and RGB-SIFT [Sande et al., 2010], and SIFT features are extracted from around the sampling points, and are quantized into a visual code book, which represents a compact representation of an image frame. Previously [Snoek et al., 2010] the visual codebook was constructed by k-means clustering and hard assignment, but recent findings suggest soft-assignment gives better results [Gemert et al., 2010]. A separate code book is constructed for every combination of feature, sampling method, and assignment approach. Training a concept detector then involves learning the optimum combination of features (codebooks) using a support vector machine classifier with a $\chi 2$ kernel, which has been shown to outperform the RBF kernel [Zhang et al., 2011] favored in earlier iterations of TRECVID. A variation on this approach by the best overall entrant in 2011 [Inoue et al., 2011] use SIFT features extracted at Harris-Affine and Hessian-Affine interest points, SIFT and hue histograms and HOG with dense sampling, and HOG from temporal subtraction images. Additionally, MFCC features are extracted from the audio channel. A Gaussian Mixture Model super vector is created as a codebook to combine the low level features prior to training with a SVM using a RBF kernel.

The relationships between concepts can also be leverage to improve detector performance. For example, the "Athlete" and "Basketball" concepts are related to "Sports", while "Hill", "Landscape", "Outdoors", and "Sky" are related to the "mountain" concepts. Various strategies are described in e.g. [Wei et al., 2009]; [Kennedy & Chang, 2007].

The semantic concepts described in this section represent an essential step for the understanding of video. In their own right, concepts provide a semantic annotation of a

video. However, concept relations can be used to infer additional- possibly more abstract- concepts due to their relationship in an ontological hierarchy.

Event recognition

Since its inception, concept detection is cast as a single image analysis problem, even if the sought after concept can be perceived as having a temporal duration (i.e. "people-marching", "music"). Event recognition is a temporal extension of concept detection, and covers a range of events, such as human activities like 'running', 'drinking' or smoking', but also longer duration sequences, such as 'baking a cake', or 'building a shelter'. These already represent two distinct fields of research: human motion analysis, where the focus is on the development of spatio-temporal features capable of capturing human motion over a sequence of frames, and event recognition, where events occur with the duration of a scene, and can be perceived as a sequence of concepts.

Human Motion Analysis

Although human actions can be readily recognized within a still image by a human observer, the distinction is more difficult for a machine. Consider the case of a figure with a hand extended. He could potentially be waving, shaking hands, punching, smoking, or reaching for a phone. Without resorting to external sources, such as a movie script as in [Cour et al., 2008], computer vision techniques for the recognition of human actions focus on the recognition of characteristic or periodic motion that describe each type of action.

In [Schüldt et al., 2004], space-time interest point features (STIP) [Laptev, 2005] are used to detect local motion patterns. These are adapted to fit the underlying image structure in space and time, and are made invariant to camera motion using the technique of [Laptev & Lindeberg, 2004]. The spatio- temporal neighbourhoods of local features contain information about the motion and the spatial appearance of events in image sequences, and are clustered using K-means clustering to form a set of primitive event descriptors. A sequence of these descriptors forms an alternate histogram descriptor. A video database was created consisting of the following human actions: walking, jogging, running, boxing, hand waving and hand clapping. Local features, histogram of local features and marginalized histograms of normalized spatio-temporal gradients are compared using a SVM classifier and a nearest neighbour classifier. STIP features trained using a SVM performed best. One of the issues with this work is that the dataset is synthetic, the actions are recorded in controlled and simplified settings. Later work in [Laptev et al., 2008] provides better features, based on spatial pyramid features [Lazebnik et al., 2006] extended temporally, significantly outperforming the earlier work of [Schüldt et al., 2004] on the 6-action dataset. Also a framework for generic action recognition is provided, where actions are automatically elicited from movie scripts and used to provide training examples. To show the efficacy of the approach, the following actions are defined in a realistic dataset consisting of Hollywood movies: AnswerPhone, GetOutCar, HandShake, HugPerson, Kiss, SitDown, SitUp, StandUp. A detailed overview of the problems faced in human motion analysis, as well as a comparison of various approaches is provided in [Aggarwal & Ryoo, 2011].

Event Recognition in TRECVID

The TRECVID 2010-11 event detection task on the other hand defines actions on a larger scale, as a longer temporal sequence over multiple shots, such as (in TRECVID 2010) "assembling a shelter", "baking a cake", and "driving a runner in" (this refers to baseball, when a batter can score for his team by hitting the ball such that a runner can reach the home plate). TRECVID 2010 defines events such as "Birthday party", "Changing a tire", "Flash mob gathering", "Getting a vehicle unstuck", "Grooming an animal", "Making a sandwich", "Parade", "Parkour", "Repairing an appliance", "Working on a Sewing project".

The best system of the TRECVID 2010 Multimedia Event Detection Task (MED) task was the one described in [Dantone et al., 2010], which used a spatio-temporal Hes-STIP detector [Willems et al., 2008] to find spatio-temporal blobs within a video, based on an approximation of the determinant of the Hessian. These were quantized into a visual vocabulary prior to training with an SVM.

Some of the 2011 TRECVID MED submissions greatly expanded the features considered. The best performing system [Natarajan et al., 2011] extracted a wide variety of multimodal features for training. These include SIFT, SURF, D-SIFT, CHoG for appearance models; for color features RGB-SIFT, OpponentSIFT, C-SIFT, STIP and D-STIP for spatio-temporal features, and MFCC, FDLP, and Audio Transients for audio. These were computed within the context of a spatial pyramid [Lazebnik et al., 2006] before being projected into a visual codebook. Additional high level features used include object and scene detectors, and salient keyword detection in ASR and OCR. Various combinations of low-and high level features are trained using an SVM to form individual subsystems, which are then combined using late fusion. The best run used a Bayesian Model Combination for late fusion, although a close second was achieved using weighted-average fusion.

[Xu & Chang, 2008] define temporal events from the LSCOM lexicon that exhibit a temporal duration, and examine the events: “Car Crash”, “Demonstration or Protest”, “Election Campaign Greeting”, “Exiting Car”, “Ground Combat”, “People Marching”, “Riot”, “Running”, “Shooting”, and “Walking”. These events are recognized via a technique called Temporally Aligned Pyramid Matching (TAPM), which combines the idea of pyramid matching [Lazebnik et al., 2006] over multiple sub-clips. Alternatively, earth-movers-distance (EMD) is used to compare a sequence of clips. Three low level features Grid Color Moment, Gabor Texture, and Edge Direction Histogram are used as input to train three indepent SVMs on 374 concepts from [Yanagawa et al., 2007] and then fused. The concepts form an intermediate feature representation for a temporal event. Analysis of the results show that single key-frame concept analysis was outperformed by EMD, which was in turn outperformed by the multi-level TAPM technique. In [Duan et al., 2011] this was extended to learn events from a collection of YouTube videos.

In a similar approach, [Xie & Chang, 2006] used sequences of intermediate concept representations from the TRECVID 2005 dataset in a hierarchical hidden Markov model, to define upper level concepts. [Bailer, 2011] instead defines a feature sequence kernel, where feature sequences are compared using the Longest Common Subsequence algorithm. Results show that sequence kernels outperform single key-frame approaches for concept classification.

Human motion analysis is a form of temporal event recognition, but is typically handled by spatial temporal features over successive frames [Schüldt et al., 2004]; [Laptev et al., 2008].

Abnormal event recognition in video is the application of computer vision to the field of surveillance, and extends basic temporal event recognition with some domain specifics. In works such as [Zhao et al., 2011]; [Si et al., 2011]; [Gupta et al., 2009]; [Cui et al., 2011]; [Zhang et al., 2011], a camera is continuously recording from a fixed viewpoint. After recording for a sufficiently long time, ordinary behavior, the typical actions of people with their environment is modelled to form an event vocabulary during a training phase. This can be considered as a kind of sequence of correct states. Anomalous behavior is characterized deviations from the learned temporal event model.

4.1.5 Semantic alignment in Video: Names and places

Once a video has been segmented into scenes, one of the more pertinent pieces of information is determining who is in a scene, and finding out where. Typically this kind of information is presented in visual, textual, and aural modalities, but in a complementary fashion. For example, in a news story, a specific location may be mentioned by a news reader in a studio, but only after some time will a shot showing that location appear on screen. In a television series, multiple characters may engage each other in dialogue. They may address each other by name, making the on-screen visual identification of a character a disambiguation problem, or they may be referring to another off-screen character, whereupon it becomes an alignment problem. The lack of synchronization between modalities when a location or person is mentioned makes their identification both an issue of temporal alignment and of disambiguation.

Determining location

By alignment

Several approaches annotate video with geographic information. Christel extracted named entities, which were geographic references, from news video transcripts and any on-screen text via OCR [Christel et al., 2000]. Morphological processing ensured that similar words resolved to the same place name, e.g. "Canadian" and "Canada". These words were then matched against a gazetteer containing 300 countries, states, and administrative entities and 17,000 cities and their associated longitude and latitude coordinates. Whenever a location reference occurred in the video, the spatial coordinates were projected on a map display. A similar attempt to associate a geographic entity mentioned in the transcript of news video and the shot that shows the location was made by [Yang & Hauptmann, 2006]. Extracted location named entities and the shots part of the news item were considered. The problem of synonymous location names, such as "Holland" and "Netherlands", was resolved by consulting a gazetteer, which merged synonymous names into a single entity. Location polysemy, when multiple locations with the same name, e.g. London, a city in Ontario, Canada, or the city in the United Kingdom, was also resolved by consulting the gazetteer for multiple location references and disambiguated using the contextual information in the surrounding sentences of the transcript. A trained SVM classifier, using various features, aligned the list of candidate locations with the shots in the news story. Temporal features were extracted to explore whether a location occurred

before, within, or after a shot (the time difference between a shot and the nearest mention of a location), and how close a location is to a shot compared with other locations in the same story. The syntactic role of a location term in a sentence was also a relevant feature. Locations mentioned in prepositional phrases were more likely displayed than when they occurred as a subject/object or as a modifier. Although OCR was unsuitable for providing candidate location names due to spelling errors, the appearance of overlay text, which often did contain the true location name, was utilized for computing the edit distance against all candidate names as a separate feature. Speaker diarization was also used to distinguish between the news reader, narrator, and reporter, versus news subjects, as it was observed that the former were more likely to mention a visible location, and thus speaker identity was also considered. Genre detection was applied and used as a feature, as certain kinds of news stories were more likely to contain shots which had a possible location labelling, such as politics, whereas others, e.g. business or health stories were less likely to contain location specific shots. A trained SVM classifier generated a probability of a candidate location appearing in a shot using all these features.

Engels et al. in contrast adopted a weakly supervised approach using a latent topic model to generate a topic distribution for the annotation of locations in the television series, "Buffy the Vampire Slayer" [Engels et al., 2010]. Episodes were split into scenes, from which topic distributions were generated using Latent Dirichlet Allocation (LDA) based on words in unstructured fan-supplied episode scripts. Terms were weighted by their probability of being a location reference, and then distributions were modified based on the visual similarity between scenes. Visual similarity was computed purely on parts of the image with people excluded, as people were common to all scenes. Location descriptions were propagated for scenes lacking accompanying text through visually matching. The multi- modal approach combining LDA and visual similarity propagation successfully provided location annotations despite the challenging domain of a television action series.

From natural images

Location can also be extracted by other means than the alignment between visual and text modalities. One innovative approach by [Wang et al., 2011] describes a technique for reading words contained in natural images. This allows for the understanding of street signs or billboards, thus providing a semantic connotation for the underlying location or building present in the scene. Examples provided describe signs such as ’Orpheum Theatre’, ’San Diego Automotive Museum’, and ’Garage’, which immediately gives the notion of the semantic class of the associated location. Recognizing the specific name of the institution potentially permits the cross-referencing of its geographic position by looking it up using a facility such as Google Maps.

Names and Faces

Detecting whether a person is present in a scene is the first step necessary towards discovering their identity. This is done by scanning every frame of a video with a frontal face detector. Over the years many methods have been proposed, however, Viola used a cascade of weak facial feature classifiers to find faces in real time [Viola & Jones, 2004]. Face detection, recognizing the presence of a face, is a simpler sub-problem of the more difficult face recognition task. In an authentication scenario, a

single known face is matched against a claimed identity, while in a recognition scenario an unknown face is matched against the set of known faces [Bowyer et al., 2006]. Recognition performance relates to illumination (various lighting conditions), pose (view point by which the head is viewed), facial expression (a frown versus a neutral expression), occlusion (a hat or glasses), and ageing effects [Abate et al., 2007]. Most facial recognition efforts focus on 2-dimensional facial images, although 3-dimensional facial models are becoming increasingly popular and it is expected that soon 3D only or joint 2D-3D models will outperform 2D only approaches. A large number of facial recognition techniques are surveyed in [Bowyer et al., 2006] [Abate et al., 2007] [Zhao et al., 2003].

Yang described an approach to find the shots where individuals named in news broadcasts appeared [Yang et al., 2004]. Transcripts formed the primary source of information, and shots in which a person was named formed likely candidates. Neighbouring shots were also included, because text and vision often function asynchronously and the person may appear in an earlier or later shot. The inclusion of neighbouring shots was modelled using a Gaussian model, where the probability of a person appearing in a shot decreased as the duration increased from when the individual was mentioned. Anchor detection was applied, on the assumption that a named person appearance was unlikely in a shot containing the news reader. Face recognition was applied, using externally obtained images of the named individual for matching against candidate shots. Linearly combining results from the facial recognition, anchor detection, and Gaussian text search provided the shots of the individuals named in the news broadcast.

Others focused on the task of identifying characters in a film or television series rather than news video [Everingham et al., 2009]. This was made more challenging because every on screen character must be labelled individually, which posed a disambiguation problem as multiple characters may be on screen at the same time. Speech recognition output aligned with fan made scripts formed the source text. The source text provided information about which character was on screen; the timing information and script lines indicated when a character was speaking. Face detection was applied to all frames for face tracking. This procedure maintained the correspondence of a detected face across consecutive frames, even if a face was not continuously detected due to variation in pose or expression. A character's clothing was also used as a feature. While characters may change their clothing within an episode, similar clothing still served as an indicator of character identity. Clothing was represented as the color histogram of a bounded region extracted relative to the detected face region. Within shot speaker identification was attempted by aligning information about who was speaking in the script with lip movement detection in face tracks. Face tracks may be associated with multiple identities due to misalignment with the transcript. The face tracks with the most probable speaker identifications were taken as exemplars for character identification. The remaining unlabelled face tracks then were assigned a character labelling based on their probabilistic similarity to the exemplars, which considered the similarity of the facial descriptors and of their clothing.

Pham learns face names occurring in news video using a semi-supervised method based on label propagation [Pham et al., 2010]. An initial seed training set was created by annotating several unlabelled faces. In a random walk process, faces were iteratively annotated, based on the visual similarity between labelled and unlabelled faces. Performance was increased by removing frequently occurring faces such as the

anchor man prior to the propagation step. The semi-supervised approach outperformed a SVM trained classifier on the initial training set.

Determining who appears in a video, and where appearance occurs, is a significant semantic aspect of video analysis. Research to-date shows that the news video- and television show broadcasts provide adequate information in the textual and visual domains that such determination can be reduced to an alignment challenge.

4.1.6 Video Fingerprinting

Digital video fingerprinting refers to techniques in which computer software identifies and extracts characteristic components of a video, ultimately being able to uniquely identify the analysed video by its resultant “fingerprint”. Video fingerprinting technology is therefore useful for identifying and comparing digital video data, and has proven effective to detect if a particular segment of video is based, totally or partially, on another original video within a reference set.

Video fingerprinting techniques have been used for various purposes (see “Video Fingerprinting Applications” below) and typically combine a wide range of visual video features (see “Video Fingerprinting Techniques” below) including key frame analysis, colour and motion changes during a video sequence, etc.

The creation of a video fingerprint involves the use of software that decodes the video data and then applies several feature extraction algorithms. One important characteristic of video fingerprints is that these are highly compressed when compared to the original source file; they can therefore be stored for later comparison in a more efficient manner. Of course, as a form of lossy video compression, video fingerprints are not suitable for reconstructing the original content.

Video Fingerprinting Applications

Video fingerprinting techniques are useful in a wide range of scenarios [Schavemaker et al., 2009]. A domain in which these come up as particularly relevant is the media domain. The most obvious media process that benefits from this technology is copy detection for finding copyright infringement, like the illegal re-encoding of movies, or even movie extracts [Law-To et al., 2007]. However, this technology can be applied for different purposes, for example for identifying known content in an unwanted open access repository. Besides, video fingerprinting is used for monitoring (in particular, for counting) the broadcasts of advertisements in TV, and for the detection of changes in those advertisements.

Besides the more straightforward applications, new manners of leveraging video fingerprinting are appearing recently. One of these novel applications is the linking of archive footage to finished television productions. For example, this method is able to show the link between a documentary and all the clips in the archive it consists of. By identifying the re-use of archive material, documentation may be applied from the complete programme to source clips or the other way round. This permits users browse through collections of video in an advanced manner, as currently the metadata is not able to provide such detailed information.

Furthermore, within the emerging trend of Smart TV applications, and in particular relating to Interactive Television, video fingerprinting is highly relevant as it is able to automatically recognize the content being featured, enabling the possibility of

interactive functionalities. In a similar manner, it is also able to support content-aware advertising.

Video Fingerprinting Techniques

Several methods have been explored for the extraction of unique features of a digital video that can be stored as a fingerprint of the content, in a way that evaluation and identification of video can be performed by comparing the fingerprints. However, it is worth pointing out that for digital video data, both the audio and the video image can be fingerprinted, having individual significance for different application areas. (Cf. Section 5.3.1 for audio fingerprinting.)

Additionally, it should be noted that the increasing number of videos currently available online due to the development of user generated content sites (for instance, in average 72 hours of video are uploaded to YouTube every minute[8]) presents video fingerprinting technologies with a scalability challenge.

With respect to the work carried out for detecting the similarity of video clips, many are the approaches that can be highlighted. The most relevant ones usually apply still image features for clustering key frames of video sequences. However, other approaches use video oriented features such as sequences of key frames and camera motion (cf. [Zhu et al., 2005]) or motion trajectories (cf. [Chang et al., 1997]). These approaches are optimized for scalability rather than for precision in terms of matching, and can be used as a pre-processing step to find candidates to be matched with further processing.

Many relevant approaches in the area of video fingerprinting can be mentioned from within the copy detection task in TRECVID 2008, the main points of which we briefly address below in order to highlight the wide range of techniques explored with respect to the topic:

[Kucuktunc et al., 2008] use different MPEG-7 visual features (e.g., ScalableColor, ColorLayout, Color-structure, Homogeneous Texture and EdgeHistogram) as signatures, comparing the signatures of every first frame each two seconds in the query video to the centre frames of every shot in the data set. The approach of [Chen & Jiang, 2008] is focused on intensity information, classifying key frames of a segment into classes with different template labels. [Gao et al., 2008] segment the video temporally by extracting features for each segment, including the global intensity histogram, intensity ordinal measurements and enhanced local colour histogram. [Zhang et al., 2008] present two approaches for video fingerprinting (MPEG-7 visual descriptors extracted from spatial-temporal elements vs. shot lengths as signature of each video) and a simple combination for video and audio fingerprints (multiplying confidence values of the two approaches), using the cumulative shot length sequence for matching. [Héritier et al., 2008] use a "bag of visual terms" for representing key frames. [Douze et al., 2008] use local features extracted from key frames, either on a regular basis or based on motion activity. [Joly et al., 2008] work on two approaches (trajectories of interest vs. a extensions on still image processing) together with a way of applying them sequentially to improve the results. [Gursoy & Gunsel, 2008] apply Nonnegative Matrix Factorization to each video frame and use dimensionality reduction to get matrices of rank 2, which are then matched. In [Le et al., 2008] a local

[8] http://www.youtube.com/t/press_statistics

feature based approach is used, extracting patches along the trajectory of tracked feature points. In [Koskela et al., 2008], a “best matching unit” signature is calculated for each frame, based on ColorLayout and EdgeHistogram and self-organizing maps (SOMs) trained on the extracted features. [Liang et al., 2008] sample key frames every 20 frames, extracting up to 512 SURF descriptors from each key frame, iteratively refining match position and range on individual frame descriptor matches to discard outliers. [Orhan et al., 2008] focus on normalized Hu moment invariants (NHMI), extracted globally from every frame, using the 2nd to 6th moments, normalized by the power of the 1st one in order to accommodate for gamma and quality changes, creating a signature of six integers per frame from those moments.

Some results from the further research carried out in this area after the aforementioned TRECVID 2008 video copy detection task should be highlighted: [Bober & Brasnett, 2009] propose an image signature based on the trace transform, extracted globally as well as locally around points that are robust in scale space. Finally, [Döhring & Lienhart, 2009] propose a system for mining broadcasts for recurring video sequences based on gradient histograms, assuming that no transformations have been applied to the clips.

4.2 Manual Annotation Techniques

Automatic annotation tools have already proven to be valuable information sources. Though in recent years, there is a lot of activity in the field of automatic semantic concept detection (see Section 4.1), the performance of these tools decreases dramatically when the number of concepts increases [Hauptmann et al., 2007] [Hanjalic et al., 2008]. Bridging the semantic gap (i.e. translating the features to semantic concepts) when the number of requested concepts is large is a task in which human annotators still outperform automatic tools.

Simple manual annotation of content can be performed quite easily. All that has to be done is looking at the content, writing information about the content in a text file and storing the text file together with the content. This kind of annotation has been in existence for decades (be it creating a text file on a computer, or writing notes on cards). However, this kind of annotation has a lot of drawbacks, these techniques are very inefficient, miss a lot of structure, usually don’t have an underlying (meta-)data model, are very dependent on practices of individual annotators, are unable to model time-coded metadata in a uniform way, etc.

Luckily, over the recent years, new annotation tools for manual annotation have been created. These tools can be categorised in several ways.

4.2.1 Repurposing existing metadata

Before diving into the manual annotation process and manual annotation tools, we first want to draw attention on metadata sources that are also manually generated and are often overlooked.

One form of manually generated metadata that is often overlooked is metadata that already exists before the content is created, or metadata that is created during the production process. In a broadcast environment, a lot of this kind of metadata is generated. Each programme comes with a script, scene descriptions exist for fiction programmes, news rundowns are created for broadcast news shows, moreover,

articles are written with the news items for use on the web portals, subtitles are generated etc… Of course, it is obvious to also include this data with the annotation of the corresponding content.

Though this metadata is a very important information source, there arise several problems when this is applied into practice.

There are however some obstacles for using pre-existing metadata in practice. The main reason is that the tools used to create this metadata are usually stand alone products, often from different vendors. Moreover they usually use different proprietary formats to store the metadata they generate. Therefore, integrating these metadata sources in the production process is a very difficult task.

4.2.2 Manual annotation

There are many different forms of manual semantic annotations. But in general the process of manual annotation consists of applying a selected set of content descriptors onto the media item that is annotated, possibly combined with a time frame in which the descriptor is applicable.

Tools for manual annotation

In completely manual annotation all the initiative for annotating content is left to the user. The computer system is restricted to displaying the content. The simplest form of a manual annotation tool is a text editor. However, this is not an ideal annotation tool.

In the next paragraph, we present a brief overview of several state-of-the-art manual annotation tools. For more extensive information about these tools, we refer to [Dasiopoulou et al., 2011], or to the referenced websites of the tools themselves.

- **Anvil**[9]**.** Anvil is a free video annotation tool, supporting multi layered annotation based on a user-defined coding scheme in XML format. Development has started in 2001 and it has been further extended as part of the PhD thesis of M. Kipp [Kipp, 2005]. It supports both descriptive and structured metadata. The annotation can be applied to temporal segments or to the entire video. It also supports spatiotemporal annotations and annotations attached to specific points in the video. Annotations are performed on hierarchical user-defined levels. The temporal segments on which an annotation applies have to be manually selected by the user. Other features include integration of data coming from phonetic tools such as PRAAT.
- **VIA**[10]**.** VIA stands for Video Image and Annotation tool. Annotations are based on an OWL ontology created by the user. At the start of the annotation process, the selected OWL ontology is loaded. Annotations are possible in several granularities: entire video, video segments, moving regions, entire frames and still regions within a frame. The annotation interface consists of three main panels: “Regions”, “Shots” and “Video”. In the region mode, the annotator manually draws rectangular regions in the video and tags the region with concepts from the ontology. In “shots” mode, the annotator manually selects the boundaries of the shot and then selects the applicable concepts from the

[9] http://www.anvil-software.de

[10] http://mklab.iti.gr/project/via

ontology. In the "video" mode, it is possible to apply concepts to the entire video.

- **VideoAnnEx**[11]**.** VideoAnnEx is an annotation tool developed by IBM. It supports annotation based on concepts. These concepts are obtained from an XML lexicon, that can be predefined or user generated. VideoAnnEx supports annotation of frames, shots and regions. The shots can be obtained from an automatic shot detection tool, or can be user defined. Support for the VideoAnnEx tool has ended.
- **Ontolog**[12]**.** Ontolog has been developed by Jon Heggland as part of his PhD research [Heggland, 2005]. It is a tool for annotating video and audio resources using ontologies. It contains some default description schemes, but it also supports user generated ones. It has a straightforward way to create simple ontologies. The interface contains a preview window and a list of so called "annotation strata". These strata contain concepts that can be further divided in subconcepts. On the time axis, segments can be selected manually to connect the concept with the specific time interval in the video.
- **Advene**[13]**.** Advene stands for "Annotate Digital Video, Exchange on the Net" and is created by Aubert et al. at Université Claude Bernard [Aubert & Prié, 2007]. The goal of the tool is to provide a model and a format to share annotations of video documents. It also supports editing and visualising what they call "hypervideos" (videos together with the annotation). The annotation is performed according to user defined annotation schemes. A distinction is made between schemes that define concepts and schemes that define relationships between concepts. Several MIME types to store the annotations are supported, including XML.

  A particularity in this tool is that annotations can be displayed in several forms. Next to the standard timeline view, a "tree-view" and "transcription-view" can be used. Though annotation can be performed using Advene, it is not its main task. The main task is displaying annotations that already exist.
- **ELAN**[14]**.** Elan is a professional tool for the creation of complex annotations of video and audio resources, developed at Max Planck Institute. Unlike many other recent annotation tools, it supports only descriptive annotations, i.e. sentences, words or a feature description. It provides also support for vocabularies to facilitate the user to enter annotations. The annotation can be performed on multiple layers called "tiers", which can be hierarchically connected. Different tiers can for instance be used for different languages.
- **EXMARaLDA**[15]**.** EXMARaLDA stands for Extensible Markup Language for Discourse Annotation. As can be derived from the name, it is mainly meant for annotating dialogues and transcriptions.
- **SVAS**[16]**.** SVAS stands for Semantic Video Annotation Suite. It has been developed by Joanneum Research [Schallauer et al., 2008]. It consists of two

[11] http://www.research.ibm.com/VideoAnnEx/

[12] http://www.idi.ntnu.no/~heggland/ontolog/

[13] http://liris.cnrs.fr/advene

[14] http://tla.mpi.nl/tools/tla-tools/elan/

[15] http://www.exmaralda.org/

[16] http://www.joanneum.at/en/digital/products-solutions/semantic-video-annotation.html

main components. The first one automatically extracts shots and keyframes, and the second one is the annotation tool itself. The annotation tool can be used to edit the result of the automatic analysis and to add descriptive metadata to the annotation. These annotations are stored in the MPEG-7 ISO standard. This standard enables to tag several concepts, such as persons, places, events and objects. These tags can be applied to entire shots or to specific regions within the shot. One specific feature of this tool is object re-detection, which automatically detects recurring appearances of already tagged objects.

- **MacVisSTA[17].** MacVisSTA stands for Macintosh Visualisation for Situated Temporal Analysis. This system focuses on a very specific type of annotations in video, namely behavioral annotation (speech, gaze, gesture,…). It allows annotators to annotate segments with the type of user interaction that can be observed in that segment.

Tool	Descriptive metadata	Structured metadata	Shot detection / keyframe extraction	Time-coded annotations	Spatial annotations	Format
Anvil	X	X		X		
VIA	X	X		X	X	XML
VideoAnnEx	X	X	X	X	X	XML MPEG7
Ontolog	X	X		X		RDF
Advene		X	X	X		XML
ELAN	X			X		XML
EXMARaLDA	X			X		
SVAS	X	X	X	X	X	XML MPEG-7
MacVisSTA		X		X		

Table 1: Overview of Manual Annotation Tools

4.2.3 Collaborative annotation

With collaborative annotation, we mean that several annotators work together in manually annotating content. Examples include systems where annotations added by one annotator are verified by a second one. In another workflow, several annotators could work on different aspects of the annotation synchronously or asynchronously.

[17] http://sourceforge.net/projects/macvissta/

An example of such a system is the Co-Annotea system by [Hunter & Schroeter, 2008]. Their annotation system is not limited to videos but can also be applied to audio, text, html, and many other data types.

Some of the manual annotation approaches described above also support asynchronous collaborative annotation to some extent. Examples are Eva and VideoAnnEx.

4.2.4 Hybrid annotation

One of the main disadvantages of manual annotation is the speed at which annotation can be made. Although the software systems described have improved this efficiency a lot, they do not use any existing technique for automatic annotation, except for some low-level features such a shot detection.

Recently, new hybrid approaches have been proposed by several researchers. In these systems the annotator and computer system work together to obtain annotations better and faster. This hybrid process is sometimes also called semi-automatic annotation or synergistic annotation.

The idea of semi-automatic annotation is not new. E.g. in 1997, Carrer et al. proposed an annotation tool in which some tasks were performed automatically [Carrer et al., 1997]. Their system is called Vane. It performs some annotation tasks automatically, such as shot detection and keyframe extraction, and detection some low-level visual descriptors in the content. Based on this information a human annotation could then proceed by including high-level semantic features in the annotation. Note that shot detection and keyframe extraction, though quite novel at that time, are now used in many manual annotation tools.

Zhu et al proposed a system in which annotations entered by an annotator can be automatically applied to video segments that are similar in nature [Zhu et al., 2002].

Another example of such an annotation system is the CASAM system introduced by Creed et al. [Creed et al., 2010]. Their proposed system is meant for journalists and other people working at a news redaction. First, the system attempts to automatically annotate everything it can in the video (people, objects, events, emotions,…). When it encounters concepts that it does not understand, it poses questions to the user.

4.2.5 Crowdsourcing Approaches

While automatic annotation techniques for multimedia content have largely improved during recent years (cf. Section 0), arguably the quality of annotations performed through manual labor still ranks higher (cf. Section 0). However, manual annotation techniques do typically require the involvement of expert users in tasks such as the ones explained above. The problem is that having users fully devoted to annotation tasks is not usually possible, as they are effort- and time-consuming, and thus it is necessary to devise new methods of involving users on these.

Several methods have been explored to encourage users collaborate in the process of annotating, around the concept of “crowdsourcing” [Howe, 2006], which we address in this subsection. Crowdsourcing, a portmanteau of the terms "crowd" and "outsourcing", refers to a distributed problem-solving and production model, where tasks typically performed by single individuals are split into smaller tasks and then sourced to a community of users via different techniques.

Arguably, the term itself has become some sort of buzzword after its use on many different contexts and the hype built around it. Still, the philosophy behind the approach is worth considering, and extremely valuable for gathering semantic annotations and new knowledge, as demonstrated in vertical domains such as Cultural Heritage [Oomen & Aroyo, 2011].

We cover below different approaches towards crowdsourcing, with a particular emphasis on the benefits for semantic annotation.

Human Computation

One of the most outstanding examples of crowdsourcing at a Web scale is the reCAPTCHA [Ahn et al., 2008] project, eventually acquired by Google in 2009. The motivation behind reCAPTCHA is that humans collectively spend a huge amount of time solving CAPTCHAs (Completely Automated Public Turing tests to tell Computers and Humans Apart), which are security measures that websites use to prevent automated programs from abusing online services by asking humans to perform a task that computers cannot perform, such as deciphering distorted characters.

reCAPTCHA channels these human actions (otherwise wasted time of "human computation") into a useful purpose: helping to digitize old printed material; users are asked to decipher scanned words from books that computerized optical character recognition have failed to recognize.

This way, while users are performing small tasks for a particular interest (in this case, prove they are human in order to access a particular website), at the same time their human computation actions are leveraged for different purposes (in this case, for digitizing books).

Crowdsourcing Marketplaces

A different approach of getting users performing tasks in order to leverage human intelligence is through way of crowdsourcing systems [Surowiecki, 2004], marketplaces where "requesters" can programmatically post tasks that "workers" can perform for a small amount of money.

The most prominent example of this approach is Amazon's Mechanical Turk[18] (MTurk), originally developed for in-house use, and launched publicly in 2005. Tasks in MTurk are referred to as HITs (Human Intelligence Tasks), and can be posted to the marketplace by requesters by making use of the API provided by the service. Workers can select which HITs they are interested in, and receive rewards on completion.

Another example of an enterprise crowdsourcing platform is CrowdFlower[19], which has a contributor base of around 1.5 million users worldwide and offers the ability to distribute tasks to them such as product categorization, business listing verification, and SEO content creation, with automated management and quality control.

Tasks performed through such crowdsourcing Internet marketplaces range from identifying performers on a music album, to performing a drawing (e.g., a sheep[20]). This approach has also been adopted for semantic annotation and management

[18] http://www.mturk.com

[19] http://crowdflower.com

[20] http://www.thesheepmarket.com/

purposes, such as in [Simperl et al., 2011]. Additionally, recent studies [Parent & Eskenazi, 2011] show an increasing use of crowdsourcing techniques for speech tasks, especially in terms of speech acquisition, speech labeling (transcription and annotation) and assessment of dialog systems. Examples of recent research in the area include reports on efficiently eliciting speech transcriptions [Kunath & Weinberger, 2010] and speech-accent ratings [Novotney & Callison-Burch, 2010].

The main concern with respect to the use of such crowdsourcing techniques, and indeed a challenge for such approaches, is the risk of obtaining low quality output as a consequence of the purely economic motivation for users, who may find financial incentives to cheat on tasks. In addition, it can be argued that speech tasks may become expensive due to the high volume of data [Parent & Eskenazi, 2011].

Games With a Purpose (GWAP)

One particular approach towards human computation is the concept of Games With a Purpose (GWAP), initially proposed by Louis von Ahn [Ahn, 2006]. This revolves around the idea of having users performing those actions that computers typically are not good at, such as identifying concepts in a picture or a video, or identifying a song, while they are playing a game.

Several of the following examples are available at gwap.com:

- **ESP Game**[21] – Users play this game in pairs, but they cannot communicate with each other. The system presents them both with the same image and they need to type words that describe the image; when they write the same word, they earn points. The game takes advantage of the human computation as it collects the tags that the users have implicitly agreed upon as descriptive for the given image.
- **Peekaboom** – Two players play this game: the “boom” is presented with an image and a description of an object in it, so he selects the area of the image where the object is, making it visible for the other player, the “peek”, who has to guess the object so both gain points. The game leverages the human participation to generate maps that relate object with regions of images representing them.
- **Squigl**[22] – Two players are presented with the same image and a word, and they need to trace the object described by the word in the image, earning more points the closer both traces are. This game is then also able to generate object-region maps.
- **Phetch** – One player, the “describer”, is presented with an image that he can describe with entire sentences, so the rest of players (the “seekers”) are able to find the image through the use of search engines. This way, the game is able to collect descriptive captions for labeling images suitable to assist sight impaired readers.
- **Verbosity**[23] – One player, the “narrator”, is assigned a word that he needs to describe to the other user, the “guesser”, filling some sentence templates. The

[21] http://www.gwap.com/gwap/gamesPreview/espgame/

[22] http://www.gwap.com/gwap/gamesPreview/squigl/

[23] http://www.gwap.com/gwap/gamesPreview/verbosity/

game is thus able to collect common-sense knowledge about the described objects.

- **TagATune**[24] – Two players listen to the same clip of a music tune, and they need to figure out if they are listening to the same song or not by describing it. The game is then able to collect descriptive tags related to the audio files.
- **Matchin**[25] – Two players see the same image, and decide which one they like most. If they agree, they earn points. Through this game, the system is able to collect subjective opinions on images.
- **FlipIt** – A user is presented with sixteen images, although "turned down" so he doesn't see them unless he flips them in turns, trying to find similarly looking images. Through the user's interactions with the game, the system is able to identify images that look similar.
- **PopVideo** – Two players watch the same video and identify the objects that show up on it; when they agree at a close time, they gain points. The concept is similar to the ESP Games for single images, only that in this case, for the collection of annotations on a video over time.
- **Fold.it**[26] – Users try to fold proteins, scoring based on how well each of the proteins is packed; the smaller, the better. The eventual goal of the game is to have human folders work on proteins that do not have a known structure. Humans are better than computers at certain aspects of protein structure prediction, thus the game is able to provide more efficient protein detection and discovery.

Semantic GWAPs

The idea of leveraging the interaction of users for semantic annotation purposes has been explored in research trends that include the use of GWAP approaches to create and curate semantic content [Siorpaes & Hepp, 2008]. A survey on GWAPs for knowledge acquisition [Thaler et al., 2011] covers most of the games addressed above as well as many others. Also, those listed below are examples of GWAPs with a focus on semantic technologies:

- **OntoTube** – Two players watch the same video, and then the system asks them questions based on tags and descriptions found in YouTube. If they agree on the answers, they earn points. Based on the responses, the game is then able to gather insightful annotations for the video.
- **OntoPronto** – Two players are presented with the first paragraph of a Wikipedia article, and some related concepts of the Proton ontology; they have to decide which one suits better to the article, and gain points if they agree. This way, the game is able to annotate the article with relevant concepts.
- **SpotTheLink** – Two players are presented with an entity from DBpedia, and try to select fitting concepts from the Proton ontology, and the relations amongst them, gaining points if they agree. The game is able to align ontologies through the interactions of the players.

[24] http://www.gwap.com/gwap/gamesPreview/tagatune/

[25] http://www.gwap.com/gwap/gamesPreview/matchin/

[26] http://fold.it

- **BetterRelations**[27] - Two players are presented with a concept from DBpedia, along with two different facts about them; they need to decide which one is more important, and if there is agreement they gain points. The game is able to perform and ordering of relations of a concept based on their importance, what they call “common sense ordering”.

Particularly relevant for TOSCA-MP are the games that involve the annotation of images (identifying concepts inside them) and videos (identifying concepts inside them along their duration). As we have covered, they typically require at least two players, and therefore the knowledge acquisition is performed based on consensus.

Gamification

A recent trend for engaging users and encouraging them to perform certain desired tasks is the “Gamification” approach, not to be confused with GWAP. Gamification refers the use of game design techniques and mechanics, such as the handling of achievement “virtual badges” or leader boards based on the actions of the users.

Gamification techniques foster user participation by taking advantage of the psychological predisposition of humans to engage in gaming, encouraging them to perform certain tasks that they would otherwise would consider boring. The most prominent (and earlier) examples of the gamification approach being applied can be found on location-based platforms such as Foursquare[28] or Gowalla[29] (aquired by Facebook[30] in 2011, and recently discontinued as a standalone service), where users receive incentives for willingly providing information on their location and the places they go to.

[27] http://lodgames.kl.dfki.de/betterRelations/

[28] http://www.foursquare.com

[29] http://www.gowalla.com/

[30] http://www.facebook.com/

5 Semantic Retrieval of AV Content

In this section, different approaches towards semantic retrieval of audiovisual content are discussed. Concept-based (semantic) video retrieval is well represented in literature; for instance, [Snoek & Worring, 2009] review 300 references arguing that current text-only solutions for video search engines are unsatisfactory and showing concept-based promising alternatives.

Due to the nature of the semantic search paradigm, proper annotations over multimedia content are necessary in order to retrieve the desired results. As described in Section 4, the annotation and indexing methods can be either automatic or manual. Importantly enough, the Linked Data paradigm has also been explored within the area of video semantic retrieval. [Waitelonis & Sack, 2009] present a prototype implementation of exploratory video search and show how traditional keyword-based search can be augmented by the use of Linked Open Data.

The section is structured as follows: In subsection 5.1, we cover speech-oriented retrieval of audiovisual content, which is based mostly on speech-to-text transformations. In subsection 5.2, multimodal approaches are discussed, covering “known-instance search" (KIS) and "instance search" (INS) topics of TRECVID. Finally, in subsection 5.3, we discuss audio-similarity approaches towards the retrieval of audiovisual content.

5.1 Speech-Oriented Retrieval of AV Content

The most direct approach towards automatically annotating and indexing audiovisual content is the use of the speech (when available in multimedia content) in order to obtain the words speoken along the duration of a given multimedia asset. In this case, the extracted text can be used to automatically index the assets with respect to an ontological model by semantically analysing the text and performing Named Entity Recognition [Nadeau, 2007] on it. (The complementary approach, i.e., natural language queries by applying speech-to-text techniques on the searcher speech, is also possible but not necessary.)

Based on this approach, we cover below the most important aspects of semantic retrieval of content (subsection 5.1.1), followed by the state of the art on automatic speech recognition, regarding the speech-to-text step (subsection 5.1.2), and finally particular approaches that combine both areas (subsection 5.1.3).

5.1.1 Semantic Information Retrieval

Automatic indexing is able to relate particular parts along the duration of a multimedia asset with different semantic terms, hence facilitating the retrieval of the exact part of the speech where the searched terms, or related ones, are spoken. Retrieval of such content is then to be performed though the same means as with other type of semantic annotations in nonspeech scenarios.

Semantic search is able to deliver more precise answers than traditional Information Retrieval (IR) technologies that follow lightweight syntax-centric models such as the predominant keyword paradigm (i.e., queries based on keywords, which are matched against a bag-of-words document representation), because of the more expressive models used that enable more complex queries. Database (DB) and Knowledge-

based Expert (KB) systems go beyond the mere retrieval of documents (document retrieval), delivering more precise answers (data retrieval).

The technical differences between the traditional IR paradigm and DB and KB systems fall into three main dimensions: the “query model”, as the representation of the user needs; the “data model”, as the underlying resources; and the matching technique. By using more expressive models for the representation of the user needs in the query model, and through the structure and semantics inherent in the data model, DB and KB systems allow for more complex queries, facilitating the retrieval of concrete answers that match them.

Semantic search can be considered as an information/content retrieval paradigm, which can be triggered by a query in natural language, to which semantic technologies (via an ontological model) are applied. Semantic search systems can combine a wide range of techniques, ranging from statistics-based IR methods for ranking, DB methods for efficient indexing and query processing, up to complex reasoning techniques for making inferences.

5.1.2 Automatic Speech Recognition

The identification of language in speech, i.e., mapping the words being spoken to their textual representation, falls under the research area of Automatic Speech Recognition (ASR), which has been deeply studied over since the 1970s [Rabiner & Juang, 1993], and such technologies are usually referred to as “speech-to-text”. However, despite the many achievements in the field, ASR still remains far from being a solved problem [Baker et al., 2009].

During the past decades, there has been a considerable growth in terms of availability regarding infrastructure and research tools supporting ASR, including Carnegie Mellon University Language Model (CMU LM) toolkit, Hidden Markov Model Toolkit (HTK), Sphinx, and Stanford Research Institute Language Modeling (SRILM). This has enabled a growing common speech corpora for speech training, development, and evaluation which has permitted automated systems to increase proficiency. Additionally, the character of recorded speech has evolved from limited and constrained speech materials to great quantities of progressively more realistic and spontaneous speech.

In terms of algorithms, the usual ASR paradigm since the early 1970s (and still predominating nowadays) has been focused around the use of statistical methods [Poor, 1988], especially through stochastic processing with Hidden Markov Models (HMMs) [Baker, 1975][Jelinek, 1976]. Different approaches have been taken within this paradigm, including algorithms like expectation maximization (EM) [Dempster et al., 1977] and forward-backward or Baum-Welch [Baum, 1972] to train HMMs from data, or decision trees [Breiman et al., 1984] to categorize sets of features, such as pronunciations from training data. Deterministic approaches include corrective training [Bahl et al., 1993] and some neural network techniques [Beaufays et al., 2002] [Lippman, 1987].

In addition to ASR, the complementary approach of “voice recognition” refers to the functionality of identifying the speaker (the act of finding the identity of the person saying the words, and not what the speaker is saying), which is also helpful metadata to consider about the multimedia content.

One particular subarea of speech recognition worth highlighting is Keyword Spotting, which deals with the identification of keywords in utterances. Depending on the scope

of the keyword spotting, two different approaches are to be considered. Either keyword spotting in unconstrained speech, and in isolated word recognition. In particular, unconstrained speech appears when keywords may not be separated from other words, and no grammar is enforced on the sentence containing them. Some algorithms used for this task include sliding window and garbage model, K-best hypothesis, and iterative Viterbi decoding [Silaghi et al., 2000].

5.1.3 Semantic Technologies Applied to Speech Retrieval

Efforts have been taken in the direction of providing an ontology-based retrieval of human speech. [Tejedor et al., 2007] present a semantic based search model for human speech corpora, stressing the search for meanings rather than words. Their proposed framework covers the complete recognition/retrieval cycle, from word spotting to semantic annotation, query processing, and search result presentation.

Some successful techniques applied for the mentioned approach include dependency parsing and spell correction (as an attempt at correcting the miss-spelt words), custom stemming and inflectional normalization (a term normalization process to reduce words to their linguistic root is usually needed in an Information Retrieval system) and salient term/phrase selection [Chaisorn et al., 2010].

However, issues in the speech-based retrieval of video have been highlighted [Nock et al., 2003], including the relationship between speech transcription accuracy and retrieval performance, query processing schemes and the critical problem of mapping between cues in speech and the relevant video shots. More sophisticated schemes involving up-front video ranking and other multimodal approaches, which are addressed in depth in the subsection 5.2 below, show more promising results.

5.2 Multimodal Approaches

5.2.1 Video Search, Exploration, and Navigation

A video search and indexing system is largely characterized by a number of underlying component modules; a feature extraction module, a semantic indexing module which applies machine learning and fusion strategies to the extracted features, and a search interface. The premier conference where video indexing systems can be compared against each other is TRECVID. As such, the best performing participants of the TRECVID 2010 and 2011 known item search task are described. These searches can be performed automatically, returning a large ranked list of relevant shots, or interactively, with the aid of a human operator. As such, there are various aspects which all influence the overall retrieval performance. For automatic searches, these can be the choice of query to concept mapping, or fusion strategy of the various concept detectors and modalities. In the interactive search, the features used, query method, the browsing interface, and the expertise of the searcher also play a role.

Originally started as the video track of TREC, (text retrieval and extraction conference) in 2001, it became an independent event in 2003 and has been held yearly since then. Semantic indexing or tagging refers to the automatic recognition of certain concepts, typically scenes, objects, events, and people. This task was originally known as the "high-level feature extraction" task, but was renamed to "semantic indexing" in 2010. The concepts used are a subset of the Large Scale Concept Ontology for Multimedia (LSCOM) ontology, which was developed by archivists based on an analysis of

Internet query logs [Kennedy et al., 2006]. Video search similarly was known as the "search" task, and split into two separate tasks, "known-instance search" (KIS), and "instance search" (INS). The known instance search task simulates the scenario where a searcher believes a relevant video exists in the archive, but can only formulate a textual description of its contents. The system takes this textual description as input, and returns the desired video either automatically within a top-100 confidence score ranked list of videos, as in a web search engine, or by way of an interactive search process with a human operator [Smeaton & Over, 2010]. The instance search task, on the other hand, starts with a visual example, marked in the image, of a (possibly unique) person, character, location, or object, and requires the automatic retrieval of a ranked list of up to 1000 shots, by likelihood of containing the example query. Characteristic of this task is the lack of examples (a single video clip), an emphasis on real-time behaviour, and an emphasis on retrieving specific (named) entities. The instance search task involves retrieving a visual exemplar in related video, and tends to rely heavily on visual only techniques[Kraaij & Over, 2010]. In contrast, the known item search task involves formulating a textual query to retrieve a ranked list of video results. Salient developments in the KIS task are described [Smeaton & Over, 2011].

Most TRECVID groups used an open source text retrieval system, Lemur[31], Lucene[32], or Terrier [Ounis et al., 2006] to index and query text sources. Some groups also performed their own ASR, and fused this with the provided ASR, and some others used OCR to find additional text on screen.

Known Item Search

For this task, a textual description was given describing the topic sought in a video collection. These consisted of some word or phrases describing the target video, and a list of words describing visible people, places and objects in the video. Some example topic queries follow:

- 0001 KEY VISUAL CUES: man, clutter, headphone QUERY: Find the video of bald, shirtless man showing pictures of his home full of clutter and wearing headphone
- 0002 KEY VISUAL CUES: Sega advertisement, tanks, walking weapons, Hounds QUERY: Find the video of an Sega video game advertisement that shows tanks and futuristic walking weapons called Hounds.
- 0003 KEY VISUAL CUES: Two girls, pink T shirt, blue T shirt, swirling lights background QUERY: Find the video of one girl in a pink T shirt and another in a blue T shirt doing an Easter skit with swirling lights in the background.
- 0004 KEY VISUAL CUES: George W. Bush, man, kitchen table, glasses, Canada QUERY: Find the video about the cost of drugs, featuring a man in glasses at a kitchen table, a video of Bush, and a sign saying Canada.
- 0005 KEY VISUAL CUES: village, thatch huts, girls in white shirts, woman in red shorts, man with black hair QUERY: Find the video of a Asian family visiting a village of thatch roof huts showing two girls with white shirts and a woman in

[31] http://www.lemurproject.org/

[32] http://lucene.apache.org/

red shorts entering several huts with a man with black hair doing the commentary.

The evaluation used three metrics: (a) mean inverted rank (MIR) of known items found (0 if not found), if the run was interactive, rather than automatic as above, the mean inverted rank was the fraction of topics found, (b) mean elapsed time (mins) and (c) user satisfaction (interactive) (1-7(best)). MIR is also known as mean reciprocal rank. The reciprocal rank of a query response is the multiplicative inverse of the rank of the first correct answer. The mean reciprocal rank is the average of the reciprocal ranks of results for a sample of queries, Q.

Two categories existed for this task, one based on an interactive search and one where the entire search was automated [Smeaton & Over, 2011].

Sjöberg et al. performed two automated runs and two interactive runs. For text search, the Lucene search engine was used, with ASR text and production metadata as input [Sjöberg et al., 2010]. The MediaMill concept detectors were used as they had proven to the best overall in the previous year. Various search index approaches were tried; with or without stemming, stop words, and WordNet synonyms. Using sample queries, the best performance settings used stemming, no stop words, and without synonyms. A comparison of a single search index for all textual data versus separate indexes for the ASR and production metadata showed that a single index had better performance. The metadata contributed significantly towards good retrieval performance. A word list was generated for all concepts learned for the semantic indexing task; the concept name itself was used as an initial word and then expanded using WordNet synonyms. The obtained list of synonyms were then manually edited; words with too broad meaning were removed, e.g. the concept "people marching" was set to be activated for the word "march" but not for "people" in the textual query. The search process then became a matter of matching textual queries to the expanded concept word lists which in turn activated the appropriate detectors. This approach was very successful in text search. Two approaches to improve the text-only baseline were adopted; one approach combined detector (visual and textual) scores in a weighted average, the second re-ranked the text baseline based on the visual concept detector scores.

For the first approach, the weights in the geometric mean of concept detector scores were computed analogous to the inverse document frequency,

$$\mathrm{wiidf} = \log \frac{\#\mathrm{shots}}{\#\mathrm{shots\ with\ concept\ i}}$$

To guard against videos with many concepts in them, a different run used tf-idf type weights where the term frequency was computed as the

$$\mathrm{w(i,j)tf-idf} = \frac{r_{(i,j)}}{\sum_k r_{(k,j)}}$$

where $r(i,j)$ is the detector score for concept i in video j, computed over all the concept scores in video j. The final score then was the arithmetic mean of the visual concept scores and the Lucene (text) scores, weighted 15% and 85% respectively.

The second approach re-ranked the text search results using visual concept scores. The six best performing systems on the semantic indexing task each provided a ranked list of shots for every concept. These lists are merged by a heuristic approach which goes through the system specific lists, in the order of best performance, and takes the maximum TRECeval rank over its shots in the first list where the video is

encountered. The total number of shots from that particular video in all lists is added to that video specific score. Then, when a query is given and matching concepts identified, the relevant lists are merged with the text only search results with the final video score r_j, where $r_{(j,text)}$ is the text search outcome, and s_j, is the rank of the video in the video list merged over concepts and w an experimentally determined weight of 0.1.

$$r_j = r_{(j,text)} + w\frac{1}{s_j^2}$$

The mean inverted rank scores were computed as follows: for the first approach which used a idf-weighted geometric mean of concept scores and the Lucene search results, 0.260 (6th overall). When the concepts were weighted using tf-idf, the result was 0.265. The second approach, using concept based reranking gave a score of 0.264 (4th overall). The text-only baseline scored 0.266, and a concept only search score 0.003.

For the interactive search, the user was presented with the results from the automatic run using the weighted concept scores (approach 1). The basic interface presented each video as a 4x4 collage of keyframes. Two interfaces were investigated; the first (normal) interface allowed for a detailed inspection of a video, for example to inspect the production metadata, view enlarged keyframes, or to play the video. The second (fast) interface did not allow a detailed view. As the time allotted 5 minutes, the expectancy was that the fast interface would allow the user to process more searches. The fast interface successfully found 11 out of 24 queries (MIR scores of 0.455), while the normal interface only found 7.6 out of 24 queries (0.318). The user satisfaction was 5 out of 7 for the fast interface, and 6 out of 7 for the normal interface. The expectation that the fast interface would permit the processing of more search results was not supported; on the failed queries, i.e. where the searcher could not find the video in the allotted time, the average number of videos was approximately 500 videos for both interfaces [Sjöberg et al., 2010].

The best performing system in the KIS task was by [Chaisorn et al., 2010; Wan et al., 2011]. They focused heavily on applying natural language processing and information retrieval techniques to improve query formulation and to fix mistakes in the various text sources. They also derived additional text information by combining the existing ASR text with the output from their own ASR system, and by using an OCR system to capture on screen text. The OCR text at times proved a useful contributor to the description of the video contents when the ASR and metadata failed. Lucene was used to perform indexing and searching of text data. Only the information contained in the title, subject, and description fields of the production metadata was used. For pre-processing, query statements and metadata are checked for spelling errors, and the three most likely alternative words were appended to the query or document. A part of speech tagger is used to ensure that named entities are not altered, although this can only be done for the query, which is guaranteed to be well formed. As part of the spelling correction, words such as "talenthotel" have also been split into constituent sub words, i.e. "talent hotel", provided that the sub-words are also correct words. These improvements gave a 20% increase over a regular text only baseline. A further 20% improvement was realized by using a better inflection normalization system than Porter stemming. The approach adopted here extracted all nouns and verbs in the query statements and stemmed them using a dictionary of common English inflectional words. The resultant terms and their inflectional forms were consulted during the

indexing process. A further 10% increase in retrieval was realized by identifying salient terms in the query and weighting those accordingly. This was done because the meta data text, as it tends to be short, and so there were only few relevant keywords to match with. The next step aimed to find the most salient terms in a query, as per [Bendersky & Croft, 2008]. The motivation for this step was the example query “Find the video talking about George Bush and Patriot Act showing people walking on street with tape over their mouths.” In this query, the main objects of search are “George Bush”, “Patriot Act” and “people with taped mouths”. The phrases that describe the (visual) motion characteristics of the people are secondary. The assumption made is that a user would typically not describe such a motion description in the metadata. Instead, they suspect that a typical user would likely describe the video using high level concept words such as “George Bush” and “Patriot Act” or “Silent protest”, etc. As such, the entire phrase describing the motion characteristics of the people could be effectively removed from the query statement, with little risk of reducing retrieval performance. Salient terms are identified and favored by weighing proper nouns (named entities)-identified through the POS tagging- more heavily when issuing requests of the retrieval engine. It is also likely that the these terms/phrases identified as salient would also appear under the metadata <title> or <subject> fields. The retrieval of these salient phrases were enhanced by the placement of a dedicated index on the associated meta data fields. Queries involving spoken phrases were issued to a dedicated ASR index when phrases like ’narrated’, ’saying’ were encountered. In the presented video retrieval system, the query-weighting rules were hand-coded based on provided example queries. The system described above, only indexes the metadata. When ASR and OCR text is included, retrieval performance increases further 15%. Two ASR system outputs were concatenated together, to form a single text source, and a separate index created. Emphasis was made that OCR text provided significant contribution, independent of the ASR and metadata text sources.

Further improvements were realized by including the output of 130 visual concept detectors provided by Columbia University. To compensate for varying detector performance, the detector precision at rank 100 of retrieved shots was used as a measure of their accuracy and became the resultant fusion weight of the detector. A rule based approach was used to add the necessary concept detectors, e.g. if the query contained the term ’car’, the ’car’ or ’road’ detectors would be activated. Adding semantic concept detectors gave a further 3% performance increase. The system score achieved using concept detectors and text ranked second overall, with a MIR score of 0.454.

A silence/music/other classifier was also used to index each video [Tan et al., 2010] so that if a query contained the term ’singing’ or ’music’, the audio index would be activated. Classifying metadata by language[33] and including this information as an index also improved performance. One example for this proved helpful was the query "Find the video of man with dark hair speaking Spanish and wearing dark shirt". Including both language and audio indexes gave the best overall retrieval performance, with a MIR score of 0.454. The user interface for the interactive retrieval task was based on a storyboard format, displaying a sequence of keyframes in chronicle order. Video and audio play was supported. This system also ranked first overall for interactive search, with a MIR score of 0.727 with expert users. It was also

[33] http://code.google.com/apis/ajaxlanguage/

first in terms of retrieval speed (least minutes required for a user to find the desired video, 1.4 minutes here), and had a user satisfaction score of 6 out of 7. A second run using novice users scored a MIR score of 0.682.

[Guo et al., 2010] ranked third overall in the automatic known item search, with a MIR of 0.296 using Lucene to index the provided ASR and production metadata. The approach re-ranked a text-based search using the following content-based detectors: face detection, color detection, sound detection, OCR text, sound detection, and black and white video detection. While no information is given about the exact fusion strategy, they suggest that they used the content based detectors to eliminate videos from the retrieval list which could not match the query. Although this information is also not given, this presupposes some sort of query analysis strategy which activates the necessary detectors when relevant, for example the face detector if the query involves a person, the audio detector if it involves music, etc. A ranked list of 100 videos was returned, of which the top 10 were provided by the text only retrieval system, while the following 11 to 100 results were re-ranked based on the combined text and concept results.

[Lux et al., 2010] had runs ranked 5th, 7th, 9th and 10th overall in the automatic search. Their best run, 5th overall, MIR score of 0.265, was based on a text only approach. All available metadata fields, such as subject, creator, source, color, notes, keywords were merged into one text. ASR text was then added, but only words whose recognition rate exceeded 50%. Speaker and language recognition was performed, and used to add terms to the document based on pre-defined rules. For example, a male speaker would cause adding ’male’ ’guy’ ’man’ terms to the associated document; when French language was spoken, the terms ’French’ and ’France’ were added to the document. All text was merged into a single bag of words under a single tf-idf index managed by the Lucene engine. The top ranked results from a text based search were used to select exemplars for a feature based search; using a Color and Edge Directivity Descriptor, a Local Feature Histogram, and a Global Motion Histogram feature. Each feature returned a ranked list of videos, and the final list returned was based on the average inverted rank of a video in all lists (one list from text-search, up to three from content-based search). Although the search through feature space using various combinations resulted in an overall performance decrease, error analysis revealed that searching the feature space revealed relevant videos not identified by the text-only search.

The availability of the metadata makes text-based retrieval the most effective solution. [Li et al., 2010] investigation of query classes used different weights for text sources, and expanded the textual and visual sources by using Flickr as an external data source. Lemur was used for indexing. Five text fields were used, three of which, ’description’, ’keywords’ and ’title’ came from the production metadata. ASR formed a fourth, and OCR a fifth field. Query expansion was performed by augmenting query terms with keywords appearing in the Flickr tag [Bao et al., 2010]. The top 10 related tags were also added to the query. Queries were mapped to a query class, as per [Yan & Hauptmann, 2006] and [Yan et al., 2004] who define them as follows: "Named person", "Named object", "General object" and "Scene". 5 classes were defined in total.

Experiments aimed to find the optimal weights for the text fields versus query class. The best run contributed, 11th overall in automatic search, mean inverted rank (MIR) score of 0.253, used a single experimentally determined optimal weighting for the five

text fields, with queries mapped to one of five query classes. The MIR score was 0.243 when a one-query class approach was used. Varying the text field weights per query class caused a drop in MIR score to 0.214, 22nd overall which was attributed to over training.

A late fusion of text and visual detectors was also attempted. Google Images was used to retrieve training images using the visual cues provided with each query. The 130 concept detectors from the semantic indexing task were used [Li et al., 2010], as well as 12 color specific detectors specially trained for this task. It was noted that the 12 color concept detectors improved the concept-based retrieval from 0.0043 to 0.0061 on the evaluation queries. In addition, a content based retrieval approach was also attempted, using Latent Dirichlet Allocation (LDA) with 200 topics, and SIFT bag-of-word features, trained using topics retrieved from Google. Bipartite Graph Propagation was then used to find the relevance score between a query and the latent topics, in order to determine the relevant videos. The MIR score on the evaluation queries was 0.0047. A final LDA model described the joint distribution of text bag of words and the SIFT bag of words feature in a frame. The MIR score on the evaluation queries was 0.0032. The optimal fusion weights combining the visual and text subsystems remains to be determined; a simple fusion was used instead combining all modalities and giving a MIR score of 0.231, which was below the text-only approach.

Ngo et al. achieved in their best run the 8th position overall with a MIR score of 0.260 [Ngo et al., 2010] using only the provided ASR and metadata (“title”, “description”, “subject”, “keywords”, and “shotlist fields) using the Okapi [Robertson et al., 1998] indexes built by the Lemur engine. The mediocre performance of the visual semantic concept detectors was attributed to a poor query-to-concept mapping, where the items defined in the queries were too specific to be mapped properly to the 130 pre-defined concepts, as well as a decrease in overall concept detector performance. The approach adopted by Jumas et al. involved the use of WordNet to find synonyms and hypernyms for their query terms [Daróczy et al., 2010]. Synonyms, however, were not strong enough to find related features, while hypernyms frequently drifted, i.e. "man is a human being". They use a custom text retrieval engine [Daróczy et al., 2009] based on Okapi BM25 [Robertson et al., 1998] with proximity weights [Rasolofo & Savoy, 2003] [Büttcher et al., 2006] to achieve a best automatic run with a MIR score of 0.218 on metadata and custom ASR output.

[Snoek et al., 2010] MediaMill is a traditionally strong performer at TRECVID. For the automatic KIS task, pre-processing of stop word removal and Porter stemming was used followed by Lemur(Indri)indexing system for text retrieval, with indexes on the ASR transcripts and on the production meta-data. The best run, called BA, 14th overall with a MIR score of 0.238, used a single index combining meta-data and ASR. A second run, called Face, ranked 16th overall with a MIR score of 0.233, used only the meta-data. Both runs also incorporated concept detector scores. At query time, the query was passed to an index and retrieval was done using a language modelling approach to retrieval, with the Dirichlet smoothing method [Zhai & Lafferty, 2004]. Each concept detector was matched to a list of synonyms via WordNet, with manual inspection of the list of synonyms to avoid topic drift. This then allowed the matching of concepts to queries. Once a set of concept detectors was selected for a query, the score for each shot was the unweighed average of the detector scores. The shot-level scores were then combined for the entire video clip using the methodology described in [Huurnink et al., 2010]. The detector based search results were combined with the

text based search results using CombSUM fusion [Wilkins, 2009] after Borda normalization of both result lists, with weighting parameters determined from a set of training queries. The error analysis showed that for 44 queries, the BA run outperformed Face, and performed worse for 61 queries, leading to "including transcripts, … hurts more often than it helps. However, when transcripts help, they help a lot." A detector only run scored close to 0 MIR score, but improved when combined with ASR and meta-data retrieval results. Concept detectors are seemingly more suited for re-ranking text retrieval results rather than to serve as a direct source for retrieval results. Future improvements can be realized by utilizing a query-dependent approach, which in advance identifies the most promising modalities and lends the most credence to the related concept detector scores. MediaMill achieved 6th and 8th overall on the interactive search, with a MIR score of 0.409 and 0.364 respectively. Previous successful video navigation interfaces used the ForkBrowser [Rooij et al., 2008] and CrossBrowser visualizations [Snoek et al., 2007]. MediaMill used an interface based on Mediatable [Rooij et al., 2010] and investigated whether this categorization based approach was suitable for known item search. It displayed shots in a of each video across multiple columns, and also provided mechanisms to search through the meta-data and ASR transcript in multiple languages using a rich query language, play a video at multiple speed settings, and fields describing a video in more detail (e.g. title, description, detected semantic concepts).

Chen et al. system ranked 20th and 21st overall with a MIR score of 0.215 and 0.213 for the automatic search task [Chen et al., 2010]. The system used Lucene as retrieval engine for the meta-data and ASR. The focus on a query expansion strategy used a complex query to retrieve relevant videos from YouTube for the collection of the associated tags and comments that express high mutual information with the key words of the query. The YouTube tags formed a separate index. Combined with retrieval on the meta-data/ASR index, this approach resulted in a MIR score of 0.213. Concept detectors were also included by identifying the relevant detectors at query time based on the morphological analysis followed by selective expansion using the WordNet of both concept detector descriptions and the KIS queries, as per [Neo et al., 2006]. The retrieval performance using YouTube, metadata, and concept detectors was 0.215 in terms of a MIR score. The NUS group system had the 2nd best retrieval performance in the interactive search task, with a MIR score of 0.682 which was partly due to an efficient user interface and a video representation using clustered shots for a quick overview. Extensive analysis was performed to provide a feedback mechanism to refine the search results by allowing the user to select shots similar to the query, or to exclude shots from the query. As per [Chang et al., 2006], a query can be learned as a set of related concept detectors. The positive and negative examples provided by the user can be used to modify the concept detector weights [Yuan et al., 2010]. To prevent an unbalanced training set, which can happen when users only indicate positive (similar) shots, a mechanism was developed to infer the exclusive negative set based on the user selected similar shots, as per [Shuicheng YA et al., 2010]. The resultant negative shots could then also be used for training purposes. Analysis of the interactive and automatic runs shows that using YouTube to expand the query terms had a negligible effect. One would suspect that the radical increase in interactive search performance, which started from a moderate automated search baseline, was due to the extensive feedback mechanism implemented in the search interface.

In conclusion, this overview of a significant number of TRECVID 2010 KIS entrants reveals a number of strategies.

- Most groups attempt a query expansion based on an external knowledge source such as WordNet, Wikipedia, Flickr, YouTube, or Google, although with minimal impact. Query expansion to enable better matching with concept detectors also had negligible impact, which might relate to the lack of sufficient concept detectors to cover the semantic space, and to the lack of robustness of the detectors themselves.
- Nearly all groups experienced a detrimental effect when fusing concept detectors with retrieval results from text-only sources; retrieval based on ASR and production meta-data performed best. The consensus seems to be that concept detectors at best could be used to positively re-rank the results of the text-only searches. OCR text complemented ASR and production meta-data. Both top 2 performers in the automatic search task had an OCR module.
- The groups that combined multiple ASR systems achieved some an improvement over a single ASR source as errors in both systems might have cancelled out each other. The systems [Chaisorn et al., 2010; Wan et al., 2011] that much more extensively sought to revise the ASR transcript, saw a much greater increase in retrieval performance. In addition, they discovered that Porter stemming was suboptimal. Extensive use of natural language and information retrieval techniques on the text sources led to their top performance in the automatic search task.
- The use of content-based detectors to direct the search towards particular indexes based on matches with the query seems promising, i.e. [Chaisorn et al., 2010; Wan et al., 2011] use of a language detector, matching with the query type <speaking> language, and audio detectors, matching with the query type <sing>) [Guo et al., 2010] use of an audio detector, black and white detector, color detector, face detector [Lux et al., 2010] use of a language detector, gender detector, speech detector- given the placement of these groups in the overall ranking. The [Li et al., 2010] system attempted to find visual exemplars using Google Images for a given query, thus transforming the task into a search of the feature space, similar to that of the Instance Search task. The [Lux et al., 2010] made a similar attempt by using the top-ranked exemplars from a text-only search. The sophisticated scheme by [Chen et al., 2010] incorporating positive and negative user feed-back to derive a concept classifier for the desired user query also inspires comparison to the Instance Search idea of refining the feature space and inferring a new classifier based on the provided examples. One suggestion for future work by [Snoek et al., 2010] was to pursue a query-dependent choice, where text and visual concept detector weights vary based on the kind of query. This was partially investigated by [Li et al., 2010] who found an improvement when splitting queries into a variety of classes.

From the high interactive retrieval results in general one may conclude that a good interface and video traversal scheme will remedy most automatic retrieval baselines. The top two performers who stood out exceeded in both user retrieval time (1.5 minutes and 2.5 minutes vs. 3+ minutes for the rest) and MIR scores (0.727 and 0.682 vs. 0.591 for 3d place). The top performance, by [Chaisorn et al., 2010; Wan et al., 2011], probably can be attributed to a good interface starting with an extremely

competitive automatically retrieved result list; it takes little time and effort for the user to refine the initial results and find the relevant video. The second top performer [Chen et al., 2010] shows that a sophisticated model for refining the result list interactively compensates for a simple automatic baseline.

5.3 Audio Similarity Approaches

5.3.1 Audio fingerprinting

Audio fingerprinting, also called content-based audio identification, is a technology that is frequently used to link unlabelled audio fragments to corresponding metadata. An example is finding the title of a song on the base of a piece of audio data.

In fact, an audio fingerprint is a very compact content-based signature that summarizes an audio recording [Cano, 2006].

The technique has two main stages: preprocessing and matching.

- **Preprocessing**: The preprocessing consists of generating a database of audio fingerprints. For each audio file in the reference collection, a set of fingerprints is generated that uniquely identifies (part of) an audio file. The set of fingerprints that is generated in this way is then inserted into a database.
- **Matching**: In the matching stage, an unlabelled audio file is processed and a fingerprint is extracted from it, using the same extraction technique used for preprocessing. This new fingerprint is then compared with the set of reference fingerprints in the database. In case a match occurs, the information corresponding with the match is extracted from the database.

A lot of techniques for generating fingerprints have been researched over the past decades. Such fingerprints commonly consist of features ranging from low level features to high level descriptors. Especially these high-level descriptors are very useful for audio navigation, search by similarity and content-based processing [Cano, 2006].

In general, three components have to be implemented in order to build a successful audio fingerprinting system. A feature extractor that extracts useful features from the audio signal, a matching algorithm that computes the similarity of two fingerprints (or feature vectors) and a search method to compare a fingerprint with a database that possibly contains million entries.

Feature extraction

Feature extraction is a very important part of an audio fingerprinting system, as is the case for most audio processing methods. One widely used set of features in the audio domain are Mel Frequency Cepstral Coefficients (MFCC). These are so-called low-level features (i.e. they do not have a direct semantic meaning). An example of a fingerprinting system that uses MFCC-based features for retrieval in broadcast audio is researched by Cano et al [Cano et al., 2002]. Other fingerprinting systems make use of features that are much more high level, such as the estimated beats per minute (in music), rhythm or pitch estimation.

Similarity measures

A second research area for audio fingerprinting systems is similarity measures. If the features derived by the feature extraction can be quantized, a Hamming or Manhattan distance is frequently used. However, lots of other similarity measures have been proposed. Other static measures have been proposed, e.g. based on cross entropy estimation [Sukittanon & Atlas, 2002] or pseudo norm [Miçak & Venkatesan, 2001].

The more complex measures are usually some sort of probabilistic similarity measure. In this case, the reference fingerprints are stored in a probabilistic model such as a Hidden Markov Model, or a Gaussian Mixture Model. Examples of this kind of techniques can be found in [Allamanche et al., 2001] and [Batlle et al., 2002].

Matching

The third challenge is matching the extracted fingerprint with a huge database of fingerprints. This can be seen as finding the nearest neighbour of a fingerprint in the database using the defined similarity measure. However, finding a nearest neighbour is a hard problem, and most algorithms that are efficient in low-dimensional spaces have a time complexity that is exponential in the dimension. This is what is called the "curse of dimensionality". Most techniques work with creating a data structure with the computed distances between some or all fingerprint pairs in the fingerprint database. For an example, see e.g. [Kimura et al., 2001], or [Chávez et al., 2001].

Another technique is to first prune the fingerprint database using an easy to compute similarity measure, and then use a more complex measure to find the best match in the remaining set. An example of this technique is presented in [Kastner et al., 2002]. Other techniques for matching include inverted file indexing and suffix trees.

5.3.2 Audio watermarking

Audio watermarking is a different technique for audio matching. In audio watermarking, an inaudible audio signal is added to the audio data. The audio watermarking approach finds its source in copyright detection. In copyright detection, a watermark containing the copyright information is added to the audio signal.

There are several requirements that can be identified for watermarking:

- The watermark should be inaudible. The audio signal should not degrade after addition of the watermark.
- The watermark should be robust. I.e. transformations applied to the audio signal such as transcoding, copying, loudness modification,… should not destroy the watermark.
- The bitrate of the watermarked signal should be high enough; this usually results in a trade-off between robustness and capacity.
- The watermark should be retrievable in a reliable way, with an error rate that is not too high.
- Watermarking algorithms should be efficient (of low complexity).

For the inaudibility of the watermarked signal, the psychoacoustic features of the human ear are exploited. For example, Zwicker et al. have shown that if two tones simultaneously present in the signal are close to each other in frequency, then frequency masking occurs. I.e. if one tone is significantly louder than another one, the second tone is masked [Zwicker & Fastl, 1990].

5.3.3 Watermarking versus fingerprinting

Audio watermarking for content identification has several advantages over fingerprinting. First, the decoding of the watermark can be performed client side. This as opposed to fingerprinting, where the fingerprint match algorithm (the most expensive step) has to be performed server side. Moreover, audio watermarking allows recognising different versions of the same content (e.g. several recordings or broadcasts of the same song etc…). Due to this, there is an independence between the audio signal and the watermark.

Of course, audio watermarking also has a few disadvantages. First, the audio signal itself has to be modified to encode the watermark in it. In some practical situations however, this is not always feasible. Moreover, watermarking is much less robust than fingerprinting. It is much more sensitive to channel perturbations than fingerprinting.

5.3.4 Music Similarity

A particular topic within the area of audio similarity is that which relates to music. State of the art technologies in this respect have reached mainstream audiences by the use of free smartphone apps like Shazam[34] and SoundHound[35], or even API-providing opensource alternatives like Echoprint[36], all of which are covered below.

Music similarity approaches like those covered by the Music Information Retrieval Evaluation eXchange (MIREX), a community-based framework for the formal evaluation of Music Information Retrieval (MIR) systems and algorithms [Downie, 2008], have expanded traditional approaches based on fingerprinting techniques (watermarks) towards novel techniques like QbHS (“query by humming / singing”) [Song et al., 2011].

The changing landscape produced by the ubiquity of mobile devices and smartphones has also been considered, by analysing audio fingerprinting schemes in a common framework with short query probes captured from cell phone [Chandrasekhar et al., 2011], aiming at reducing latency in practical mobile audio retrieval applications.

Shazam

The Shazam app relies on an algorithm capable of identifying a short segment of music captured through a cellphone microphone by using a hashed time-frequency constellation analysis of the audio, thus based on fingerprinting, extracting reproducible hash tokens, analyzing music based on its spectrogram [Wan, 2003].

In more detail, the way Shazam works is by first of all fingerprinting a comprehensive catalog of music (i.e., each audio file in it) and storing the fingerprints in a database, a hash table, where the key is the frequency. The fingerprinting algorithm generates a time-frequency-intensity graph, where frequencies of “peak intensity” are identified; both the frequency and the time from the beginning of the track are stored.

Then, when a song is “tagged” through a mobile device, a fingerprint out of a short sample of audio (10 seconds) is created, and the app uploads the fingerprint to the service, which runs a search for a matching fingerprint in the database. So, in this

[34] http://www.shazam.com/

[35] http://www.soundhound.com/

[36] http://echoprint.me/

case, “sample” audio files are fingerprinted too, and the frequency is used for the search of matching songs. When a specific song is matched multiple times based on the individual frequencies, the algorithm checks to see if the frequencies correspond in time. If a match is found, the song info is returned to the user, otherwise an error is returned.

Soundhound (QbHS):

The SoundHound smartphone app relies on a proprietary technology named Sound2Sound (S2S) Search Science, capable of recognizing various sound inputs including music and speech through feature extraction, which are then matched against a database [Mohajer et al., 2011].

The recognition engine by SoundHound also offers singing and humming functionalities (QbHS), by matching multiple aspects of the user's voice input including melody, rhythm, and lyrics, with a huge database of user recordings from midomi.com. Their matching technology is capable of finding matches regardless of the key or tempo of the user input, and takes advantage of lyrics when they are available in the sample by the user.

Echoprint (fingerprint)

Echoprint [Ellis et al., 2011] is an open source music identification service based on fingerprinting backed by a huge database of music in partnership with Musicbrainz[37]. Unlike the smartphone apps covered above, Echoprint works by exposing an API to which third-party applications can connect to in order to exploit the music recognition functionalities provided by the service.

Echoprint works on three main blocks: a “code generator” in charge of converting audio into codes; the server that stores and indexes codes; and the data itself coming from partners and other users. The code generator computes (time, hash) pairs from an audio signal using signal processing. The process starts from an 11kHz mono signal, which is decomposed into 8 different subbands. The decomposition is hashed into a 20-bit space and stored alongside the time of the onset. The server code then indexes each onset in an inverted index, and the code material for each track is also stored. The process of querying is a lookup of all the query codes in the inverted index, and the score returned is the number of overlaps of query onsets between the query and each target track.

[37] http://musicbrainz.org/

6 Conclusions

This deliverable describes state of the art in the area of semantic retrieval of multimedia (audiovisual) content beyond text resources.

It shows that the nature of the content itself and its characteristics are important for retrieval aspects. Thus, the document addresses three different blocks that are deeply connected:

1. Data, metadata and semantic aspects of multimedia content, as the base models upon which semantic retrieval solutions are built.
2. Multimedia content annotation techniques, both manual and automatic.
3. Semantic retrieval of multimedia content based on the data and annotation aspects.

It is intended that the deliverable covering the different aspects mentioned above acts as an important reference for the design and implementation of the networked media search engine to be developed within TOSCA-MP WP3.

7 Glossary

Partner Acronyms

DTO	Technicolor, DE
EBU	European Broadcasting Union, CH
FBK	Fondazione Bruno Kessler, FBK
HHI	Heinrich Hertz Institut, Fraunhofer Gesellschaft zur Förderung der Angewandten Forschung e.V., DE
IRT	Institut für Rundfunktechnik GmbH, DE
K.U.Leuven	Katholieke Universiteit Leuven, BE
JRS	JOANNEUM RESEARCH Forschungsgesellschaft mbH, AT
PLY	Playence KG, AT
RAI	Radiotelevisione Italiana S.p.a., IT
VRT	De Vlaamse Radio en Televisieomroeporganisatie NV, BE

8 References

[Abate et al., 2007] Abate, A. F., Nappi, M., Riccio, D., & Sabatino, G. (2007). 2D and 3D face recognition: A survey. Pattern Recogn. Lett., 28 , 1885--1906.

[Aggarwal & Ryoo, 2011] Aggarwal, J. K. & Ryoo, M. S. (2011). Human activity analysis: A review. ACM Comput. Surv., 43 (3), 16:1--16:43.

[Alatan et al., 2001] Alatan, A. A., Akansu, A. N., & Wolf, W. (2001). Multi-Modal Dialog Scene Detection Using Hidden Markov Models for Content-Based Multimedia Indexing. Multimedia Tools Appl., 14 , 137--151.

[Allamanche et al., 2001] Allamanche, E., Herre, J., Hellmuth, O., & Fröba, B. (2001). Content-based Identification of Audio Material Using MPEG-7 Low Level Description. ,

[Allen, 1984] Allen, J. F. (1984). Towards a general theory of action and time. Artificial Intelligence, 23 , 123--154.

[Amir et al., 2004] Amir, A., Argill, J. O., Berg, M., Chang, S.-f., Franz, M., Hsu, W., Iyengar, G., Kender, J. R., Kennedy, L., Lin, C.-y., Naphade, M., Natsev, A. p., Smith, J. R., Tesic, J., Wu, G., Zhang, D., Watson, I. T. J., & Watson, I. T. J. (2004). Ibm research trecvid-2004 video retrieval system. ,

[Aubert & Prié, 2007] Aubert, O. & Prié, Y. (2007). Advene: an open-source framework for integrating and visualising audiovisual metadata. , 1005--1008.

[Babaguchi & Nitta, 2003] Babaguchi, N. & Nitta, N. (2003). Intermodal collaboration: A strategy for semantic content analysis for broadcasted sports video. ,

[Bahl et al., 1993] Bahl, L. R., Brown, P. F., Souza, P. V. d., & Mercer, R. L. (1993). Estimating hidden Markov model parameters so as to maximize speech recognition accuracy. IEEE Trans. Speech Audio Processing, 1 , 77--83.

[Bailer, 2011]Bailer, W. (2011). A feature sequence kernel for video concept classification. , 359--369.

[Baker, 1975] Baker, J. K. (1975). Stochastic modeling for automatic speech recognition. Speech Recognition, ,

[Baker et al., 2009] Baker, J. M., Deng, L., Glass, J., Khudanpur, S., Lee, C.-H., Morgan, N., & O'Shaughness, D. (2009). Research Developments and Directions in Speech Recognition and Understanding, Part 1. IEEE Signal Processing Magazine, 26 , 75-80.

[Bao et al., 2010] Bao, L., Cao, J., Zhang, Y., Li, J., Chen, M.-y., & Hauptmann, A. G. (2010). Explicit and implicit concept-based video retrieval with bipartite graph propagation model. , 939--942.

[Batlle et al., 2002] Batlle, E., Masip, J., & Guaus, E. (2002). {Automatic Song Identification in Noisy Broadcast Audio}. ,

[Baum, 1972] Baum, L. (1972). An inequality and associated maximization technique occurring in statistical estimation for probabilistic functions of a Markov process. Inequalities, 3 , 1-8.

[Beaufays et al., 2002] Beaufays, F., Bourlard, H., Franco, H., & Morgan, N. (2002). Handbook of Brain Theory and Neural Networks. ,

[Bendersky & Croft, 2008] Bendersky, M. & Croft, W. B. (2008). Discovering key concepts in verbose queries. , 491--498.

[Berners-Lee, 2006]Berners-Lee, T. (2006). Linked Data - Design Issues. ,

[Berners-Lee et al., 2006] Berners-Lee, T., Hall, W., Hendler, J. A., O'Hara, K., Shadbolt, N., & Weitzner, D. J. (2006). A Framework for Web Science. Foundations and Trends in Web Science, 1 ,

[Bertini et al., 2005] Bertini, M., Bimbo, A. D., & Nunziati, W. (2005). Common Visual Cues for Sports Highlights Modeling. Multimedia Tools Appl., 27 , 215--218.

[Besacier et al., 2004] Besacier, L., Quénot, G., Ayache, S., & Moraru, D. (2004). Video story segmentation with multi-modal features: experiments on TRECvid 2003. , 221-227.

[Bober & Brasnett, 2009] Bober, M. & Brasnett, P. (2009). MPEG-7 Visual Signature Tools. ,

[Bowyer et al., 2006] Bowyer, K. W., Chang, K., & Flynn, P. (2006). A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition. Comput. Vis. Image Underst., 101 , 1--15.

[Bürger & Hausenblas, 2011] Bürger, T. & Hausenblas, M. (2011). Handbook of Metadata, Semantics and Ontologies. ,

[Büttcher et al., 2006] Büttcher, S., Clarke, C. L. A., & Lushman, B. (2006). Term proximity scoring for ad-hoc retrieval on very large text collections. , 621--622.

[Cano, 2006] Cano, P. (2006). Content-Based Audio Search: from Fingerprinting to Semantic Audio Retrieval. ,

[Cano et al., 2002] Cano, P., Batlle, E., Mayer, H., & Neuschmied, H. (2002). Robust Sound Modeling for Song Detection in Broadcast Audio. , 1--7.

[Carrer et al., 1997] Carrer, M., Ligresti, L., Ahanger, G., & Little, T. D. C. (1997). An Annotation Engine for Supporting Video Database Population. Multimedia Tools Appl., 5 (3), 233--258.

[Chaisorn et al., 2003] Chaisorn, L., Chua, T.-S., Koh, C.-K., Zhao, Y.-L., Xu, H., Feng, H., & Tian, Q. (2003). A two-level multi-modal approach for story segmentation of large news video corpus. ,

[Chaisorn et al., 2003] Chaisorn, L., Chua, T.-S., & Lee, C.-H. (2003). A Multi-Modal Approach to Story Segmentation for News Video. World Wide Web, 6 , 187-208.

[Chaisorn et al., 2010] Chaisorn, L., Wan, K.-w., Zheng, Y.-t., Zhu, Y., Kok, T.-s., Tan, H.-l., Fu, Z., & Bolling, S. (2010). TRECVID 2010 Known-item Search (KIS) Task by I2R. ,

[Chandrasekhar et al., 2011] Chandrasekhar, V., Sharifi, M., & Ross, D. (2011). Survey and Evaluation of Audio Fingerprinting Schemes for Mobile Query-By-Example Applications. 12th International Society for Music Information Retrieval Conference (ISMIR), ,

[Chang et al., 1997] Chang, S.-F., Chen, W., Meng, H. J., Sundaram, H., & Zhong, D. (1997). VideoQ: an automated content based video search system using visual cues. , 313-324.

[Chang et al., 2006] Chang, S.-F., Jiang, W., Hsu, W., Kennedy, L., Xu, D., Yanagawa, A., & Zavesky, E. (2006). Columbia University TRECVID-2006 Video Search and High-Level Feature Extraction. ,

[Chang et al., 2005] Chang, S.-F., Manmatha, R., & Chua, T.-S. (2005). Combining text and audio-visual features in video indexing. 5 , v/1005 - v/1008 Vol. 5.

[Chasanis et al., 2009] Chasanis, V., Kalogeratos, A., & Likas, A. (2009). Movie segmentation into scenes and chapters using locally weighted bag of visual words. , 35:1--35:7.

[Chen & Jiang, 2008] Chen, J. & Jiang, J. (2008). University of Bradford at TRECVID 2008: Content Based Copy Detection Task. ,

[Chen et al., 2008] Chen, L.-H., Lai, Y.-C., & Liao, H.-Y. M. (2008). Movie scene segmentation using background information. Pattern Recognition, 41 (3), 1056 - 1065.

[Chen et al., 2010] Chen, X., Yuan, J., Nie, L., Zha, Z.-J., Yan, S., & Chua, T.-S. (2010). TRECVID 2010 Known-item Search by NUS. ,

[Christel et al., 2000] Christel, M. G., Olligschlaeger, A. M., & Huang, C. (2000). Interactive Maps for a Digital Video Library. IEEE MultiMedia, 7 , 60--67.

[Chávez et al., 2001] Chávez, E., Navarro, G., Baeza-Yates, R., & Marroquín, J. L. (2001). Searching in metric spaces. ACM Comput. Surv., 33 (3), 273--321.

[Cour et al., 2008] Cour, T., Jordan, C., Miltsakaki, E., & Taskar, B. (2008). Movie/Script: Alignment and Parsing of Video and Text Transcription. , 158-171.

[Cox et al., 2006] Cox, M., Tadic, L., & Mulder, E. (2006). Descriptive Metadata for Television. ,

[Creed et al., 2010] Creed, C., Lonsdale, P., Hendley, R., & Beale, R. (2010). Synergistic Annotation of Multimedia Content. , 205--208.

[Cui et al., 2011] Cui, X., Liu, Q., Gao, M., & Metaxas, D. N. (2011). Abnormal detection using interaction energy potentials. , 3161--3167.

[Dantone et al., 2010] Dantone, M., Sullivan, K., & Tesic, J. (2010). Multimedia Event Detection (MED) Evaluation Task. ,

[Daróczy et al., 2009] Daróczy, B., Fekete, Z., Brendel, M., Rácz, S., Benczúr, A., Siklósi, D., & Pereszlényi, A. (2009). SZTAKI @ ImageCLEF 2008: visual feature analysis in segmented images. , 644--651.

[Daróczy et al., 2010] Daróczy, B., Nemeskey, D., Pethes, R., Petrás, I., Benczúr, A. A., Falavigna, D., & Gretter, R. (2010). JUMAS @ TRECVID 2010. ,

[Dasiopoulou et al., 2011] Dasiopoulou, S., Giannakidou, E., Litos, G., Malasioti, P., & Kompatsiaris, Y. (2011). Knowledge-driven multimedia information extraction and ontology evolution. , 196--239.

[Dempster et al., 1977] Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, 39 , 1-21.

[Douze et al., 2008] Douze, M., Gaidon, A., Jegou, H., Marszałek, M., & Schmid, C. (2008). INRIA-LEAR's Video Copy INRIA-LEAR's Video Copy Detection System. ,

[Downie, 2008] Downie, J. S. (2008). The music information retrieval evaluation exchange (2005-2007): A window into music information retrieval research. Acoustical Science and Technology, 29 , 247-255.

[Duan et al., 2011] Duan, L., Xu, D., Tsang, I., & Luo, J. (2011). Visual Event Recognition in Videos by Learning from Web Data. Pattern Analysis and Machine Intelligence, IEEE Transactions on, PP (99), 1.

[Döhring & Lienhart, 2009] Döhring, I. & Lienhart, R. (2009). Mining TV Broadcasts for Recurring Video Sequences. ,

[Ellis et al., 2011] Ellis, D., Whitman, B., & Porter, A. (2011). Echoprint - An Open Music Identification Service. ,

[Engels et al., 2010]Engels, C., Deschacht, K., Becker, J. H., Tuytelaars, T., Moens, S., & Van Gool, L. (2010). Automatic annotation of unique locations from video and text. , 115.1--115.11.

[Everingham et al., 2009] Everingham, M., Sivic, J., & Zisserman, A. (2009). Taking the bite out of automated naming of characters in TV video. Image Vision Computing, 27 , 545--559.

[Fleischman et al., 2007] Fleischman, M., Evans, H., & Roy, D. (2007). Unsupervised content-based indexing for sports video retrieval. ,

[Fleischman et al., 2007] Fleischman, M., Roy, B., & Roy, D. (2007). Temporal feature induction for baseball highlight classification. , 333--336.

[Franz et al., 1999] Franz, M., Mccarley, J. S., Ward, T., & Zhu, W.-j. (1999). Segmentation and Detection at IBM : Hybrid Statistical Models and Two-tiered Clustering. ,

[Franz et al., 2011] Franz, T., Troncy, R., & Vacura, M. (2011). Multimedia Semantics: Metadata, Analysis and Interaction. ,

[Gao et al., 2008] Gao, Z., Zhao, Z., Liu, T., Nan, X., Mei, M., Zhang, B., Liu, X., Peng, X., Zheng, H., Zhao, Y., & Cai, A. (2008). BUPT at TRECVID 2008. ,

[Goela et al., 2007] Goela, N., Wilson, K. W., Niu, F., Divakaran, A., & Otsuka, I. (2007). An SVM Framework for Genre-Independent Scene Change Detection. , 532-535.

[Gruber, 1993] Gruber, T. R. (1993). A translation approach to portable ontology specifications. Knowledge Acquisition, 5 ,

[Guo et al., 2010] Guo, X., Chen, Y., Liu, W., Mao, Y., Zhang, H., Zhou, K., Wang, L., Hua, Y., Zhao, Z., Zhao, Y., & Cai, A. (2010). BUPT-MCPRL at TRECVID 2010. ,

[Gupta et al., 2009] Gupta, A., Srinivasan, P., Shi, J., & Davis, L. S. (2009). Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. , 2012 -2019.

[Gursoy & Gunsel, 2008] Gursoy, O. & Gunsel, B. (2008). Istanbul Technical University at TRECVID2008. ,

[Hanjalic et al., 2008] Hanjalic, A., Lienhart, R., Ma, W.-Y., & Smith., J. R. (2008). The Holy Grail of Multimedia Information Retrieval: So Close or Yet So Far Away? IEEE, 98 , 541-547.

[Hauptmann et al., 2007] Hauptmann, A., Yan, R., & Lin, W.-H. (2007). How many high-level concepts will fill the semantic gap in news video retrieval? , 627--634.

[Hauptmann & Witbrock, 1998] Hauptmann, A. G. & Witbrock, M. J. (1998). Story Segmentation and Detection of Commercials in Broadcast News Video. , 168-179.

[Heath & Bizer, 2011] Heath, T. & Bizer, C. (2011). Linked Data: Evolving the Web into a Global Data Space. ,

[Heggland, 2005] Heggland, J. (2005). OntoLog: Flexible Management of Semantic Video Content Annotations. ,

[Hoashi et al., 2004] Hoashi, K., Sugano, M., Naito, M., Matsumoto, K., & Nakajima, Y. (2004). Shot boundary determination on MPEG compressed domain and story segmentation experiments for TRECVID 2004. , 109--120.

[Howe, 2006]Howe, J. (2006). The rise of crowdsourcing. ,

[Hsu et al., 2005] Hsu, W. H., Kennedy, L. S., Chang, S.-f., Franz, M., & Smith, J. R. (2005). Columbia-IBM news video story segmentation in trecvid 2004. ,

[Hunter & Schroeter, 2008] Hunter, J. & Schroeter, R. (2008). Co-Annotea: A System for Tagging Relationships Between Multiple Mixed-Media Objects. IEEE MultiMedia, 15 (3), 42--53.

[Huurnink et al., 2010] Huurnink, B., Snoek, C. G. M., de Rijke, M., & Smeulders, A. W. M. (2010). Today's and tomorrow's retrieval practice in the audiovisual archive. , 18--25.

[Héritier et al., 2008] Héritier, M., Foucher, S., & Gagnon, L. (2008). CRIM Notebook Paper - TRECVID 2008 Video CRIM Notebook Paper - TRECVID 2008 Video Copy Detection Using Latent Aspect Modeling Over SIFT Matches. ,

[Inoue et al., 2011] Inoue, N., Kamishima, Y., Wada, T., Shinoda, K., & Sato, S. (2011). TokyoTech+Canon at TRECVID 2011. ,

[Jelinek, 1976] Jelinek, F. (1976). Continuous speech recognition by statistical methods. IEEE, 64 , 532--557.

[Joly et al., 2008] Joly, A., Law-to, J., & Boujemaa, N. (2008). INRIA-IMEDIA TRECVID 2008: Video Copy Detection. ,

[Kastner et al., 2002] Kastner, T., Allamanche, E., Herre, J., Hellmuth, O., Cremer, M., & Grossmann, H. (2002). MPEG-7 Scalable Robust Audio Fingerprinting. ,

[Kennedy et al., 2006] Kennedy, L., Hauptmann, A., Naphade, M., Smith, A. H. J. R., & Chang, S. F. (2006). LSCOM Lexicon Definitions and Annotations Version 1.0, DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia. ,

[Kennedy & Chang, 2007] Kennedy, L. S. & Chang, S.-F. (2007). A reranking approach for context-based concept fusion in video indexing and retrieval. , 333--340.

[Kijak et al., 2003] Kijak, E., Gravier, G., Oisel, L., & Gros, P. (2003). Audiovisual integration for tennis broadcast structuring. , 289--312.

[Kimura et al., 2001] Kimura, A., Kashino, K., Kurozumi, T., & Murase, H. (2001). Very quick audio searching: introducing global pruning to the Time-Series Active Search. , 1429--1432.

[Kipp, 2005] Kipp, M. (2005). Gesture Generation by Imitation: from Human Behavior to Computer Character Animation. ,

[Klyne & Carroll, 2004] Klyne, G. & Carroll, J. J. (2004). Resource Description Framework (RDF): Concepts and Abstract Syntax - W3C Recommendation. ,

[Koskela et al., 2008] Koskela, M., Sjöberg, M., Viitaniemi, V., & Laaksonen, J. (2008). PicSOM Experiments in TRECVID 2008. ,

[Kraaij & Over, 2010] Kraaij, W. & Over, P. (2010). TRECVID 2010 Instance Search Task. ,

[Kraaij et al., 2004] Kraaij, W., Smeaton, A. F., Over, P., & Arlandis, J. (2004). TRECVID - An Introduction. ,

[Kucuktunc et al., 2008] Kucuktunc, O., Bastan, M., Gudukbay, U., & Ulusoy, O. (2008). Bilkent University Multimedia Database Group at TRECVID 2008. ,

[Kunath & Weinberger, 2010] Kunath, S. & Weinberger, S. (2010). The Wisdom of the Crowd's Ear: Speech Accent Rating and Annotation with Amazon Mechanical Turk. ,

[Laptev, 2005] Laptev, I. (2005). On Space-Time Interest Points. Int. J. Comput. Vision, 64 (2-3), 107--123.

[Laptev & Lindeberg, 2004] Laptev, I. & Lindeberg, T. (2004). Velocity Adaptation of Space-Time Interest Points. , 52--56.

[Laptev et al., 2008] Laptev, I., Marszalek, M., Schmid, C., & Rozenfeld, B. (2008). Learning realistic human actions from movies. ,

[Law-To et al., 2007] Law-To, J., Chen, L., Joly, A., Laptev, I., Buisson, O., Gouet-Brunet, V., Boujemaa, N., & Stentiford, F. (2007). Video copy detection: a comparative study. , 371--378.

[Lazebnik et al., 2006] Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. 2 , 2169 - 2178.

[Le et al., 2008] Le, D., Wu, X., Satoh, S., Rajgure, S., & Gemert, J. (2008). National Institute of Informatics, Japan at TRECVID 2008. ,

[Lee et al., 2012] Lee, W., Bailer, W., Bürger, T., Champin, P.-A., Evain, J.-P., Malaisé, V., Michel, T., Sasaki, F., Söderberg, J., Stegmaier, F., & Strassner, J. (2012). Ontology for Media Resources 1.0. ,

[Li et al., 2010] Li, H., Bao, L., Gao, Z., Overwijk, A., Liu, W., Zhang, L.-f., Yu, S.-i., Chen, M.-y., Metze, F., & Hauptmann, A. (2010). Informedia @ TRECVID 2010. ,

[Li et al., 2004] Li, Y., Narayanan, S., & Kuo, C. C. J. (2004). Content-based movie analysis and indexing based on audiovisual cues. Circuits and Systems for Video Technology, IEEE Transactions on, 14 (8), 1073 - 1085.

[Liang et al., 2008] Liang, Y., Liu, X., Wang, Z., Li, J., Cao, B., Cao, Z., Dai, Z., Guo, Z., Li, W., Luo, L., Meng, Z., Qin, Y., Qiu, S., Tian, A., Wang, D., Wang, Q., Zhu, C., Hu, X., Yuan, J., Yuan, P., & Zhang, B. (2008). THU and ICRC at TRECVID 2008. ,

[Lippman, 1987] Lippman, R. (1987). An introduction to computing with neural nets. IEEE ASSP Mag., 4 ,

[Liu et al., 2006] Liu, S., Xu, M., Yi, H., Chia, L.-T., & Rajan, D. (2006). Multimodal semantic analysis and annotation for basketball video. EURASIP J. Appl. Signal Process., 2006 , 182--182.

[Lux et al., 2010] Lux, M., Schoeffmann, K., Fabro, M., Kogler, M., & Taschwer, M. (2010). ITEC-UNIKLU Known-Item Search Submission. ,

[MPEG-7, 2001] MPEG-7 (2001). MPEG-7: Multimedia Content Description Interface. ,

[Masolo et al., 2001] Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A., & Schneider, L. (2001). The WonderWeb Library of Foundational Ontologies Preliminary Report. ,

[Miçak & Venkatesan, 2001] Miçak, M. & Venkatesan, R. (2001). A Perceptual Audio Hashing Algorithm: A Tool For Robust Audio Identification . . . , 51--65.

[Mohajer et al., 2011] Mohajer, K., Emami, M., Grabowski, M., & Hom, J. M. (2011). System and Method for Storing and Retrieving Non-text-based Information. ,

[Nadeau, 2007] Nadeau, David; Sekine, S. (2007). A survey of named entity recognition and classification. Linguisticae Investigationes, ,

[Natarajan et al., 2011] Natarajan, P., Natarajan, P., Manohar, V., Wu, S., Tsakalidis, S., Vitaladevuni, S. N., Zhuang, X., Prasad, R., Ye, G., Liu, D., Jhuo, I.-H., Chang, S.-F., Izadinia, H., Saleemi, I., Shah, M., White, B., Yeh, T., & Davis, L. (2011). BBN VISER TRECVID 2011 Multimedia Event Detection System. ,

[Neo et al., 2006] Neo, S.-y., Zhao, J., Kan, M.-y., & Chua, T.-s. (2006). Video retrieval using high level features: Exploiting query matching and confidence-based weighting. , 143--152.

[Ngo et al., 2010] Ngo, C.-w., Zhu, S.-a., Tan, H.-k., Zhao, W.-l., & Wei, X.-y. (2010). VIREO at TRECVID 2010: Semantic Indexing, Known-Item Search, and Content-Based Copy Detection. ,

[Nock et al., 2003] Nock, H. J., Iyengar, G., & Neti, C. (2003). Issues in Speech-Based Retrieval of Video. ,

[Novotney & Callison-Burch, 2010] Novotney, S. & Callison-Burch, C. (2010). Cheap, Fast and Good Enough: Automatic Speech Recognition with Non-Expert Transcription. ,

[Oomen & Aroyo, 2011] Oomen, J. & Aroyo, L. (2011). Crowdsourcing in the Cultural Heritage Domain: Opportunities and Challenges. ,

[Orhan et al., 2008] Orhan, O., Hochreiter, J., Poock, J., Chen, Q., Chabra, A., & Shah, M. (2008). University of Central Florida at TRECVID 2008 Content Based Copy Detection and Surveillance Event Detection. ,

[Ounis et al., 2006] Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., & Lioma, C. (2006). Terrier: A High Performance and Scalable Information Retrieval Platform. (6), 18--25.

[Over et al., 2010] Over, P., Awad, G., Fiscus, J., Antonishek, B., Smeaton, A. F., Kraaij, W., & Quenot, G. (2010). TRECVID 2010 -- An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. ,

[Parent & Eskenazi, 2011] Parent, G. & Eskenazi, M. (2011). Speaking to the Crowd: Looking at Past Achievements in Using Crowdsourcing for Speech and Predicting Future Challenges. , 3037-3040.

[Pham et al., 2010] Pham, P. T., Moens, M.-F., & Tuytelaars, T. (2010). Naming persons in news video with label propagation. , 1528--1533.

[Poor, 1988] Poor, H. (1988). An Introduction to Signal Detection and Estimation. ,

[Quénot et al., 2004] Quénot, G., Moraru, D., Ayache, S., Charhad, M., Guironnet, M., Carminati, L., Mulhem, P., Gensel, J., Pellerin, D., & Besacier, L. (2004). CLIPS-LIS-LSR-LABRI experiments at TRECVID 2004. ,

[Rabiner & Juang, 1993] Rabiner, L. R. & Juang, B.-H. (1993). Fundamentals of speech recognition. ,

[Rasheed & Shah, 2005] Rasheed, Z. & Shah, M. (2005). Detection and representation of scenes in videos. Multimedia, IEEE Transactions on, 7 (6), 1097 - 1105.

[Rasolofo & Savoy, 2003] Rasolofo, Y. & Savoy, J. (2003). Term proximity scoring for keyword-based retrieval systems. , 207--218.

[Robertson et al., 1998] Robertson, S. E., Walker, S., Jones, S., & Hancock-Beaulieu, M. M. (1998). Okapi at TREC-7. ,

[Rooij et al., 2010] Rooij, O., van Wijk, J., & Worring, M. (2010). MediaTable: Interactive Categorization of Multimedia Collections. IEEE Comput. Graph. Appl., 30 , 42--51.

[Sadlier & O'Connor, 2005] Sadlier, D. A. & O'Connor, N. E. (2005). Event detection in field sports video using audio-visual features and a support vector Machine. Circuits and Systems for Video Technology, IEEE Transactions on, 15 (10), 1225 - 1233.

[Schallauer et al., 2008] Schallauer, P., Ober, S., & Neuschmied, H. (2008). Efficient semantic video annotation by object and shot re-detection. ,

[Schavemaker et al., 2009] Schavemaker, J., Doets, P. J., Bailer, W., Stiegler, H., Lee, F., Neuschmied, H., Kraaij, W., Brandt, P., Eendebak, P., Ranguelova, E., & Thean, A. (2009). VdFP -- Video Fingerprinting Technologies for vdFP -- Video Fingerprinting Technologies for Media and Security Applications - D1: Report on Existing Technologies. ,

[Schüldt et al., 2004] Schüldt, C., Laptev, I., & Caputo, B. (2004). Recognizing Human Actions: A Local SVM Approach. , 32-36.

[Shuicheng YA et al., 2010] Shuicheng YA, N., Zhongyang HUAN, G., Qiang CHE, N., Zheng SON, G., Si LI, U., Xiangyu CHE, N., Xiaotong YUA, N., Tat-Seng CHU, A., Yang HU, A., & Shengmei SHE, N. (2010). Boosting Classification with Exclusive Context. ,

[Si et al., 2011] Si, Z., Pei, M., Yao, B., & Zhu, S.-C. (2011). Unsupervised learning of event AND-OR grammar and semantics from video. , 41-48.

[Sidiropoulos et al., 2009] Sidiropoulos, P., Mezaris, V., Kompatsiaris, I., Meinedo, H., & Trancoso, I. (2009). Multi-modal scene segmentation using scene transition graphs. , 665--668.

[Silaghi et al., 2000] Silaghi, M.-C., Bourlard, H., & Suisse, M. V. (2000). Iterative Posterior-Based Keyword Spotting Without Filler-Models: Iterative Viterbi Decoding And One-Pass Approach. ,

[Simperl et al., 2011] Simperl, E., Norton, B., & Vrandecic, D. (2011). Crowdsourcing Tasks within Linked Data Management. ,

[Siorpaes & Hepp, 2008] Siorpaes, K. & Hepp, M. (2008). Games with a Purpose for the Semantic Web. IEEE Intelligent Systems, ,

[Sjöberg et al., 2010] Sjöberg, M., Koskela, M., Chechev, M., & Laaksonen, J. (2010). {PicSOM} Experiments in {TRECVID} 2010. ,

[Smeaton & Over, 2010] Smeaton, A. & Over, P. (2010). TRECVID 2010 Known-Item Search. ,

[Smeaton & Over, 2011] Smeaton, A. & Over, P. (2011). TRECVID 2011 Known-

Item Search. ,

[Smeulders et al., 2000] Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-Based Image Retrieval at the End of the Early Years. IEEE Trans. Pattern Anal. Mach. Intell., 22 (12), 1349--1380.

[Snoek & Worring, 2005] Snoek, C. G. M. & Worring, M. (2005). Multimedia Event-Based Video Indexing using Time Intervals. {IEEE} Transactions on Multimedia, 7 (4), 638--647.

[Snoek & Worring, 2009] Snoek, C. G. M. & Worring, M. (2009). Concept-based video retrieval - Foundations and Trends in Information Retrieval. 4 ,

[Snoek et al., 2007] Snoek, C. G. M., Worring, M., Koelma, D. C., & Smeulders, A. W. M. (2007). A learned lexicon-driven paradigm for interactive video retrieval. IEEE Trans. Multimedia, 9 , 280--292.

[Snoek et al., 2006] Snoek, C. G. M., Worring, M., van Gemert, J. C., Geusebroek, J.-M., & Smeulders, A. W. M. (2006). The challenge problem for automated detection of 101 semantic concepts in multimedia. , 421--430.

[Snoek et al., 2010] Snoek, C. G. M., van de Sande, K. E. A., de Rooij, O., Huurnink, B., E.Gavves, odijk, D., de Rijke, M., Gevers, T., Worring, M., Koelma, D. C., & Smeulders, A. W. M. (2010). The MediaMill TRECVID 2010 Semantic Video Search Engine. ,

[Song et al., 2011] Song, C.-J., Lee, S.-P., Park, S.-J., Shin, S., & Jang, D. (2011). The Music Retrieval Method Based on The Audio Feature Analysis Technique with The Real World Polyphonic Music. ,

[Sukittanon & Atlas, 2002] Sukittanon, S. & Atlas, L. E. (2002). Modulation frequency features for audio fingerprinting. , 1773--1776.

[Surowiecki, 2004] Surowiecki, J. (2004). The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations. ,

[Tan et al., 2010] Tan, H. L., Zhu, Y., Chaisorn, L., & Rahardja, S. (2010). Audio onset detection using energy-based and pitch-based processing. , 3689 -3692.

[Tavanapong & Zhou, 2004] Tavanapong, W. & Zhou, J. (2004). Shot clustering techniques for story browsing. Multimedia, IEEE Transactions on, 6 (4), 517 - 527.

[Tejedor et al., 2007] Tejedor, J., García, R., Fernández, M., López-Colino, F. J., Perdrix, F., Macías, J. A., Gil, R. M., Oliva, M., Moya, D., Colás, J., & Castells, P. (2007). Ontology-Based Retrieval of Human Speech. ,

[Thaler et al., 2011] Thaler, S., Siorpaes, K., Simperl, E., & Hofer, C. (2011). A survey on games for Knowledge Acquisition. ,

[Troncy et al., 2011]Troncy, R., Mannens, E., Pfeiffer, S., & Deursen, D. V. (2011). Media Fragments URI 1.0. ,

[Viola & Jones, 2004] Viola, P. & Jones, M. (2004). Robust real-time face detection. International Journal of Computer Vision, 57 , 137--154.

[Waitelonis & Sack, 2009] Waitelonis, J. & Sack, H. (2009). Towards Exploratory Video Search Using Linked Data. ,

[Wan, 2003] Wan, A. L.-C. (2003). An Industrial-Strength Audio Search Algorithm. , 7-13.

[Wan et al., 2011] Wan, K.-W., Zheng, Y.-T., & Chaisorn, L. (2011). Known-item video search via query-to-modality mapping. , 1133--1136.

[Wang et al., 2011] Wang, K., Babenko, B., & Belongie, S. (2011). End-to-end scene text recognition. , 1457-1464.

[Wei et al., 2009] Wei, X.-Y., Jiang, Y.-G., & Ngo, C.-W. (2009). Exploring inter-concept relationship with context space for semantic video indexing. , 15:1--15:8.

[Wilkins, 2009] Wilkins, P. (2009). An Investigation Into Weighted Data Fusion for Content-Based Multimedia Information Retrieval. ,

[Willems et al., 2008] Willems, G., Tuytelaars, T., & Gool, L. (2008). An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector. , 650--663.

[Xie & Chang, 2006] Xie, L. & Chang, S. F. (2006). Pattern Mining in Visual Concept Streams. , 297--300.

[Xu & Chang, 2008] Xu, D. & Chang, S. F. (2008). Video Event Recognition Using Kernel Methods with Multilevel Temporal Alignment. 30 (11) , 1985--1997.

[Xu & Chua, 2006] Xu, H. & Chua, T.-S. (2006). Fusion of AV features and external information sources for event detection in team sports video. ACM Trans. Multimedia Comput. Commun. Appl., 2 , 44--67.

[Yan & Hauptmann, 2006] Yan, R. & Hauptmann, A. G. (2006). Probabilistic latent query analysis for combining multiple retrieval sources. , 324--331.

[Yan et al., 2004] Yan, R., Yang, J., & Hauptmann, A. G. (2004). Learning query-class dependent weights in automatic video retrieval. , 548--555.

[Yanagawa et al., 2007] Yanagawa, A., Chang, S.-F., Kennedy, L., & Hsu, W. (2007). Columbia University's Baseline Detectors for 374 LSCOM Semantic Visual Concepts. ,

[Yang et al., 2004] Yang, J., Chen, M.-y., & Hauptmann, A. (2004). Finding person x: Correlating names with visual appearances. 3115 , 270--278.

[Yang & Hauptmann, 2006] Yang, J. & Hauptmann, A. G. (2006). Annotating News Video with Locations. 4071 , 153-162.

[Yeung et al., 1998] Yeung, M., Yeo, B.-L., & Liu, B. (1998). Segmentation of Video by Clustering and Graph Analysis. Computer Vision and Image Understanding, 71 (1), 94 - 109.

[Yuan et al., 2010] Yuan, J., Zha, Z.-J., Zhao, Z., Zhou, X., & Chua, T.-S. (2010). Utilizing related samples to learn complex queries in interactive concept-based video search. , 66--73.

[Zhai & Lafferty, 2004] Zhai, C. & Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. ACM Trans. Inf. Syst., 22 , 179--214.

[Zhai & Shah, 2005] Zhai, Y. & Shah, M. (2005). A general framework for temporal video scene segmentation. 2 , 1111 -1116 Vol. 2.

[Zhang et al., 2011] Zhang, J., Hu, W., Yao, B., Wang, Y., & Zhu, S.-C. (2011). Inferring social roles in long timespan video sequence. , pp. 1456 --1463.

[Zhang et al., 2008] Zhang, Q., Tolias, G., Mansencal, B., Saracoglu, A., Aginako, N., Alatan, A., Alexandre, L. A., Avrithis, Y., Benois-Pineau, J., Chandramouli, K., Corvaglia, M., Damnjanovic, U., Dimou, A., Esen, E., Fatemi, N., Garcia, I.,

Guerrini, F., Hanjalic, A., Jarina, R., Kapsalas, P., King, P., Kompatsiaris, I., Makris, L., Mezaris, V., Migliorati, P., Moumtzidou, A., Mylonas, P., Naci, U., Nikolopoulos, S., Paralic, M., Piatrik, T., Poulin, F., Pinheiro, A. M. G., Raileanu, L., Spyrou, E., & Vrochidis, S. (2008). COST292 experimental framework for TRECVID 2008. ,

[Zhao et al., 2011] Zhao, B., Fei-Fei, L., & Xing, E. P. (2011). Online detection of unusual events in videos via dynamic sparse coding. , 3313 -3320.

[Zhao et al., 2003] Zhao, W., Chellappa, R., Phillips, P. J., & Rosenfeld, A. (2003). Face recognition: A literature survey. ACM Comput. Surv., 35 , 399--458.

[Zhu et al., 2005] Zhu, X., Elmagarmid, A. K., Xue, X., Wu, L., & Catlin, A. C. (2005). InsightVideo: toward hierarchical video content organization for efficient browsing, summarization and retrieval. IEEE Transactions on Multimedia, 7 (4), 648-666.

[Zhu et al., 2002] Zhu, X., Fan, J., Xue, X., Wu, L., & Elmagarmid, A. K. (2002). Semi-automatic Video Content Annotation. ,

[Zwicker & Fastl, 1990] Zwicker, E. & Fastl, H. (1990). {Psychoacoustics. Facts and Models (Springer Series in Information Sciences)}. ,

[Rooij et al., 2008] de Rooij, O., Snoek, C. G. M., & Worring, M. (2008). Balancing thread based navigation for targeted video search. , 485--494.

[Gemert et al., 2010] van Gemert, J. C., Veenman, C. J., Smeulders, A. W. M., & Geusebroek, J.-M. (2010). Visual Word Ambiguity. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 32 (7), 1271 -1283.

[Sande et al., 2010] van de Sande, K. E. A., Gevers, T., & Snoek, C. G. M. (2010). Evaluating Color Descriptors for Object and Scene Recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 32 (9), 1582 -1596.

[Ahn, 2006] von Ahn, L. (2006). Games with a Purpose. Computer, ,

[Ahn et al., 2008] von Ahn, L., Maurer, B., McMillen, C., Abraham, D., & Blum, M. (2008). ReCAPTCHA: Human-Based Character Recognition via Web Security Measures. ,

