Deliverable D1.1 State of the Art and Requirements Analysis for Hypervideo

Evlampios Apostolidis / CERTH
Michail Dimopoulos / CERTH
Vasileios Mezaris / CERTH
Daniel Stein / Fraunhofer
Jaap Blom / BEELD EN GELUID
Ivo Lašek / UEP
Mathilde Sahuguet / EURECOM
Benoit Huet / EURECOM
Nicolas de Abreu Pereira / RBB
Jennifer Müller / RBB

30/09/2012

Work Package 1: Intelligent hypervideo analysis

LinkedTV

Television Linked To The Web

Integrated Project (IP)

FP7-ICT-2011-7. Information and Communication Technologies

Grant Agreement Number 287911

Dissemination level	PU
Contractual date of delivery	30/09/2012
Actual date of delivery	30/09/2012
Deliverable number	D1.1
Deliverable name	State of the Art and Requirements Analysis for Hypervideo
File	LinkedTV_D1.1.tex
Nature	Report
Status & version	Reviewed & V1.0
Number of pages	92
WP contributing to the deliverable	1
Task responsible	CERTH
Other contributors	Fraunhofer, BEELD EN GELUID, UEP, EURECOM, RBB
Author(s)	Evlampios Apostolidis / CERTH Michail Dimopoulos / CERTH Vasileios Mezaris / CERTH Daniel Stein / Fraunhofer Jaap Blom / BEELD EN GELUID Ivo Lašek / UEP Mathilde Sahuguet / EURECOM Benoit Huet / EURECOM Nicolas de Abreu Pereira / RBB Jennifer Müller / RBB
Reviewer	Tomas Kliegr (UEP)
EC Project Officer	Thomas Kuepper
Keywords	shot segmentation, spatiotemporal segmentation, concept/event detection, face detection, face recognition, keyword extraction, automatic speech recognition, object re-detection, annotation tool

Abstract (for dissemination)	This deliverable presents a state-of-art and requirements analysis report for hypervideo authored as part of the WP1 of the LinkedTV project. Initially, we present some use-case (viewers) scenarios in the LinkedTV project and through the analysis of the distinctive needs and demands of each scenario we point out the technical requirements from a user-side perspective. Subsequently we study methods for the automatic and semi-automatic decomposition of the audiovisual content in order to effectively support the annotation process. Considering that the multimedia content comprises of different types of information, i.e., visual, textual and audio, we report various methods for the analysis of these three different streams. Finally we present various annotation tools which could integrate the developed analysis results so as to effectively support users (video producers) in the semi-automatic linking of hypervideo content, and based on them we report on the initial progress in building the LinkedTV annotation tool. For each one of the different classes of techniques being discussed in the deliverable we present the evaluation results from the application of one such method of the literature to a dataset well-suited to the needs of the LinkedTV project, and we indicate the future technical requirements that should be addressed in order to achieve higher levels of performance (e.g., in terms of accuracy and time-efficiency), as necessary.

Table of Contents

1 Introduction

This deliverable presents a state-of-art and requirements analysis report for hypervideo, authored as part of the WP1 of the LinkedTV project. The analysis starts with studying methods for the automatic and semi-automatic decomposition of the audiovisual content in order to effectively support the annotation process. Considering that the multimedia content comprises of different types of information (visual, textual, audio) we report various methods for the analysis of the three different streams: the visual stream; the audio stream; and the textual stream. Regarding the visual part we first list various state-of-art methods for the decomposition of the visual content into meaningful parts, which can be: temporal segments (i.e., shots or scenes); spatial segments (i.e., chromatically uniform regions in still images); and spatiotemporal segments (extracted for instance, by tracking moving parts in a sequence of frames). Subsequently, we report various algorithms for the association/labelling of these segments with a first form of semantic descriptors, with the use of concept detection and face analysis methods. Concerning the non-visual information, we describe several methods for text analysis (i.e., keyword extraction, text clustering) and audio analysis (i.e., speech recognition, speaker identification) in order to extract useful information that is complementary to the information extracted from the visual stream. By exploiting this multi-modal information we will be able to achieve improved analysis results, for instance a better decomposition of the media content into more meaningful segments from a human perspective. Subsequently we investigate several state-of-art methods for instance-based labelling and event detection in media, which will enrich the previous analysis results by associating new more descriptive labels to the content segments, using input from all the different modalities. Finally we present various annotation tools which could integrate the developed analysis results so as to effectively support users in the semi-automatic linking of hypervideo content, and report on the initial progress in building the LinkedTV annotation tool based on one of the reviewed existing tools. For each one of the different classes of techniques being discussed in the deliverable (e.g. shot segmentation, scene segmentation, etc.) we report the evaluation results from the application of one such method of the literature to a dataset well-suited to the needs of the LinkedTV project, and we indicate the future technical requirements that should be addressed in order to achieve higher levels of performance (e.g., in terms of accuracy and time-efficiency), as necessary.

The main body of the deliverable starts in section 2 with a detailed presentation of the use-case scenarios in the LinkedTV project and a brief description of the main user-side requirements that arise from them. In this section, by "user" we mean the viewer of the LinkedTV videos. Specifically, this purpose two different types of scenarios are presented, the News Show and the Documentary scenario and for each scenario three different user archetypes are described. Through the analysis of the distinctive needs and demands of each user we point out the technical requirements from a user-side perspective, which help us to define in the following sections of this deliverable the different techniques that should be utilized and cooperate in order to provide the described services to the user.

Section 3 reports on methods for the preprocessing and representation of the visual information. Since shot is the basic elementary unit of a video sequence, the first subsection discusses state-of-art techniques for shot segmentation, which is the first preprocessing step for further analysis of the video content. The next subsection is dedicated to another type of temporal segmentation methods that perform shot grouping into scenes, thus providing a higher level of abstraction. The section continues with techniques for the spatial segmentation of still images into regions that are meaningful from a human perspective (e.g., they represent individual objects or parts of a picture). Combining temporal and spatial information, the next subsection discusses algorithms for the spatiotemporal segmentation of the video that aim at the detection and tracking of moving parts in a sequence of frames. Since the efficient representation of the visual information is a basic prerequisite for the majority of the higher-level visual analysis techniques in the sequel of this deliverable, we then proceed with reviewing state-of-art methods for the description of the visual content, distinguishing between visual descriptors that represent global features of the image (e.g., color or texture) and local features (e.g., edges or salient regions). Section 3 concludes with the technical requirements for the different problems discussed above (shot segmentation etc.), as they are derived from further analysis of the user side requirements in light of some preliminary results of the related techniques. Specifically, reporting the evaluation results of the currently selected techniques on a LinkedTV dataset, this subsection describes the future technical improvements that are required in order to achieve better performance in terms of detection accuracy and time efficiency.

Section 4 deals with the problem of visual object and scene labelling and discusses state-of-art methods for concept detection and face analysis. In particular, the first subsection reports on techniques for the consecutive processing steps of a concept detection pipeline i.e., the presentation of the visual content with the use of low-level visual descriptors, the intermediate step of visual word assignment when local descriptors are utilized, and the learning and classification step for the labelling of the visual content. The section continues focusing on face analysis techniques and presenting various algorithms for face detection, face recognition, and face clustering. Moreover this part of the section lists various available tools that could be used to perform the above mentioned face analysis tasks. Consequently, based on the additional processing capabilities that modern GPUs can offer, we proceed with discussing several state-of-art visual analysis algorithms (e.g., shot segmentation, feature extraction and description etc.) that have been modified appropriately, in order to exploit this additional processing power and accelerate their performance. At the final subsection we present the evaluation results of some proposed methods and tools from the relevant literature, and for each one of them we indicate the future technical requirements that should be fulfilled to achieve higher levels of performance.

Besides the visual part there are other sources of information that could be used to enrich the overall amount of extracted information. So, the next section presents techniques for the extraction of complementary information to the visual stream, via text and audio analysis. Initially we report on methods for text analysis, by focusing particularly on keyword extraction and text clustering methods, while in the following we briefly list several available tools for these tasks. Subsequently we discuss various state-of-art techniques for audio analysis. These techniques correspond at two different types of analysis, named speech recognition and speaker identification. Finally, this section ends with a description of the evaluation results from the currently used algorithms applied on a well-suited for the LinkedTV purposes dataset, and the definition of the technical requirements that have to be addressed for further improvements of the algorithms' performance.

Section 6 survey methods for event and instance-based labelling of visual information. The analysis starts with the task of instance-based labelling, by discussing various state-of-art methods for object re-detection. By "object re-detection" we mean a two step procedure, where a system initially extracts low-level visual characteristics from a particular object presented in a user-specified image, and subsequently detects and demarcates appropriately (e.g., with a bounding box) all the following occurrences of this object in a video sequence. Section 6 proceed with reviewing state-of-art methods for event detection. Initially, we list various approaches that utilize model vectors for the representation of the visual content. Following, we describe algorithms for the reduction of the high dimensionality of these feature vectors. Finally we present techniques for associating media with events. This sections concludes with a summary of the evaluation results of some baseline approaches for object re-detection and face analysis, while some indicative results of the two techniques are also illustrated. Based on these results we outline the current problems and the technical requirements, and we discuss some plans for future improvements.

Section 7 deals with the user assisted annotation problem. Here the term "user" refers to the video editor which interacts with the automatic analysis results and produces the final video. Initially, we analyse the workflow and the particular needs of the annotation procedure within the LinkedTV scope. The sections continues with some prominent examples of available tools for audio transcription and object description. Afterwards, we point out the technical requirements and functionalities that must be supported by the annotation tool for the purposes of the LinkedTV project. This analysis will provide to us some guidelines for the selection of the most appropriate tool and will help us to indicate the future improvements that should be reached. The final subsection is a presentation of the current status of the employed annotation tool.

The deliverable concludes in section 8 with a summary of the reported multimedia analysis methods and a few concluding remarks about the current status and the performance of the tested techniques. Moreover, considering the available space for improvements, in this final part we note some future plans for improvements by integrating novel technologies and by combining input for different existing techniques, in order to achieve more accurate results from the perspective of media analysis applications and needs.

List of Figures

List of Tables

2 Scenarios overview and main user-side requirements

The audio-visual quality and content of videos found in the web, as well as their domains, are very heterogeneous. For this reason, LinkedTV has foreseen possible scenarios for inter-linkable videos in WP 6, which will serve as reference for the necessities to be expected from WP 1 regarding video content recognition techniques on the intended material (see also [SAM$^+$12a, SAM$^+$12b, SAS$^+$12]). LinkedTV envisions a service that offers enriched videos which are interlinked with the web, and targets a broad audience. For starting our work on this problem however, we have sketched three archetypal users (by “users” here we mean the consumers-viewers of the LinkedTV videos) and their motivations for two scenarios. We believe that this step in the project planning is crucial, and we try to be as concrete and detailed as possible. The use cases derived below will serve as guidelines for the identification of specific requirements and functionalities that must be sufficiently supported by the LinkedTV platform, from a user (viewer) perspective. This step will consequently help us to define the techniques that must be implemented and integrated in the multimedia analysis part, as well as synergies among them, in order to provide the desired services to each individual user. Based on this analysis, sections 3 to 7 of this deliverable will focus on state-of-art methods and the technical requirements for the implementation of the selected techniques.

2.1 News Show Scenario

The news show scenario uses German news broadcast as seed videos, provided by Public Service Broadcaster Rundfunk Berlin-Brandenburg (RBB)[1]. The main news show is broadcast several times each day, with a focus on local news for Berlin and Brandenburg area (Figure 1). For legal and quality reasons, the scenario is subject to many restrictions as it only allows for editorially controlled, high quality linking. For the same quality reason only web sources selected from a restricted white-list are allowed. This white-list contains, for example, videos produced by the Consortium of public service broadcasting institutions of the Federal Republic of Germany (ARD) and a limited number of approved third party providers.

The audio quality of the seed content can generally be considered to be clean, with little use of jingles or background music. Interviews of the local population may have a minor to thick accent, while the eight different moderators have a very clear and trained pronunciation. The main challenge for visual analysis is the multitude of possible topics in news shows. Technically, the individual elements will be rather clear; contextual segments (shots or stories) are usually separated by visual inserts and the appearance of the anchorperson, and there are only few quick camera movements.

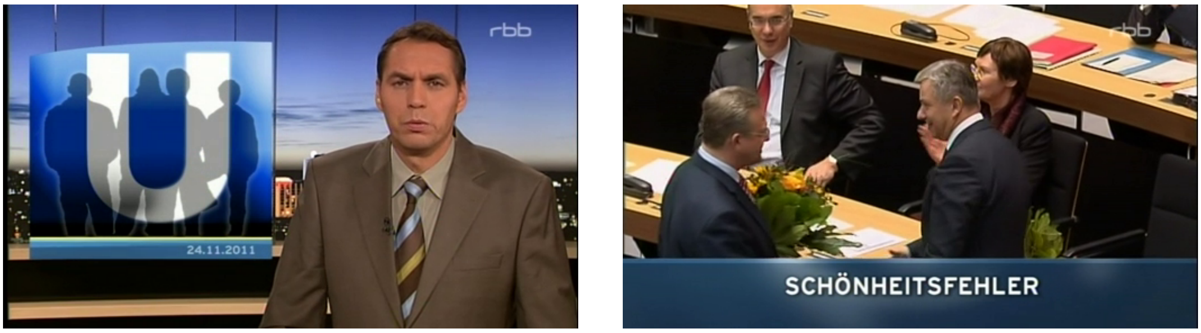


Figure 1: Screenshots from the news broadcast scenario material, taken from the German news show “rbb Aktuell”

2.1.1 Scenario Archetypes

Ralph is a 19-year old carpenter who has always lived in Prenzlau, Brandenburg (100 km from Berlin). After school he served an apprenticeship as carpenter in the neighboring village Strehlow. He does not care particularly for the latest technology, but as a “digital native” it is not difficult for him to adapt to new technologies and to use them.

[1] www.rbb-online.de

Ralph comes home from working on a building site in Potsdam, and starts watching the enhanced "rbb AKTUELL" edition. The first spots are mainly about politics and about Berlin. Ralph is not particularly interested, neither in politics nor in Berlin as he lives in a small town in Brandenburg, so he skips these parts after the first seconds. After a while there is the first really interesting news for Ralph: a spot about the restoration of a church at a nearby lake; as a carpenter, Ralph is always interested in the restoration of old buildings. Therefore, he watches the main news spot carefully and views an extra video and several still images about the church before and after its restoration. Finally, the service also offers links to a map and the website of the church which was set up to document the restoration for donators and anyone else who would be interested. Ralph saves these links to his smart phone so he can visit the place on the weekend.

Technical requirements: for skipping, scene/story segmentation is needed (Section 3.2). Other videos representing the church before and after the restoration where the object is clickable should be retrieved via object re-detection using the viewed still images (Section 6.1).

Nina, 32, is a typical inhabitant of Berlin's hippest quarter, Prenzlauer Berg. She is a well-educated and well-informed young mother, and when a subject is interesting for her, she will take the time to understand it properly.

Nina's baby has fallen asleep after feeding, so she finds some time for casually watching TV, to be informed while doing some housework. Browsing the programme she sees that yesterday's enhanced "rbb AKTUELL" evening edition is available and starts the programme. Nina watches the intro with the headlines while starting her housework session with ironing some shirts. Watching a news spot about Berlin's Green Party leader who withdrew from his office yesterday, Nina is kind of frustrated as she voted for him and feels her vote is now "used" by someone she might not have voted for. She would like to hear what other politicians and people who voted for him think about his decision to resign. She watches a selection of video statements of politicians and voters and bookmarks a link to an on-line dossier about the man and his political carrier, which she can browse later on her tablet. Eventually, the baby awakes so Nina pauses the application so she can continue later.

Technical requirements: the politician needs to be identified, either via automatic speech recognition (Section 5.2.1) and speaker identification (Section 5.2.2) or via face detection (Section 4.2.1) and face recognition (Section 4.2.2). If he is moving and must be clickable, spatiotemporal segmentation (Section 3.4) is required.

Peter, 65, is a retired but socially active widower who lives in Potsdam, a small and wealthy town near Berlin. Since his retiring he has a lot of time for his hobbies, and is involved in several activities in his neighborhood. Apart from watching TV and listening to the radio he is also very interested in new technology and likes to use new services via the internet.

Peter watches the same news show as Ralph and Nina, but with different personal interest and preferences. One of the spots is about a fire at famous Café Keese in Berlin. Peter is shocked. He used to go there every once in a while, but that was years ago. As he hasn't been there for years, he wonders how the place may have changed over this time. In the news spot, smoke and fire engines was almost all one could see, so he watches some older videos about the story of the famous location where men would call women on their table phones – hard to believe nowadays, he thinks, now that everyone carries around mobile phones! Moreover he searches for videos depicting fires in other places of Berlin city. After checking the clips on the LinkedTV service, he returns to the main news show and watches the next spot on a new Internet portal about rehabilitation centers in Berlin and Brandenburg. He knows an increasing number of people who needed such facilities. He follows a link to a map of Brandenburg showing the locations of these centers and bookmarks the linked portal website to check some more information later. At the end of the show, he takes an interested look at the weather forecast, hoping that tomorrow would be as nice as today so he could go out again to bask in the sun.

Technical requirements: to recognize a recently burned down Café, object re-detection is of little help, which is why keyword extraction is needed (Section 5.1.1). The algorithm is fed by automatic speech recognition (Section 5.2.1). The retrieval of relative videos presenting similar fire disasters, can be based on event detection (Section 6.2).

2.2 Documentary Scenario

The basis of the television content for the documentary scenario is the Dutch variant of "Antiques Roadshow", "Tussen Kunst & Kitsch"[2]. In this programme from the Dutch public broadcaster AVRO[3], people

[2] tussenkunstenkitsch.avro.nl

[3] www.avro.nl

can bring in objects of art-historical interest to be assessed on authenticity and interest by experts, who also give an approximate valuation (see Figure 2). Each show is recorded in another location, usually in a museum or another space of cultural-historical interest. The programme consists of various scenes in which a wide array of objects are discussed, from paintings to teapots and vases to toy cars.

The scenario is focused on enriching objects, locations and people in the programme with high-quality related items. Since the show is produced by an established Dutch public broadcaster, these related items are currently restricted to a curated list of sources. This list contains (a) items related directly to the Antique Roadshow, like scenes from other episodes and the programme website, (b) items from established thesauri like the United List of Artist names[4], and (c) items from established on-line sources such as the European digital library (Europeana)[5] and Wikipedia[6].

The speech of the experts is trained, but spontaneous. Dialects of local regions like Groningen might appear, but usually, only one person is talking. For all episodes, subtitle files are available. There are hardly any camera movements in the shots and the cuts from shot-to-shot and scene-to-scene are usually hard cuts. In the wide shots, there is a rather large number of visitors visible, so that many faces are in view simultaneously.

Figure 2: Screenshots from the documentary scenario material, taken from the Dutch show “Tussen Kunst & Kitsch”

2.2.1 Scenario Archetypes

Rita, 34, is an administrative assistant at the Art History department of the University of Amsterdam. She didn’t study art herself, but spends a lot of her free time on museum visits, creative courses and reading about art.

In the latest episode of “Kunst & Kitsch”, the show’s host Nelleke van der Krogt gives an introduction to the programme. Rita wonders how long Nelleke has been hosting the show and who were the hosts before her, since it’s been on for a very long time, but Rita didn’t start watching it until 2003. Rita also sees that she can select to view all segments from each specific “Tussen Kunst & Kitsch” expert, including her and her sister’s favorite: Emiel Aardewerk. She sends a link to her sister which will bring her to the overview of all segments with Emiel, before she switches off. Rita decides to visit the service again tomorrow to learn even more.

Technical requirements: specific segments of videos containing a person of interest can be identified automatically by using a combination of techniques such as shot and scene segmentation (Sections 3.1 and 3.2) and face clustering (Section 4.2.3) or speaker identification (Section 5.2.2). Note that, while their respective names appear as a text overlay in the video, their names are hardly mentioned in the subtitles.

Bert, 51, has an antiques shop in Leiden, which he has owned for the past 25 years. He studied Egyptology and Early Modern and Medieval Art at his alma mater, Leiden University. He visits many art fairs and auctions – both national and international – but also takes part in on-line auctions and actively scours the web looking for information on objects he is interested in purchasing and re-selling.

Bert has recently bought a wooden statuette depicting Christ at an antiques market for 100 €. He suspects it is quite old (17th century) and that he could re-sell the object for a nice price in his shop. He has done quite a bit of research on-line already, but also would like to see if anything like it has ever

[4] www.getty.edu/research/tools/vocabularies/ulan

[5] www.europeana.eu

[6] www.wikipedia.org

appeared on “Tussen Kunst & Kitsch”. He requests the system to find segments of various episodes which depict similar objects. After some browsing, he finds a segment about an oak statue of a Maria figure, originating in the province of Brabant, and made in the late 17th century. The value is estimated at 12,500 €.

With this rough estimate in mind, Bert enters an on-line auction and places the first bidding price for his statue into the system. While he is already at the auction portal, he uses the keywords from the “Tussen Kunst & Kitsch” session as search parameters (statue, wood, 17th century) and sees that an individual seller has put a piece for sale on a user-based platform for a mere 500 €. It is quite worn, but Bert knows that with a little restoration work, the value is likely to increase four-fold. He purchases the statue and decides he has done enough research and business for today.

Technical requirements: since Bert is looking for segments from a clear subset of videos, shot and scene segmentation (Sections 3.1 and 3.2) are needed, while similar objects can be retrieved through a Content Based Image Retrieval (CBIR) framework that is based on concept detection (Section 4.1). A strong keyword extraction (Section 5.1.1) is also required for searching relative information.

Daniel, 24, is a student of Economics and Business, a research master at the University of Groningen in the north of the Netherlands. In his free time he likes to scour flea markets and thrift stores in search of antiques and collectibles that he can re-sell for a nice profit on-line.

Daniel is bargain hunting at a flea market in Drachten in the north of Holland when he spots a nice silver box, which he thinks is roughly from the 18th or 19th century. He has seen this design before, but he can’t remember exactly where. He takes a snap of the jar and uploads it to the “Tussen Kunst & Kitsch” interface, where he is given an overview of related programme segments and other information on similar objects. He is shown an overview of all programme segments that possibly contain similar or related objects. Since there are a large number of segments, he decides to filter out any objects that fall outside the scope of his search. After refining the search results, Daniel finds a segment that is about a silver tea jar from 18th century Friesland.

It looks quite similar to the one he has found, but it is still not exactly the comparison he is looking for. He sees that there are recommendations for similar videos from not just other “Tussen Kunst & Kitsch” segments, but also from external sources. He sees a recommended object from Europeana, a silver etrog box used for the Jewish festival of Sukkot, which was made in London in 1867. Daniel decides to buy the box and heads home with a new possible treasure.

Technical requirements: similar to Bert, shot and scene segmentation (Sections 3.1 and 3.2) are required for video segmentation, while keyword extraction (Section 5.1.1) is needed to find videos on the same topic. However, since comparable videos are of interest as well, video with identical or similar objects can be retrieved via object re-detection (Section 6.1) and concept detection (Section 4.1) using the images he uploads.

3 Visual Information Preprocessing and Representation

This section is a survey of state-of-art methods and techniques for the preprocessing and representation of the visual information. The studied aspects include: temporal segmentation of a video sequence into elementary units like shots and scenes; spatial segmentation of still images to regions; spatiotemporal segmentation of video content for the detection and tracking of moving objects; and techniques for the description and representation of the visual information. Moreover in the final subsection we report some evaluation results from techniques we already tested on LinkedTV content, and we indicate the future technical requirements that arise based on these results as well as on the scenario specifications of Section 2.

3.1 Shot Segmentation

Video shot segmentation partitions a video into basic structural parts which correspond to a sequence of consecutive frames captured without interruption by a single camera. Shot boundary detection focuses on the transitions between consecutive video frames. Different kinds of transitions may occur and the basic distinction is between abrupt and gradual ones, where the first occurs when stopping and restarting the video camera and the second is caused by the use of some spatial, chromatic or spatio-chromatic effects such as fade in/out, wipe or dissolve. Various techniques have been proposed for the efficient detection of transitions between successive frames and detailed surveys can be found in [YWX$^+$07, GN08, CNP06]. In the following part, we categorize the various different methods based on their applicability to uncompressed or compressed data.

Uncompressed data domain: the simplest of these techniques is the pair-wise pixel comparison, or the alternatively named template matching. The detection is performed by evaluating the difference in intensity or color values of corresponding pixels in two successive frames and by comparing the number of changes against a threshold. An extension to N-frames, which uses some pixel intensity evolution patterns for gradual transition detection, is described in [WCK$^+$03]. A drawback of these methods is their sensitivity to small camera and object motion and thus, in order to tackle this some techniques smooth the images by applying a 3x3 average filter before performing the pixel comparison. Several algorithms (see for example [GCAL03]) increase the efficiency of this approach by sub-sampling pixels from particular positions and representing the visual content with less but more descriptive information.

To overcome the locality issues of pixel-based methods, some researchers introduced techniques based on intensity or color histograms. Histograms do not incorporate any spatial information, a fact that makes them less sensitive to local or small global movements. These methods initially compute gray or color components of the image, subsequently they arrange them into a number of bins and finally a shot boundary is detected after a bin-wise comparison between histograms of two successive frames [CGP$^+$00]. Several extensions of this technique have been proposed, including weighted differences, Chi-Square tests, or histogram intersections, combined with different color spaces such as RGB, HSV, YIQ, Lab, Luv, and Munsell. However, a drawback of these methods is that transitions between frames with different content but similar intensity or color distributions are not detectable. To overcome this, some methods propose the segmentation of the image into non-overlapping blocks and the computation of histogram differences between consecutive frames at block-level [BDBP01].

Image features describing the structural or the visual information of each frame have also been proposed. A widely applied algorithm based on edge analysis was used in [ZMM99] for detecting and classifying video production effects like cuts, fades, dissolves and wipes which are usually presented between consecutive shots. In [Lie99] Lienhart studies the effectiveness of the Edge Change Ratio (ECR) algorithm proposed in [ZMM95], for the detection of both abrupt and two types of gradual transitions, caused by dissolve and fade effects. The detection is performed by counting the number of exiting and entering pixels which correspond to edges between successive frames. Recently, Park et. al. [PPL06] proposed a method for shot change detection by using the SIFT local descriptor of Lowe [Low04] which represents local salient points of the image, while another clustering-based algorithm employing the same descriptor was introduced in [CLHA08]. Furthermore, in [TMK08] a video shot meta-segmentation framework is described, which combines two novel features, called Color Coherence and Luminance Centre of Gravity, and a Support Vector Machines (SVM) classifier.

Several other machine learning approaches have also been proposed to solve the shot boundary detection problem. Some of them implement temporal multi-resolution analysis [Ngo03, CFC03], while other methods model different types of transitions by applying Coupled Markov Chains [SB03]. Differently, Lienhart and Zaccaring [LZ01] proposed a neural network classifier for transition detection, and a training-based approach was developed in [Han02]. Another method based on supervised classification is described in [CLR07], while the authors in [LYHZ08] utilize some image features as input vectors to Support Vector Machine (SVM) classifiers.
Finally, other approaches include: adaptive thresholding by using a sliding window method [LL05, SLT$^+$05] or by utilizing the information entropy [ZyMjWn11]; graph-based techniques [LZL$^+$09, YWX$^+$07]; motion-aided algorithms [LL02]; fuzzy theory [GHJ05, XW10]; block-level statistics [LHSL98]; Singular Value Decomposition (SVD) [CKP03]; B-spline interpolation [NT05]; QR-decomposition [AF09]; and fusion techniques [MDY08].

Compressed data domain: significant efforts have also been made for the implementation of techniques that will be applicable to compressed video data. Several works try to exploit the structural characteristics of compression standards like MPEG and H.264/AVC in order to describe efficient techniques for shot boundary detection. Some early approaches are reviewed in [KC01, LS03]. In [LS03] Lelescu and Schonfeld proposed a novel real-time approach for both abrupt and gradual transition detection based on statistical sequential analysis. Their method uses stochastic processes and models the shot changes by modifications in the parameters of the stochastic process. Alternatively, the authors in [CI01] introduce a method which is based on temporal distribution of macro-block types, while in [DVZP04] the detection of gradual transitions is based on the magnitude and the direction of the calculated motion vectors. In [KGX$^+$06] a DCT-based method is described, which exploits the perceptual blockiness effect detection on each frame without using any threshold parameter. Lee and Hayes [LH01] use the number and the mean of bi-directional predicted macro-blocks within a B-frame to define an adaptive threshold for the identification of shot changes and a variation of this approach has been described few years latter in [FBOD08]. Bescos [Bes04] combined deterministic, statistical parametric and statistical non-parametric metrics and applied them to DC images to detect abrupt and gradual transitions, while in [PC02] dissolves and wipes are detected by employing macroblocks of P and B frames. An alternative method for video shot segmentation based on encoder operation is described in [VFSC06]. Finally, several techniques have also been introduced for the H.264/AVC video and some early attempts are described in [LWGZ04].

3.2 Scene Segmentation

Scenes are higher-level temporal segments covering either a single event or several related events taking place in parallel. By segmenting a video to scenes, one organizes the visual content in higher levels of abstraction, thus contributing to the more efficient indexing and browsing of the video content. Mainly, scene segmentation techniques use as input the shot segments of a video and try to group them into sets according to their semantic similarity. Several methods have been proposed for this task and reviews can be found in [ZL09, Pet08]. In our analysis we define four major categories for scene segmentation based on visual analysis only; (a) graph-based methods; (b) methods using inter-shot similarity measurements; (c) clustering-based methods; and (d) other methods. Additionally, we identify an extra group of methods based on audio-visual analysis.

One of the most efficient graph-based approaches called Scene Transition Graph (STG) was described by Yeung et. al. [YYL98]. The authors calculate the visual similarity of shots, by measuring the similarity between extracted keyframes for each shot, and introduce some temporal constraints in order to construct a directed graph. By representing the inter-connections of shots with this graph, the scene boundaries are extracted by finding the cut edges of this graph. This graph-based approach has been adopted by other researchers and several years later Rasheed and Shah [RS05] proposed a weighted undirected graph called Shot Similarity Graph (SSG), while Ngo introduced another similar approach algorithm called Temporal graph in [NMZ05].

Similarity or dissimilarity measures at shot-level have also been proposed as a technique for grouping shots into scenes. Rasheed et al. [RS03] proposed a two-pass algorithm which calculates a color similarity measure and involves the scene dynamics. Truong [TVD03] implemented inter-shot similarity measures satisfying the film grammar, which is a set of production rules about how the movies or TV shows should be composed. Rui et. al. [RHM99] define a function of color and activity of shots in order to measure the inter-shot similarity and construct a table-of-contents. A similar approach was proposed by Zhu in [ZEX$^+$05]. Finally, other works [CKL09] define dissimilarity and similarity measures respectively, using a "bag-of-words" representation for each shot, extracted by histogram analysis. We will describe the "bag-of-words" representation in detail in Section 4.1.1.3, since it is a basic processing step of many concept detection algorithms as well.

A group of clustering-based methods have also been proposed for scene segmentation. Some of them describe hierarchical clustering [HLZ04, GPLS02], while other approaches describe an unsupervised clustering algorithm by combining multi-resolution analysis and Haar wavelet transformations [LZ07]. Okamoto et al. [OYBK02] used spatiotemporal image slices for clustering, while in [ZLCS04] the segmentation is performed via spectral clustering after the representation of the video with K-partite graphs. Moreover, another spectral method has been proposed in [OGPG03] and is based on visual similarity and temporal relationships.

Several different kinds of methods have also been introduced. Hoashi et al. [SHMN04] parse video scenes by utilizing a statistical model and Support Vector Machines (SVM). Hsu et. al. [HC04] proposed an approach, called BoostMe, which is based on discriminative models, while Zhai et. al. [ZS06] exploited the Markov Chain Monte Carlo (MCMC) algorithm to determine the boundaries between two video scenes. Other statistical-based proposals perform scene boundary detection using Hidden Markov Models (HMM) [XXC$^+$04] and Gaussian Mixture Models (GMM) [LLT04]. Alternatively, Zhao et. al. [ZWW$^+$07] determined the optimal scene boundaries by addressing this task as a graph partition problem and by using the N-cut algorithm, while other methods consider the spatiotemporal coherence [ZL09], or implement a competition analysis of splitting and merging forces between shots. Moreover, in [AK02] a novel method based on the so called "mosaic" is described, which utilizes the information of specifically physical settings and camera locations, while in [CLML08] the authors defined a similarity measure based on the background information obtained by the "mosaic" technique. Finally, some rule-based methods have also been described in the relevant literature and indicative examples can be found in [CRzT03, WC02].

Besides visual content, information extracted from the audio channel could also be useful as a second criterion for scene change detection and for this, many approaches have been presented. In [AKT03] the authors exploit the audio-visual features and find the correspondences between two sets of audio scenes and video scenes using a nearest neighbour algorithm. Iurgel [IMER01] proposed a method for news videos by applying Hidden Markov Models (HMMs) to audio and video features, while authors in [SMK$^+$11] defined a Generalized STG approach that jointly exploits low-level and high-level features automatically extracted from the visual and the auditory channel. Several other methods for video scene segmentation that incorporate methods for audio and visual data analysis, can be studied in [CTKO03, VNH03, CMPP08].

3.3 Spatial Segmentation

Image segmentation is the process of separating an image into different parts which correspond to something that humans can easily separate and view as individual objects or image parts (e.g., sky or sea). The segmentation process is based on various image features like color (see [LM01b]), pixel position (see [JBP10, SGH98]), texture (see [MKS04a]) etc. Several examples of recent efforts are reported in [Cat01, Bha11, SL10] and others.

Thresholding is definitely one of the most simple, popular and effective approaches used in image segmentation. This technique is generally used for gray scaled images and aims to partition an input image into pixels of two (background-foreground) or more values via comparison with a predefined threshold value. However, the lack of spatial information leads to false segmentation in cases of objects with low contrast or noisy images with varying background, and hence additional tools are required for coloured or synthetic images. Some methods focusing on efficient choice of the threshold value are reported in [Tra11].

Besides thresholding on pixel intensity values, another widely applied technique is the thresholding of histograms. For gray-level images, peaks and valleys in 1D histograms can be easily identified as objects and backgrounds, while for the case of color images either 2D or 3D histograms must be used to locate the N-significant clusters and their thresholds. However, noise could lead to segmentation ambiguities and thus some smoothing provisions are usually adopted. Comprehensive studies of histogram-based methods can be found in [LM01b, Sha00, LM01b]. Some proposed techniques employ HUE histograms, while other approaches combine Fisher LDA, entropy-based thresholding, or B-splines with adaptive thresholding algorithms. Edge-based approaches partition the image on the basis of abrupt changes either in intensity or in texture. Image edges are detected and linked into contours that represent boundaries of image objects. Edge detectors can be gradient-based, zero crossing, Laplacian/Gaussian and coloured. The most broadly used detectors are the Sobel operator, the Prewitt operator, the Roberts' cross operator and the Canny edge detector [Can86]. Several approaches can be found in the literature, including boundary analysis algorithms, heuristics for adaptive thresholding, active contours, or clustering-based approaches. Detailed surveys can be found in the previously mentioned works [LM01b, Sha00].

These studies also report some early region-based techniques. These techniques can be divided into two groups; methods using region growing algorithms; and methods applying region splitting and merging. Region growing technique sets some initial seeds and group pixels or sub-regions into large regions based on predefined homogeneity/similarity criteria. Several early approaches have been proposed including techniques exploiting gradient information and fuzzy logic. Region split and merge divides the image into disjoint regions and then either merges and/or splits them to satisfy the pre-specified constraints. Numerous variations have been investigated, like Watershed transformations, fuzzy expert systems, Voronoi's diagrams and pyramidal segmentation [Mis11].

Another class of methods represents images using graphs and the image segmentation is performed via graph partitioning into a set of connected components that correspond to image regions. The various different methods can be divided in four categories: (a) minimal spanning tree methods, where the clustering and grouping of pixels are performed on the minimal spanning tree [FH04]; (b) graph cut with cost function methods, where relevant methods include Minimal cut, Normalized-cut [SM00], Ratio cut, Minmax cut, Mean cut, and graph cut on Markov random fields models [SK10, LS10, DB08, LVS08]; (c) shortest path-based methods [BS07]; and (d) other methods that they do not fit to the above categories, such as the random walker method [Gra05] and the dominant set method [PP03].

Image segmentation can also be performed effectively by clustering image pixels into meaningful subgroups. Clustering either requires some initial seeds defined by the user or uses non-parametric methods for finding the salient regions without the need for seed points. The most common clustering method widely applied in image segmentation techniques is the K-means [CLP98], while a fuzzy version called Fuzzy C-means (FCM) has been proposed in [TP98]. A more sophisticated approach has been introduced in [MKS04a] where the authors use the K-Means-with-connectivity-constraints (KMCC) algorithm, proposed in [KS00]. Their unsupervised region-based segmentation technique uses a combination of conditional filtering by a moving average filter and pixel classification by means of the KMCC algorithm, in order to form connected regions that correspond to the objects contained in the image. Alternatively, a robust Mean-Shift algorithm for image segmentation is presented in [CM02], while other clustering-based techniques use the Expectation-Maximization (EM) algorithm and the Gibbs Random Field (GRF).

Neural networks have also been widely applied in image segmentation procedures. Relative surveys can be found in [Cat01, EPdRH02]. According to [EPdRH02], several different types of Artificial Neural Networks (ANNs) have been trained to perform both pixel- and feature-based image segmentation such as: Feed-Forward ANNs, Self-Organizing Maps (SOMs)/Self-Organizing Feature Maps (SOFMs), Hopfield Networks and a parallel approach called Convolutional Neural Networks (CNN). Approaches using only pixel data include Probabilistic ANNs, Radial Basis Function (RBF) Networks, Constraint Satisfaction ANNs, and Pulse-Coupled Neural Networks (PCNN). Developed techniques based on image features are Recursive Networks, variants of RBF Networks, Principal Component Networks, Neural Gas Networks and Dynamic ANNs.

Moreover, other spatial segmentation approaches exploit the fuzzy set theory. We have already reported the FCM clustering algorithm. Several approaches which combine fuzzy logic with neural networks are reviewed in [Bha11]. Examples include: a Fuzzy Min-Max Neural Network [EFP05]; a Fuzzy Labelled Neural Gas (FLNG) [VHS$^+$06]; and a Neuro-Fuzzy system, called Weighted Incremental Neural Networks (WINN) [Muh04].

Finally, a number of methods follow a different approach. Some alternatives of the methods mentioned above incorporate Genetic algorithms for the optimization of relevant parameters, while others are combined with wavelet transformations [LST00, NSK00]. Moreover, Mishra et. al. [Mis11] reviewed some Agent-based image segmentation techniques, while several probabilistic and Bayesian methods are described in Sharma's thesis [Sha00]. Finally, a new algorithm using Particle Swarm Optimization (PSO) was proposed in [Moh11] and a novel approach of active segmentation with fixation has been introduced by Mishra et. al. in [MAC09].

3.4 Spatiotemporal Segmentation

Spatiotemporal segmentation refers to the task of dividing video frames into regions that may correspond to moving objects in the scene. This higher-level information about the objects represented in a sequence of frames can be obtained via the combination of information from the temporal segmentation of the sequence of frames and the spatial segmentation of each frame into regions. For instance, by using the time boundaries defined from temporal segmentation and by grouping pixels using appropriate criteria (e.g., color uniformity, motion uniformity) one can detect parts of the image sequence that correspond to the same object. At the following subsections we report methods for moving object detection and tracking.

3.4.1 Motion Extraction and Moving Object Detection

The goal of motion extraction is to detect image regions that present movement and segment them from the rest of the image. Subsequently, moving object detection is a process highly dependent on this, since only the extracted regions need to be considered. The various approaches for motion extraction and detection of moving objects can be categorized into techniques which process uncompressed data, and techniques that are designed to work in the compressed data domain.

Uncompressed data domain: the most widely applied technique is called Background Subtraction and is a region-based method which detects moving regions by taking the difference between the current image and a reference background image, in a pixel-by-pixel fashion. However, this method is sensitive to illumination changes and small movements in the background, and thus many improvements have been proposed to tackle this problem. Some proposals model the recent history of each pixel's intensity by using Mixtures of Gaussian distributions [VSLB10, BBPB10]. Other approaches model the pixel intensities by employing the Kalman filter [ZS03], or Hidden Markov Models (HMMs) [SRP$^+$01]. A completely different methodology is introduced in [MP04], where the background distribution is modelled by a non-parametric model, based on Kernel Density Estimation (KDE). Another approach [CGPP03] is based on the temporal filtering of a set of frames by using a median value, while a recent approach which combines the use of a median value for background initialization with some updating rules is proposed in [ZL10]. In addition, some works [SWFS03, LHGT02] exploit the spatial co-occurrence of image variations in order to model the background object and detect the foreground moving regions, while in [ZHH11] the authors combine affine region detectors with a simple model for background subtraction, in order to detect moving objects for real-time video surveillance. Moreover, background modelling using edge features has also been considered [MD01], while an alternative method which utilizes a spatiotemporal texture called Space-Time Patch has been proposed in [YMF11].

Several other region-based techniques have also been introduced, proposing different solutions. Meyer et. al. [MDN97] and Wixon [WH99] use the information from flow vectors of moving objects over time to detect moving regions in an image sequence even in the presence of camera motion, while in [LWYW07] the optical flow is integrated with a double background filtering method. Furthermore, many researchers proposed block-based algorithms. In this case, the image is divided into blocks and after the calculation of some block-specific image features the motion detection is performed based on block matching algorithms. In [MOH00] the block correlation is measured using the Normalised Vector Distance (NVD) measure, while in [MD01] an edge histogram calculated over the block area is used as a feature vector describing the block. Moreover, two block-based techniques which utilize the image noise distribution are described in [ALK05, DHC07], and motion feature extraction using block matching algorithms is performed in [IXM02, MKS04c].

Besides the methods mentioned above, alternative approaches have also been reported in the relevant literature. For example, some entropy-based methods are described in [CC07, JSR04], while in [SHS08] multiple moving regions from the image sequences are detected by applying particle filters. Moreover, a novel approach which analyzes the statistical characteristics of sequences of 3D image histograms is introduced in [ITP05], and another unconventional method which combines the strengths of multiscale and variational frameworks has been presented in [FJKD06]. Finally, a time-sequential approach for the extraction of motion layers from image sequences by utilizing an interactive graph cut-based method has been described in [FPR08].

Compressed data domain: a number of methods for the detection of motion in compressed videos has been proposed after the appearance of compression standards like MPEG and H.264. Useful studies of early attempts can be found in [ZGZ03, ZDGH05]. In MPEG compressed domain, Yu et. al. [YDT03] proposed a robust object segmentation algorithm using motion vectors and discrete cosine transform (DCT) coefficients jointly, while similar approaches have been proposed in [Por04, JH02]. In [Por04] the DCT coefficients are used for the extraction of some frequency-temporal features, which are further exploited for volume growing from homogeneous blocks. The information from motion vectors is then used to estimate an affine motion model for each volume, and moving objects are defined by iterative merging of volumes with similar motion via hierarchical clustering. In [JH02] translational motion vectors are accumulated over a number of frames and the magnitude of the displacement is calculated for each macroblock; macroblocks are subsequently assigned to regions by uniformly quantizing the magnitude of the displacement. In [SR00], segmentation is performed using ac/dc discrete cosine transform (DCT) coefficients only; foreground/background classification is based on thresholding the average temporal change of each region, while the macroblock motion vectors are not used. Alternatively, Zeng et. al. [ZGZ03] introduced a method for moving object detection based on inter-frame differences of the DC image, while another real-time unsupervised spatiotemporal segmentation technique which utilizes motion vectors and DC coefficients is described in [MKBS04]. Color and motion information is directly extracted from the I- and P-frames of the MPEG-2 compressed stream and an iterative rejection scheme based on the bilinear motion model is used to effect foreground/background segmentation. Then, the meaningful foreground spatiotemporal objects are formed by initially examining the temporal consistency of the output of iterative rejection, subsequently clustering the resulting foreground macroblocks to connected regions and finally performing region tracking. Another alternative method has been introduced in [BRS04] where the authors proposed an accumulation of motion vectors over time, followed by a combination of a K-Means clustering algorithm which defines the number of the objects, and the Expectation-Minimization algorithm for object segmentation. Furthermore, Wang et al. [WYG08] described an extension of the Gaussian Mixture background model to MPEG compressed domain and used it in a way similar to its pixel-domain for the segmentation of moving object, while Manerba et al. [MBPLM08] proposed a combination of motion information and region-based color segmentation.

In addition, several algorithms have been introduced focusing on the H.264 compression standard. Zeng et al. [ZDGH05] described a block-based Markov Random Field (MRF) model to segment moving objects from the sparse motion vector field, while Liu et al. [LLZ07] used accumulated motion vectors to enhance the salient motion and identify background motion model and moving objects, by employing MRF to model the foreground field. In [KN08] blocks whose motion vectors are not fitted to a global motion model are regarded as outliers, and a temporal filter is used to remove the noise in them. Subsequently, motion history images are employed to detect moving object from the outlier mask. A real-time algorithm which is based on motion vectors and decision modes is introduced in [SCBG+09]. This algorithm uses fuzzy logic and describes the position, velocity and size of the detected moving regions in a comprehensive way, in order to work with low-level information but also to be able to manage highly comprehensive linguistic concepts. Moreover, two recent approaches are reported in [NL09, WW10]. In [NL09] the motion vectors are firstly refined by spatial and temporal correlation of motion and a first segmentation is produced from the motion vector difference after global motion estimation. This segmentation is then improved by using intra prediction information in intra-frame and afterwards it is projected to subsequent frames where expansion and contraction operations are following. In [WW10] the authors applied the approach of Babu et. al. [BRS04] to the H.264 compressed domain. Firstly, vector median filtering and forward block motion vectors accumulation are used to obtain more dense motion field and define the block characteristic, and subsequently the moving object is extracted using a mixed and hierarchical clustering algorithm based on improved K-Means and Expectation-Minimization (EM) algorithm.

3.4.2 Moving Object Tracking

Object tracking aims to locate particular objects between consecutive frames in image sequences. Many algorithms have been proposed and implemented to overcome difficulties that arise from noise, occlusion, clutter, and changes in the foreground/background environment. Similarly, the methods for object tracking are varying between algorithms which work with uncompressed data and techniques which are based on using information coming from the compressed stream.

Uncompressed data domain: early attempts were based on the extraction of image features such as pixel's intensity, colors, edges or contours and used them for the establishment of the correspondence between model images and target images. For example the feature-based method in [CRM03] uses a color histogram-based representation of the target image and a Bhattacharyya coefficient for similarity measurement, while tracking is performed by the mean-shift algorithm. This algorithm due to its simplicity and robustness is also used in [Col03, HDS04], while another variation of this approach is described in [YDD05]. In addition, some contour-based tracking algorithms have also been described. These methods track objects by representing their outlines as bounding contours and update these contours dynamically in successive frames. Paragios et al. [PD00] proposed a geodesic active contour objective function and a level set formulation scheme for object tracking, while Peterfreund [Pet99] explores a new active contour model based on a Kalman filter. Another Kalman filter-based approach was employed in [SM01] where an active shape model is used to model the contour of a person in each video frame, while an extended Kalman filter formulation was employed for the extraction of motion trajectories in [BSK04]. Yokoyama and Poggio [YP05] use gradient-based optical flow and edge detectors to construct lines of edges in each frame. Then, background lines of the previous frame are subtracted and contours of objects are obtained by using an active contour model to the clustering lines. Each detected and tracked object has a state for handling occlusion and interference. Optical flow and graph representation have also been used by Cohen et. al. in [MCB+01], while another approach described in [HS05] predicts the objects' contours using block motion vector information and updates them via occlusion/dis-occlusion detection procedures.

Other uncompressed domain approaches use particle filters. The authors in [RK10], dealing with the fact that particle filter is a time-consuming tracking technique, introduced a GPU-based implementation of a tracking method which employs the Particle Swarm Optimization (PSO) algorithm. Alternatively, a region tracking model using Markov Random Fields (MRF) is proposed in [BDRB08]. After motion-based segmentation, a spatiotemporal map is updated using a Markov Random Field segmentation model in order to maintain consistency in object tracking.

Compressed data domain: several methods for object tracking that are directly applicable to compressed data have also been proposed. Dong et. al. [DXZF11] summarize many techniques compatible with the H.264 compression standard. They categorize various approaches according to the required degree of partial decoding, and they distinguish them to entropy decoding level methods [LLZ07, PDBP+09], macroblock decoding level methods [YSK07] and frame decoding level methods [KBNM09]. Moreover, in this work they propose a real-time entropy decoding level algorithm combining the prediction mode information with the information from DCT coefficients and motion vectors. Besides the H.264 standard, a lot of effort has been made in the field of MPEG compression standards. Early attempts are described in [LC01, CZQ01]. Lien and Chen [LC01] handle object tracking as a macroblock-linking problem, combining motion vectors with DCT AC coefficient energies of intra-coded macroblocks, while a similar approach is described in [AKM02]. The information from motion vectors is also exploited in [PL03] but in this case, tracking is implemented using the mean-shift algorithm. Finally, Chen et. al. proposed an alternative method utilizing DC difference images and directed graphs in [CZQ01].

3.5 Content Description

A widely studied area of image processing focuses on methods and techniques that aim at the effective description of the visual content. As presented in the previous sections, this step is a basic prerequisite for many processing tasks like e.g., the estimation of similarity among images or the segmentation of an image into regions. For that purpose, many descriptors have been introduced for the representation of various image features and can be broadly divided into two main groups, the global and the local descriptors. The criterion for this categorization is the locality of the feature that is represented. Specifically, the first group of descriptors use global characteristics of the image such as color or texture histograms, while the descriptors of the second group represent local salient points or regions, for example edges or corners. In the following subsections we survey various descriptors from the relevant literature, based on this categorization.

3.5.1 Global Descriptors

The descriptors of this group represent the visual information by exploiting general image characteristics such as: color, texture and shape. Some color and texture descriptors presented in [MOVY01] have been proposed by the well know MPEG-7 standard, as described in the MPEG-7 Overview[7].

For the color information, the MPEG-7 standard defines the following four descriptors. The Dominant Color Descriptor (DCD) consists of the values, the percentages and the variances of dominant predefined colors, along with the spatial coherency which represents the overall spatial homogeneity of the dominant colors in the image. The Color Structure Descriptor (CSD) is capable of distinguishing between images with the same color information but different structure, by capturing both global color information and the local spatial structure of the color with a color structure histogram. The Color Layout Descriptor (CLD) is a compact and resolution-invariant descriptor which has been implemented for the representation of the spatial distribution of color in the YCbCr color space, and it can be used globally in an image or in an arbitrary-shaped region of interest. The Scalable Color Descriptor (SCD) is based on an HSV color histogram that measures the color distribution across the entire image and a Haar transform on the histogram values, which provides scalability when the full resolution is not required. Finally, the Group of frame (GoF) or Group-of-pictures (GoP) Descriptor extends the previous one for the description of color information in sequences or groups of images.

Additionally, for the representation of the texture information, the MPEG-7 standard introduces the following two descriptors. The Homogeneous Texture Descriptor (HTD) provides a quantitative characterization of the image's texture and is calculated by filtering the image with a bank of orientation- and scale-sensitive filters and computing the mean and standard deviation of the filters' outputs in the frequency domain. The Edge Histogram Descriptor (EHD) is useful in image matching tasks and represents the spatial distribution and orientation of five types of edges, namely four directional edges and one non-directional edge in the image, by utilizing histograms for the representation of the local-edge distribution within 16 equally divided regions of the image. The Texture Browsing Descriptor (TBD) distinguishes the texture information in terms of regularity, coarseness and directionality in a way similar to the human perception.

Besides the MPEG-7 color and texture descriptors, other approaches have also been proposed. Recently, Chatzichristofis et. al. [CB08a, CB08b, CBL09] introduced some compact composite descriptors. The Color and Edge Directivity Descriptor (CEDD) [CB08a] incorporates color and texture information in a histogram. Two fuzzy systems are used to map the colors of the image in a 24-color custom palette and fuzzy versions of five digital filters proposed by the EHD descriptors are employed to form 6 texture areas. Another descriptor called Fuzzy Color and Texture Histogram (FCTH) [CB08b] uses the same color information and the high frequency bands of the Haar wavelet transform in a fuzzy system to form 8 texture areas. Compact versions of these descriptors with smaller feature vectors are available as CCEED and CFCTH, while a combination of CEDD and FCTH, called Joint Composite Descriptor (JCD), has been introduced in [CBL09]. Moreover, the authors in [HTL10] propose two MPEG-7 based color descriptors named HMMD Color Histogram and Spatial Color Mode. Other color descriptors include the Color histograms and Color Moments [ZC04, QMC05] and the Grid Color Moment (GCM)

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

Descriptor [ZJN06]. Similarly, other representations for the texture features include the Gabor Texture (GBR) Descriptor, the Wavelet Texture Descriptors and the Local Binary Patterns (LBP) [OPM02].

Regarding the representation of shape information, MPEG-7 introduces three descriptors named Region-based Shape Descriptor (RSD), Contour-based Shape Descriptor (CSD) and 3D Shape Descriptor (3DSD), while other proposed shape descriptors are the Edge Orientation Histogram (EOH) [GXTL08], the Edge Direction Histogram (EDH) [ZHLY08], and the Histogram of Oriented Gradients (HOG) [DT05]. Extensions of the latest descriptor have been proposed by many researchers. In [BZM08] a Pyramidal HOG (PHOG) is presented, in [CTC$^+$12] a Compressed HOG (CHOG) descriptor is described, while an extension that integrates temporal information, called 3DHOG, is presented in [KMS08]. Moreover for the description of motion information that refers to objects within the image sequence or the used camera, the MPEG-7 standard defined the four following descriptors: Motion Activity Descriptor (MAD), Camera Motion Descriptor (CMD), Motion Trajectory Descriptor (MTD), Warping and Parametric Motion Descriptor (WMD and PMD). Alternatively, the authors in [LMSR08] proposed a global descriptor called Histograms of Optical Flow (HOOF) for the representation of the motion information.

Finally, a totally different global descriptor which exploits several important statistics about a scene is called GIST and is introduced in [SI07]. It is calculated by using an oriented filter at several different orientations and scales. This descriptor can encode the amount or strength of vertical or horizontal lines in an image, which can contribute to matching images with similar horizon lines, textures or building represented in them.

3.5.2 Local Descriptors

Local descriptors have been proposed in order to overcome the global descriptors' inability to handle image transformations like changes in viewpoint and lighting conditions. The basic idea is to define local image features (points or regions) that are invariant to a range of transformations and use these features for the calculation of invariant image descriptors. These descriptors could exploit different image properties like pixel color/intensities, edges or corners. Their locality makes them suitable for the cases where image clutter, partial visibility or occlusion takes place, while their high degree of invariance provides robustness to various geometric and photometric transformations. Numerous local descriptors have been introduced and comprehensive studies can be found in [LA08, MS05, BB11, GHT11].

Regarding the definition of suitable image points or regions, which are referred as "points of interest" or "regions of interest" in the relevant literature, various techniques have been proposed. Edges within the image could be good candidates for points of interest, and thus any edge detector algorithm could be utilized for the determination of these points. Examples include the previously mentioned (see Section 3.3) Sobel operator, the Prewitt operator, Roberts' Cross operator and the well known Canny edge detector [Can86]. Another method is the Harris-Laplace point detector presented in [TM08]. For the determination of salient points, this detector relies on a Harris corner detector and for each corner selects a scale-invariant point if the local image structure under a Laplacian operator has a stable maximum. Alternatively, Jurie and Triggs [JT05] proposed a dense sampling strategy for the definition of points of interest, while a random approach was described in [LP05] by Li and Perona.

After the detection of salient points or regions, local descriptors are used for the representation of these image features. The most well-known and widely used descriptor has been introduced by Lowe [Low04] and is called SIFT. This descriptor is invariant to changes in scale and rotation and describes the local shape of a region, based on local gradient histograms sampled in a square grid around the key-point. The gradient of an image is shift-invariant, while the gradient direction and the relative gradient magnitude remain the same under changes in illumination. The effectiveness of this descriptor urged many researchers to implement more efficient versions of it. In [KS04, MS05] the authors applied Principal Component Analysis (PCA) in order to reduce the high dimensionality of the SIFT descriptor, introducing the PCA-SIFT and GLOH descriptors respectively. Mortensen et. al. [MHS05] presented a feature descriptor that augments SIFT with a global context vector (SIFT+GC) which adds curvilinear shape information from a much larger neighborhood. Lazebnik et. al. [LSP05] defines a rotation invariant descriptor called RIFT, while other approaches combine SIFT descriptors with kernel projections, such as KPB-SIFT [ZCCY10] and CDIKP [TWY08]. Moreover, in order to incorporate the rich color information some researchers presented colored versions of the SIFT descriptor. Examples of this approach include the HSV-SIFT descriptor [BZMn08], the HUE-SIFT descriptor [vdWGB06], the OpponentSIFT descriptor, the rgSIFT descriptor, the Transformed ColorSIFT descriptor, the C-SIFT descriptor [GvdBSG01] and the RGB-SIFT descriptor [vdSGS10a].

Partially inspired by the SIFT descriptor, Bay et. al. [BETVG08] proposed a new one called SURF. This descriptor is based on sums of 2D Haar wavelet responses and exploits efficiently the integral image. Moreover, it is considered as faster than SIFT and provides more robustness against different image transformation than SIFT. Indicative extensions of this descriptor are the Dense-SURF variation that has been proposed in [Tao11] and a colored version presented in [BG09]. In addition based on SIFT and GLOH descriptors, Tola et. al. [TLF08] introduced the DAISY descriptor. This descriptor depends on histograms of gradients like SIFT and GLOH, but uses a Gaussian weighting and circularly symmetrical kernel, providing speed and efficiency for dense computations. Alternatively, Belongie et. al. [BMP02] described another distribution-based descriptor called Shape Context, which is based on an edge histogram, sampled from a log-polar grid, while Lazebnik in [LSP03] extracted affine regions from texture images and introduced an affine-invariant descriptor based on the spin images proposed in [JH99].

Moreover, some filter-based descriptors have also been proposed. An early approach was Steerable Filters as a set of basis filters which can synthesize any filter with an arbitrary orientation. Alternatively, Gabor filters were used for image representation in [Lee96], while some geometric filters were proposed in [CJ07]. A novel method proposed in [VZ05] called Textons, is based on the idea that image textures can be characterized by vector quantized responses of a linear filter bank. Based on the approach of Textons, the authors in [SJC08] proposed a pixel-wise descriptor called Semantic Texton Forest. This descriptor does not utilize the responses of the computationally expensive filter banks but instead it uses ensembles of decision trees. Other approaches include: local derivatives, generalized moment invariants, multi-scale phase-based local feature [CJ03], multiple support regions [CLZY08] and covariant support regions [YJX09].

Finally, after the extraction of the low-level feature descriptors many authors propose a “Bag-of-Words” (BoW) approach for a higher-level representation of the visual content. According to this approach the calculated descriptors are grouped using some clustering algorithm (like for example the K-Means algorithm) to create a codebook of visual words, and based on this visual vocabulary each image is represented by the histogram of visual words which are present in the image. This procedure is performed in many concept detection algorithm and thus is described in more a detailed way in Section 4.1.1.3.

3.6 Technical Requirements

Shot segmentation: according to the mentioned user-side requirements in Section 2, the decomposition of the visual content to elementary segments like shots or scenes is a crucial step for the further processing and analysis of the video content. The definition of time boundaries corresponding to video shots allows skipping or searching specific parts of the visual content based on time, while providing reasonable timestamps for other analysis techniques, such the spatiotemporal segmentation and the tracking of moving objects. Moreover, the description of the visual content of each shot using state-of-art low- and mid-level descriptors can be used for measuring the visual similarity among different shots of the video and, combined with temporal information, for grouping them to form higher level segments like scenes.

Currently, we segment the video into shots based on the approach proposed in [TMK08]. The employed technique detects both abrupt and gradual transitions. Specifically, this technique exploits image features such as color coherence, Macbeth color histogram and luminance center of gravity, in order to form an appropriate feature vector for each frame. Then, given a pair of selected successive or non-successive frames, the distances between their feature vectors are computed forming distance vectors, which are then evaluated with the help of one or more Support Vector Machines (SVM) classifiers. The use of these classifiers eliminates the need for threshold selection, as opposed to what would be the case if any of the proposed features were used by themselves and as is typically the case in the relevant literature. Moreover, in order to further improve the results in LinkedTV, we augmented the above technique with a baseline approach to flash detection. Using the latter we minimize the number of incorrectly detected shot boundaries due to camera flash effects.

Our preliminary evaluation based on the LinkedTV News Show and the Documentary scenarios indicates that shot segmentation performs remarkably well. For a more detailed presentation of the evaluation results we refer the reader to [SAM$^+$12a, SAM$^+$12b]. The detection accuracy based on human defined ground-truth data is around 90%, while a small number of false positives and false negatives is caused due to rapid camera zooming operations and shaky or fast camera movements. Hence, our shot segmentation method can be a reliable tool for the annotation procedure, since it just needs little attention in order to tackle some minor issues. These findings are consistent with the findings of the relevant Task of the TRECVID benchmarking activity, which run for several years during the previous decade.

The technical requirements, in order to increase the detection accuracy of the shot segmentation algorithm, include a more efficient flash detector which will minimize the effect of the camera flashlights presented especially in videos of the news show scenario, and further improvements that will solve the problems occurred from fast movement and rapid zooming operations of the camera. Additionally, a requirement regarding the time-efficiency of the algorithm will be addressed through a faster method for the classification step. This method is expected to be based on a GPU-based implementation for SVM classifiers, which will lead to a significant decrease of the required processing time for this task.

Scene segmentation: a higher level decomposition of the visual content into segments that are more meaningful from a human perspective, like topics in a news show or different chapters of a documentary, is also desired and for this reason we segment the video into scenes using a scene segmentation algorithm. This algorithm is based on grouping shots into sets which correspond to individual scenes of the video, and was proposed in [SMK$^+$11]. This method builds upon the well-known technique of Scene Transition Graph (STG) [YYL98] and introduces two extensions of it. The first one, called Fast STG, aims at reducing the computational cost of shot grouping by considering shot linking transitivity and the fact that scenes are by definition convex sets of shots. Based on these facts, the proposed approximation limits the number of shot pairs whose possible linking needs to be evaluated, thus allowing the faster detection of scene boundaries while achieving the same performance levels with the original STG algorithm. The second one builds on the former to construct a probabilistic framework towards multiple STG combination. The latter allows for combining STGs built by examining different forms of information extracted from the video (i.e., low level audio or visual features, visual concepts, audio events), while at the same time alleviating the need for manual STG parameter selection.

In our preliminary experiments we employed only the low-level visual features of the video frames, and by evaluating the effectiveness of our scene segmentation technique on the documentary scenario we have seen that the results are mixed. Roughly half of the scenes detected were deemed unnecessary by the annotators. The reason for this is mostly that the scene does not change based on a semantic view point but, according to this algorithm, a scene is declared only considering the similarity of the visual content based on low-level descriptors. Hence, based on the results of the evaluation process and the findings of the work in [SMK$^+$11] we should consider if the incorporation of high-level visual and audio features to the scene boundary detection algorithm, either those discussed in this work or other such features considered within LinkedTV, could enhance the detection accuracy.

Furthermore, a more sophisticated approach for segmenting video into scenes in terms of different topics would be more efficient in applications such as news video analysis. For this, taking advantage of the available techniques for text analysis and speech recognition would be beneficial. Based on the findings in [SMK$^+$11] the main requirement is to design and implement a new topic segmentation method which will effectively combine the output of each method, aiming at a more suitable segmentation of video in a topic or story level that corresponds to the human perception in a more efficient way. In addition, we should note that the used algorithm is a general approach for the temporal segmentation of any kind of videos into scenes. Therefore it does not exploit any knowledge about possible characteristics that a video may have, like the structure of a news video where the anchorperson usually appears between the different presented topics of the show. Therefore, another technical requirement is to consider if a domain-specific technique designed for videos from the news or the documentary domains could achieve better results, and thus is more appropriate within the LinkedTV scope.

Spatial segmentation: concerning the spatial segmentation method and based on the specified scenarios and the user-side requirements, we concluded that the analysis of the visual content with this technique is not necessary at this phase of the project. Since we deal primarily with video content, other methods like spatiotemporal segmentation for object tracking (as described in Section 3.4) and object re-detection for instance-based labelling (as described in Section 6.1) are more appropriate for the needs of the video annotation system.

Spatiotemporal segmentation: as mentioned within scenario descriptions in Section 2, a video may contain static (a church) or moving (a car) objects of interest from the user's perspective, which should be detected and tracked so that they can be clicked on for further information. The case of static objects is addressed by an object re-detection algorithm which is analysed in Section 6.3, while for the detection and tracking of moving objects, a spatiotemporal segmentation algorithm can be employed. In our preliminary experiments the spatiotemporal segmentation of a video shot into differently moving objects is performed as in [MKBS04]. This compressed domain method overcomes the high computational complexity and the need for full video decoding of pixel domain techniques, achieving in theory real-time performance. The unsupervised method uses motion and color information directly extracted from the MPEG-2 compressed stream. Only I- and P-frames are examined, since they contain all the information necessary for the proposed algorithm; this is also the case for most other compressed domain algorithms. The bilinear motion model is used to model the motion of the camera (equivalently, the perceived motion of static background) and, wherever necessary, the motion of the identified moving objects. Then, an iterative rejection scheme and temporal consistency constraints are employed for detecting differently moving objects, accounting for the fact that motion vectors extracted from the compressed stream may not accurately represent the true object motion.

The evaluation of the spatiotemporal segmentation technique was based on human observation and some preliminary results have been published in [SAM$^+$12a, SAM$^+$12b]. For each video we examined the cases where (a) the moving object was correctly detected, (b) the moving object was detected with over 2 seconds delay after its appearance, and (c) the moving object was erroneously not detected. By detection we mean that the algorithm successfully marked the moving object with a bounding box in one or more consecutive frames where the object appears, without taking under consideration in our evaluation the case where a spatial over-segmentation of a moving object into two or more bounding boxes is observed. The algorithm performed reasonably well, detecting the moving objects correctly in most cases. Indicative results of successful detection are shown in Figure 3. More specifically, the rate of correctly detected moving objects was over 83% in our preliminary experiments, while the rate of the delayed detections was around 3%. However, in some cases the employed technique exhibited over-sensitivity, detecting still regions as moving objects. In over the half of these cases, this response was due to the presence of camera movement or zooming in/out operations, while gradual transitions between successive frames lead to misleading results in about 10% of these cases. In the remaining cases (about 35%), such response was observed in chromatically uniform regions of the video, as a result of the MPEG encoder's inability to correctly estimate the true motion in these uniform regions. Figure 4 illustrates two examples of incorrect detection, due to gradual transition effect and erroneous motion information from the MPEG encoder.

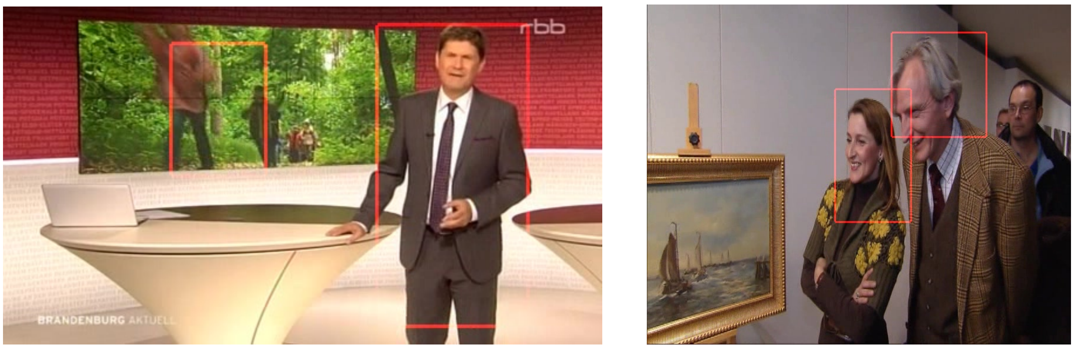


Figure 3: Well detected moving objects in the news show and the documentary material

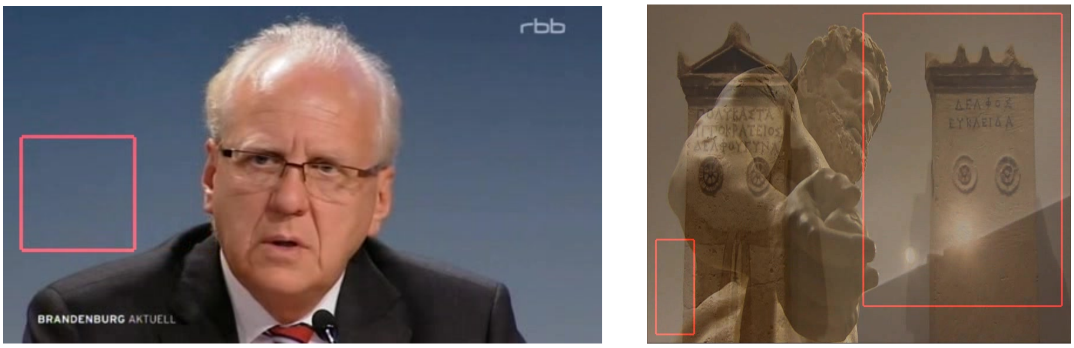


Figure 4: Incorrect moving object detection due to gradual transition effect (dissolving) and erroneous motion vectors from the MPEG encoder

Aiming at further improvement of the algorithm's performance concerning its effectiveness, the first requirement is to address the false positives due to over-sensitivity of spatiotemporal segmentation by introducing criteria that relate to the homogeneity of the different parts of each video frame. Moreover, a second requirement which combines the conflicting needs about time efficiency and accuracy is going to be tackled through the extension of the current method in order to be directly applicable on videos in MPEG-4 format, which is widely used in LinkedTV. This will alleviate the need for MPEG-2 to MPEG-4 transcoding prior to the application of the segmentation algorithm, which introduces significant computational overhead.

Content description: as mentioned before, the efficient description of the visual information is a basic prerequisite and is among the initial steps of many visual analysis algorithms. In our implemented techniques both global and local descriptors have been employed. Shot segmentation uses three global features for the description of the content of each frame, as mentioned previously. Scene segmentation groups shots into scenes based on the visual similarity of the keyframes of each shot, which for that purpose are represented by an HSV histogram. Differently, the object re-detection algorithm described in Section 6.3 exploits the effectiveness of the local SURF descriptor [BETVG08] for matching purposes among pairs of images, while the technique for concept detection discussed in Section 4.4 employs the well known SIFT descriptor [Low04].

Concerning the technical requirements for the content description part, our future plans for the representation of visual information in a more efficient way include two goals which correspond to the decrease of the calculation time and the improvement of the algorithm's accuracy. Regarding the first one, we intend to accelerate the feature extraction and description part using GPU-based implementations of the SIFT and SURF algorithms, by exploiting the parallelism of these techniques and the power of the modern GPUs. Concerning the second one, we plan to obtain more robust and reliable results through the combination of existing and new local and global descriptors and feature extraction techniques, where appropriate.

4 Visual Object and Scene Labelling

This section describes techniques for the efficient labelling of visual objects and scenes within a video sequence. First, we report some state-of-art techniques about all the processing steps of the concept detection algorithm's pipeline. For this purpose, we first list methods for the description of the visual content that have been used explicitly in various concept detection approaches (some of them were also reported in Section 3.5, where general-purpose visual content descriptors were discussed). Subsequently, we describe techniques for an intermediate step called "visual word assignment" which is performed when local descriptors are used in combination with the popular "bag-of-words approach". Then we present several learning algorithms for the final classification and labelling step of concept detection. Besides concept detection, Section 4 reports on state-of-art methods for face analysis, focusing on three tasks: face detection; face recognition; and face clustering. Moreover, we discuss several GPU-based implementations of known algorithms that could be used in the image labelling procedure as well as in other image processing tasks, in order to accelerate their performance. Finally, we draw conclusion regarding the performance of a concept detection algorithm and a tool for face analysis from the reported literature, and we define the future challenges that should be addressed in order to achieve higher levels of performance.

4.1 Concept Detection

The detection of high-level concepts within a multimedia document is one of the most important tasks of multimedia analysis. Due to the rapidly growing amount of audiovisual content, this appealing research problem has attracted a lot of interest within the multimedia research community, since high-level concepts detected from videos could be used for facilitating tasks such as video similarity evaluation and multimedia indexing and retrieval. However, the mapping of low-level features to high-level concepts is still a pretentious and difficult problem, due to the well-known "semantic gap".

Many techniques have been introduced to address this task and they are mainly based on the following procedure. First, various low-level features from the visual (and possibly also the audio) channel are extracted, in order to form a robust description and represent the multimedia content in a meaningful way. Then, usually when local descriptors are employed to represent the low-level visual features, a clustering algorithm is applied to form a vocabulary of "visual words" or "regions", e.g., according to the well-known "bag-of-words" approach. The representation of the visual content is subsequently followed by a learning and classification step which assigns this low-level visual features to high-level visual concepts, thus performing the image labelling. Moreover, many researchers aiming at more reliable and accurate results implement an additional learning step where rule and association mining techniques are applied to the detected concepts, in order to exploit the semantic relations and the temporal consistence among them and achieve more efficient labelling of the image content. Focusing mainly on the visual part of the multimedia content, this section reports on various state-of-art methods that have been introduced for each one of the above mentioned phases of the concept detection pipeline.

4.1.1 Content Representation

4.1.1.1 Global Descriptors

From the previously reported global descriptors (see Section 3.5.1), many of them have been employed for the representation of visual content in several concept detection techniques. These techniques can be distinguished between algorithms that describe features at the image-level, and methods that are designed to represent arbitrarily shaped regions within the image.

Regarding the first category, Spyrou et. al. [STMA09, STA08, MSAK09] have recently proposed an approach that exploits image's color and texture information. Based on their prior work [SA07] they use color and texture information to perform concept detection. In order to improve the detection accuracy, they initially split the image in a predefined number of homogeneous regions by applying the K-Means clustering algorithm on the RGB color values of the image and then, they extract the color and the texture descriptors for the centroid of each calculated cluster. For the representation of the low-level color and texture features of each region they employ the color and texture descriptors [MOVY01] from the well-known MPEG-7 standard [CSP01]. More specifically, Dominant Color Descriptor (DCD), Color Structure Descriptor (CSD), Color Layout Descriptor (CLD) and Scalable Color Descriptor (SCD) are extracted to capture the color properties, while Homogeneous Texture Descriptor (HTD) and Edge Histogram Descriptor (EHD) are used for the texture properties. Moreover, in order to obtain a single description for each defined region, they merge color and texture descriptors by using an “early fusion” approach as proposed in [SLBM$^+$05]. In [HTL10] the authors address the image classification task by utilizing two color descriptors named HMMD Color Histogram and Spatial Color Mode. The first descriptor calculates the dominant colors within the image, while the second one describes their spatial occurrence over the entire image. Moreover, texture information is described according to the previously mentioned MPEG-7 Edge Histogram Descriptor (EHD) and Homogeneous Texture Descriptor (HTD) [MOVY01]. In [HCC$^+$06] the authors employ the Grid Color Moments (GCM) and Gabor Texture (GBR) [ZHLY08] descriptors in combination with the SIFT local descriptor [Low04], while a similar approach which utilizes the same color and texture descriptors has been proposed in [CEJ$^+$07]. Grid Color Moments (GMM) in combination with Local Binary Patterns (LBP) have also been employed in [LS09] to represent color and texture information respectively. Moreover, an alternative approach presented in [SI07] is based on the GIST descriptor to determine image properties like spatial frequency, color and texture.

Besides the approaches described above, many attempts try to perform concept detection by focusing at the characteristics of arbitrarily shaped regions of the image. In [SMH05], images are first segmented into homogeneous regions that will be considered as the smallest entity describing the content, i.e., words. Then, color and texture information for each region are represented by HSV histograms and Gabor Filters respectively. Moreover, several approaches use the previously discussed color and texture descriptors in combination with some region-based shape descriptors. For example in [LS09] shape information is represented by Edge Orientation Histograms (EOH) [GXTL08], while in [CEJ$^+$07] and in [ZHLY08] the authors describe shapes with the Edge Direction Histogram (EDH). Moreover, another global descriptor that has been introduced in various works for the representation of the shape information is the so-called Histogram of Oriented Gradients (HOG) [DT05]. Finally, a region-based approach for semantic image analysis via exploiting object-level spatial contextual information has been introduced in [PSE$^+$11]. In this work, the authors initially segmented the image to regions utilizing a modified K-Means-with-connectivity-constraints (KMCC) algorithm [MKS04b], and subsequently the MPEG-7 descriptors Scalable Color (SCD), Homogenous Texture (HTD), Region Shape (RS) and Edge Histogram (EHD) were used for the representation of each segment. Finally a region-concept assignment procedure was implemented by using the aforementioned visual features in a classification algorithm.

4.1.1.2 Local Descriptors

A number of local descriptors has also been used to represent the visual features of the image. As mentioned in Section 3.5.2 the descriptors that fall within this category are locally extracted, based on the appearance of the object at particular interest points. So, the first step is the definition of the points of interest in the image. Techniques that have been employed in various concept detection algorithms include: the Harris-Laplace point detector [TM08]; a dense sampling strategy [JT05]; and a random sampling approach [LP05]; the Maximally Stable Extremal Regions [MCUP02]; and the Maximally Stable Color Regions [For07]. After the detection of feature points, several feature descriptors have been proposed for their representation. The most broadly used is the well-known SIFT descriptor [Low04] (see for example [JZN06, MDG$^+$10]) and its extensions. For example the University of Surrey within the scope of ImageCLEF2010 [TYB$^+$10] used the HSV-SIFT [BZMn08] and HUE-SIFT [vdWGB06], as well as some colored extensions such as OpponentSIFT, rgSIFT, C-SIFT, and RGB-SIFT that have been presented in [vdSGS10a].

Similarly the University of Amsterdam in ImageCLEF2009 [VDSGS10b] utilized the SIFT descriptor and its extensions OpponentSIFT, RGB-SIFT and C-SIFT, while another used and computationally efficient SIFT-variant was presented in [USS10]. All these descriptors have specific invariance properties with respect to common changes in illumination conditions and have been shown to improve visual classification/labelling accuracy. Moreover, in a more sophisticated approach [MDK10] the authors used the SIFT descriptor for the construction of the proposed feature tracks. These feature tracks are sets of local interest points found in different frames of a video shot that exhibit spatio-temporal and visual continuity defining a trajectory in the 2D+Time space. By using the calculated feature tracks, the authors generate a “bag-of-spatiotemporal-words” model for the shot which facilitates capturing the dynamics of video content.

Besides SIFT, another local feature descriptor that has been adopted in many concept detection approaches due to its good performance is the SURF descriptor [BETVG08]. Recent works on concept detection that utilize this descriptor include [MSV$^+$11, SLZZ10, YXL$^+$09]. Moreover, a Dense-SURF variation has been used in [Tao11], while a colored version of this descriptor has been introduced in [BG09]. Finally, two different approaches for the representation of local image features include the DAISY descriptor introduced by Tola et. al. [TLF10] and the pixel-wise descriptor called Semantic Textons proposed by Shotton et. al. [SJC08]. Last but not least, there is a number of techniques which attempt to enrich the image feature representation using a combination of global and local descriptors. For indicative examples we refer the reader to [JRY08, CEJ$^+$07]. In [JRY08] the SIFT local descriptor is combined with a global color descriptor, while in [CEJ$^+$07] the authors propose a multi-modal framework for concept detection and they investigate different approaches using both global and local low-level visual features combined with low-level audio features, by employing ensemble fusion techniques with multiple parameter sets.

In the above discussed approaches, the local descriptors are calculated for interest points extracted over the entire image. However, as with the global descriptors some approaches utilize local descriptors to represent arbitrarily shaped regions of the image. In a previously referred work (see [PSE$^+$11] in Section 4.1.1.1), where the authors proposed an approach for semantic image labelling, after the segmentation of the image into regions besides global descriptors they also examined the performance of the SIFT local descriptor. In particular, a set of key-points is estimated for every resulting region using dense sampling and a SIFT descriptor is calculated at each key-point. By applying a “bag-of-word” technique (as it will be described in the following Section 4.1.1.3) each region is represented by a histogram of visual words that it contains, and this region feature vector will be used by a classification algorithm for the final region-concept assignment procedure.

4.1.1.3 Visual Word Assignment

As mentioned in the introduction of this section, this intermediate step is performed when local descriptors are used for the representation of the visual content and is referred as “visual word assignment”. By this procedure local descriptors are transformed to a “bag-of-words” representation. First, a visual vocabulary is created via grouping similar keypoints into a large number of clusters and treating each cluster as a visual word. Then, the previously calculated local descriptors are assigned to this visual vocabulary in a manner that each descriptor is mapped to a visual word. By following this procedure, an image can be represented by a histogram of visual words that it contains and this representation can be used as input for the subsequent learning or classification step of the concept detection algorithm. This method has been proposed in [LP05] and for indicative examples of several works that utilize this approach we refer the reader to [AM10, JYNH10, MSHvdW07, ZMLS07, vdSGS08, LM01a, JT05, vGVSG10, vdSGS10a, JNY07, LSP06, NJT06].

The most common method for the construction of the visual vocabulary is the K-Means clustering algorithm which shows good performance (see examples in [AM10, MSHvdW07, ZMLS07, vdSGS08]). Following the application of K-Means, the next step is the assignment of the descriptors to the visual words. Typically this procedure was implemented using the Nearest Neighbor algorithm (like in [ZMLS07, vdSGS08, WAC+04]). However, assigning each descriptor to multiple visual words by using the so-called soft-assignment presented in [vGVSG10], is beneficial to performance at the expense of extra calculation time. According to [vGVSG10], in soft assignment one needs to obtain the closest visual words and calculate a posterior probability over these. In order to make word assignment even more efficient many authors proposed the use of several tree-based assignment algorithms [MM07, MTJ06, NS06]. These algorithms allow for a logarithmic rather than a linear assignment time. One of the most interesting approaches is the “Extremely Randomized Clustering Forests” introduced by Moosmann et. al. in [MTJ06]. A similar approach called “Texton Forest” has been proposed by Shotton et. al. in [SJC08]. The difference here is that each decision node in the tree works on multiple values of the descriptor instead of one. Moreover, Lazebnik et. al. [LSP06] introduced the “Spatial Pyramid” exploiting spatial information by increasingly subdividing the image and obtaining a visual word frequency histogram for each region separately. Extensions of this work were described in [HCX08, JCL07].

Besides the “bag-of-word” approach various different approaches have been introduced, that try to create similar in nature intermediate-level representation of the initially extracted low-level descriptors before the learning/classification step. Spyrou et. al. [STMA09, STA08, MSAK09, SA07] apply a subtractive clustering method in order to construct a region thesaurus, containing all the region types which may or may not represent the concepts that are chosen to be detected. Each region type of the region thesaurus contains the appropriate merged color and texture description. Latent Semantic Analysis (LSA) [DDF+90] is applied subsequently in order to exploit co-occurrences between the defined regions. By measuring the distances of the regions of an image to the region types, a model vector is formed that captures the semantics of an image. Another region-based approach is presented in [SMH05], where Latent Semantic Indexing (LSI) is used for the construction of a visual dictionary with “visual terms” based on the grouping of regions with similar content. Similarly, a “bag-of-regions” technique is described in [GA07], where images are partitioned into regions and subsequently the regions are clustered to obtain a codebook of region types for scene representation. Moreover, in [SA06] low-level features are extracted from segmented regions of an image, utilizing a mean-shift algorithm in the process. Finally, other techniques that fall into this category are: a hybrid thesaurus approach in [BFGS04]; a “bag-of-keypoints” algorithm in [CDF+04]; a dictionary of curve fragments called “shape alphabet” in [OPZ06]; and a lexicon-driven approach in [SWKS07].

4.1.2 Learning Methods for Object and Scene Labelling

For the training and classification step several different machine learning approaches have been introduced. All theses approaches try to define an efficient function which maps appropriately the image features to high-level concepts. State-of-art methods include the well known Support Vector Machines (SVMs) [Vap98], the Regularized Least Squares (RLS) approach [WSH09], the Logistic Regression (LS) technique [YH06a], the K-Nearest Neighbor (K-NN) algorithm [SDI06] and the Kernel Discriminant (KDA) Analysis [LGL01], just to name a few.

Support Vector Machines are feed-forward networks that can be used for pattern classification and non-linear regression. They construct a hyperplane that acts as a decision space in such a way that the margin of separation between positive and negative examples is maximized. Several approaches utilize SVM classifiers for the learning and detection of the high-level concepts. In [LS09] SVM classifiers are trained with simple color, edge and texture features. In [STMA09, STA08, MSAK09, SA07] the low-level features are initially assigned to a higher-level representation through the construction of a region thesaurus. Subsequently, the image is represented by a new model vector whose values are the elements of the defined region thesaurus. Finally, an SVM is employed to estimate the best association between the calculated model vectors and the high-level concepts. In [CEJ+07] three different global features are defined at the first step for the representation of the visual content. Consequently two types of SVM classifiers are learned for the concept detection. The first one is trained over each one of these three features individually and the second one is trained over an overall vector which is constructed by concatenating the three features into one. The final score is obtained by averaging the detection scores of all SVM classifiers. The work in [AM10] presents the concept detection system of MRIM-LIG that was used for the ImageCLEF 2010 VCDA task. For the learning step the authors applied SVM classifiers with Radial Basis Function (RBF) kernel. Another approach that employs SVM classifiers is presented in [SG10]. Besides the RBF kernel, two different types of kernels that have also been utilized in SVMs are the Chi-Square kernel [VDSGS10b, ZMLS07, JNY07] and the Histogram Intersection kernel [Tao11, MBM08]. Moreover, an extension that combines the SVM classifiers with a new modular approach called

Matrix Modular classifier is described in [ZG09], while several other SVM-based approaches include: the Linear SVMs (LSVM) [Vap98]; the Mixtures of Linear SVM experts (MLSWVM) [FRKZ10]; and the Linear Subclass SVM (LSSVM) [GMKS12a];

Apart from the SVM-based techniques, other kernel-based learning methods are also considered as a good choice for learning from large-scale visual codebooks. An approach that combines KDA with Spectral Regression, called SRKDA, has been introduced by Cai et. al. [CHH07] for large scale image and video classification problems. In their work, they demonstrate that this method can achieve a significant speed-up over eigen-decomposition while obtaining smaller error rate compared to state-of-art classifiers like SVMs. This approach has been adopted in [TYB+10, TKM+09]. Other proposals include: a K-NN classifier in [JRY08]; a Gaussian Mixture Model (GMM) in [ABC+03]; a Hidden Markov Model (HMM) in [PGKK05]; some graph-based semi-supervised learning approaches in [WHH+09, TLQC10, THY+11]; a Multiple Instance Learning (MIL) algorithm in [YDF05] and a neural network-based approach that uses the Self-Organized Map (SOM) algorithm along with MPEG-7 features in [LKO02].

In contrast to the above mentioned techniques, that try to map low-level features to high-level concepts in a discriminative manner, some methods aiming at reliable and efficient concept detection by exploiting contextual information have also been introduced. A useful amount of information can be obtained by taking into consideration the spatial relations among objects or regions presented in the scene. An ontology of spatial relations is proposed in [HAB08] aiming to facilitate image interpretations, while in [YLZ07] some spatial constraints and the topological relationships between regions of the image are utilized for automated image region annotation. Other examples exploiting this kind of information can be studied in [YMF07, KH05, TCYZ03].

A different source of information for improving the concept detection results can be the relations among concepts. For exploiting the latter, classification with association rules has been proposed in recent studies in data mining to achieve higher classification accuracy than traditional rule-based classifiers [YH03, LHP01]. Cao et al. [CLL+06] constructed fusion rules based on intuition and human knowledge. Other approaches use graphical models for the representation of inter-concept relationships in probabilistic structures (see for example [NS03, YCH06]). Following a different approach, the authors in [WTS04] created an ontology hierarchy by considering the possible influences between concepts and proved that the ontology-based concept learning improves the accuracy of the detection algorithm. A similar approach is presented in [FGL08], where Fan et. al. built a concept ontology using both semantic and visual similarity, trying to exploit the inter-concept correlations and organize the image concepts hierarchically. In [FGLJ08] they extend this approach by modelling the contextual relationships between image concepts and several patterns of the relevant salient objects, with which they co-appear.

In addition, many proposals exploit the temporal information considering the dependence between consecutive images in a short period of time [BBL04]. A general probabilistic temporal context model has been proposed in [BLB05], where a first-order Markov property is utilized to integrate temporal context sequences. Moreover, in [LWT+08] the authors use inter-concept associations and temporal rules to enhance the performance of semantic concept detection for video data. Yang and Hauptmann in [YH06b] employed some active learning techniques with temporal sampling strategies, to improve the accuracy of concept detectors. Naphade et al. [RNKH02] used inter-concept relations and temporal relationships to learn a probabilistic Bayesian network, while some multi-cue fusion approaches which utilize contextual correlations and temporal dependencies have also been proposed in [WC08].

Moreover, regarding the fact that an image can be assigned to more than one visual concept (e.g., an urban scene with a mountain in the background), several multi-label classification approaches have been recently proposed [BLSB04, ZZ07, YKR07, QHR+07]. According to the works of Tsoumakas et. al. [TK07, TKV11] and Prajapati et. al. [PTG12] the algorithms for multi-label classification can be divided into (a) problem transformation methods and (b) algorithm adaptation methods. Methods of the first group transform the multi-label classification task into one or more single-label classification, regression or ranking tasks. These methods include: the Binary Relevance algorithm; the Ranking via Single Label technique; the Ranking via Pair-wise Comparison method [HFCB08] and the Calibrated Label Ranking extension [FHLMB08]; the Label Powersets; the Pruned Sets [Rea08]; and the Random k-Labelsets (RAkEL) method [TV07, TKV11]. Methods in the second group extend specific learning algorithms in order to handle multi-label data directly. Approaches that fall into this category include: the Multi-Label kNN (MLkNN) algorithm [ZZ07]; a conjunction of Binary Relevance algorithm with k-NN called BR-kNN [STV08]; the MMAC technique [TCP04]; a Back-Propagation Multi-Label Learning (BP-MLL) method [ZZ06]; Decision Trees [CK01]; and Boosting algorithms [SS00].

Finally, some approaches known as cross-domain learning techniques aim to address the problem of adapting concept detectors among different domains. For more information on this topic we refer the reader to [JZCL08, YYH07, DTXM09, JLC11].

4.2 Face analysis in videos

This section focuses on the process of analyzing human faces in digital content. When dealing with video content, two approaches are used: video-image based methods (analysis from frames extracted from the video, compared with still images); and video-video methods (video inputs compared against a video database). We will review the first approach, as the goal is to apply it on keyframes extracted from each shot of the videos. Indeed, we aim here the recognition of persons that are presented at the shot level (within a shot or scene) rather than at the frame level.

Actually, three components of analysis are of interest: detection; recognition; and clustering. Face detection is used as a prior tool to perform the other tasks, which both stem from the calculation of the similarity between detected faces. While face recognition aims at matching a face with an ID using a known database of faces, face clustering takes as input a set of unknown faces and groups them together according to their similarity. Several sources of variability in face appearance (such as lighting, occlusion, scale, orientation, expression, etc.) affect the performance of the classifiers for such tasks.

We will now review the three components individually: Face detection; Face recognition; and Face clustering. The last part focuses on some available tools for face analysis purposes.

4.2.1 Face Detection

Face detection is the process of detecting a human face in a picture and extracting it from the background. As stated in [HL01], we can divide face detection methods into two categories, namely feature-based methods and image-based methods. Unlike the former that exploits faces properties (such as skin color, geometry, etc.), the latter does not perform feature derivation and analysis but relies on a training phase over intensity images. For more information see [ZZ10, HL01].

The main contribution in this field was the Viola-Jones face detector [VJ01], which became the most widely used framework for face detection. The general idea is to use a boosted cascade of weak classifiers based on Haar-like features. It presents the advantage to be very rapid, and thus it is suitable for real-time detection. However, a major drawback is that it does not perform well for sideway poses, while in addition it requires an intensive training phase. This work motivated numerous research works, and was the base for further improvement. In particular, Lienhart and Maydt [LM02] extended the classifier by introducing rotated axial features.

4.2.2 Face Recognition

Face recognition aims at comparing a face against a set of other faces. Similarly as face detection, face recognition approaches can be distinguished between local approach and holistic approach. The holistic approach studies the face as a whole and can itself be divided between the statistic approaches (the popular Eigenfaces and Fisherfaces for instance) and the AI approaches (that use AI methods such as Neural Networks, Hidden Markov Models or Support Vector Machines). Comparing faces can be very computationally expensive, so several methods have been designed to reduce the dimensionality of the data. Eigenfaces are based on Principal Component Analysis (PCA) [KS90] for this purpose; the so-called eigenfaces refers to the eigenvectors corresponding to the greatest eigenvalues. Linear Discriminant Analysis (LDA) techniques have also been studied, and among them is the Fisherfaces proposed by Belhumeur et. al. [BHK97]. Several variations to those methods have been proposed [JA09, TEBEH06, ZCPR03].

On the opposite, the local approach makes use of the different local features and measurements to compute similarity between faces. The earlier methods employ pure geometric calculations between face features [Kan73]; graph matching methods such as Elastic Bunch Graph Matching (EBGM) [WFKvdM97] store faces as graphs, with each node standing for a different facial feature. Local Binary Patterns (LBP) is another local appearance-based method. In [AHP04], Ahonen et. al. extract LBP histograms to describe local features of images that take both shape and texture into account, and calculate the distance using the Chi-square measure.

Global approaches are not robust against pose, illumination or viewpoint changes, so an alignment phase has to be performed prior to analysis. On the contrary, feature-based methods are less sensitive to such change but the main issue is the accuracy of feature extraction techniques.

4.2.3 Face Clustering

Diverse techniques have been used for face clustering from video input. They are usually built in two steps: first is the choice of a dissimilarity matrix to characterize distances between faces; then comes the choice of a clustering algorithm.

The face images are described through features such as Eigenfaces [BL09], SIFT features [ANP07], intensity histograms [VSP06], cloth color information [BL09], or a combination of several of them. A dissimilarity matrix is then computed with various methods. In [BL09], Begeja et. al. use a simple Euclidean distance for this task, while Vretos et. al. [VSP06] introduce the notion of mutual entropy to compare faces. Other measures include Chi-square distance [SZS06]. The clustering algorithms are also presenting variety, and depend on the application and the features that are used. Examples include: hierarchical agglomerative clustering [BL09, ANP07]; fuzzy c-means algorithm [VSP06]; associative chaining [RM03] etc.

One of the main problems of this task is caused due to pose variation: partitioning the input face images into groups of faces with the same pose is used in several works to enhance the capacity of the clustering algorithm. Indeed, as pointed out in [GMC96], face images of different persons in the same pose are more similar to each other than images of the same person in different poses. Gong et. al. represent faces using Gabor wavelet transform and construct a pose Eigenspace to visualize those face poses, using PCA [GMC96]. In [HWS08], Huang et. al. detect the eyes location with Gabor filters, then cluster into poses using the distance between eyes.

Other works exploit properties of videos (for instance, some temporal relations between extracted keyframes), while the aforementioned works mostly deal with images, thus losing some information in the process. Sequence of images introduce some domain knowledge that is used by [TT08] to incorporate some constraints to the algorithm (e.g., two faces in the same image must be in different clusters). An interesting approach for television shows is described in [YYA10]. In this paper, the authors prefer a low processing time over the accuracy of clustering; in order to do so, they use similarity between shots (shot-based clustering) and frame sequences, instead of calculating and comparing face features.

4.2.4 Available Face Analysis Tools

For those tasks, two kinds of tools are of interest: public APIs and the OpenCV library for C++.

Public APIs provide black-box services for face analysis. In particular, Face.com [fac] is a research team that offers face detection and recognition services through its public REST API. It provides very accurate results as acknowledged by the study made on the Labelled Faces in the Wild database (LFW, University of Massachusetts) [aUoM12]. The main advantage is that it is very simple of use, very powerful and scalable to the web. However, Face.com was bought by Facebook during June 2012 [aqu12] and made an announcement on the 7th of July that they would wind down their services. Therefore, it is not possible to use it any more but it served as a proof of concept of face analysis on videos.

Several other tools exist that mainly focus on identification for security programs [neu, acs] or device login [key]. Thus, most of them require frontal face pictures or pictures in controlled environment for training, and they are not adapted for face recognition in news or entertainment videos. Interesting on-line APIs include the following:

Betaface [bet] is a company specialized in face detection and recognition on digital media and aims “to open up new navigation, categorization and search possibilities of rich media assets”. It also has a web API that is free of charge for non-commercial projects.

Luxland’s FaceSDK [lux] offers face detection, face recognition and facial feature detection solutions, that focuses mainly on biometrical login systems. It is delivered as a cross-platform library which performs face detection, face matching (returning a face similarity level), facial features detection and allows maximizing performance through multi-core support. Unlike Face.com, this tool is a commercial service and therefore not free of charge.

Lambdal Labs A free face recognition API is planned by Lambdal Labs [lam12]. So far, the API only performs face detection and some face features extraction (eyes, nose, mouth). At the time of writing, a face recognition tool is to be delivered in the next future.

Ayonix Ayonix is a Japan-based IT company that is specialized in image recognition technology [ayo]. They offer three different commercial applications for face analysis, namely face detection, face recognition and face matcher.

We discarded some APIs that only perform face detection (SNFaceCrop [snf] and Visage Technologies [vis]). A state-of-art technologies that are relevant to this context, are some photo management software that can be used for indexing pictures according to faces they contain. Picasa, Apple's iPhoto and Micosoft's Window Live Photo Gallery or Fotobounce all have different approaches for photo management and organization. For instance, Picasa automatically clusters detected faces and requires user's approval to save and name the clusters, and then refines its results.

OpenCV [Bra00] is the reference library for computer vision, started by Intel in 1999. It has an implementation of the Viola-Jones algorithm (improved by Lienhart-Maydt) for face detection, while from June 2012 it includes a face recognition class (implementing Eigenfaces, Fisherfaces and Local Binary Patterns Histograms).

4.3 GPU-based Processing

Many of the tasks that were discussed in the previous sections can be accelerated by using the computational power of modern General Purpose Graphic Processing Units (GPGPU). The basis for the enhancement of the efficiency in each case is the identification of parts within the algorithms that can be processed simultaneously by the GPGPU, exploiting its parallel architecture. In the following paragraphs, various GPU-based implementations are categorized according to the task they are used for.

Temporal segmentation: many researchers studied the problem of video temporal segmentation (at shot/scene level) focusing on procedures that could be performed in parallel, in order to reduce the computation time by utilizing the GPUs' processing power. An early attempt is described in [KS07] where the authors introduce a straightforward shot boundary detection technique which detects hard cuts and is based on the differences of color histograms between successive frames of the decompressed video. A couple of years latter, Gomez-Luna et. al. [GLGLBG09] implemented an algorithm for both abrupt and gradual transition detection based on luminance and contour information from frames of decompressed MPEG stream, using multi-layer perceptron neural networks for classification purposes. Zernike moments have also been used as feature descriptors in GPU-based shot boundary detection implementations, and Ujaldon [Uja09] studied some improvements for their calculation on GPU, working with grayscale images. Additionally, Toharia et. al. [TRS$^+$12] used a shot boundary detection application, which is based on Zernike moments descriptors, to present a performance analysis ranging from a single CPU and a single GPU scenario to a Multi-CPU Multi-GPU environment.

Spatial segmentation: several examples of early works on GPU-based spatial segmentation methods can be found in studies by Hadwiger et al. [HLSB04] and Owens et. al. [OLG$^+$07]. The recent GPU-based approaches range between graph-based techniques, watershed transforms, neural networks implementations and others. For example Barnat et. al. [BBBC11] proposed a CUDA implementation for computing strongly connected components of a directed graph, while in [FXZ11] an improved version of the graph-based segmentation algorithm originally proposed in [FH04] is presented. An earlier CUDA implementation of a graph-based segmentation technique can be found in [VN08]. Apart from graph-based implementations, some researchers suggest the use of neural networks as fast mechanisms for image segmentation. Martinez-Zarzuela et. al. [MZDPAR$^+$11] introduced a neural network architecture for multiple scale color image segmentation on a GPU. This architecture, called BioSPCIS, is inspired from the mammalian visual system and provides robustness to lighting variations. Similarly, the authors in [XMW$^+$11] describe another GPU-based implementation which simulates the spiking neural networks, a powerful computational model inspired by the human neural system. Other GPU-based approaches include: watershed transform implementations [KP08]; the Mumford-Shah piecewise constant multiphase segmentation technique [GLCC$^+$11]; an image contour detection method combined with optical flow algorithm [SK11]; and a technique [AKWD10] which involves super-paramagnetic clustering using the Metropolis algorithm.

Spatiotemporal segmentation: some GPU-based implementations for spatiotemporal segmentation have also been proposed in the literature. Huang et. al. [HPP$^+$08] implemented a computation and data-intensive algorithm for motion vector extraction, called Vector Coherence Mapping (VCM), on various GPUs and compared the performance against a state-of-art CPU. In [FIW08] the authors described another GPU-based algorithm for extracting moving objects in real time. The method is robust both to noise and to intensity changes caused by scene illumination changes or by camera function, using background subtraction, while it reduces the time for transferring calculation results from GPU to CPU and vice-versa. In [LX09] a GPU-based CUDA-implemented mean-shift tracking algorithm is introduced. The authors are dealing with the fact that the mean-shift algorithm needs a large number of color histograms by employing K-Means clustering to partition the object color space, and representing the color distribution with a quite small number of bins. Alternatively, Rymut and Kwolek [RK10] presented a CUDA implementation of a tracking algorithm based on adaptive appearance models, using a Particle Swarm Optimization (PSO) algorithm. Grundmann et. al. [GKHE10] introduced a hierarchical graph-based algorithm for spatiotemporal segmentation, by utilizing the GPU-based dense optical flow of Welberger et. al. [WTP$^+$09], in order to improve the segmentation quality. Finally, the authors in [RHPA10] developed a GPU-based version for the motion estimator proposed by Bruno and Pellering in [BP02] in order to implement a faster approach of the spatiotemporal visual saliency model introduced by Marat et. al. in [MHPG$^+$09].

Visual feature extraction and description: interest points or regions within the image can be effectively detected by utilizing an edge detector algorithm. Among them, Canny detector [Can86] is the most preferable and thus various attempts for GPU-based implementations of this detector have been described. For example a simple CUDA implementation is described by Fung et. al. at [8], while OpenGL and Cg versions have been presented by Ruiz et. al. [RUG07]. Another parallel implementation is introduced in [NYWC11], while a more detailed study on Canny detector is presented by Luo and Duraiswami in [LD08]. Alternatively, other detectors have also been implemented such as the Prewitt operator by Kong et. al. [KDY$^+$10] and the Harris corner detector by Xie et. al. [XGZ$^+$10]. Image transforms can also be employed to detect shape-based features. For instance, Hopf and Ertl [HE00] proposed an early OpenGL implementation of the wavelet transform. In addition a number of feature detection methods have implemented by Fung et. al. [FM05] in their OpenVIDIA library.

Moreover, as mentioned in Section 3.5.2, two of the most widely used methods for visual feature extraction and description are the SIFT [Low99] and SURF [BETVG08] descriptors. Based on the initial definitions of these descriptors, many researchers tried to accelerate their performance by implementing some GPU-based approaches. Sinha et. al. [SFPG06] presented a GPU-SIFT implementation, while some years later they combined this work with a GPU-based version of the KLT feature tracker [SFPG11]. Heymann et. al. [HFM$^+$07] also exploited the parallelism of modern GPUs to speed-up some parts of the SIFT algorithm. On the other hand, parallel implementations of the SURF descriptor have been proposed by Cornelis and Van Gool in [CVG08] and Timothy et. al. in [TFH08]. Other GPU-based approaches for image feature description include: the multi-size local descriptors which utilize orientation maps [TLF10], introduced by Ichimura [Ich11]; a more scalable SIFT-variant called Eff2 descriptor [DLJ$^+$10]; the Zernike moments [Uja09]; and a parallel implementation of the Histogram of Oriented Gradients (HOG) algorithm [PR09].

Learning and classification: several machine learning approaches have been used for learning and classification purposes, both in temporal segmentation algorithms (as described in Section 4.1.2) and concept detection algorithms (as presented in Section 4.1.2). The reduction of the needed learning time within these approaches could accelerate their performance, and thus many researchers have focused on this task. Catanzaro et. al. [CSK08] presented an SVM classifier that works on GPUs using the Sequential Minimal Optimization (SMO) for single precision floating point arithmetics. Carpenter [Car09] presented another implementation that uses the same SMO solver but mixes double and single precision arithmetic to enhance the accuracy, while in [ADMK11] the authors described a modification of the LIBSVM using GPUs to accelerate parts of the procedure, by porting the calculation of the RBF kernel matrix elements to the GPU, in order to significantly decrease the processing time for SVM training without altering the classification results compared to the original LIBSVM.

Image Labelling: since image labelling is based on the concept detection pipeline as described in Section 4.1, it is obvious that an overall acceleration of the labelling procedure can be based on two improvements: the computational time for the feature extraction and description; and the time at the learning and classification step. Regarding the first part it is obvious that any GPU-based implementation from the above mentioned could be employed to reduce the processing time. Concerning the second part, GPU-based versions of classifiers like the previously mentioned [CSK08, Car09, ADMK11] could also be used to improve the time efficiency of the learning and classification step. Moreover, a general GPU-based framework for the acceleration of the different parts of the visual labelling procedure using the CUDA programming language, has been described in [vdSGS11].

4.4 Technical Requirements

Concept detection: associating media with appropriate labels (concepts) that best describe the visual content is a useful and necessary functionality that has to be implemented by a system for multimedia content analysis. Such functionality enables us to support several new tasks based on the information

[8] http://http.developer.nvidia.com/GPUGems2/gpugems2_chapter40.html

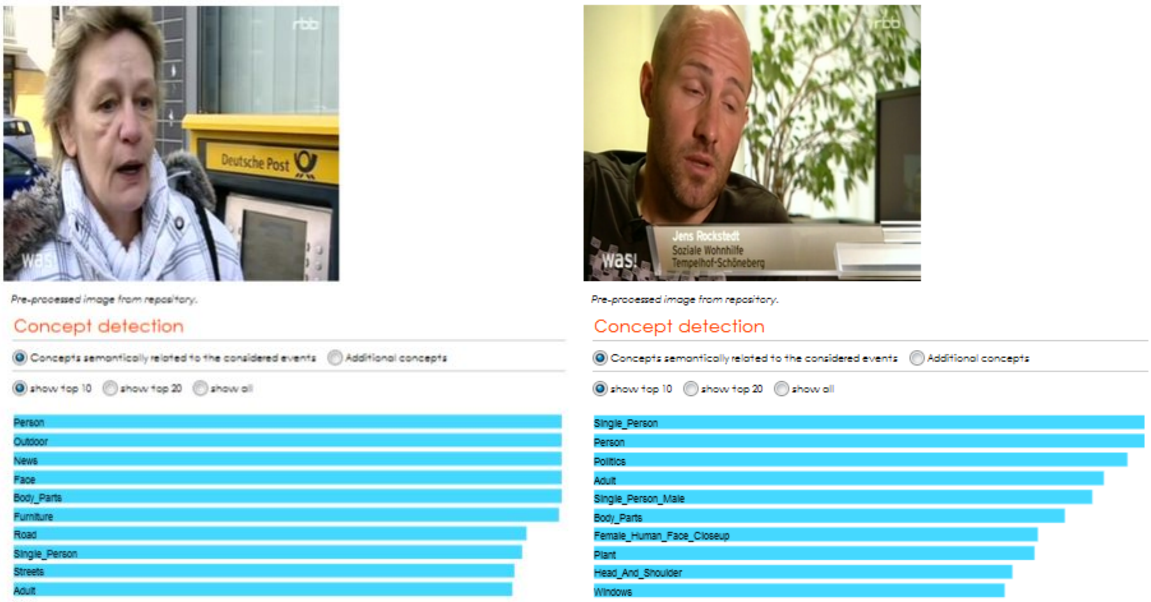
from the content labels. For example, finding thematically similar videos or making video recommendation could be done through estimating video similarity, by taking under consideration the labels that describe the videos. This functionality can be supported by concept detection and face analysis techniques.

Currently, concept detection is performed by using a baseline approach adopted from [MSV$^+$11]. Initially, 64-dimension SURF descriptors [BETVG08] are extracted from video keyframes by performing dense sampling. These descriptors are then used by a Random Forest implementation in order to construct a "bag-of-words" representation (including 1024 elements) for each one of the extracted keyframes. Following the representation of keyframes by histograms of words, a set of Linear SVMs is used for training and classification purposes, and the responses of the classifiers for different key frames of the same shot are appropriately combined. The final output of the classification for a shot is a value in the range [0, 1], which denotes the Degree of Confidence (DoC) with which the shot is related to the corresponding concept. Based on these classifier responses, a shot is described by a 323-element model vector, the elements of which correspond to the detection results for 323 concepts defined in the TRECVID 2011 SIN task. As a first simple improvement, these 323 concepts were selected among the 346 concepts originally defined in TRECVID 2011, after discarding a few that are either too generic (e.g., "Eukaryotic Organism") or irrelevant to the current data being considered in LinkedTV. Moreover, we exploited the relations between the concepts. When a concept implies another concept (e.g., "Man" implies "Person"), then the confidence level of the second concept is reinforced with the help of an empirically set factor a. When a concept excludes another concept (e.g., "Daytime Outdoor" excludes "Nightime") and if the confidence score of the first concept is higher than the second, the first one is enhanced and the second one is penalized accordingly.

For the evaluation of the algorithm's performance we opted to take the ten top-scoring concepts for each keyframe and mark whether each of them is true for the image or not. Further, the human evaluator could mark concepts that he found particularly useful in describing the image, and he could mark concepts where he was uncertain if they applied. A coarse description of the evaluation results can be found in [SAM$^+$12a, SAM$^+$12b]. In summary, concept detectors often succeed in providing useful results (as can be seen in Figure 5). In this small-scale preliminary evaluation, the percentage of correctly detected concepts (among the top-10 ones) for the news show scenario is 64.2%, and out of these, 24.4% were marked as particularly useful. However, for the documentary scenario, it soon became apparent that the training material (i.e., TRECVID videos) was quite different and therefore the classifiers trained on it did not extend well to the "Antique Roadshow" material. Moreover, in the news show scenario even in the lower-ranked concepts useful information could be gathered, while even after filtering out some TRECVID concepts that were too generic for our purposes, many concepts like "professional video" or "civilian person" were still found to be of little help by the annotators. Hence, there is significant room for improvement.

Two counterbalancing technical requirements should be addressed for the concept detection algorithm: the reduction of the processing time; and the improvement of the detection's reliability. Regarding the first requirement, two steps of the concept detection algorithm could be accelerated by the use of GPU-based implementations, in line with what was presented in Section 4.3. The first step is the image feature extraction and description, and the second step is the classification process. Hence, some novel GPU-based techniques that implement feature detection (like Harris-Laplace detector or Canny edge detector), feature description (like SURF or SIFT descriptors) and classification (like SVM classifiers) could decrease significantly the time consumed at these processing steps. Moreover, the classification step could be significantly accelerated by restricting the list of the used concepts. Currently we employ 323 TRECVID concepts and many of them still seem to be of little use, and thus the definition of a new limited list with the most suitable and useful concepts for our needs will result in further improvement of the time efficiency of the classification step. This list will be created by filtering out the inappropriate and useless concepts from the TRECVID list and by adding some new trained concepts that best fit to the documentary scenario.

At the same time, we can benefit from the use of this new list of concepts in order to obtain more reliable results, addressing the accuracy requirement. Further improvement could be obtained by exploiting the relations between concepts in a more efficient way. Moreover, this requirement could also be addressed by incorporating and combining new ways for feature extraction and description. Specifically, two extensions of the SIFT descriptor called RGB-SIFT and Opponent-SIFT are currently being incorporated into the algorithm, while the interest point detection is based on both dense sampling and Harris/Laplace detector. The enriched information from the combination of different ways of describing the image content should improve the overall accuracy of the algorithm. In addition, concerning the



Pre-processed image from repository.

Concept detection

● Concepts semantically related to the considered events ○ Additional concepts

● show top 10 ○ show top 20 ○ show all

Person
Outdoor
News
Face
Body_Parts
Furniture
Road
Single_Person
Streets
Adult

Pre-processed image from repository.

Concept detection

● Concepts semantically related to the considered events ○ Additional concepts

● show top 10 ○ show top 20 ○ show all

Single_Person
Person
Politics
Adult
Single_Person_Male
Body_Parts
Female_Human_Face_Closeup
Plant
Head_And_Shoulder
Windows

Figure 5: Top 10 TREC-Vid concepts detected for two example screenshots. In the left picture 8 concepts have been detected correctly, one concept ("news") is ambiguous and one concept ("furniture") is wrong. In the right picture 7 detected concepts are correct, whereas there is one ambiguous concept ("politics") and two wrong concepts ("female human face closeup" and "windows").

learning and classification step, we may also examine the possibility of improving the algorithm's accuracy by using a novel formulation which extends the Multiclass SVM formulation. We refer to this method as Linear Subclass SVMs [GMKS12a] and for the efficient implementation of this technique we exploited the Sequential Dual Method (SDM) described in [KSC$^+$08].

Nevertheless, considering the conflicting relation of the mentioned requirements, it is obvious that we have to look into the advantages and disadvantages of each approach and make the best decision which will lead to the best compromise between computational complexity and accuracy of results.

Face analysis: the scenarios also show the need to derive information from faces in a video: grouping faces of people across a video and identifying them, is a key step of the process that can be used to enrich the content. Face analysis is performed on keyframes extracted from the video (we used shot segmentation results to extract three keyframes per shot) and is made up of two steps: first we perform grouping of faces based on recognition results; then the clusters are identified either manually (by an annotator) or through a recognition phase when possible.

As first work, we used Face.com API for detection and recognition purposes, before it closed down. Clustering of faces was performed on our side, based on the results of recognition by the API, using an unsupervised and automated-learning method. Accurate groupings were made according to the following process: the training is initialized with the first detected face in the video and then for each subsequent picture, the detected faces are matched against the initial face. If the recognition confidence level is higher than a threshold (80% performed well in our experiments), both faces are associated with the same ID, otherwise a new face ID is created. After every assignment, the corresponding face model is retrained before performing recognition on the next candidate face.

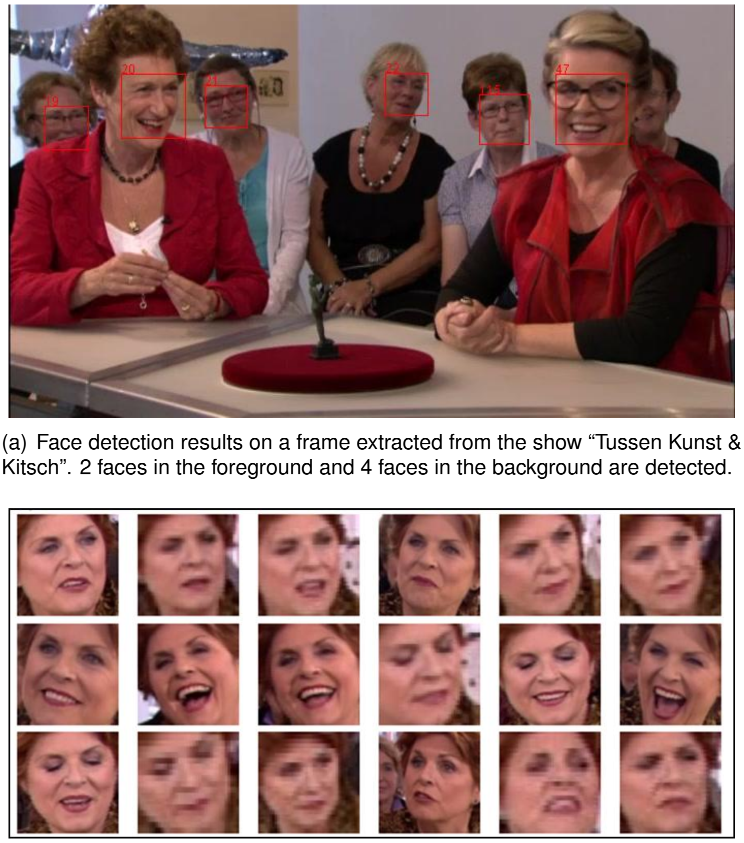
We provide two kinds of evaluation on the clusters: the percentage of pure clusters (i.e., clusters that contain the face from one person only) on one side; and the percentage of misclassified faces over the whole number of classified faces on the other. If there are faces of different persons in the cluster, we assume the person whose face has the higher number of appearances to be correct and count the others as mismatches. See Table 1 for an overview of the results.

The results clearly draw a distinction between two types of behavior, one for the content of each provider: documentary videos present a percentage of pure clusters around 85% and a higher percentage of rightly classified faces (around 92%), while the news show content is processed with very high accuracy (all results above 93%). See Figure 6 for two example screenshots of the output. The main difference between those contents lies in the number of faces detected and above all, on their definition. Indeed, the documentary videos have a mean of 1343 faces clustered (i.e., detected and matched to

Table 1: Results from the face clustering, based on pure clusters and faces that are correctly classified.

Video	pure clusters	correctly classified faces
Documentary 1	84.6%	94.8%
Documentary 2	86.9%	92.6%
Documentary 3	80.7%	90.8%
News 1	95.4%	97.8%
News 2	93.4%	98%
News 3	100%	100%

other faces), while the news show content has a mean of 254 faces. This can be explained easily. In the seed video for the documentary scenario, “Tussen Kunst & Kitsch”, a lot of faces are detected, since most of the scenes occur in front of an audience and thus many faces appear in the background. On the opposite, the news show videos taken from “rbb Aktuell” mostly display two kinds of scenes: close (mostly frontal) views of the presenters, reporters or people interviewed on which detection and recognition is straightforward; and news reports usually featuring only a few people, more rarely a crowd. Faces in a crowd or an audience may introduce some noise because the displayed faces are small and unclear, and therefore it is more likely to be misclassified.



(a) Face detection results on a frame extracted from the show “Tussen Kunst & Kitsch”. 2 faces in the foreground and 4 faces in the background are detected.

(b) Faces extracted from a pure cluster from the show “Tussen Kunst & Kitsch”, featuring the show host.

Figure 6: Example screenshots of the face analysis methods processing videos from the documentary scenario.

The closing down of Face.com determined our first technical requirement and enhanced the need to get our own implementation of such a tool that wouldn’t use a closed API such as Face.com. For face detection we intend to adopt the openCV implementation of the well-known Viola-Jones algorithm, and future work will include a mean to discard the “noisy” faces (of people in the background/audience), for instance by rejecting faces under a certain size threshold. Regarding the face recognition, openCV offers the choice between different methods: Eigenfaces, Fisherfaces and Local Binary Patterns Histograms (LBP). The choice of the most appropriate algorithm to use relies on the training data. Eigenfaces and Fisherfaces are both interesting approaches when using an important training data set with pictures taken in a constrained environment. However they are not suitable for the needs of the project, where our goal is to compare faces one to another which possibly originate from different scenes/shots and therefore have various backgrounds. On the contrary, local approaches partition the image into local areas, so they are more robust against various changes (such as illumination, pose change or facial expression). LBP appears as an interesting approach as it combines high recognition rate and a significantly lower processing time when compared to other local methods, as highlighted in [RdSVC09]. Hence, an LBP-based algorithm would be a good choice for the face recognition task required by LinkedTV. An idea here is to make a search on face images likely to appear in a video (known thanks to the available metadata, for instance) in order to get material of "candidate faces" for recognition. For face clustering, we need to implement a new strategy that directly uses the similarity measures between all the faces (which we could not access in Face.com). Aiming at a more efficient identification of the clusters we consider making a database of reappearing faces such as the anchors or reporters for the news show, which will be enriched when manually annotating a group of faces. Moreover, we also aim at performing a second round of the analysis in order to achieve more reliable results by grouping some clusters together.

5 Complementary Text and Audio Analysis

This section is dedicated to text and audio analysis techniques that can offer information complementary to the information extracted by the visual analysis methods. The first subsection presents various state-of-art techniques for text analysis, such as keyword extraction and text clustering, while in addition it reports on several available tools for performing this analysis. The next subsection refers to state-of-art methods for the analysis of the audio channel. These methods are oriented to various types of analysis, such as speech recognition and speaker identification or clustering. Finally, the last subsection presents the evaluation results of the currently used algorithms and states the main technical requirements for further improvement of their performance.

5.1 Text Analysis

There are several sources, where we can retrieve textual information about a particular video. These include subtitles, annotations of videos (done by an author or an editor) or transcripts obtained as a result of automatic speech recognition. These texts are a valuable source of information about the video itself. There are two main directions on how we can cope with texts in order to facilitate further automatic processing.

Keyword Extraction: this task refers to the identification of important words mentioned in the video or in accompanying texts. The words are used to tag videos. These tags, labels or keywords serve as descriptors for quick orientation in video content, easier filtering during the search and categorizing videos with same tags.

Clustering: this is an approach to group similar videos together based on textual representation of their content.

5.1.1 Keyword Extraction

The objective of keyword extraction or glossary extraction is to identify and organize words and phrases from documents into sets of "glossary-items" or "keywords".

Keywords are often used to help document categorization [BM06]. In many systems they are referenced as tags. Collaboratively created tags form folksonomies. They became popular on the Web around 2004 [VW12] as a part of social software applications, such as social bookmarking and photograph annotation. Folksonomies have been used also for enhancing personalized search [XBF$^+$08].

Keywords identified in a document can be used to cluster documents [TPG03] too. The selection of features of a document is crucial for good clustering results. In this case keywords serve as features of documents. The idea is that keywords represent the essential content of the given document. Another example of keyword phrases being used for document clustering is given in [LCH08a].

However, in already created document collections the identification of keywords can be a time demanding task. Therefore the automatic approach is desirable [CCNH07]. In this subsection, we examine methods used for automatic keywords extraction, while a list of available tools is provided in Section 5.1.3. In the following, basic approaches to keyword identification are introduced.

Statistical Approaches: Basic statistical approach is the usage of classical *tf-idf* measure. A possible application is described in [MFGSMV04]. The definition of *tf-idf* comprises of two parts. Term Frequency is a normalized count of how often a given term appears in the document. The inverse document frequency is a measure of the general importance of the term:

$$idf(t,D) = \log \frac{|D|}{|d \in D : t \in d|}$$

Where $|D|$ is the count of all documents and $|d \in D : t \in d|$ is the count of documents that contain the given term. Thus:

$$tf\text{-}idf(t,d,D) = td(t,d) \times idf(t,D)$$

The intuition here is that a keyword has a high *tf-idf* word value. Several methods to extract keywords based on term frequency, document frequency, etc. can be also found among works focusing on query expansion [Eft95, XC96].

The drawback of *tf-idf* is that we need to have a comprehensive corpus to count *idf* values for examined terms. This is not always the case. An alternative approach is to extract keywords from a single document using word co-occurrence statistical information [MI03]. The proposed algorithm takes into account the bias between a given term and frequent terms in the article. If the probability distribution of co-occurrence between a term and the frequent terms is biased towards a particular subset of frequent terms, then the term is likely to be a keyword. The degree of bias of the distribution is measured by the χ^2 measure.

The approach introduced in [WYX07] analyses the occurrence of salient words in salient sentences and vice versa, and identifies the salient sentences by searching for salient words in them. It aims to fuse the ideas of PageRank [BP98] and HITS [Kle99] algorithms in a unified framework for keyword extraction and document summarization. In [SB93] mutual information statistics are used to discover two-word phrases. The co-occurrence statistics over the entire document collection are utilized in [KC99] to identify related words. Many subsequent metrics have been developed to assess term relationship levels, either by narrowing the analysis for only short windows of text [GWR99], or broadening it towards topical clusters [WT06]. An unsupervised, pattern-oriented approach to keyword extraction with respect to a given ontology is proposed in PANKOW system [CHS04]. PANKOW uses Google to obtain occurrence statistics of discovered key phrases. The work is further extended as C-PANKOW (Context-driven PANKOW) [CLS05], which improves several shortcomings of PANKOW. By off-line processing, the generation of large number of linguistic patterns and correspondingly large number of Google queries is avoided. Also the annotation context is used in order to distinguish the significance of a pattern match for the given annotation task.

Alternatively, in [SZL+08] the tagged training documents are treated as triplets (of words, docs, tags), and represented in two bipartite graphs, which are partitioned into clusters by Spectral Recursive Embedding. A two-way Poisson Mixture Model is proposed to model the document distribution into mixture components within each cluster and aggregate words into word clusters simultaneously. A new document is classified by the mixture model, based on its posterior probabilities, so that keywords are selected according to their ranks. Authors claim the ability of the system to perform real-time tagging with the average performance of processing one document per second.

Linguistic Approaches: Performing linguistic analysis of sentences can improve results of keyword extraction [Hul03]. Linguistic approaches help us to better understand the exact structure of the sentence and thus filter out the not important words. For example, often it can be desirable to extract only nouns or certain combination of parts of speech as keywords. In [LAJ01] summarization techniques are used to extract informative sentences from documents.

Famous part of speech taggers include Stanford Part of Speech Tagger [TKMS03], LingPipe [Car] and Apache OpenNLP [Bal]. An algorithm using some of previously mentioned part of speech taggers (Stanford Part of Speech Tagger [TKMS03] and Apache OpenNLP [Bal]) to distinguish subject-verb-object triples in English sentences is provided in [RDF+07]. A preprocessing step requiring a part of speech tagging is required also to build a text graph in TextRank model [MTT04], based on co-occurrences of links between words. Moreover, a wide variety of linguistic analysis techniques are enabled in GATE system [CMBT02, MBC03], which allows not only NLP-based entity recognition, but also for identifying relations between such entities. GATE powers many annotation systems (e.g., KIM [PKO+04] or Artequakt [AKM+03]).

Machine Learning Approaches: Keyword extraction can be seen as supervised learning problem. First a set of training documents is provided to the system. Each document comes with a human-chosen keywords set. A supervised learning algorithm is used to learn a model. Then, the gained knowledge is applied to find keywords in new documents.

Supervised machine learning algorithms have been proposed to classify a candidate phrase into either a keyphrase or not. GenEx [Tur00] and Kea [WPF+99] are two typical systems, and the most important features for classifying a candidate phrase are the frequency and location of the phrase in the document. In [Hul03] supervised learning is used together with a linguistic analysis combining lexical and syntactical features. In [Mun97] an unsupervised learning algorithm based on Adaptive Resonance Theory neural networks is used to discover two-word keyphrases. Moreover, in [CH98] a clustering algorithm is applied over the input collection of documents and keywords are then extracted as cluster digests. Finally, the authors in [MFW09] propose a variety of features that indicate significance of a word in the sentence. Some of the features exploit also Wikipedia as a backing knowledge base. Bagged [Bre96] decision trees are then used to provide classification models.

5.1.2 Text Clustering

Clustering in general is a method of organizing data into classes (i.e., into clusters), such that there is:

- High intra-class similarity
- Low inter-class similarity

Thus generated clusters contain groups of similar items. There are various possible ways of using generated clusters. The clustering is often applied during the filtering of search results on the web [LC03, GNPS03]. A comprehensive survey of web clustering search engines is provided in [CORW09]. We see two main applications in the context of LinkedTV project:

- Expanding answers to user queries (e.g., find similar videos to the video a user is currently watching). Videos from the same cluster might be used as a recommendation for a user.
- Coupling videos in related groups for additional filtering.

Flat Clustering: K-Means is the most important flat clustering algorithm and is also very common for text clustering applications. Its objective is to minimize the average squared distance of documents from cluster centers, where a cluster center is defined as the mean or centroid $\vec{\mu}$ of the documents in a cluster ω:

$$\vec{\mu}(\omega) = \frac{1}{|\omega|} \sum_{\vec{x} \in \omega} \vec{x}$$

K-Means algorithm belongs to the group of linear time clustering algorithms, therefore it is often used on large data sets. It has $O(nkT)$ time complexity where k is the number of desired clusters and T is the number of iterations [Roc66]. K-Means is most effective when the desired clusters are approximately spherical with respect to the similarity measure used. However, documents (under the standard representation as weighted word vectors and some form of normalized dot-product similarity measure) usually do not form spherical clusters. Anyway incremental K-Means remains the most popular incremental clustering algorithm. A K-Means type subspace clustering algorithm has been proposed in [JNY+08]. In this algorithm, a new step is added in the K-Means clustering process to automatically calculate the weights of keywords in each cluster, so that the important words of a cluster can be identified by the weight values.

Examples of co-clustering (simultaneous clustering of words and documents) are given in [Dhi01]. A similar approach aiming at enhancing document clustering with more meaningful interpretation is proposed in [WZL+08], while a new language model is proposed to simultaneously cluster and summarize the document. Also in [ST00] first word-clusters that capture most of the mutual information about the set of documents are constructed, and then document clusters that preserve the information about the word clusters are defined. An alternative proposal in [SKK00] – Bisecting K-Means – provides better results in the textual documents domain. It begins with a set containing one large cluster consisting of every element and iteratively picks a cluster in the set, splits it into two clusters and replaces it by the split clusters. Splitting a cluster consists of applying the basic K-Means algorithm n times with $k = 2$ and keeping the split that has the highest average element-centroid similarity.

Hierarchical Clustering: Hierarchical clustering outputs a hierarchy or differently a structure. It is not necessary to prespecify the number of clusters and most hierarchical algorithms that have been used in information retrieval are deterministic[9]. These advantages of hierarchical clustering come at the cost of lower efficiency. The produced structure is formed by a single, all-inclusive cluster at the top and singleton clusters of individual items at the bottom. Each intermediate level can be viewed as combining two clusters from the next lower level (or splitting a cluster from the next higher level). The resulting hierarchy forms a tree called dendrogram. This tree graphically displays the merging process and the intermediate clusters. For document clustering, this dendrogram provides a taxonomy, or hierarchical index.

There are two basic approaches to generating a hierarchical clustering:

– Agglomerative: Starts with all items as individual clusters and, at each step, merge the most similar or closest pair of clusters. This requires a definition of cluster similarity or distance.

– Divisive: Starts with one, all-inclusive cluster, and at each step split a cluster until only singleton clusters of individual points remain. In this case, at each step we need to decide which cluster to split and how to perform the split.

IROCK (Improved RObust Clustering using Links) [SLL06] belongs to the class of agglomerative hierarchical clustering algorithms. This method starts by placing each object in its own cluster, and then merges these atomic clusters into larger clusters until a certain termination condition is satisfied. The merging strategy is to choose the pair of objects with highest goodness values.

Another kind of agglomerative hierarchical text clustering algorithms is introduced in [LCH08b]. The two ways of clustering are Clustering based on Frequent Word Sequences (CFWS) and Clustering based on Frequent Word Meaning Sequences (CFWMS). These algorithms treat a text document as a sentence of words, instead of bag-of-words. The closeness between the documents is measured using the words that are common among the documents.

Many hierarchical clustering as well as flat clustering techniques produce hard clusters. That means that each document is assigned to one and only cluster. However, the border between particular topics is sometimes not so clear. In order to handle the fuzziness, a modified fuzzy C-Means [MS04] algorithm is proposed in [RSS04]. In this modification text documents are clustered based on the cosine similarity coefficient, rather than on the euclidean distance. The modified algorithm works with normalized k-dimensional data vectors.

Zhao and Karypis [ZKF05] recommended a hybrid approach known as Repeated Bisections. This overcomes the main weakness with partitional approaches, which is the instability in clustering solutions due to the choice of the initial random centroids. Repeated Bisections starts with all instances in a single cluster. At each iteration it selects one cluster whose bisection optimizes the given criteria function. The cluster is bisected using standard K-means method with K=2, while the criteria function maximizes the similarity between each instance and the centroid of the cluster to which it is assigned. As such, this is a hybrid method that combines a hierarchical divisive approach with partitioning. Another example of a hybrid approach involving both K-Means and agglomerative hierarchical clustering, is provided in [CKPT92]. K-Means is used because of its run-time efficiency and agglomerative hierarchical clustering is used because of its quality.

The bag-of-word representation of a document is limited as it only counts the term frequencies in the document, while ignoring the important information of the semantic relationships between key terms. As the clustering performance is heavily relying on the distance measure of document pairs, finding an accurate distance measure which can break the limitation of bag-of-words is important. Thus, several approaches were proposed to use an external knowledge base to enrich the document representation [HZL$^+$09, BRG07, HFC$^+$08].

WordNet [Mil95] is used as a knowledge base supporting clustering in [HSS03, dBRHDA97, UnLBG01]. However, the coverage of WordNet is limited. Thus recent research uses more live datasets such as Wikipedia. Alternatively, in [HZL$^+$09] the document representation is enriched with Wikipedia concept and category information. Text documents are clustered based on a similarity metric which combines document content information, concept information as well as category information. A similar approach using concepts from Wikipedia to enrich clustered texts has been also introduced in [BRG07] and [HFC$^+$08].

Multi-modal Clustering: Multi-modal clustering techniques enable to cluster items based on various types of features (modalities). It is not limited only to textual document domain. For example in [CRH03]

[9] In contrast to K-Means, which is typically non deterministic.

multi-modal clustering is used to analyse the usage of a particular web site. A similar application for web usage mining is described in [HCC01]. Here the content of a web page is represented in multiple modalities in order to group related web pages together. In the textual document domain multi-modal clustering was used in [BR08]. The presented system thus allows users to interactively choose clustering criteria (e.g., document's genre or the author's mood). Finally, an inspiring example of multi-modal clustering applied on multimedia collections can be found in [BJ07]. Multimedia are inherently multi-modal with information about images, sounds and contained texts.

5.1.3 Available Text Analysis Tools

Bellow, we briefly list available tools enabling text analysis - particularly keyword identification and text clustering:

- Gnesim (http://radimrehurek.com/gensim/intro.html) – Gensim is a free Python framework designed to automatically extract semantic topics from documents. Gensim aims at processing raw, unstructured digital texts (plain text). The algorithms in Gensim, such as Latent Semantic Analysis, Latent Dirichlet Allocation or Random Projections, discover semantic structure of documents, by examining word statistical co-occurrence patterns within a corpus of training documents.

- Apache Mahout (http://mahout.apache.org/) – The Apache Mahout machine learning library is implemented on top of Apache Hadoop using the map/reduce paradigm. The core libraries are highly optimized to allow for good performance also for non-distributed algorithms. Implemented algorithms include Latent Dirichlet Allocation, Singular value decomposition, Dirichlet process clustering, and others.

- CLUTO (http://glaros.dtc.umn.edu/gkhome/views/cluto) – CLUTO is a software package for clustering low- and high-dimensional datasets and for analysing the characteristics of the various clusters. CLUTO is well-suited for clustering data sets arising in many diverse application areas, including information retrieval.

- Apache Solr Clustering Component (http://wiki.apache.org/solr/ClusteringComponent) – Apache Solr is mainly a full text search platform from the Apache Lucene project. However, it provides also a component dedicated to document clustering.

- GraphLab (http://graphlab.org/) – GraphLab is a graph-based, high performance, distributed computation framework written in C++. It provides among others a clustering functionality (currently it supports K-Means++ [AV07]).

- Carrot^2 (http://project.carrot2.org/) – Carrot^2 is an Open Source Search Results Clustering Engine. It can automatically organize small collections of documents (search results but not only) into thematic categories. Carrot^2 implements Lingo [OW05] and STC [SW03] algorithm.

- LingPipe (http://alias-i.com/lingpipe/) – LingPipe is a tool kit for processing text using computational linguistics. Among others enables also text clustering (flat and hierarchical clustering) and key phrases extraction.

- Natural Language Toolkit (http://nltk.org/) – NLTK is a suite of open source Python modules, data and documentation for research and development in natural language processing. It includes various linguistic tools and also frequent words extractor and clustering package.

- Yahoo! Term Extraction Service (http://developer.yahoo.com/search/content/V1/termExtraction.html) – The Term Extraction Web Service provides a list of significant words or phrases extracted from a larger content. The request can be sent using Yahoo! Query Language (YQL).

- Topia.termextract (http://pypi.python.org/pypi/topia.termextract/) – Topia.termextract is a package that determines important terms within a given piece of content. It uses linguistic tools, such as "Parts-Of-Speech" (POS) tagging, and some simple statistical analysis to determine the terms and their strength.

- TerMine (http://www.nactem.ac.uk/software/termine/) – TerMine is a web based service providing a key phrases extraction functionality. It uses the C-value/NC-value method [FAT09] that combines linguistic and statistical information.

– GATE (`http://gate.ac.uk/`) – GATE is a text processing software suite that enables to use many plugins solving various text processing problems, such as document clustering and key phrase extraction.

– Scikit-learn (`http://scikit-learn.org/`) – Scikit-learn is rather a general purpose machine learning Python module.

– Weka (`http://www.cs.waikato.ac.nz/ml/weka/`) – Weka is a general purpose machine learning suite written in Java. It can be used to text clustering as well.

5.2 Audio Analysis

5.2.1 Automatic Speech Recognition

5.2.1.1 Speech search: from broadcast news to spontaneous conversational speech

The use of speech recognition technology to exploit the linguistic content that is available as spoken content in videos, also referred to as "speech search", has proven to be helpful to bridge the semantic gap between low-level media features and conceptual information needs and its use has been advocated since many years. The potential of speech-based indexing has been demonstrated most successfully in the broadcast news (BN) domain. Broadcast news involves relatively clean, planned and well-structured speech and the domain was studied in depth along benchmarks focusing on both speech recognition (HUB-4) and spoken document retrieval (TREC SDR) [GAV00].

Speech search is currently moving beyond the conventional domain of broadcast news, into areas in which speech is not pre-scripted, but instead is produced spontaneously and often in the context of a conversation or a natural communication setting. Such domains include: interviews (cultural heritage), lectures (education), meetings (business), debates (public life), consumer/professional internet media, especially podcasts (education, entertainment), telephone conversations and voice mail (enterprise) and spoken annotations, for example, for photo archives (personal media). These domains offer multiple challenges that spoken content search must address in order to offer users effective solutions. Spontaneous conversational speech is well known to be highly unpredictable. The variability arises from a range of sources including, individual speaker style, speaker accent, articulation, topic and also differences in channel conditions. Searching spontaneous conversational speech is made even more challenging by the fact that humans produce speech in an unstructured matter, meaning that the decision of where to place the boundaries of a document or a speech recognition result is critical if a retrieval system is to be truly effective. Creating appropriate surrogates for time-continuous media like audio or video is also important. Good surrogates allow users to review results and chose items for further listening or viewing in a time-efficient manner. Finally, spoken audio contains a bounty of information that is encoded in structure, prosody and non-lexical audio. New methods must be developed in order to fully exploit these additional information sources.

5.2.1.2 Speech recognition

Speech recognition systems convert an acoustic signal into a sequence of words via a series of processes. Processing of an audio file typically starts with speech activity detection (SAD) [HWO07], in order to filter out the audio parts that do not contain speech. After SAD, often speaker diarization [ABE$^+$11] methods are applied: the speech fragments are split into segments that only contain speech from one single speaker with constant audio conditions. Each segment is labelled with a speaker ID. Next, features are extracted from the segmented audio and can be normalized for speaker and audio variations, followed by a decoding pass and optionally, after adapting the acoustic model for each speaker cluster, a secondary decoding pass that uses the adapted models for producing the final transcription.

5.2.1.3 Out-of-vocabulary

Reducing out-of-vocabulary (OOV) words remains to be an important topic in the news domain given the domain's named entity dynamics. To support efficient recognition, it is crucial that the speech recognition system can adapt to the linguistic variations in the target collections, to reduce the number of words unknown to the recognition system. The properties of spoken content in other domains than broadcast news (e.g., corporate, cultural heritage), may focus on a specific period such as a war, and as a result are packed with names, keywords, or euphemisms specific to that topic, but generally absent in models built from current news texts. To limit the number of out-of-vocabulary (OOV) words, the ASR engine employed in a multimedia retrieval environment should use models that can deal with linguistic variation. First, a query consisting of an OOV word, a so-called QOV (query-out-of-vocabulary), will never match terms in an ASR transcript, even if the QOV occurs in the speech. Second, the word occurring in the transcript at the position of the OOV may induce a false alarm to another query.

Several solutions to this problem have been proposed, ranging from the use of larger recognition vocabularies, to dynamic adaptation of vocabularies based on temporal information in parallel information sources or metadata that is available for a document [Ros95, AGFM00, AG05, HG09]. Another strategy for dealing with QOV words is to avoid speech recognition vocabulary restrictions, by creating audio document representations based on phones or sub-word units, instead of words [Ng00, SMQS98].

5.2.1.4 Alignment

The Webster on-line English dictionary defines collateral as “parallel, coordinate, or corresponding in position, order, time, or significance”. We use the term here to refer to data that is somehow related to the primary content objects, but that is not regarded as metadata. The term metadata will be used to refer to the description of documents or collections as found in a catalog or index. Metadata may consist of content descriptors that reflect the coverage of the audiovisual document, such as summaries and keywords, and of contextual descriptors, also called surface features, that specify e.g., document length, the document’s location, and its production date. In contrast to metadata, collateral data are not describing a primary media object. They can be documents by themselves, produced either as byproduct in the pre-production or post-production stage (e.g., scripts, program guide summaries, reviews), or independently of the primary object (e.g., related newspaper articles).

Despite the fact that metadata and collateral text data can be formally separated, collateral text data may show great overlap with content descriptions that are part of the metadata. They may also be used to generate metadata descriptions, but once these have been created, the multimedia documents and those collateral data sources become separate objects again. Take, for instance, subtitling information for the hearing-impaired (e.g., CEEFAX pages 888 in the UK) that is available for the majority of contemporary broadcast items, at least for news programs. Subtitles contain a nearly complete transcription of the words spoken in the video items, and provide an excellent information source for automatic indexing. However, they are usually not part of the archival description of broadcast material.

Collateral data in the broadcast news domain can be found in the form of subtitling information for the hearing impaired. It is an obvious and cheap solution for indexing. The number of words in teletext subtitling transcripts is typically cut down drastically due to a minimum of available space on a screen, mixing up phrases in an attempt to convey the same message with less and often other words. Nevertheless, already in 1995 the feasibility of using subtitling for indexing was demonstrated [BFJ$^+$95]. In that case, subtitles where recorded using a teletext decoder (nowadays available with most video capture TV cards) and synchronized with the video by adding the time at which each line of text appeared during the broadcast. This was accurate to within a few seconds.

If the collateral data at hand more or less follow the speech in the collection, making it available for indexing requires the synchronization of the text data to the audiovisual source. Except for broadcast-related sources such as subtitling, collateral data usually does not contain time information. This process of labelling the text with time-codes is called the ‘alignment’ of text and speech, a well-known procedure used frequently in ASR, for example when training acoustic models (see e.g., [YKO$^+$00]). The alignment of collateral data holds for surprisingly low text-speech correlation levels, especially when some additional deception is applied. However, when the collateral data only correlates with the speech on the topic level, full-blown speech recognition must be called in, using the collateral data as a strong prior (‘informed speech recognition’) or source for extensive domain tuning.

If the textual content does not match the speech too well the alignment procedure may fail to find a proper alignment. In order to produce a suitable, time-coded index, a two-pass strategy such as proposed in [MJTG98] can be applied. A baseline large vocabulary ASR system[10] is used to generate a relatively inaccurate transcript of the speech with word-timing labels. This transcript is referred to as ‘hypothesis’. Next, the hypothesis is aligned to the minutes at the word level using a dynamic programming algorithm. At the positions where the hypothesis and the minutes match so called ‘anchors’ are placed. Using the word-timing labels provided by the speech recognition system, the anchors are used to generate segments. Individual segments of audio and text are accurately synchronized using forced alignment.

[10] Optionally the speech recognition is adapted to the task, for example by providing it with a vocabulary extracted from the minutes

Table 2: Results on the NIST 2001 Speaker Recognition Evaluation, taken from [RAC+03]

System	EER (%)
Acoustic baseline (GMM-UBM cepstral features)	0.7
Pitch and energy distributions	16.3
Pitch and energy slopes + durations + phone context	5.2
Prosodic statistics	8.1
Phones n-grams	4.8
Phones binary trees	3.3
Phone cross-stream + temporal	3.6
Pronunciation modeling	2.3
Word n-grams	11.0
Combined with single-layer perceptron	0.2

5.2.2 Speaker Identification

The goal of speaker identification (SID) is to classify a speaker, based on information derived from an audio signal. In [KL10], commonly used features are listed and organized based on the following classes:

Short-term spectral features Short-term spectral features are features that focus on the “raw” audio signal. According to [KL10], the most predominant short-term spectral features are mel-frequency cepstral coefficients, linear predictive cepstran coefficients, line spectral frequencies, and perceptual linear prediction coefficients.

Voice source features Voice source features relate to properties of glottal sounds, e.g. a throaty or husky voice quality. They can be modelled by inverse filtering of the original audio signal with a linear predictor.

Spectro-temporal features These features aim at modelling spectral features over time, since e.g. the energy level of a voice signal or the speed of formant transitions might contain useful information that are specific to a certain speaker.

Prosodic features Prosody refers to the rhythm, stress, and intonation of speech. The most important features extracted are the fundamental frequency, pause statistics, the duration of phones, speaking rate and energy distribution.

High-level features High-level features include the speaker-specific choice of vocabulary, typical catchphrases and the like. Possible ways to measure them is by using customized n-grams/language models or bag-of-words.

[RQD00] is a classic example for a spectral-based speaker recognition system and is still used in many applications today. It assumes a text-independent system (i.e. having no prior knowledge of what the speaker will say), single-person (i.e. only one speaker per audio signal) speech recognition task. The authors make use of Gaussian Mixture Models (GMMs) of increasing model order, using spectral energies over mel-filters, cepstral coefficients and delta cepstra of range 2. A single, universal background model (UBM) is used. The system employs a 2048 mixture UBM, consisting of 2 gender-dependent models. Performance is tested on the NIST 1999 Speaker Recognition Evaluation (SRE) set.

SID for clean speech data can be considered to be well-studied. [RAC+03] summarizes the findings of the Johns Hopkins University Workshop in 2002, which aimed at incorporating high-level features into the recognition system. The GMM-UBM already showed an accuracy of 0.7% EER, while all systems based on higher-level features were severely deterioated (see Table 2). The individual results have been fused by a single layer perceptron with sigmoid outputs, which gave a EER boost from 0.7% to 0.2%.

In [BSMK10], subword units are included as an additional feature. [Bau12] further enriches spectral features with information of the topic that a speaker talks about. In debates from, e.g., the German parliament, topic classification proves to be a valuable additional input. Other recent articles include [Vin12], which investigates speaker recognition rates for source separation tasks, i.e., when many different speakers are speaking simultaneously.

5.3 Technical Requirements

Text analysis: scenarios most closely related to keyword extraction and clustering are those about Bert and Daniel (see Section 2.2). Both of them are looking for a well defined subset of videos or for related videos from the same group. Keyword extraction and textual clustering helps to fulfil these needs – to group videos with similar content and classify them into groups characterized by the same keyword.

Technical requirements, in order to extract most appropriate keywords and provide efficient grouping of videos, include selection of the most suitable textual sources and evaluation of their balancing. For our scenarios, we have access to several sources of textual information about a particular video. These include subtitles, manual annotations of videos, and finally the transcripts obtained from the ASR. In the future there is also the possibility to exploit textual information presented directly in video images. However, preliminary experiments on optical character recognition (OCR) had a rather mixed quality, and thus an improvement is needed in order to be usable for keyword extraction.

Further technical requirements deal with the problem of multilingualism. According to our scenarios, LinkedTV has to deal at least with three different languages (English, German and Dutch), while performing text analysis. In order to build possibly universal keyword extraction system – to some extent language independent, we experimented with statistical approaches to keyword extraction. However, it seems that a support of linguistic analysis leads to better results. For example, limiting the set of possible keywords only to nouns increases precision significantly. In the future, we consider the use of linguistic analysis – at least "Part of Speech" (POS) tagging as a glue for proper keyword identification. It may also serve as one of the words features for machine learning approaches to keyword extraction. Additionally, we plan to focus on key phrases identification [SB93] improvement.

By clustering, we consider rather flat clustering techniques as there have not been any special requirements on making hierarchies of videos based on their textual content. On the other side flat clustering meets technical requirements posed by the UTA tool developed within WP4 and can exploit the outcomes of video clustering into a predefined number of groups. They can directly help to improve word disambiguation performed within WP2. This is the subject of our future evaluation.

Final technical requirements deal with the selection of a proper representation of video content. An important decision by clustering is what kind of features do we use. In LinkedTV scenarios, we identified the following representations of textual video content:

- Bag of words representation of texts extracted with the help of ASR, from subtitles and provided as manually created meta data.
- Identified keywords.
- Named entities extracted in collaboration with WP2 (as described in D2.3).
- Soft entity classification to a predefined set of classes done in collaboration with WP2 (as described in D2.3).
- Combination of above mentioned representations.

Apart from pure textual representation, concepts identified in video during visual analysis (Section 4.1) can serve as one of the features. In this case concepts identified on the shot level need to be aggregated. In order to incorporate visual concepts, multi-modal clustering seems to be a promising approach. In relation to this, the evaluation of the influence of individual features is the subject of our future work.

Audio analysis - Speech recognition: speech recognition is commonly measured as the word error rate (WER), which is defined by the Levenshtein distance [Lev66] (i.e., the minimum number of substitutions, deletions, and insertions necessary to transform the hypothesis into the reference), divided by the reference length. On the news material, we annotated one video of half an hour length. The German ASR system had an overall WER of 37.3%, with the largest error source being substitutions (25.6%).

Introducing 500 local pronunciation variants for Berlin dialect gave 1% absolute improvement for the relevant parts. However, a proportion of locals speaking with a dialect with heavy background noise (Berlin tavern visitors talking about a local soccer team) is absolutely not intelligible. However, we believe that for this particular case, background noise is the main factor for the quality deterioration.

We conclude that, apart from further developing a Berlin dialect model, we need to strengthen our acoustic model for local outdoor interview situations, and we need to strengthen our language model for spontaneous speech.

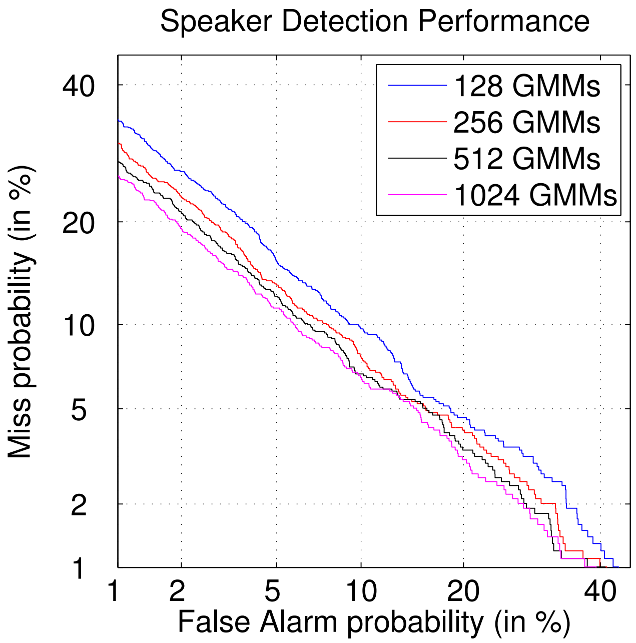


Figure 7: Speaker identification for German politicians: DET Curve for different mixture sizes of the GMM, on a withheld test corpus of 994 audio files from the German parliament.

For Dutch, we analysed in how far the subtitles of the text can be used for forced alignments. In order to assess the closeness of the subtitles to what is actually spoken, we annotated 52 sentences from a video, and treated the subtitles as hypothesis. The WER is at 26.9% for this segment, while the largest error source are the insertions (18.8%), i.e., the words missing in the subtitles, so that the superfluous speech could be collected by a garbage model. The Dutch ASR performance for this part of the text is at 51.9% due to unoptimized models, and at this stage not usable for our purposes. The next step will be to adopt the models onto the material, and also to see in how far the forced alignment algorithm can cope with the discrepancies of the subtitles with respect to what is actually spoken.

Audio analysis - Speaker recognition: in selected videos from the news show, no German parliament speaker was present. Since we are looking for reliable results on a large data set, we took a distinct set of 994 audio files taken from German parliament speeches to evaluate the quality of the models. Speaker Recognition evaluation is given as the equal error rate (EER), i.e., the error for the rejection threshold which produces an equilibrium of false positive and false negative matches. We also depict the Detection Error Trade-Off (DET) curves as described in [MDK+97]. A GMM with 128 mixtures has an Equal Error Rate (EER) of 9.86, whereas using 1024 mixtures leads to an improvement of 8.06 EER. See Figure 7 for Detection Error Trade-Off (DET) curves.

Our main problem, however, is the collection of speaker labelled material from the local context of both the news show and the documentary scenario. Face (re-)detection is only a weak indication of a person speaking, since for example in interview situations the camera will often focus on the face of the conversational partner to show his or her emotion to the viewer. Also, from our experience NER is of little help here because a person seldom says his own name in interview situations. Thus, as future work we plan to analyze in how far banner detection can help obtaining decent training material. This should be taken with a grain of salt as well, as the name will only be shown once, but the speaker might continue in subsequent shots.

6 Event and Instance-based Labelling of Visual Information

This section presents methods for event and instance-based labelling of the visual information. Specifically, the first subsection is dedicated to algorithms and techniques for the different processing steps that contribute to object re-detection, which is the core of our instance-based labelling efforts, while the following subsection refers to state-of-art algorithms corresponding to the individual parts of the event detection pipeline. The final subsection presents the evaluation results of a preliminary, baseline implementation of object re-detection and points out the necessary technical requirements and our future plans in order to meet them.

6.1 Object Re-detection for Instance-based Labelling

Object re-detection aims at finding occurrences of specific objects in a single video or a collection of still images and videos. The samples for learning an object are taken directly from the video on which the object re-detection is performed. Therefore this problem can be interpreted as an image matching problem. However, the different viewing and lighting conditions under which the object's re-appearance takes place, make this task more difficult and complicated.

The most popular approach for the estimation of the similarity between pairs of images is based on the definition of regions of interest. This framework consists of three steps. Firstly, the image feature extraction step, where the salient points and regions are defined either by some edge detection algorithm (Harris detector [HS88], Canny edge detector [Can86], SUSAN detector [SB97], CSS detector [MS98], etc.), or by their center and an associated circular or elliptical region around them (using, for instance, SIFT [Low04], SURF [BETVG08], DoH [Lin94], HLSIFD [YHT10] algorithms). Secondly, the image feature description step, where the predefined regions are represented by global color, structure or texture descriptors (see Section 3.5.1) or by feature descriptors with edge orientation information that provide robustness to scale and rotation transformations, such as SIFT [Low04, AWRDG08], SURF [BETVG08] and HOG [DT05]. Finally, at the matching step, for each described region of the query image the most appropriate corresponding region at the compared image is determined, using the Nearest Neighbor algorithm. Afterwards, erroneous matches can be filtered out by applying some geometric constraints, obtained for example from the RANSAC algorithm [FB81, RDGM10].

However, for extreme changes in illumination and pose (rotation, scaling, occlusion) of the matched image the number of matched pairs is extremely low. To tackle this problem some approaches define a valid range of horizontal and vertical angle [MY09, YM09] or a valid range of angle and illumination [YHCT12] and simulate the image to every possible view using these poses to the matching procedure. Nevertheless, the creation of different views from the query image and the brute force matching with the candidate image is time-consuming, and thus inappropriate for real-time operation. To tackle this, Ta et. al. [TCGP09] proposed an efficient algorithm called SURFTrac which combines SURF descriptors and motion information in order to predict the position of the interesting points at the subsequent frame, aiming at the restriction of the search area and the reduction of the computation time. Sivic and Zisserman [SSZ04] based on their previous work [SZ03] presented a technique for fast video object retrieval, in a manner similar with the text retrieval used by Google. They represent the image with a set of viewpoint and affine invariant SIFT descriptors, applying a temporal smoothing over a group of contiguous frames in a shot for better results. Finally, a vector quantization step creates a visual vocabulary from the descriptors and employs the technology of text retrieval at run time.

A different class of approaches that achieve RST-invariance (rotation, scaling and translation) using a prior segmentation/binarization step has also been proposed. These algorithms firstly convert the grayscale images to a binary form using some thresholding procedure, afterwards they compute some RST-invariant feature for each connected component of the image, and finally compare these features between the pair of tested images. An example of such a method is described in [TMRSSG00]. The most commonly used rotation-invariant features include Hu's seven moments [Hu62] and Zernike moments [TC88]. Recently many other rotation-invariant features have been developed and for some examples we refer the reader to [LPC^{+}04, FS06, TIT01]. Moreover, histograms of oriented gradients (HOG) [DT05] have also been introduced by many authors for rotation discriminating template matching [UK04]. In [ME07] a much faster technique was presented, where the speed-up is mainly due to the use of integral histograms. Finally, a technique which combines histograms of oriented gradients with binarization techniques is described in [Sib11].

Other techniques use circular projections for the rotation-invariant template matching, based on the fact that features computed over circular or annular regions are intrinsically rotation-invariant. Choi and Kim [CK02] accelerate circular projection-based rotation-invariant template matching via low-frequency complex Fourier coefficients. Based on this idea, Kim [Kim10] developed another RST-invariant template matching based on Fourier transform of the radial projections named Forapro. Kim and Arajo [KdA07], based on the idea that the circular projection followed by radial projection could lead to scale and rotation-discriminating template matching, proposed a new method called Ciratefi. This method consists of three cascaded filters that gradually reject pixels with no chance of matching the query image from further processing.

All the above mentioned techniques are applied on gray-scale images. However, color provides high discriminative power and some approaches for image matching by exploiting the color information, have also been proposed. Tsai and Tsai [TT02] presented a technique for matching colored objects, called color ring-projection. Araujo and Kim in their later works [AK10, dAK11] extended their initial algorithm and proposed the Color-Ciratefi, using a new similarity metric in the CIELAB space to achieve invariance to brightness and contrast changes. Finally, many other approaches use a combination of color and texture features in order to make them more descriptive. Geusebroek et al. [GvdBSG01] developed a set of color invariant features based on Gaussian derivatives, which have been embedded in SIFT descriptor by Burghouts and Geusebroek, yielding a powerful color invariant descriptor [BG09]. Many other similar attempts that combine color invariants with SIFT descriptor have been proposed, generating color-based SIFT descriptors like CSIFT [AHF06], Transformed color SIFT [vdSGS10a], SIFT-CCH [AB07], and C-color-SIFT [BG09]. Moreover, in other works, SIFT descriptors have been used in combination with photometric invariant color histograms [vdWS06], Luv color moments [QO06] and MPEG-7 color descriptors [SSBT07].

The object matching problem has also been extensively studied as a graph matching task. Leordeanu and Hebert [LH05] proposed a spectral method where the correspondences are obtained by finding the principal eigenvector of a matrix, while a similar rotation invariant approach has also been used in a point matching method [ZD06]. Alternatively, Cour et. al. [CSS06] proposed a spectral relaxation method that incorporates one-to-one or one-to-many mapping constraints, and presented bi-stochastic normalization of the compatibility matrix to improve the overall performance. Moreover, graph-based approaches that involve feature points have also been introduced. Feature points are modelled as graph nodes, and geometric relations between pairs of feature points as graph edges. However, a major limitation of this approach is that order-2 edges can only provide rotational invariance. To tackle this, Zass and Shashua [ZS08] and Duchenne et. al. [DBKP11] extended ordinary graphs to hyper-graphs, whose high-order edges can encode more complex geometric invariance.

Finally, another group of methods models the image matching task as a mathematical programming problem. Chui and Rangarajan [CR03] converted the image matching to a mixed variable optimization problem. Berg et. al. [BBM05] modelled the matching problem as a quadratic integer programming problem using pairwise relationships between feature points and penalizing both rotation and scaling differences. Recently, linear programming has also been used in object matching and indicative examples of such approaches can be found in [JDL07, JY09, LKHH10].

6.2 Event Detection

Nowadays, the rapid growth of the available video data makes clear the great need of advanced techniques for more effective ways of video indexing, summarization, browsing, and retrieval. As an essential step to facilitate automatic video content manipulation, video event detection has attracted great attention from the research community. In the following subsections we report the main steps of this demanding task and list the state-of-art techniques for each step.

6.2.1 Model Vector Presentation

Model vectors were originally proposed for the task of image and video retrieval, where they have been used for the representation of high-level semantics. However many researchers extended this idea and employed model vectors for the detection and description of high-level visual events. Regarding the efforts for indexing and retrieval of multimedia content, Smith et. al. [SNN03] used model vectors as a semantic signature for multimedia documents, where each dimension of the model vector represents the confidence score by which a concept from a pre-defined lexicon was detected. A similar approach was described in their later work [NNS04] for semantic content-based multimedia retrieval, classification and mining purposes. Rasiwasia et. al. [RMV07] introduced a “Query-by-Semantic-Example” (QBSE) approach, where images were labelled based on a vocabulary of visual concepts and were represented by a vector of posterior concept probabilities, called “semantic multinomial”. At the retrieval procedure, a “semantic multinomial” was computed for each query image and was matched to those in a database. A more sophisticated approach has been introduced in [EXCS06], where the authors proposed a semantic model vector representation for modelling the dynamic evolution of semantics within video shots. Moreover, Xu et. al. [XC08] used the Columbia374-baseline semantic concept classifiers [YCKH07] for video event classification, while Torresani et. al. [TSF10] represented images based on an ontology of visual concept and they trained weak object classifiers for object category recognition. Several other efficient uses of model vectors have been described by Natsev et. al. [MHX$^+$10] and

Mezaris et. al. [MSDK10, SMK$^+$11, GMK11a]. In [MHX$^+$10] the authors extended their previous work (see [SNN03, NNS04]) to model and detect complex events in unconstrained real-world videos, such as those from YouTube. In [MSDK10] Mezaris et. al. described a method for video temporal segmentation to scenes, based on a shot semantic similarity measure. They used a large number of non-binary detectors for the definition of a high-dimensional semantic space, where each shot was represented by the vector of detector confidence scores in the range 0 to 1. The similarity between two shots was evaluated by defining an appropriate shot semantic similarity measure. Moreover, in [SMK$^+$11, GMK11a] model vectors are constructed from the responses of trained visual concept detectors and are used as high-level visual features, which are in turn used for effective event detection in video.

6.2.2 Dimensionality Reduction Techniques

Regardless of whether we consider low-level visual features or model vectors or a combination of them, the high dimensionality of image descriptors is an important drawback for their efficient manipulation towards complex event detection. This problem becomes particularly pronounced when considering that for learning an event detector, typically only a few positive samples of the event in question are available. For this reason, dimensionality reduction often becomes a crusial part of the event detection pipeline. Several methods for dimensionality reduction have been proposed by the research community. The most widely used methods are: Principal Component Analysis (PCA); Singular Value Decomposition (SVD); and Multidimensional Scaling (MDS). The first two implement axes rotation of the original feature vector space, finding a subspace that best preserves the variance of the original distribution. Dealing with the dimensionality of the SIFT descriptor the authors in [KS04] proposed a more distinctive and robust PCA-SIFT descriptor. Other PCA-based approaches were introduced in [GD05, GGVS08, SZZ07]. In MDS the low-dimensional representation is found by minimizing certain cost functions and a weighted version of the MDS algorithm has been proposed in [WMS00]. Random Projections were used by Chuohao et. al. [YAR08] on SIFT descriptors in order to build binary hashes. However, the main drawback of these approaches is that they only characterize linear subspaces in the data. In order to resolve the problem of dimensionality reduction in non-linear cases, various other techniques have been described. Isomap [TdSL00] is a non-linear generalization of classical MDS. The main idea is to perform MDS, not in the input space, but in the geodesic space of the non-linear data manifold. Laplacian Eigenmaps [BN03] compute the low-dimension representation of a high-dimension dataset that most faithfully preserves proximity relations, mapping nearby input patterns to nearby outputs. Other non-linear approaches include Self-Organizing Maps (SOMs) and Locally Linear Embedding (LLE).

Additionally, an overview of supervised discriminant projection methods is reported by Cai et. al. in [CMM11]. Linear Discriminant Analysis (LDA) is one of the most widely applied techniques. Its goal is to maximize the discrimination between different classes, while at the same time the within class distance is minimized. Linear Discriminant Embedding (LDE) was proposed in [CCL05], combining the information of nearest neighbors and class relations between data points, while a graph-based alternative, called Marginal Fisher Analysis (MFA) was described in [YXZ$^+$07]. Moreover, in [HBW07] the authors introduced a global version of the LDE and presented a method named Linear Discriminant Projections, while an alternative approach has been proposed in [MM07]. Other reported techniques in [CMM11] are: the Mahalanobis Distance Metric Learning (MDML) [YJ06]; the Global Distance Metric Learning (GDML) [XNJR02]; the Relevant Component Analysis (RCA) [BHHSW03]; the Neighborhood Component Analysis (NCA) [GRHS04]; the Large Margin Nearest Neighbor (LMNN) [WS09]; and a distance metric learning method, using Support Vector Machines (SVM) and Relative Comparisons (SVM-RC) [SJ04].

Different approaches that have also been introduced in the relevant literature include: a machine learning algorithm called Similarity Sensitive Coding (SSC) [TFW08] and a similar approach in [Sha05]; a Subclass Discriminant Analysis (SDA) method [ZM06] and an extension of it called Mixture SDA (MSDA) [GMK11b, GMKS12b]; a building blocks approach in [BGW11]; and a new descriptor called Vector of Locally Aggregated Descriptors (VLAD) in [JaDSP10] which aggregates SIFT descriptors in order to produce a more compact representation.

6.2.3 Associating Media with Events

During the past few years significant research has been devoted to the detection and recognition of events in several application areas such surveillance, multimedia indexing etc. In [XC08], a “bag-of-words” approach is combined with a multilevel sub-clip pyramid method to represent a video clip in the temporal domain, and the Earth Mover’s Distance (EMD) is then applied for recognizing events defined in the TRECVID 2005 challenge dataset. Similarly, a "bag-of-words" technique is used in [LL03], where the authors extract spatiotemporal interest points by focusing on salient changes in both spatial and temporal domains in order to detect spatiotemporal events in the video.

Another approach has been described in [GMK11a, TGD$^+$11] and has been used in the TRECVID 2010 and 2011 MED Task. The authors propose a model vector-based approach, where visual concept detectors are used to automatically describe the shots of a video sequence in a concept space (as mentioned in Section 6.2.1), and subsequently event detection is based on the analysis of the temporal evolution of the visual concept patterns. Moreover they invoke a discriminant analysis method, which is called Mixture Subclass Discriminant Analysis (MSDA) [GMKS12b] (as reported in Section 6.2.2), in order to identify the semantic concepts that best describe the event, thus defining a discriminant concept subspace for each event. Initially automatic techniques are used for the temporal segmentation of videos to shots [TMK08] and their description with low-level visual features (using SIFT descriptors [Low04]). Then, the descriptors are clustered to create a vocabulary of visual words which maps them to a new higher-level feature space. The new feature vectors are subsequently used as input to each one of the 231 SVM-based concept detectors trained on the MediaMill and the TRECVID SIN Task datasets and the resulting vectors, with values expressing the Degree of Confidence about the presence of a specific concept, are representing the corresponding shot. Subsequently they try to define a discriminant subspace for each event including the concepts that most appropriately describe the event by applying a discriminant analysis method proposed in [GMK11b, GMKS12b] (as mentioned in Section 6.2.2). In the resulting discriminant subspace, the Nearest Neighbor classifier (NN) along with the median Hausdorff distance are used to recognize an event.

The exploitation of motion information has also been intensively studied for event detection. In [WJN08] the authors introduced a new motion feature called Expanded Relative Motion Histogram of Bag-of-Visual-Words (ERMH-BoW) to employ motion relativity and visual relatedness for event detection. Similarly to [XC08], the EMD and Support Vector Machines (SVMs) are used to recognize video events in the TRECVID 2005 challenge dataset. Some attempts have also been made for motion-based event recognition in the compressed domain [ACAB99]. In [HN07], motion vectors from the MPEG stream are compressed to form a motion image, and event recognition is performed using SVMs. Another technique is described in [CNZ$^+$07], where motion is employed for the detection of some event-based topics. Eight directions of motion vectors and intensities are efficiently extracted from motion vectors in MPEG compressed video and exploited for the final ranking. An alternative approach was presented in [EXCS06] by Ebadollahi et. al., where visual events are viewed as stochastic temporal processes in the semantic space. The dynamic pattern of an event is modelled through the collective evolution patterns of the individual semantic concepts in the course of the visual event, and then Hidden Markov Models (HMMs) are employed for event modelling and recognition. Finally, in [BBDBS10], knowledge embedded into ontologies and concept detectors based on SVMs are used for recognizing events in the domains of broadcast news and surveillance.

Moreover, recent approaches for event detection are often based on the integration of multi-modal information, since it captures the video content in a more comprehensive manner. In [BKK02], after the segmentation of the video to shots, the textual channel is analyzed for the occurrence of specific keywords that are related to a semantic event in American football. This results in a time interval and specific shots where a possible event has taken place. In [CCSW06] the event detection is based on temporal pattern analysis and multi-modal data mining. In [JZY$^+$10], three types of features, namely, spatiotemporal interest points, SIFT features, and a bag-of-MFCC-audio-words, are used to train SVM-based classifiers for recognizing the three events of the TRECVID 2010 MED Task. In [HHH$^+$10], a wide range of static and dynamic features are extracted, such as SIFT and GIST features, 13 different visual descriptors on 8 granularities, histograms of gradients and flow, MFCC audio features, and other. These features are used for training 272 SVM-based semantic detectors and, thus, represent videos with model vector sequences. Subsequently, hierarchical Hidden Markov Models (HMMs) are applied to recognize the TRECVID 2010 MED events. In [Nat11] the novel Raytheon BBN VISER system for event detection is presented, combining visual with audio words, thus resulting in a bi-modal word codebook. Initially, low level visual and audio features are extracted. Visual features are divided into appearance (SIFT, SURF, dense SURF and compressed HoG), colour (RGB-SIFT, Opponent-SIFT, Color-SIFT) and motion (spatiotemporal interest points (STIP) and dense STIP). Audio features include Mel Scale Cepstral Coefficients (MFCC), frequency domain linear prediction (FDLP) and audio transients. Then, the "bag-of-words" representation maps the low-level features to a high-dimensional feature space. High level features are divided into visual and text features. Visual features utilize objects and scene concepts through detectors. Automatic Speech Recognition (ASR) and Video Object Character Recognition

(OCR) belong to the text features. At the final step of classification both early and late fusion schemes are tested. A similar approach is described in [BYL$^+$11] where the Informedia event detection system is proposed. Both low- and high-level visual and audio features as well as text features are extracted. Along with the “bag-of-words” algorithm, Spatial-Pyramid Matching technique is used to represent the low-level visual features. Moreover, in the detector training part, besides the traditional SVM, a Sequential Boosting SVM classier is proposed to deal with the large-scale unbalanced classification problem. In the fusion part, three different methods, namely early fusion, late fusion and double fusion are tested. Another multi-modal technique is demonstrated in [IKW$^+$11] where the event detection system of Canon is described. Feature extraction includes low-level features of the visual and audio domain. SIFT is based on Harris and Hessian affine regions while the Histogram of Oriented Gradients (HOG) and the Histogram of Optical Flow (HOF) are based on STIP. Dense HOG is calculated based on a dense sampling of the image. Audio is represented by MFCCs including the delta and acceleration coefficients along with delta and double delta energy. Principal Component Analysis (PCA) is then applied to the visual features for dimensionality reduction. Subsequently, all features are fed into pre-trained GMMs leading to the so-called GMM super-vectors which are finally fed into an RBF-kernel SVM for event detection.

Finally, an audio-based event detection technique is proposed in [MLG$^+$11]. Specifically, an acoustic data-driven event detection framework for large scale event detection is presented. The approach does not rely on speech processing, is language-independent and can be used in complement with video analysis. Two low-level audio features, called Linear Frequency Cepstral Coefficients (LFCC) and Modulation Filtered Spectrogram (MSG), are used as discriminative features. Firstly a mean and standard deviation normalization is applied and then a Universal Background Gaussian Mixture Model (UBM-GMM) is trained using all data. Event-dependent GMMs are also trained by MAP adaptation from the UBM. The training process works in parallel for both features and a late fusion scheme is applied.

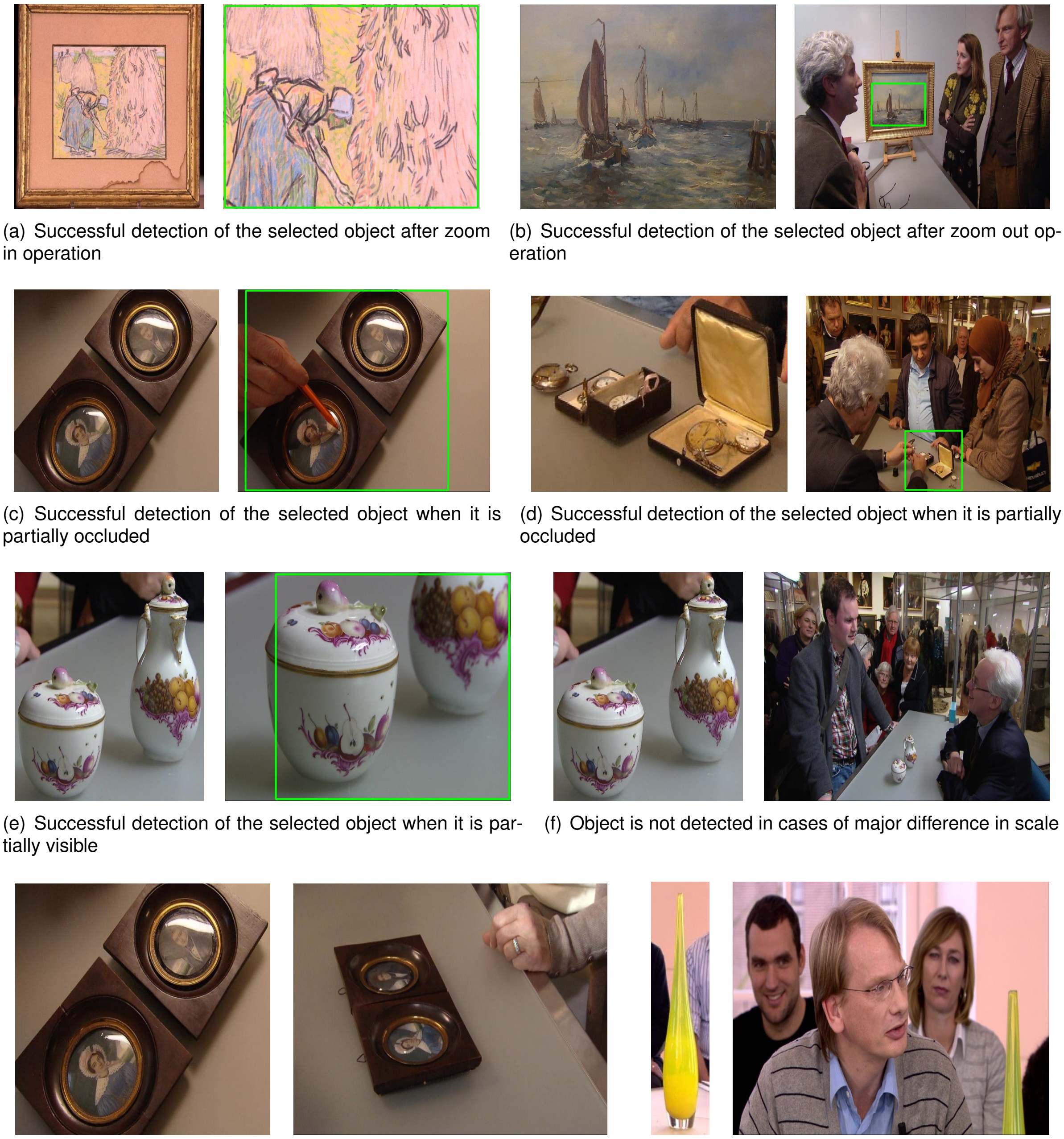
6.3 Technical Requirements

Object re-detection: according to the scenario specifications and as already mentioned in Section 3.6, the detection and tracking of objects of interest within a video, in order to be associated with further linkable information is one of the functionalities of LinkedTV annotation system. The problem of moving object detection and tracking was discussed in detail in Sections 3.4.1 and 3.4.2. However, a user may be interested in particular parts of the video where a specific static object appears; this is mainly the case for the documentary scenario. For detecting occurrences of static objects of interest in consecutive or non-consecutive video frames we experimented with a semi-automatic approach for object re-detection based on a baseline OpenCV implementation. Initially the user will manually specify the object of interest by marking a bounding box on one frame of the video, and then additional instances of the same object in subsequent frames will be automatically detected via object matching. For each pair of images, feature vectors are extracted using the SURF algorithm [BETVG08] and are compared. False matches are filtered out using a symmetrical matching scheme between the pair of images, and the remaining outliers will be discarded by applying some geometric constraints calculated from the RANSAC method [FB81].

We evaluated this algorithm by using some manually selected objects of the documentary scenario; for indicative examples we refer the reader to Figure 8. In most cases, the selected object is successfully detected when zoom in or zoom out operation is taking place (see Fig. 8(a), 8(b)), when the object is partially visible or partially occluded (see Fig. 8(c), 8(e)) or at the extreme case when both scaling and occlusion are taking place (see Fig. 8(d)). However, the technique exhibits sensitivity to major changes in scaling or rotation and in some cases fails to correctly detect the selected object (see Fig. 8(f), 8(g)). It is important to notice that in many cases rotation may lead to significant change of the background information, and thus detection failure (see Fig. 8(h)). In conclusion, the baseline implementation already offers promising functionality, but there is also room for improvement, both in terms of accuracy and also in terms of computational efficiency.

The first requirement regarding the accuracy of the detection results can be addressed via a wiser and more effective selection of the object of interest. In our experiments we used simple bounding boxes to mark the area, while a tight selection with more versatile geometric shapes could possibly eliminate detection failures due to background changes. Concerning the second requirement about the time efficiency of the algorithm, the most time consuming part of the algorithm is the image feature extraction and description part. Hence, we plan to meet this requirement by accelerating these processing steps with the use of a GPU-based implementation for the calculation of the SURF descriptors (see Section 4.3).

Event detection: in Section 4.1 we highlighted that an important functionality based on the defined



(a) Successful detection of the selected object after zoom in operation

(b) Successful detection of the selected object after zoom out operation

(c) Successful detection of the selected object when it is partially occluded

(d) Successful detection of the selected object when it is partially occluded

(e) Successful detection of the selected object when it is partially visible

(f) Object is not detected in cases of major difference in scale

(g) Object is not detected in cases of significant rotation

(h) Object is not detected when rotation changes significantly the background information

Figure 8: Object detection examples for the documentary scenario. In each of these figures the left image represents the object of interest and the green rectangle in the right image demarcates the detected object.

scenarios is the estimation of similarity among videos. After watching a video of interest, a user may need to be linked to a comparable new video or to make a group of them in a kind of personal collection. Apart from identifying and assigning labels (concepts) to the media content, such functionality can be based on similarities between media events associated to the visual content. By events we do not refer to elementary actions like "stand up" or "shake hands", but to higher-level actions like "airplane landing" or "making a cake". In this way, event detection should enable us to interlink a video with other videos that contain a similar set of visual information, thus offering a nice guideline for video recommendation.

Our current approach to detect events in the video is described in [GMK11a]. An event is detected and recognized by the temporal evolution of specific visual concept patterns. A model vector-based approach is proposed, where visual concept detectors are used to automatically describe a video sequence in a concept space. After representing each shot with an appropriate model vector, a novel Discriminant analysis (DA) technique is invoked for identifying the semantic concepts that best describe the event, thus defining a discriminant concept subspace for each event. This method extends the recently proposed Subclass Discriminant Analysis (SDA) technique [ZM06], to further improve recognition accuracy and degree of dimensionality reduction. In the resulting discriminant subspace, the nearest neighbor classifier (NN) along with the median Hausdorff distance are used to recognize an event.

Similarly with other mentioned cases, we have to address the two conflicting demands for more detection accuracy and less processing time. Regarding the first requirement we intend to incorporate the information for the audio channel. Specifically, by enriching the event detection algorithm with some audio features, we expect to obtain a significant improvement in detection accuracy. Moreover, the event detection technique should benefit from improving the accuracy of the concept detection algorithm, since the output data of this algorithm are used as input to the event detection technique. Hence, by meeting the accuracy requirements of the concept detection algorithm (as mentioned in Section 4.4), we will contribute to the further improvement of the event detection method. Concerning the needs for limited processing time, we will focus on dimensionality reduction techniques which will help us to dramatically restrict the discriminant concept subspace for each event. For this purpose, we will avoid employing the time consuming Kernel-based Discriminant Analysis (KDA) techniques and instead, we intend to use other SDA extensions, like the Mixture Subclass Discriminate Analysis (MSDA) algorithm that have we already implemented and described in [GMKS12b].

Furthermore, extending the problem of event detection to the detection of social events (by “social events” meaning here events that are planned by people, attended by people and that the media illustrating the events are captured by people), our results indicate that using a rich set of metadata (e.g., for web images, a combination of visual information, user-assigned tags, time and geo=location information etc.) is beneficial in comparison to using a more limited set of information sources for affecting Social event detection [PSM$^+$12]. To this end, we plan to develop methods that jointly exploit a wealth of available information for the videos under consideration in LinkedTV in order to further enhance the effectiveness of our event detection techniques.

7 User Assisted Annotation

This section studies the core issues of the annotation process, starting with a description of the basic characteristics and functionalities that an annotation tool must support for the LinkedTV purposes. It should be clarified that in this section with the term “user” we refer to the video producer which interacts with the automatic analysis results and makes the necessary corrections, additions etc. in order to produce the final annotated video. Since there is an ever increasing variety of available tools with varying degrees of annotation depth and interchangeabilty, we then report some state-of-art available tools for the annotation of digital audio-video data material. Afterwards, aligned with the demands of the LinkedTV annotation procedure, we identify the technical requirements that must be met by the annotation tool, like the annotation format and its fine-granularity in order to decide on which existing tool(s) the further work will be carried out. Finally we present the current status of the annotation procedure within the LinkedTV.

7.1 Workflow

In LinkedTV, we will need to store information bound by time frames as well as (possibly moving) areas within a video. We need the possibility for parallel and hierarchical annotation layers. Further, we need layers for additional information, e.g., semantic concept, named entity. The format should be interchangeable and well-established, i.e., work for the the annotation tool(s), the semantic procession, the player, and can be easily wrapped in from the statistical recognition results. More precisely, annotation is actually needed on three different levels:

- assessment and ground truth creation for textual data. Here, we may need hierarchical dependencies, like, e.g., keyword layers that further annotate automatic speech recognition transcriptions. Also, concept detections for the whole video frame, e.g., “bike” or “indoor” should be annotatable.
- assessment and ground truth creation for visual data. This is needed for the annotation of, e.g., faces within frames, static and/or moving objects of interest.

– editoring hyperlinked material. Here, the links offered to the end-user can be checked and corrected by an editor. This is especially important for the RBB's News Show scenario, where the links are somewhat restricted and have to be controlled with respect to white lists and accuracy. Also, manual high-level information as to the nature for the link could be inserted here.

While in theory these functions could be merged, we believe that this would produce too much unnecessary work overhead. The textual information, only bound by time frames, can be easily edited by existing language labelling and transcription tools, so that existing technology can be used with suitably tailored format exchange algorithms. Similar, the annotation of visual information, as will be shown below, is well-advanced in terms of available open-source technology. The editing of the hyperlinking material, however, should be performed on the same interface level as the actual player for the end-user operates. In daily usage, an editor is not interested in the vast amount of data derived in the A/V analysis step, but only in the links offered on the seed material.

See Figure 9 for a graphical representation of the overall workflow.

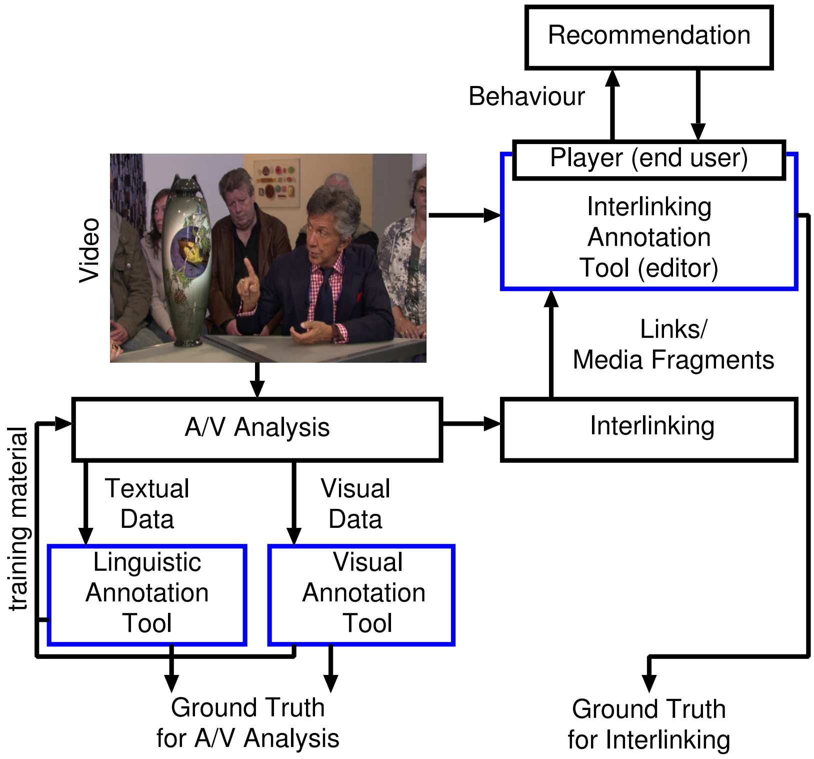


Figure 9: Workflow of the data within LinkedTV, with blue rectangles marking various needs for annotation within the project.

7.2 Available Annotation Tools

Nowadays, a huge amount of different annotation tools is available, suitably tailored for various purposes like linguistic annotation, audio transcription or object annotation. While an exhaustive list would be out of scope, we proceed to list some prominent examples.

7.2.1 Audio transcription

ELAN [WBR+06][11] (Version 4.1.2) is developed by the Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands. It is designed with special focus on sign language annotation, where the video is typically at a frame rate of 50 fps. Annotations are on a time-basis, in the current state, however, spatio-temporal parts of the video cannot be annotated. They can be grouped in layers, which can be independent of each other, embedded or aligned. Typically, the in-house format ".eaf" is used, but import and export are provided for Shoebox, Toolbox, Flex, Chat, Transcriber,

[11] `http://www.lat-mpi.eu/tools/elan/`

Praat, as well as simple csv files. Support is given for Windows, Mac and Linux. Sources are available for non-commercial applications.

Transcriber [BGWL00][12] is developed mainly for linguistic research on speech signals. It supports multiple hierarchical layers of segmentation, named entity annotation, speaker lists, topic lists, and overlapping speakers. Unicode encoding is provided, and the main architecture is in TCL/TK.

ANVIL [Kip08][13] (Version 5) was developed by Michael Kipp at the DFKI in Saarbrücken, Germany. Its main application are multi-modality studies in general. Annotations can be imported from Praat, ELAN, and others. Recently, spatio-temporal annotation has been added, so that one element does not necessarily have to refer to the whole frame but instead to screen regions (e.g. face, hand direction, etc.) ANVIL is free for non-commercial usage, but the source code is not open.

EXMARaLDA [SWHL11][14] was developed within the German project "Computergestützte Erfassungs- und Analysemethoden multilingualer Daten" (computer assisted annotation and analysis of multilingual data), University of Hamburg, Germany. Since Juli 2011, it is maintained at the Hamburger Zentrum für Sprachkorpora. Its main users are students and researchers in discourse or conversation analysis as well as language acquisition studies. A video panel is provided as well, though EXMARaLDA's main use is the annotation of multi-lingual spoken corpora. It is implemented in Java and based on the MIT License.[15]

7.2.2 Annotation tools used for object description

Kat [SSS08][16] Kat is an open source framework for semi-automatic annotation of multimedia content, which was developed within the K-Space Network of Excellence. Formal model based on the Core Ontology on Multimedia (COMM)[17].

Label Me [RTMF08][18] LabelMe is open object annotation tools written in Javascript for on-line image labelling. The source code can be downloaded to set up an own server for labelling an arbitrarily shaped simplex. Support for Amazon Mechanical Turk[19] is given.

M-OntoMat 2.0 [PAS$^+$06][20] M-OntoMat is an annotation framework tool which builds upon the OntoMat-Annotizer and supports automatic segmentation and annotation at segment and image level. It annotates domain ontology concepts like geographic objects or people to image segments or images.

7.3 Technical Requirements

We identified the following key questions for the annotation tools, which are mainly taken from [RLD$^+$06].

Source Does the tool run under any OS, or is it web-based (e.g., java application)? Is the source code available and may it be used for commercial applications? Does it come with a fee?

Video Format While conversion between video formats is usually feasible, support for a variety of common input formats is nice to have.

Interchangebility Does the annotation tool foresee import from other standardized formats, and does it support exporting its own format to other tools? Is it human readable in an XML annotation scheme?

Controls Is the view time-aligned, with reliable frame-by-frame view, ideally at a video speed at the users discretion? Can the audio be played at slower speeds? Is a reliable search functionality included?

[12] http://trans.sourceforge.net/en/features.php

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

[14] http://www.exmaralda.org/index.html

[15] http://www.opensource.org/licenses/mit-license.php

[16] http://www.uni-koblenz-landau.de/koblenz/fb4/AGStaab/Research/systeme/kat

[17] http://comm.semanticweb.org/

[18] http://labelme.csail.mit.edu/

[19] http://mturk.com/

[20] http://mklab.iti.gr/m-onto2

Annotation Does the tool provide multiple layers? Are hierarchical annotations allowed? Are the segments marked only time-wise, or is a mark-up of areas in the video (which might be moving over time) provided as well? Does it allow for search functionality?

Multiple Users Is there a support for multiple users working on the same data simultaneously? Does it allow for e.g. repository management, for crowd sourcing?

Analysis How easy is it to integrate analysis functionality like automatic segmenting, automatic speech recognition, detection of visual hot spots?

Maintenance Is it easy to learn for new users? Will it be maintained/patched/upgraded in the near future? Are there many users working with it?

For the prototype annotation tool(s) delivered in D 1.3, we will use these requirements as guidelines for the evaluation of the usefulness. Some of the requirements are nice-to-have, some are mandatory: for example, the editing of the hyperlinking should be web-based, and the tool should support multiple layers since we expect a rather large amount of parallel data. Another crucial requirement is an open-source implementation so that it can be further adapted to the LinkedTV needs. Finally, a huge priority should be that the data exchange format is interchangeable.

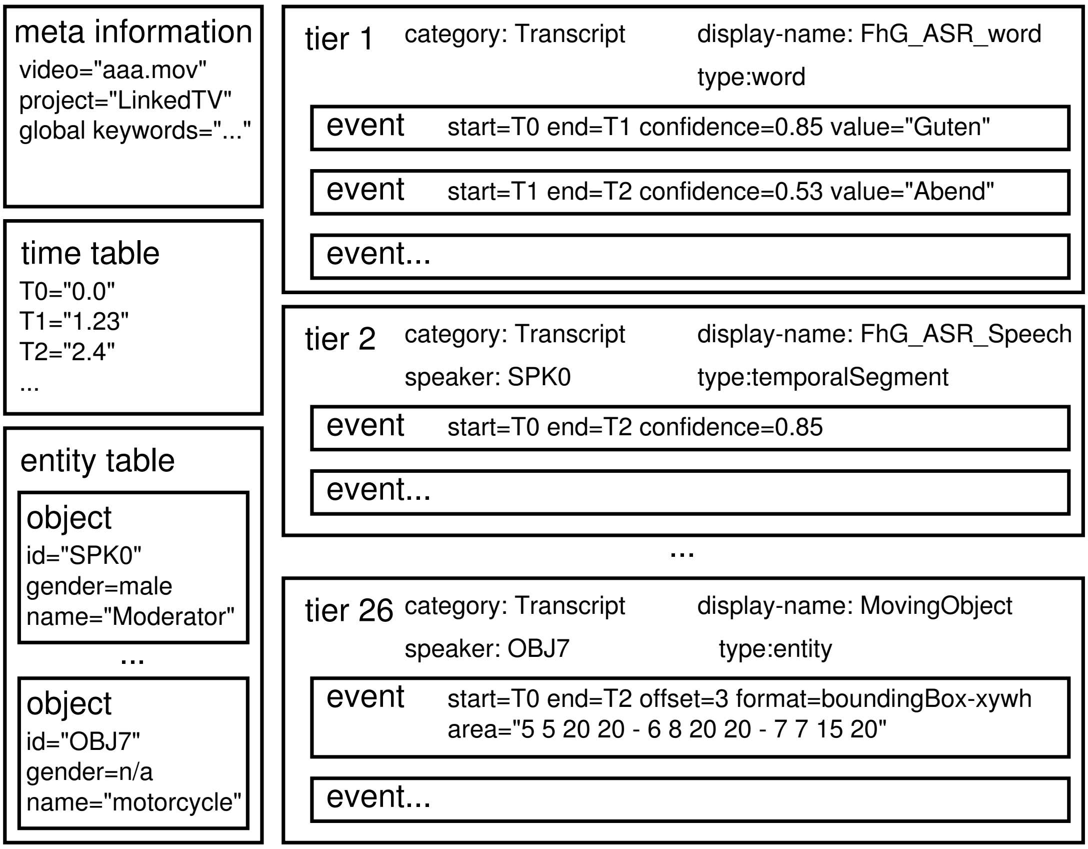


Figure 10: Format of the XML file that contains all the automatically derived data for one video. It is readable by EXMARaLDA.

7.4 Current Status within LinkedTV

In preliminary work, the following workflow has already been achieved: all the raw data derived in the various analysis techniques can be joined in a single XML file per video. The XML file stores general meta data like the name of the video file and the transcription version, global time stamps produced by the analysis techniques (e.g., start and end time of a shot), and an entity table for speakers and objects. For each information layer, a “tier” is produced which contains the single events like recognized words or objects. All events contain a start and an end point; if they are extracted from a single key frame, we still list the time span of the appropriate shot. The tiers and events can contain further information, like x-y-w-h bounding boxes, confidence scores or other, possibly arbitrary information. See Figure 10 for an overview of the format, and Figure 11 for a graphical representation of the workflow.

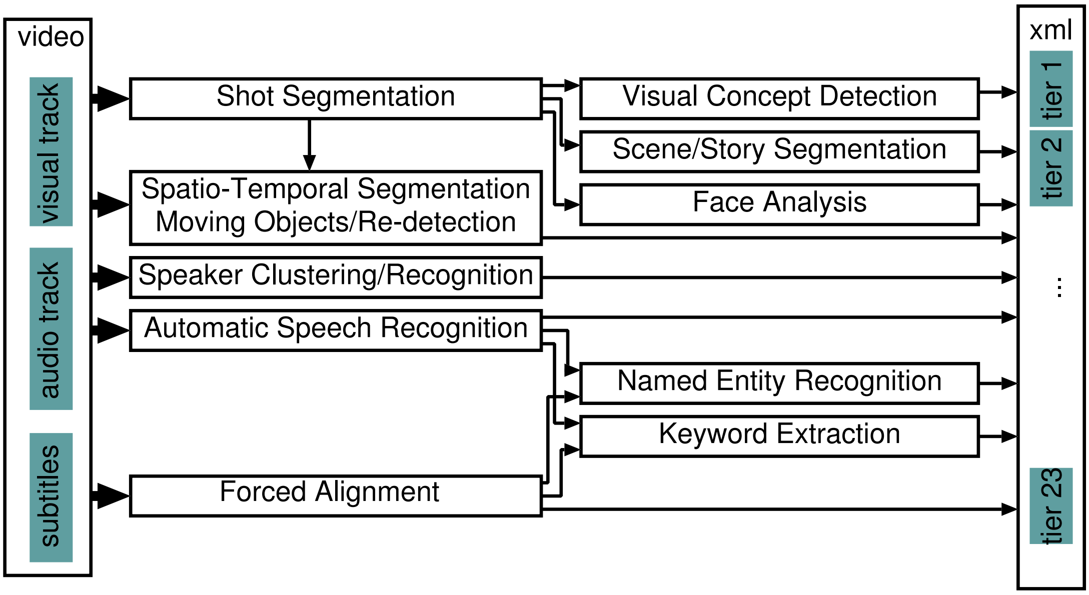
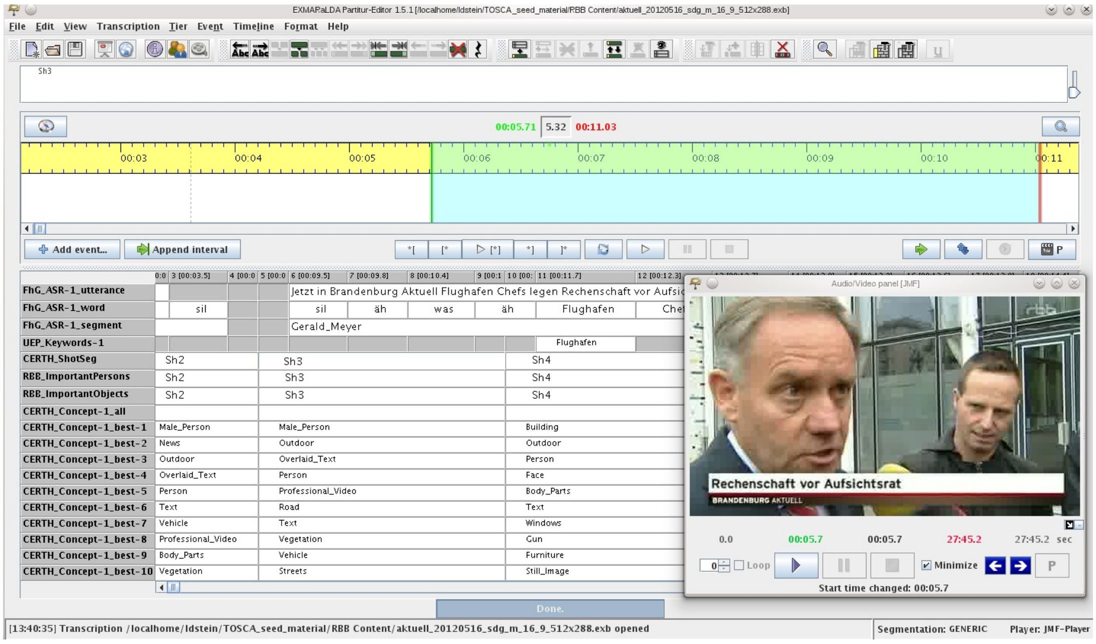


Figure 11: Workflow of the raw a/v analysis data as derived in WP 1.

For the annotation of textual data that does not contain spatial information, we currently selected the “Extensible Markup Language for Discourse Annotation” (EXMARaLDA) toolkit [SDE+09], as described above. EXMARaLDA bears the benefit that it is open source, compiles easily on different platforms and has a relatively easy data exchange format. See Figure 12 for a snapshot. Since EXMARaLDA does not support spatial annotation, this information is currently stored as additional information within the XML (i.e., when the XML file is edited with the annotation tool and saved thereafter, no information loss of spatial data occurs). In the next step, we are searching for a loss-free way to convert the XML into a format a spatial annotation tool can process.



	0:0	3 [00:03.5]	4 [00:0	5 [00:0	6 [00:09.5]	7 [00:09.8]	8 [00:10.4]	9 [00:1	10 [00:	11 [00:11.7]	12 [00:12.3]
FhG_ASR-1_utterance					Jetzt in Brandenburg Aktuell Flughafen Chefs legen Rechenschaft vor Aufsic						
FhG_ASR-1_word		sil			sil	äh	was	äh		Flughafen	Che
FhG_ASR-1_segment					Gerald_Meyer						
UEP_Keywords-1										Flughafen	
CERTH_ShotSeg	Sh2			Sh3					Sh4		
RBB_ImportantPersons	Sh2			Sh3					Sh4		
RBB_ImportantObjects	Sh2			Sh3					Sh4		
CERTH_Concept-1_all											
CERTH_Concept-1_best-1	Male_Person			Male_Person					Building		
CERTH_Concept-1_best-2	News			Outdoor					Outdoor		
CERTH_Concept-1_best-3	Outdoor			Overlaid_Text					Person		
CERTH_Concept-1_best-4	Overlaid_Text			Person					Face		
CERTH_Concept-1_best-5	Person			Professional_Video					Body_Parts		
CERTH_Concept-1_best-6	Text			Road					Text		
CERTH_Concept-1_best-7	Vehicle			Text					Windows		
CERTH_Concept-1_best-8	Professional_Video			Vegetation					Gun		
CERTH_Concept-1_best-9	Body_Parts			Vehicle					Furniture		
CERTH_Concept-1_best-10	Vegetation			Streets					Still_Image		

Start time changed: 00:05.7

Done.

[13:40:35] Transcription /localhome/ldstein/TOSCA_seed_material/RBB Content/aktuell_20120516_sdg_m_16_9_512x288.exb opened

Segmentation: GENERIC Player: JMF-Player

Figure 12: Screenshot of the EXMARaLDA GUI, with automatically derived information extracted from an rbb video.

8 Conclusions

The goal of this deliverable was to report on the state-of-art methods and the arising technical requirements for hypervideo in the LinkedTV framework. For that purpose, we initially presented two of its envisioned scenarios: the News Show scenario taking material from the "rbb aktuell" German news show, and the Documentary scenario taking material from the Dutch "Tussen Kunst & Kitsch" show. For each scenario we described in detail three user archetypes. We believe that the provision of detailed archetype descriptions is a very basic step, so that all the various video analysis techniques have a clear purpose in the readers' mind. Based on this, we can clearly specify the user-side requirements and we can both realize which method should be incorporated or must be improved and estimate whenever we have reached a level of maturity for a technique, that offers all the functionality we need.

After the scenarios specification, in the next two sections we reported various state-of-art methods about several visual, textual and audio analysis techniques, along with a requirement analysis part for each one of them. Concerning on the visual information processing, we described a set of methods that perform temporal decomposition of the media content into elementary parts (shots, scenes), while other methods implement spatial and spatiotemporal segmentation of still images and sequences of frames, in order to detect static and moving objects of interest from the viewer's perspective. Moreover we presented various methods for associating the specified content segments with appropriate labels, focusing on concept detection and face analysis techniques. In the following section we investigated various techniques for text and audio analysis, such as keyword extraction and speech recognition, which could provide complementary information, thus contributing to a more accurate definition of the previously defined content segments and the associated labels to them. State-of-art methods for event detection and object re-detection are presented in the next section, since this functionality could further improve the labelling procedure, providing information about high-level concepts in the video. Finally, we reported some available tools for media annotation and we described the functionality and the specifications that the LinkedTV annotation tool must meet in order to effectively support the user annotation procedure of the hypervideo.

In addition to the review of the state-of-art, for most classes of techniques (e.g., shot segmentation, concept detection) we made preliminary experiments with an implementation of such a technique that is already available to the LinkedTV consortium partners, on LinkedTV content. We report the results of these experiments in the present deliverable, following the relevant state-of-art review, and based on them and the scenarios we outlined in Section 2, we extract the technical requirements and the future challenges that relate to each analysis process. We are aware that not all the used techniques are "bleeding-edge" technology but sometimes already well-established. Through the evaluation we pinpointed several aspects of the analysis techniques that show much room for improvement. Since we are at an early stage within the LinkedTV project, we will use the findings of this deliverable as a strong basis from which to proceed. Furthermore, in addition to the challenges already addressed in the individual technical requirements subsections, we realized that we could incorporate additional techniques such as Optical Character Recognition (OCR) for the banner information, in order to obtain a database for face/speaker recognition. Another challenge will be to interweave the single results into refined multimodal information and to find synergies among existent methods in order to improve the accuracy of the current methods or to obtain new types of information. For example a topic segmentation technique could be based on information from both the audio and visual parts, while a person detection algorithm could gain information from automatic speech recognition, speaker recognition and face recognition. Also, in order to find reasonable story segments in a larger video, we could draw knowledge both from speech segments, topic classification, and video shot segments. As a final example, video similarity can be estimated with feature vectors carrying information from the concept detection, the keywords extraction, the topic classification and the entities detected within the video.

Therefore, we see the work reported here as a concrete basis and useful knowledge for future improvements aiming to extend state-of-art technologies in order to meet the high requirements of multimedia analysis for the hypervideo linking procedure within the LinkedTV project.

Bibliography

[AB07] C. Ancuti and P. Bekaert. SIFT-CCH: Increasing the SIFT distinctness by color co-occurrence histograms. In *5th International Symposium on Image and Signal Processing and Analysis, ISPA '07*, pages 130–135, September 2007.

[ABC+03] A. Amir, M. Berg, S.-F. Chang, G. Iyengar, C.-Y. Lin, A. Natsev, C. Neti, H. Nock, M. Naphade, W. Hsu, J. R. Smith, B. Tseng, Y. Wu, D. Zhang, and I. T. J. Watson. IBM research TRECVID-2003 video retrieval system. In *NIST TRECVID-2003*, 2003.

[ABE+11] X. Anguera, S. Bozonnet, N. W. D. Evans, C. Fredouille, O. Friedland, and O. Vinyals. Speaker diarization : A review of recent research. *IEEE Transactions On Audio, Speech, and Language Processing (TASLP), special issue on New Frontiers in Rich Transcription, February 2012, Volume 20, ISSN: 1558-7916*, 05 2011.

[ACAB99] E. Ardizzone, M. L. Cascia, A. Avanzato, and A. Bruna. Video indexing using MPEG motion compensation vectors. In *Proceedings of the 1999 IEEE International Conference on Multimedia Computing and Systems - Volume 2*, ICMCS '99, pages 725–729, Washington, DC, USA, 1999. IEEE Computer Society.

[acs] Acsys Biometrics. http://www.acsysbiometrics.com/.

[ADMK11] A. Athanasopoulos, A. Dimou, V. Mezaris, and I. Kompatsiaris. GPU acceleration for Support Vector Machines. In *12th International Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS '11*, Delft, The Netherlands, 04/2011 2011.

[AF09] A. Amiri and M. Fathy. Video shot boundary detection using QR-decomposition and Gaussian transition detection. *EURASIP J. Adv. Sig. Proc.*, 2009, 2009.

[AG05] A. Allauzen and J.-L. Gauvain. Open vocabulary ASR for audiovisual document indexation. In *Proceedings of ICASSP*, pages 1013–1016, April 2005.

[AGFM00] C. Auzanne, J. S. Garofolo, J. G. Fiscus, and W. M.Fisher. Automatic Language Model Adaptation for Spoken Document Retrieval. In *Proceedings of RIAO 2000*, pages 132–141, 2000.

[AHF06] A. E. Abdel-Hakim and A. A. Farag. CSIFT: A SIFT descriptor with color invariant characteristics. In *CVPR '06*, pages 1978–1983. IEEE Computer Society, 2006.

[AHP04] T. Ahonen, A. Hadid, and M. Pietikainen. Face Recognition with Local Binary Patterns Computer Vision. In Tomás Pajdla and Jiří Matas, editors, *Proceedings on Computer Vision, ECCV 2004*, volume 3021 of *Lecture Notes in Computer Science*, chapter 36, pages 469–481. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2004.

[AK02] A. Aner and J. R. Kender. Video summaries through mosaic-based shot and scene clustering. In *Proceedings of the 7th European Conference on Computer Vision-Part IV*, ECCV '02, pages 388–402, London, UK, 2002. Springer-Verlag.

[AK10] S. A. Araújo and H. Y. Kim. Color-Ciratefi: A color-based RST-invariant template matching algorithm. *17th Int. Conf. Systems, Signals and Image Processing*, 2010.

[AKM02] R. S. V. Achanta, M. S. Kankanhalli, and P. Mulhem. Compressed domain object tracking for automatic indexing of objects in MPEG home video. In *ICME*, pages 61–64. IEEE, 2002.

[AKM+03] H. Alani, S. Kim, D. E. Millard, M. J. Weal, W. Hall, P.I H. Lewis, and N. R. Shadbolt. Automatic ontology-based knowledge extraction from web documents. *IEEE Intelligent Systems*, 18(1):14–21, January 2003.

[AKT03] Y. Ariki, M. Kumano, and K. Tsukada. Highlight scene extraction in real time from baseball live video. In *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, MIR '03, pages 209–214, New York, NY, USA, 2003. ACM.

[AKWD10] A. Abramov, T. Kulvicius, F. Wörgötter, and B. Dellen. Real-time image segmentation on a GPU. pages 131–142. Springer-Verlag, Berlin, Heidelberg, 2010.

[ALK05] T. Alexandropoulos, V. Loumos, and E. Kayafas. Block-based change detection in the presence of ambient illumination variations. *JACIII*, pages 46–52, 2005.

[AM10] R. Albatal and P. Mulhem. MRIM-LIG at ImageCLEF 2010 Visual Concept Detection and Annotation task. In *CLEF (Notebook Papers/LABs/Workshops)*, 2010.

[ANP07] P. Antonopoulos, N. Nikolaidis, and I. Pitas. Hierarchical Face Clustering using SIFT Image Features. In *IEEE Symposium on Computational Intelligence in Image and Signal Processing, CIISP '07*, pages 325–329, April 2007.

[aqu12] face.com aquisition. http://face.com/blog/facebook-acquires-face-com, June 2012.

[aUoM12] Computer Vision Laboratory at University of Massachusetts. Labeled Faces in the Wild - Results. http://vis-www.cs.umass.edu/lfw/results.html, July 2012.

[AV07] D. Arthur and S. Vassilvitskii. K-means++: the advantages of careful seeding. In *SODA '07: Proceedings of the 19th annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035, 2007.

[AWRDG08] F. Alhwarin, C. Wang, D. Ristić-Durrant, and A. Gräser. Improved SIFT-features matching for object recognition. In *Proceedings of the 2008 international conference on Visions of Computer Science: BCS International Academic Conference*, VoCS'08, pages 179–190, Swinton, UK, 2008. British Computer Society.

[ayo] Ayonix. http://ayonix.com/.

[Bal] J. Baldridge. Apache opennlp. http://opennlp.apache.org/.

[Bau12] D. Baum. Recognising speakers from the topics they talk about. *Speech Communication*, 54(10):1132–1142, 2012.

[BB11] P. Bilinski and F. Bremond. Evaluation of local descriptors for action recognition in videos. In *Proceedings of the 8th international conference on Computer vision systems*, ICVS'11, pages 61–70, Berlin, Heidelberg, 2011. Springer-Verlag.

[BBBC11] J. Barnat, P. Bauch, L. Brim, and M. Ceska. Computing Strongly Connected Components in Parallel on CUDA. In *IEEE International Parallel Distributed Processing Symposium, IPDPS '11*, pages 544–555, May 2011.

[BBDBS10] L. Ballan, M. Bertini, A. Del Bimbo, and G. Serra. Video annotation and retrieval using ontologies and rule learning. *IEEE Multimedia*, 17(4):80–88, 2010.

[BBL04] M. Boutell, C. Brown, and J. Luo. Learning spatial configuration models using modified Dirichlet priors. In *Workshop on Statistical Relational Learning*, 2004.

[BBM05] A. C. Berg, T. L. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondences. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1*, pages 26–33, Washington, DC, USA, 2005. IEEE Computer Society.

[BBPB10] P. L. M. Bouttefroy, A. Bouzerdoum, S. L. Phung, and A. Beghdadi. On the analysis of background subtraction techniques using gaussian mixture models. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP '10*, pages 4042–4045, 2010.

[BDBP01] M. Bertini, A. Del Bimbo, and P. Pala. Content-based indexing and retrieval of TV news. *Pattern Recogn. Lett.*, 22(5):503–516, April 2001.

[BDRB08] O. Brouard, F. Delannay, V. Ricordel, and D. Barba. Spatio-temporal segmentation and regions tracking of high definition video sequences based on a Markov Random Field model. In *15th IEEE International Conference on Image Processing, ICIP '08*, pages 1552 –1555, October 2008.

[Bes04] J. Bescos. Real-time shot change detection over online MPEG-2 video. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(4):475 – 484, April 2004.

[bet] Betaface. http://betaface.com.

[BETVG08] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-Up Robust Features (SURF). *Comput. Vis. Image Underst.*, 110(3):346–359, June 2008.

[BFGS04] N. Boujemaa, F. Fleuret, V. Gouet, and H. Sahbi. Visual content extraction for automatic semantic annotation of video news. In *IS&T/SPIE Conference on Storage and Retrieval Methods and Applications for Multimedia, part of Electronic Imaging Symposium*, San José, CA, USA, January 2004.

[BFJ$^+$95] M. G. Brown, J. T. Foote, G. J. F. Jones, K. Sparck Jones, and S. J. Young. Automatic Content-based Retrieval of Broadcast News. In *Proceedings of the 3rg ACM international conference on Multimedia*, pages 35–43, San Francisco, November 1995. ACM Press.

[BG09] G. J. Burghouts and J.-M. Geusebroek. Performance evaluation of local colour invariants. *Comput. Vis. Image Underst.*, 113(1):48–62, January 2009.

[BGW11] M. Brown, H. Gang, and S. Winder. Discriminative learning of local image descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):43–57, January 2011.

[BGWL00] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman. Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication (special issue on Speech Annotation and Corpus Tools)*, 33 (1–2), 2000.

[Bha11] S. Bhattacharyya. A brief survey of color image preprocessing and segmentation techniques. *Journal of Pattern Recognition Research*, 6(1):120–129, 2011.

[BHHSW03] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning distance functions using equivalence relations. In Tom Fawcett and Nina Mishra, editors, *Machine Learning, Proceedings of the Twentieth International Conference (ICML), August 21-24, 2003, Washington, DC, USA",*, pages 11–18. AAAI Press, 2003.

[BHK97] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(7):711–720, July 1997.

[BJ07] R. Bekkerman and J. Jeon. Multi-modal clustering for multimedia collections. In *CVPR*, 2007.

[BKK02] N. Babaguchi, Y. Kawai, and T. Kitahashi. Event based indexing of broadcasted sports video by intermodal collaboration. *IEEE Transactions on Multimedia*, 4(1):68–75, March 2002.

[BL09] L. Begeja and Z. Liu. Searching and Browsing Video in Face Space. In *11th IEEE International Symposium on Multimedia, ISM '09*, pages 336 –341, December 2009.

[BLB05] M. Boutell, J. Luo, and C. Brown. A generalized temporal context model for classifying image collections. *Multimedia Systems*, 11:82–92, 2005. 10.1007/s00530-005-0202-7.

[BLSB04] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004.

[BM06] C. H. Brooks and N. Montanez. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 625–632, New York, NY, USA, 2006. ACM.

[BMP02] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(4):509–522, April 2002.

[BN03] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6):1373–1396, June 2003.

[BP98] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Computer Networks and ISDN Systems*, pages 107–117. Elsevier Science Publishers B. V., 1998.

[BP02] E. Bruno and D. Pellerin. Robust motion estimation using spatial Gabor-like filters. *Signal Process.*, 82(2):297–309, 2002.

[BR08] R. Bekkerman and H. Raghavan. Interactive clustering of text collections according to a user-specified criterion. 2008.

[Bra00] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[Bre96] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996. 10.1007/BF00058655.

[BRG07] S. Banerjee, K. Ramanathan, and A. Gupta. Clustering short texts using wikipedia. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 787–788, New York, NY, USA, 2007. ACM.

[BRS04] R. V. Babu, K. R. Ramakrishnan, and S. H. Srinivasan. Video object segmentation: a compressed domain approach. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(4):462 – 474, April 2004.

[BS07] X. Bai and G. Sapiro. A geodesic framework for fast interactive image and video segmentation and matting. In *11th IEEE International Conference on Computer Vision, ICCV '07*, pages 1–8, October 2007.

[BSK04] D. Buzan, S. Sclaroff, and G. Kollios. Extraction and clustering of motion trajectories in video. In *Proceedings of the 17th International Conference on Pattern Recognition, ICPR '04*, volume 2, pages 521 – 524, August 2004.

[BSMK10] D. Baum, D. Schneider, T. Mertens, and J. Köhler. Constrained subword units for speaker recognition. In *Odyssey 2010: The Speaker and Language Recognition Workshop*, Brno, Czech Republic, 2010.

[BYL$^+$11] L. Bao, S.-I Yu, Z.-Z. Lan, A. Overwijk, Q. Jin, B. Langner, M.I Garbus, S. Burger, F. Metze, and A. Hauptmann. Informedia @ trecvid2011. In *TRECVID*, 2011.

[BZM08] A. Bosch, A. Zisserman, and X. Munoz. Image Classification using ROIs and Multiple Kernel Learning. *IJCV 2008.*, 2008.

[BZMn08] A. Bosch, A. Zisserman, and X. Muñoz. Scene classification using a hybrid generative/discriminative approach. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(4):712–727, April 2008.

[Can86] J Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(6):679–698, June 1986.

[Car] B. Carpenter. Lingpipe. http:/www.alias-i.com/.

[Car09] A. Carpenter. cuSVM: a CUDA implementation of SVM. 2009.

[Cat01] A. Catalin. A review on neural network-based image segmentation techniques. 2001.

[CB08a] S. A. Chatzichristofis and Y. S. Boutalis. CEDD: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval. In *Proceedings of the 6th international conference on Computer vision systems*, ICVS'08, pages 312–322, Berlin, Heidelberg, 2008. Springer-Verlag.

[CB08b] S.A. Chatzichristofis and Y.S. Boutalis. FCTH: Fuzzy color and texture histogram - a low level feature for accurate image retrieval. In *9th International Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS '08.*, pages 191–196, May 2008.

[CBL09] S.A. Chatzichristofis, Y.S. Boutalis, and M. Lux. Selection of the proper compact composite descriptor for improving content based image retrieval. In *Signal Processing, Pattern Recognition and Applications*. ACTA Press, 2009.

[CC07] M.-C. Chang and Y.-J. Cheng. Motion detection by using entropy image and adaptive state-labeling technique. In *IEEE International Symposium on Circuits and Systems, ISCAS '07*, pages 3667–3670, May 2007.

[CCL05] H.-T. Chen, H.-W. Chang, and T.-L. Liu. Local discriminant embedding and its variants. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, pages 846–853, Washington, DC, USA, 2005. IEEE Computer Society.

[CCNH07] P.-A. Chirita, S. Costache, W. Nejdl, and S. Handschuh. P-tag: large scale automatic generation of personalized annotation tags for the web. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 845–854, New York, NY, USA, 2007. ACM.

[CCSW06] M. Chen, S.-C. Chen, M.-L. Shyu, and K. Wickramaratna. Semantic event detection via multimodal data mining. *IEEE Signal Processing Magazine*, 23(2):38–46, March 2006.

[CDF$^+$04] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[CEJ$^+$07] S.-F. Chang, D. Ellis, W. Jiang, K. Lee, A. Yanagawa, A. C. Loui, and J. Luo. Large-scale multimodal semantic concept detection for consumer video. In *Proceedings of the 2007 International Workshop on Multimedia Information Retrieval*, MIR '07, pages 255–264, New York, NY, USA, 2007. ACM.

[CFC03] T.-S. Chua, H.M. Feng, and A. Chandrashekhara. An unified framework for shot boundary detection via active learning. In *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '03*, volume 2, pages 845–848, April 2003.

[CGP$^+$00] P. Chiu, A. Girgensohn, W. Polak, E. Rieffel, and L. Wilcox. A genetic algorithm for video segmentation and summarization. In *IEEE International Conference on Multimedia and Expo, ICME 2000*, volume 3, pages 1329 –1332, 2000.

[CGPP03] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati. Detecting moving objects, ghosts, and shadows in video streams. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(10):1337–1342, October 2003.

[CH98] C.-H. Chang and C.-C. Hsu. Integrating query expansion and conceptual relevance feedback for personalized web information retrieval. *Computer Networks and ISDN Systems*, 30(17):621 – 623, 1998.

[CHH07] D. Cai, X. He, and J. Han. Efficient Kernel Discriminant Analysis via Spectral Regression. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, ICDM '07, pages 427–432, Washington, DC, USA, 2007. IEEE Computer Society.

[CHS04] P. Cimiano, S. Handschuh, and S. Staab. Towards the self-annotating web. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 462–471, New York, NY, USA, 2004. ACM.

[CI01] J. Calic and E. Izquierdo. Towards real-time shot detection in the MPEG-compressed domain. In *Proceedings of WAIMIS*, pages 1390–1399, 2001.

[CJ03] G. Carneiro and A.D. Jepson. Multi-scale phase-based local features. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 736–743, June 2003.

[CJ07] G. Carneiro and A.D. Jepson. Flexible spatial configuration of local image features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2089–2104, December 2007.

[CK01] A. Clare and R. D. King. Knowledge discovery in multi-label phenotype data. In *Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery*, PKDD '01, pages 42–53, London, UK, 2001. Springer-Verlag.

[CK02] M.-S. Choi and W.-Y. Kim. A novel two stage template matching method for rotation and illumination invariance. *Pattern Recognition*, 35(1):119–129, January 2002.

[CKL09] V. Chasanis, A. Kalogeratos, and A. Likas. Movie segmentation into scenes and chapters using locally weighted bag of visual words. In *Proceedings of the 2009 ACM International Conference on Image and Video Retrieval*, CIVR '09, pages 35:1–35:7, New York, NY, USA, 2009. ACM.

[CKP03] Z. Cernekova, C. Kotropoulos, and I. Pitas. Video shot segmentation using singular value decomposition. In *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '03*, volume 3, pages 181–184, April 2003.

[CKPT92] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey. Scatter/gather: a cluster-based approach to browsing large document collections. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '92, pages 318–329, New York, NY, USA, 1992. ACM.

[CLHA08] Y. Chang, D.-J. Lee, Y. Hong, and J. Archibald. Unsupervised video shot segmentation using global color and texture information. In *Proceedings of the 4th International Symposium on Advances in Visual Computing*, ISVC '08, pages 460–467, Berlin, Heidelberg, 2008. Springer-Verlag.

[CLL$^+$06] J. Cao, Y. Lan, J. Li, Q. Li, X. Li, F. Lin, X. Liu, L. Luo, W. Peng, D. Wang, H. Wang, Z. Wang, Z. Xiang, J. Yuan, B. Zhang, J. Zhang, L. Zhang, X. Zhang, and W. Zheng. Intelligent multimedia group of tsinghua university at TRECVID 2006l. In *Proc. of TRECVID*, 2006.

[CLML08] L.-H. Chen, Y.-C. Lai, and H.-Y. Mark Liao. Movie scene segmentation using background information. *Pattern Recogition.*, 41(3):1056–1065, March 2008.

[CLP98] C. W. Chen, J. Luo, and K. J. Parker. Image segmentation via adaptive k-mean clustering and knowledge-based morphological operations with biomedical applications. *IEEE Transactions on Image Processing*, 7(12):1673–1683, 1998.

[CLR07] M. Cooper, T. Liu, and E. Rieffel. Video segmentation via temporal pattern classification. *IEEE Transactions on Multimedia*, 9(3):610–618, April 2007.

[CLS05] P. Cimiano, G. Ladwig, and S. Staab. Gimme' the context: context-driven automatic semantic annotation with c-pankow. In *Proceedings of the 14th international conference on World Wide Web*, WWW '05, pages 332–341, New York, NY, USA, 2005. ACM.

[CLZY08] H. Cheng, Z. Liu, N. Zheng, and J. Yang. A deformable local image descriptor. In *CVPR '08*, 2008.

[CM02] D. Comaniciu and P. Meer. Mean Shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analalysis and Machine Intelligence*, 24(5):603–619, May 2002.

[CMBT02] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. Gate: an architecture for development of robust hlt applications. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 168–175, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[CMM11] H. Cai, K. Mikolajczyk, and J. Matas. Learning linear discriminant projections for dimensionality reduction of image descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(2):338–352, 2011.

[CMPP08] A. Chianese, V. Moscato, A. Penta, and A. Picariello. Scene detection using visual and audio attention. In *Proceedings of the 2008 Ambi-Sys workshop on Ambient media delivery and interactive television*, AMDIT '08, pages 1–7, ICST, Brussels, Belgium, Belgium, 2008. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[CNP06] C. Cotsaces, N. Nikolaidis, and I. Pitas. Video shot detection and condensed representation. A review. *Signal Processing Magazine, IEEE*, 23(2):28–37, March 2006.

[CNZ$^+$07] T.-S. Chua, S.-Y. Neo, Y. Zheng, H.-K. Goh, X. Zhang, X. Zhang, S. Tang, Y.-D. Zhang, J.-T. Li, J. Cao, H.-B. Luan, Q.-Y. He, X. Zhang, and X. Zhang. TRECVID 2007 search tasks by NUS-ICT. In *TRECVID'07*, 2007.

[Col03] R. T. Collins. Mean-Shift blob tracking through scale space. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR '03, 16-22 June 2003, Madison, WI, USA*, pages 234–240. IEEE Computer Society, 2003.

[CORW09] C. Carpineto, S. Osiński, G. Romano, and D. Weiss. A survey of web clustering engines. *ACM Comput. Surv.*, 41(3):17:1–17:38, 2009.

[CR03] H. Chui and A. Rangarajan. A new point matching algorithm for non-rigid registration. *Comput. Vis. Image Underst.*, 89(2-3):114–141, 2003.

[CRH03] E. Chi, A. Rosien, and J. Heer. Lumberjack: Intelligent discovery and analysis of web user traffic composition. In Osmar Zaiane, Jaideep Srivastava, Myra Spiliopoulou, and Brij Masand, editors, *WEBKDD 2002 - Mining Web Data for Discovering Usage Patterns and Profiles*, volume 2703 of *Lecture Notes in Computer Science*, pages 1–16. Springer Berlin / Heidelberg, 2003. 10.1007/978-3-540-39663-5_1.

[CRM03] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(5):564–575, May 2003.

[CRzT03] L. Chen, S. J. Rizvi, M. Tamer zsu, and M. Tamer. Incorporating audio cues into dialog and action scene extraction. In *Proceedings of SPIE Conference on Storage and Retrieval for Media Databases*, pages 252–264, 2003.

[CSK08] B. Catanzaro, N. Sundaram, and K. Keutzer. Fast support vector machine training and classification on graphics processors. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pages 104–111, New York, NY, USA, 2008. ACM.

[CSP01] S.-F. Chang, T. Sikora, and A. Purl. Overview of the MPEG-7 standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):688 –695, June 2001.

[CSS06] T. Cour, P. Srinivasan, and J. Shi. Balanced graph matching. In *NIPS '06*, 2006.

[CTC$^+$12] V. Chandrasekhar, G. Takacs, D. M. Chen, S. S. Tsai, Y. Reznik, R. Grzeszczuk, and B. Girod. Compressed histogram of gradients: A low-bitrate descriptor. *Int. J. Comput. Vision*, 96(3):384–399, February 2012.

[CTKO03] Y. Cao, W. Tavanapong, K. Kim, and J.H. Oh. Audio-assisted scene segmentation for story browsing. In *Proceedings of the 2nd international conference on Image and video retrieval*, CIVR'03, pages 446–455, Berlin, Heidelberg, 2003. Springer-Verlag.

[CVG08] N. Cornelis and L. Van Gool. Fast scale invariant feature detection and matching on programmable graphics hardware. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPRW '08*, pages 1–8, June 2008.

[CZQ01] H. Chen, Y. Zhan, and F. Qi. Rapid object tracking on compressed video. In *Proceedings of the Second IEEE Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing*, PCM '01, pages 1066–1071, London, UK, 2001. Springer-Verlag.

[dAK11] S. A. de Araújo and H. Y. Kim. Ciratefi: An RST-invariant template matching with extension to color images. *Integr. Comput.-Aided Eng.*, 18(1):75–90, 2011.

[DB08] A. Delong and Y. Boykov. A scalable graph-cut algorithm for N-D grids. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR '08*, pages 1–8, June 2008.

[DBKP11] O. Duchenne, F. Bach, I.-S. Kweon, and J. Ponce. A tensor-based algorithm for high-order graph matching. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(12):2383–2395, 2011.

[dBRHDA97] M. de Buenaga Rodríguez, J. M. Gómez Hidalgo, and B. Díaz-Agudo. Using wordnet to complement training information in text categorization. 1997.

[DDF$^+$90] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, 41(6):391–407, 1990.

[DHC07] M. A. Dewan, M. J. Hossain, and O. Chae. A block based moving object detection utilizing the distribution of noise. In *Proceedings of the 1st KES International Symposium on Agent and Multi-Agent Systems: Technologies and Applications*, KES-AMSTA '07, pages 645–654, Berlin, Heidelberg, 2007. Springer-Verlag.

[Dhi01] I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the 7th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '01, pages 269–274, New York, NY, USA, 2001. ACM.

[DLJ$^+$10] K. Daason, H. Lejsek, A. Jhansson, B. Jnsson, and L. Amsaleg. GPU acceleration of Eff2 descriptors using CUDA. In *ACM Multimedia*, pages 1167–1170. ACM, 2010.

[DT05] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR '05*, volume 1, pages 886–893, June 2005.

[DTXM09] L. Duan, I. W.-H. Tsang, D. Xu, and S. J. Maybank. Domain transfer SVM for video concept detection. In *CVPR '09*, pages 1375–1381. IEEE, 2009.

[DVZP04] C. Doulaverakis, S. Vagionitis, Michalis E. Zervakis, and E. G. M. Petrakis. Adaptive methods for motion characterization and segmentation of MPEG compressed frame sequences. In *ICIAR '04*, pages 310–317, 2004.

[DXZF11] P. Dong, Y. Xia, L. Zhuo, and D. Feng. Real-time moving object segmentation and tracking for h.264/avc surveillance videos. In *18th IEEE International Conference on Image Processing, ICIP '11*, pages 2309–2312, September 2011.

[EFP05] P.A. Estevez, R.J. Flores, and C.A. Perez. Color image segmentation using fuzzy min-max neural networks. In *Proceedings of the 2005 IEEE International Joint Conference on Neural Networks, IJCNN '05*, volume 5, pages 3052–3057, July-August 2005.

[Eft95] E. N. Efthimiadis. User choices: A new yardstick for the evaluation of ranking algorithms for interactive query expansion. *Information Processing & Management*, 31(4):605 – 620, 1995.

[EPdRH02] M. Egmont-Petersen, D. de Ridder, and H. Handels. Image processing with neural networks - A review. *Pattern Recognition*, pages 2279–2301, 2002.

[EXCS06] S. Ebadollahi, L. Xie, S.-F. Chang, and J.R. Smith. Visual event detection using multi-dimensional concept dynamics. In *IEEE International Conference on Multimedia and Expo*, pages 881–884, July 2006.

[fac] face.com. http://face.com.

[FAT09] K. Frantzi, S. Ananiadou, and J. Tsujii. The C-value/NC-value method of automatic recognition for multi-word terms. In *Research and Advanced Technology for Digital Libraries*, pages 520–520, 2009.

[FB81] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981.

[FBOD08] M. A. Fouad, F. M. Bayoumi, H. M. Onsi, and M. G. Darwish. Shot transition detection with minimal decoding of MPEG video streams, 2008.

[FGL08] J. Fan, Y. Gao, and H. Luo. Integrating concept ontology and multitask learning to achieve more effective classifier training for multilevel image annotation. *Trans. Img. Proc.*, 17(3):407–426, March 2008.

[FGLJ08] J. Fan, Y. Gao, H. Luo, and R. Jain. Mining multilevel image semantics via hierarchical classification. *Trans. Multi.*, 10(2):167–187, February 2008.

[FH04] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, September 2004.

[FHLMB08] J. Fürnkranz, E. Hüllermeier, E. Loza Mencía, and K. Brinker. Multilabel classification via calibrated label ranking. *Mach. Learn.*, 73(2):133–153, November 2008.

[FIW08] S. Fukui, Y. Iwahori, and R.J. Woodham. GPU based extraction of moving objects without shadows under intensity changes. In *IEEE Congress on Evolutionary Computation, CEC '08 (IEEE World Congress on Computational Intelligence)*, pages 4165–4172, June 2008.

[FJKD06] L. Florack, B. Janssen, F. Kanters, and R. Duits. Towards a new paradigm for motion extraction. In *Proceedings of the 3rd International Conference on Image Analysis and Recognition - Volume 1*, ICIAR'06, pages 743–754, Berlin, Heidelberg, 2006. Springer-Verlag.

[FM05] J. Fung and S. Mann. OpenVIDIA: parallel GPU computer vision. In *Proceedings of the 13th annual ACM international conference on Multimedia*, MULTIMEDIA '05, pages 849–852, New York, NY, USA, 2005. ACM.

[For07] P.-E. Forssn. Maximally stable colour regions for recognition and matching. In *CVPR '07*. IEEE Computer Society, 2007.

[FPR08] M. Fradet, P. Perez, and P. Robert. Time-sequential extraction of motion layers. In *15th IEEE International Conference on Image Processing, ICIP '08*, pages 3224–3227, October 2008.

[FRKZ10] Z. Fu, A. Robles-Kelly, and J. Zhou. Mixing Linear SVMs for Nonlinear Classification. *Neural Networks, IEEE Transactions on*, 21(12):1963 –1975, December 2010.

[FS06] J. Flusser and T. Suk. Rotation moment invariants for recognition of symmetric objects. *IEEE Transactions on Image Processing*, 15(12):3784 –3790, December 2006.

[FXZ11] W. Feng, H. Xiang, and Y. Zhu. An improved graph-based image segmentation algorithm and its GPU acceleration. In *Workshop on Digital Media and Digital Content Management, DMDCM '11.*, pages 237–241, May 2011.

[GA07] D. Gokalp and S. Aksoy. Scene classification using bag-of-regions representations. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR '07*, pages 1–8, June 2007.

[GAV00] J.S. Garofolo, C.G.P. Auzanne, and E.M Voorhees. The TREC SDR Track: A Success Story. In *Eighth Text Retrieval Conference*, pages 107–129, Washington, 2000.

[GCAL03] S. J. F. Guimaraes, M. Couprie, A. de Albuquerque Araújo, and N. J. Leite. Video segmentation based on 2D image analysis. *Pattern Recogn. Lett.*, 24(7):947–957, April 2003.

[GD05] K. Grauman and T. Darrell. The pyramid match kernel: discriminative classification with sets of image features. In *10th IEEE International Conference on Computer Vision, ICCV '05*, volume 2, pages 1458–1465, October 2005.

[GGVS08] J. C. Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. Smeulders. Kernel codebooks for scene categorization. In *Proceedings of the 10th European Conference on Computer Vision: Part III*, ECCV '08, pages 696–709, Berlin, Heidelberg, 2008. Springer-Verlag.

[GHJ05] X. Gao, B. Han, and H. Ji. A shot boundary detection method for news video based on rough sets and fuzzy clustering. In *Proceedings of the 2nd international conference on Image Analysis and Recognition*, ICIAR'05, pages 231–238, Berlin, Heidelberg, 2005. Springer-Verlag.

[GHT11] S. Gauglitz, T. Höllerer, and M. Turk. Evaluation of interest point detectors and feature descriptors for visual tracking. *Int. J. Comput. Vision*, 94(3):335–360, September 2011.

[GKHE10] M. Grundmann, V. Kwatra, Mei Han, and I. Essa. Efficient hierarchical graph-based video segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR '10*, pages 2141–2148, June 2010.

[GLCC$^+$11] S. Gorthi, A. Le Carvennec, H. Copponnex, X. Bresson, and J.-P. Thiran. GPU-accelerated Convex Multi-phase Image Segmentation. Technical report, 2011.

[GLGLBG09] J. Gómez-Luna, J. M. González-Linares, J. I. Benavides, and N. Guil. Parallelization of a video segmentation algorithm on CUDA-enabled Graphics Processing Units. In *Proceedings of the 15th International Euro-Par Conference on Parallel Processing*, Euro-Par '09, pages 924–935, Berlin, Heidelberg, 2009. Springer-Verlag.

[GMC96] S. Gong, S. McKenna, and J.J. Collins. An investigation into face pose distributions. In *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition, 1996*, pages 265–270, October 1996.

[GMK11a] N. Gkalelis, V. Mezaris, and I. Kompatsiaris. High-level event detection in video exploiting discriminant concepts. In *9th International Workshop on Content-Based Multimedia Indexing, CBMI '11*, pages 85–90, June 2011.

[GMK11b] N. Gkalelis, V. Mezaris, and I. Kompatsiaris. Mixture subclass discriminant analysis. *IEEE Signal Processing Letters*, 18(5):319 –322, May 2011.

[GMKS12a] N. Gkalelis, V. Mezaris, I. Kompatsiaris, and T. Stathaki. Linear subclass support vector machines. *IEEE Signal Processing Letters*, 19(9):575 –578, September 2012.

[GMKS12b] N. Gkalelis, V. Mezaris, I. Kompatsiaris, and T. Stathaki. Mixture subclass discriminant analysis link to restricted gaussian model and other generalizations. *IEEE Transactions on Neural Networks and Learning Systems*, 2012. accepted for publication.

[GN08] P. Geetha and V. Narayanan. A survey of content-based video retrieval, 2008.

[GNPS03] F. Giannotti, M. Nanni, D. Pedreschi, and F. Samaritani. Webcat: Automatic categorization of web search results. In *SEBD*, pages 507–518, 2003.

[GPLS02] D. Gatica-Perez, A. Loui, and M.-T. Sun. Finding structure in consumer videos by probabilistic hierarchical clustering. Idiap-RR Idiap-RR-22-2002, IDIAP, 0 2002. IEEE Transactions on Circuits and Systems for Video Technology, accepted for publication.

[Gra05] L. Grady. Multilabel random walker image segmentation using prior models. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR '05*, volume 1, pages 763–770, June 2005.

[GRHS04] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. Neighbourhood components analysis. In *Advances in Neural Information Processing Systems 17*, pages 513–520. MIT Press, 2004.

[GvdBSG01] J.-M. Geusebroek, R. van den Boomgaard, A.W.M. Smeulders, and H. Geerts. Color invariance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(12):1338 –1350, dec 2001.

[GWR99] S. Gauch, J. Wang, and S. M. Rachakonda. A corpus analysis approach for automatic query expansion and its extension to multiple databases. *ACM Trans. Inf. Syst.*, 17(3):250–269, July 1999.

[GXTL08] X. Gao, B. Xiao, D. Tao, and X. Li. Image categorization: Graph edit distance+edge direction histogram. *Pattern Recogn.*, 41(10):3179–3191, October 2008.

[HAB08] C. Hudelot, J. Atif, and I. Bloch. Fuzzy spatial relation ontology for image interpretation. *Fuzzy Sets Syst.*, 159(15):1929–1951, August 2008.

[Han02] A. Hanjalic. Shot-boundary detection: unraveled and resolved? *IEEE Transactions on Circuits and Systems for Video Technology*, 12(2):90–105, February 2002.

[HBW07] G. Hua, M. Brown, and S. Winder. Discriminant embedding for local image descriptors. In *11th IEEE International Conference on Computer Vision, ICCV '07*, pages 1–8, October 2007.

[HC04] W.H.-M Hsu and S.-F. Chang. Generative, discriminative, and ensemble learning on multi-modal perceptual fusion toward news video story segmentation. In *IEEE International Conference on Multimedia and Expo, ICME '04*, volume 2, pages 1091–1094, June 2004.

[HCC01] J. Heer, E. Chi, and E. Chi. Identification of web user traffic composition using multi-modal clustering and information scent. In *Proc. of the Workshop on Web Mining, SIAM Conference on Data Mining*, pages 51–58, 2001.

[HCC+06] A. G. Hauptmann, M.-Y. Chen, M. Christel, D. Das, W.-H. Lin, R. Yan, J. Yang, G. Backfried, and X. Wu. Multi-lingual broadcast news retrieval. In *Proc. of TRECVID*, 2006.

[HCX08] J. He, S.-F. Chang, and L. Xie. Fast kernel learning for spatial pyramid matching. In *CVPR'08*, 2008.

[HDS04] G.D. Hager, M. Dewan, and C.V. Stewart. Multiple kernel tracking with SSD. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR '04*, volume 1, pages 790–797, June-July 2004.

[HE00] M. Hopf and T. Ertl. Hardware accelerated wavelet transformations. In *Proceedings of EG/IEEE TCVG Symposium on Visualization VisSym 00*, pages 93–103, 2000.

[HFC+08] J. Hu, L. Fang, Y. Cao, H.-J. Zeng, H. Li, Q. Yang, and Z. Chen. Enhancing text clustering by leveraging wikipedia semantics. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 179–186, New York, NY, USA, 2008. ACM.

[HFCB08] E. Hüllermeier, J. Fürnkranz, W. Cheng, and K. Brinker. Label ranking by learning pairwise preferences. *Artif. Intell.*, 172(16-17):1897–1916, November 2008.

[HFM+07] S. Heymann, B. Frhlich, Fakultt Medien, K. Mller, and T. Wiegand. SIFT implementation and optimization for general-purpose GPU. In *WSCG 07*, 2007.

[HG09] B.-J. Hsu and J. Glass. Language model parameter estimation using user transcriptions. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP '09, pages 4805–4808, Washington, DC, USA, 2009. IEEE Computer Society.

[HHH+10] M. Hill, G. Hua, B. Huang, M. Merler, A. Natsev, J. R. Smith, L. Xie, H. Ouyang, and M. Zhou. IBM research TRECVID-2010 video copy detection and multimedia event detection system, 2010.

[HL01] E. Hjelmas and B. K. Low. Face Detection: A Survey. *Computer Vision and Image Understanding*, 83(3):236 – 274, 2001.

[HLSB04] M. Hadwiger, C. Langer, H. Scharsach, and K. Bhler. State of the art report 2004 on GPU-based segmentation. Technical report, 2004.

[HLZ04] X.-S. Hua, L. Lu, and H.-J. Zhang. Optimization-based automated home video editing system. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(5):572–583, May 2004.

[HN07] A. Haubold and M. Naphade. Classification of video events using 4-dimensional time-compressed motion features. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, CIVR '07, pages 178–185, New York, NY, USA, 2007. ACM.

[HPP$^+$08] J. Huang, S. P. Ponce, S. I. Park, C. Yong, and F. Quek. GPU-accelerated computation for robust motion tracking using the CUDA framework. In *5th International Conference on Visual Information Engineering, VIE '08*, pages 437–442, August 2008.

[HS88] C. Harris and M. Stephens. A Combined Corner and Edge Detection. In *Proceedings of The Fourth Alvey Vision Conference*, pages 147–151, 1988.

[HS05] K. Hariharakrishnan and D. Schonfeld. Fast object tracking using adaptive block matching. *IEEE Transactions on Multimedia*, 7(5):853– 859, October 2005.

[HSS03] A. Hotho, S. Staab, and G. Stumme. Wordnet improves text document clustering. In *Proc. of the SIGIR 2003 Semantic Web Workshop*, pages 541–544, 2003.

[HTL10] M. J. Huiskes, B. Thomee, and M. S. Lew. New trends and ideas in visual concept detection: the mir flickr retrieval evaluation initiative. In *Proceedings of the 2010 International Conference on Multimedia Information Retrieval*, MIR '10, pages 527–536, New York, NY, USA, 2010. ACM.

[Hu62] M.-K. Hu. Visual pattern recognition by moment invariants. *Transactions on Information Theory, IRE*, 8(2):179–187, February 1962.

[Hul03] A. Hulth. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03, pages 216–223, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[HWO07] M.A.H. Huijbregts, C. Wooters, and R.J.F. Ordelman. Filtering the unknown: Speech activity detection in heterogeneous video collections. In *Proceedings of Interspeech 2007*, page 4, Antwerp, 2007. International Speech Communication Association. ISSN=1990-9772.

[HWS08] P. Huang, Y. Wang, and M. Shao. A New Method for Multi-view Face Clustering in Video Sequence. In *IEEE International Conference on Data Mining Workshops, ICDMW '08*, pages 869 –873, December 2008.

[HZL$^+$09] X. Hu, X. Zhang, C. Lu, E. K. Park, and X. Zhou. Exploiting wikipedia as external knowledge for document clustering. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 389–396, New York, NY, USA, 2009. ACM.

[Ich11] N. Ichimura. Extracting multi-size local descriptors by GPU computing. In *IEEE International Conference on Multimedia and Expo, ICME '11*, pages 1–6, July 2011.

[IKW$^+$11] N. Inoue, Y. Kamishima, T. Wada, K. Shinoda, S. Sato, and N. Shimura. Tokyotech+canon at trecvid 2011. In *Proc. TRECVID Workshop 2011*, 2011.

[IMER01] U. Iurgel, R. Meermeier, S. Eickeler, and G. Rigoll. New approaches to audio-visual segmentation of tv news for automatic topic retrieval. In *Proceedings of the 2001 IEEE International Conference on the Acoustics, Speech, and Signal Processing*, volume 3 of *ICASSP '01*, pages 1397–1400, Washington, DC, USA, 2001. IEEE Computer Society.

[ITP05] P. Iliev, P. Tzvetkov, and G. Petrov. Motion detection using 3D image histograms sequences analysis. In *Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS '05, IEEE*, pages 596–601, September 2005.

[IXM02] E. Izquierdo, J. Xia, and R. Mech. A generic video analysis and segmentation system. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '02*, volume 4, pages 3592–3595, May 2002.

[JA09] R. Jafri and H. R. Arabnia. A Survey of Face Recognition Techniques. *JIPS*, 5(2):41–68, 2009.

[JaDSP10] H. Je andgou, M. Douze, C. Schmid, and P. Peandrez. Aggregating local descriptors into a compact image representation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR '10*, pages 3304–3311, June 2010.

[JBP10] A. Joulin, F. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR '10*, pages 1943–1950, June 2010.

[JCL07] W. Jiang, S.-F. Chang, and A.C. Loui. Kernel sharing with joint boosting for multi-class concept detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR '07*, pages 1–8, June 2007.

[JDL07] H. Jiang, M. S. Drew, and Z.-N. Li. Matching by linear programming and successive convexification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(6):959–975, 2007.

[JH99] A. E. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3d scenes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(5):433–449, May 1999.

[JH02] M. L. Jamrozik and M. H. Hayes. A compressed domain video object segmentation system. In *Proceedings of the 2002 International Conference on Image Processing*, volume 1, pages 113–116, 2002.

[JLC11] W. Jiang, A. Loui, and S.-F. Chang. Cross-domain learning for semantic concept detection. In *TV Content Analysis: Techniques and Applications (CRC Press)*. Taylor and Francis, 2011.

[JNY07] Y.-G. Jiang, C.-W. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, CIVR '07, pages 494–501, New York, NY, USA, 2007. ACM.

[JNY$^+$08] L. Jing, M.I K. Ng, X. Yang, J. Z. Huang, and J. Z. Huang. A text clustering system based on k-means type subspace clustering and ontology. 2008.

[JRY08] J. Jiang, X. Rui, and N. Yu. Feature annotation for visual concept detection in ImageCLEF 2008. In *Workshop: Working Notes for the CLEF 2008 Workshop*, 2008.

[JSR04] G. Jing, C. E. Siong, and D. Rajan. Foreground motion detection by difference-based spatial temporal entropy image. In *TENCON 2004 IEEE Region 10 Conference*, volume 1, pages 379–382, November 2004.

[JT05] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *10th IEEE International Conference on Computer Vision, ICCV '05*, volume 1, pages 604–610, October 2005.

[JY09] H. Jiang and S. X. Yu. Linear solution to scale and rotation invariant object matching. In *CVPR '09*, pages 2474–2481. IEEE, 2009.

[JYNH10] Y.-G. Jiang, J. Yang, C.-W. Ngo, and A.G. Hauptmann. Representations of keypoint-based semantic concept detection: A comprehensive study. *IEEE Transactions on Multimedia*, 12(1):42 –53, January 2010.

[JZCL08] W. Jiang, E. Zavesky, S.-F. Chang, and A. Loui. Cross-Domain Learning Methods for High-Level Visual Concept Classification. In *IEEE International Conference on Image Processing*, San Diego, California, U.S.A, October 2008.

[JZN06] Y.-G. Jiang, W.-L. Zhao, and C.-W. Ngo. Exploring semantic concept using local invariant features. In *Proceedings of Asia-Pacific Workshop on Visual Information Processing*, 2006.

[JZY$^+$10] Y.-G. Jiang, X. Zeng, G. Ye, S. Bhattacharya, D. Ellis, M. Shah, and S.-F. Chang. Columbia-UCF TRECVID2010 Multimedia Event Detection: Combining Multiple Modalities, Contextual Concepts, and Temporal Matching. In *NIST TRECVID Workshop*, 2010.

[Kan73] T. Kanade. Picture Processing System by Computer Complex and Recognition of Human Faces. In *doctoral dissertation, Kyoto University*. December 1973.

[KBNM09] C. Kas, M. Brulin, H. Nicolas, and C. Maillet. Compressed domain aided analysis of traffic surveillance videos. In *3rd ACM/IEEE International Conference on Distributed Smart Cameras, ICDSC '09*, pages 1–8, September 2009.

[KC99] M.-C. Kim and K.-S. Choi. A comparison of collocation-based similarity measures in query expansion. *Information Processing & Management*, 35(1):19 – 30, 1999.

[KC01] I. Koprinska and S. Carrato. Temporal video segmentation: A survey captions on MPEG compressed video. *DSignal Processing Image Communication*, pages 477–500, 2001.

[KdA07] H. Y. Kim and S. A. de Araújo. Grayscale template-matching invariant to rotation, scale, translation, brightness and contrast. In *Proceedings of the 2nd Pacific Rim conference on Advances in image and video technology*, PSIVT'07, pages 100–113, Berlin, Heidelberg, 2007. Springer-Verlag.

[KDY$^+$10] J. Kong, M. Dimitrov, Y. Yang, J. Liyanage, L. Cao, J. Staples, M. Mantor, and H. Zhou. Accelerating MATLAB Image Processing Toolbox functions on GPUs. In *Proceedings of the 3rd Workshop on General-Purpose Computation on Graphics Processing Units*, GPGPU '10, pages 75–85, New York, NY, USA, 2010. ACM.

[key] KeyLemon. http://www.keylemon.com.

[KGX$^+$06] H. Koumaras, G. Gardikis, G. Xilouris, E. Pallis, and A. Kourtis. Shot boundary detection without threshold parameters. *J. Electronic Imaging*, pages 1–3, 2006.

[KH05] S. Kumar and M. Hebert. A hierarchical field framework for unified context-based classification. In *Proceedings of the 10th IEEE International Conference on Computer Vision*, volume 2 of *ICCV '05*, pages 1284–1291, Washington, DC, USA, 2005. IEEE Computer Society.

[Kim10] H. Y. Kim. Rotation-discriminating template matching based on fourier coefficients of radial projections with robustness to scaling and partial occlusion. *Pattern Recogn.*, 43(3):859–872, 2010.

[Kip08] M. Kipp. Spatiotemporal coding in anvil. In *Proceedings of the 6th international conference on Language Resources and Evaluation (LREC-08)*, 2008.

[KL10] T. Kinnunen and H. Li. An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, 52:12–40, 2010.

[Kle99] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *JOURNAL OF THE ACM*, 46(5):604–632, 1999.

[KMS08] A. Klser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In Mark Everingham, Chris J. Needham, and Roberto Fraile, editors, *BMVC*. British Machine Vision Association, 2008.

[KN08] C. Käs and H. Nicolas. An approach to trajectory estimation of moving objects in the H.264 compressed domain. In *Proceedings of the 3rd Pacific Rim Symposium on Advances in Image and Video Technology*, PSIVT '09, pages 318–329, Berlin, Heidelberg, 2008. Springer-Verlag.

[KP08] C. Kauffmann and N. Piche. Cellular automaton for ultra-fast watershed transform on GPU. In *19th International Conference on Pattern Recognition, ICPR '08*, pages 1–4, December 2008.

[KS90] M. Kirby and L. Sirovich. Application of the Karhunen-Loeve procedure for the characterization of human faces. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(1):103 –108, January 1990.

[KS00] I. Kompatsiaris and G. M. Strintzis. Spatiotemporal segmentation and tracking of objects for visualization of videoconference image sequences. *Circuits and Systems for Video Technology, IEEE Transactions on*, 10(8):1388–1402, December 2000.

[KS04] Y. Ke and R. Sukthankar. PCA-SIFT: a more distinctive representation for local image descriptors. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR '04*, volume 2, pages 506–513, June-July 2004.

[KS07] P. Kehoe and A.F. Smeaton. Using graphics processor units (GPUs) for automatic video structuring. In *8th International Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS '07*, page 18, June 2007.

[KSC+08] S. S. Keerthi, S. Sundararajan, K.-W. Chang, C.-J. Hsieh, and C.-J. Lin. A sequential dual method for large scale multi-class linear SVMs. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 408–416, New York, NY, USA, 2008. ACM.

[LA08] J. Li and N. M. Allinson. A comprehensive review of current local features for computer vision. *Neurocomput.*, 71(10-12):1771–1787, June 2008.

[LAJ01] A. M. Lam-Adesina and G. J. F. Jones. Applying summarization techniques for term selection in relevance feedback. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 1–9, New York, NY, USA, 2001. ACM.

[lam12] Lambda Labs API. http://lambdal.com/free-face-recognition-api.html, July 2012.

[LC01] W.-N. Lie and R.-L. Chen. Tracking moving objects in MPEG-compressed videos. *IEEE International Conference on Multimedia and Expo*, 0, 2001.

[LC03] D. J. Lawrie and W. B. Croft. Generating hierarchical summaries for web searches. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 457–458, New York, NY, USA, 2003. ACM.

[LCH08a] Y. Li, S. M. Chung, and J. D. Holt. Text document clustering based on frequent word meaning sequences. *Data & Knowledge Engineering*, 64(1):381 – 404, 2008.

[LCH08b] Y. Li, S. M. Chung, and J. D. Holt. Text document clustering based on frequent word meaning sequences. *Data Knowl. Eng.*, 64(1):381–404, January 2008.

[LD08] Y. Luo and R. Duraiswami. Canny edge detection on NVIDIA CUDA. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPRW '08*, pages 1–8, June 2008.

[Lee96] T. S. Lee. Image representation using 2D gabor wavelets. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(10):959–971, October 1996.

[Lev66] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707–710, February 1966.

[LGL01] Y. Li, S. Gong, and H. Liddell. Recognising trajectories of facial identities using kernel discriminant analysis. In *Proceedings of the British Machine Vision Conference*, page 2003, 2001.

[LH01] S. Lee and M. H. Hayes. Scene change detection using adaptive threshold and submacroblock images in compressed seqeunces. *IEEE International Conference on Multimedia and Expo*, 2001.

[LH05] M. Leordeanu and M. Hebert. A spectral technique for correspondence problems using pairwise constraints. In *Proceedings of the 10th IEEE International Conference on Computer Vision*, volume 2 of *ICCV '05*, pages 1482–1489, Washington, DC, USA, 2005. IEEE Computer Society.

[LHGT02] L. Li, W. Huang, I. Y. H. Gu, and Qi Tian. Foreground object detection in changing background based on color co-occurrence statistics. In *Proceedings of the 6th IEEE Workshop on Applications of Computer Vision, WACV 2002*, pages 269 – 274, 2002.

[LHP01] W. Li, J. Han, and J. Pei. Cmar: Accurate and efficient classification based on multiple class-association rules. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, ICDM '01, pages 369–376, Washington, DC, USA, 2001. IEEE Computer Society.

[LHSL98] M.-S. Lee, B.-W. Hwang, S. Sull, and S.-W. Lee. Automatic video parsing using shot boundary detection and camera operation analysis. In *Proceedings of the 4th International Conference on Pattern Recognition*, volume 2, pages 1481 –1483 vol.2, August 1998.

[Lie99] R. Lienhart. Comparison of automatic shot boundary detection algorithms. pages 290–301, 1999.

[Lin94] T. Lindeberg. *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers, Norwell, MA, USA, 1994.

[LKHH10] H. Li, E. Kim, X. Huang, and L. He. Object matching with a locally affine-invariant constraint. In *CVPR '10*, pages 1641–1648. IEEE, 2010.

[LKO02] J. Laaksonen, M. Koskela, and E. Oja. PicSOM-self-organizing image retrieval with MPEG-7 content descriptors. *Trans. Neur. Netw.*, 13(4):841–853, July 2002.

[LL02] W.-K. Li and S.-H. Lai. A motion-aided video shot segmentation algorithm. In *Proceedings of the 3rd IEEE Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing*, PCM '02, pages 336–343, London, UK, 2002. Springer-Verlag.

[LL03] I. Laptev and T. Lindeberg. Space-time interest points. In *Proceedings of the 9th IEEE International Conference on Computer Vision*, volume 1, pages 432 –439, October 2003.

[LL05] S. Li and M.-C. Lee. An improved sliding window method for shot change detection. In *SIP'05*, pages 464–468, 2005.

[LLT04] H. Lu, Z. Li, and Y.-P. Tan. Model-based video scene clustering with noise analysis. In *Proceedings of the 2004 International Symposium on Circuits and Systems, ISCAS '04*, volume 2, pages 105–108, May 2004.

[LLZ07] Z. Liu, Y. Lu, and Z. Zhang. Real-time spatiotemporal segmentation of video objects in the H.264 compressed domain. *J. Vis. Comun. Image Represent.*, 18(3):275–290, June 2007.

[LM01a] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *Int. J. Comput. Vision*, 43(1):29–44, June 2001.

[LM01b] L. Lucchese and S. K. Mitra. Color image segmentation: A state-of-the-art survey, 2001.

[LM02] R. Lienhart and J. Maydt. An extended set of Haar-like features for rapid object detection. In *Proceedings of the 2002 International Conference on Image Processing*, volume 1, pages 900–903, 2002.

[LMSR08] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR '08*, pages 1–8, June 2008.

[Low99] D. G. Lowe. Object recognition from local scale-invariant features. In *The Proceedings of the 7th IEEE International Conference on Computer Vision*, volume 2, pages 1150 –1157, 1999.

[Low04] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, November 2004.

[LP05] F.-F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR '05*, volume 2 of *CVPR '05*, pages 524–531, Washington, DC, USA, 2005. IEEE Computer Society.

[LPC+04] J.-H. Li, Q. Pan, P.-L. Cui, H.-C. Zhang, and Y.-M. Cheng. Image recognition based on invariant moment in the projection space. In *Proceedings of 2004 International Conference on Machine Learning and Cybernetics*, volume 6, pages 3606–3610, August 2004.

[LS03] D. Lelescu and D. Schonfeld. Statistical sequential analysis for real-time video scene change detection on compressed multimedia bitstream. *IEEE Transactions on Multimedia*, 5(1):106–117, March 2003.

[LS09] D.-D. Le and S. Satoh. Efficient concept detection by fusing simple visual features. In *Proceedings of the 2009 ACM symposium on Applied Computing*, SAC '09, pages 1839–1840, New York, NY, USA, 2009. ACM.

[LS10] J. Liu and J. Sun. Parallel graph-cuts by adaptive bottom-up merging. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR '10*, pages 2181–2188, June 2010.

[LSP03] S. Lazebnik, C. Schmid, and J. Ponce. Sparse texture representation using affine-invariant neighborhoods. In *International Conference on Computer Vision & Pattern Recognition*, volume 2, pages 319–324, 2003.

[LSP05] S. Lazebnik, C. Schmid, and J. Ponce. A sparse texture representation using local affine regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1265–1278, August 2005.

[LSP06] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2169–2178, 2006.

[LST00] S. Liapis, E. Sifakis, and G. Tziritas. Color and/or texture segmentation using deterministic relaxation and fast marching algorithms. In *ICPR 2000*, pages 3621–3624, 2000.

[lux] FaceSDK by Luxland. http://www.luxand.com/facesdk/.

[LVS08] X. Liu, O. Veksler, and J. Samarabandu. Graph cut with ordering constraints on labels and its applications. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR '08*, pages 1–8, June 2008.

[LWGZ04] Y. Liu, W. Wang, W. Gao, and W. Zeng. A novel compressed domain shot segmentation algorithm on H.264/AVC. In *International Conference on Image Processing, ICIP '04.*, volume 4, pages 2235 – 2238, October 2004.

[LWT+08] K.-H. Liu, M.-F. Weng, C.-Y. Tseng, Y.-Y. Chuang, and M.-S. Chen. Association and temporal rule mining for post-filtering of semantic concept detection in video. *IEEE Transactions on Multimedia*, 10(2):240 –251, February 2008.

[LWYW07] N. Lu, J. Wang, L. Yang, and H. Wu. Motion detection based on accumulative optical flow and double background filtering. In *Proceedings of the 2007 World Congress on Engineering*, pages 2–4, 2007.

[LX09] P. Li and L. Xiao. Mean Shift parallel tracking on GPU. In *Proceedings of the 4th Iberian Conference on Pattern Recognition and Image Analysis*, IbPRIA '09, pages 120–127, Berlin, Heidelberg, 2009. Springer-Verlag.

[LYHZ08] X. Ling, O. Yuanxin, L. Huan, and X. Zhang. A method for fast shot boundary detection based on SVM. In *Congress on Image and Signal Processing, CISP '08*, volume 2, pages 445–449, May 2008.

[LZ01] R. Lienhart and A. Zaccarin. A system for reliable dissolve detection in videos. In *Proceedings of the International Conference on Image Processing*, volume 3, pages 406–409, 2001.

[LZ07] J. Liao and B. Zhang. A robust clustering algorithm for video shots using haar wavelet transformation. In *Proceedings of SIGMOD2007 Ph.D. Workshop on Innovative Database Research (IDAR2007)*, Beijing, China, June 2007.

[LZL+09] L. Li, X. Zeng, X. Li, W. Hu, and P. Zhu. Video shot segmentation using graph-based dominant-set clustering. In *Proceedings of the 1st International Conference on Internet Multimedia Computing and Service*, ICIMCS '09, pages 166–169, New York, NY, USA, 2009. ACM.

[MAC09] A. Mishra, Y. Aloimonos, and L. F. Cheong. Active segmentation with fixation. In *12th IEEE International Conference on Computer Vision*, pages 468 –475, October 2009.

[MBC03] D. Maynard, K. Bontcheva, and H. Cunningham. Towards a semantic extraction of named entities. 2003.

[MBM08] S. Maji, A.C. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR '08*, pages 1–8, June 2008.

[MBPLM08] F. Manerba, J. Benois-Pineau, R. Leonardi, and B. Mansencal. Multiple moving object detection for fast video content description in compressed domain. *EURASIP Journal of Advanced Signal Processing*, 2008, January 2008.

[MCB+01] G. Medioni, I. Cohen, F. Bremond, S. Hongeng, and R. Nevatia. Event detection and analysis from video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(8):873–889, August 2001.

[MCUP02] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *British Machine Vision Conference*, volume 1, pages 384–393, 2002.

[MD01] M. Mason and Z. Duric. Using histograms to detect and track objects in color video. In *30th Applied Imagery Pattern Recognition Workshop, AIPR 2001*, pages 154–159, October 2001.

[MDG+10] A. Moumtzidou, A. Dimou, N. Gkalelis, S. Vrochidis, V. Mezaris, and I. Kompatsiaris. ITI-CERTH participation to TRECVID 2010. In *TRECVID 2010 Workshop*, 2010.

[MDK+97] A. F. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET Curve in Assessment of Detection Task Performance. In *Proc. Eurospeech*, volume 4, pages 1899–1903, Rhodes, Greece, 1997.

[MDK10] V. Mezaris, A. Dimou, and I. Kompatsiaris. On the use of feature tracks for dynamic concept detection in video. In *17th IEEE International Conference on Image Processing, ICIP '10*, pages 4697–4700, September 2010.

[MDN97] D. Meyer, J. Denzler, and H. Niemann. Model based extraction of articulated objects in image sequences for gait analysis. In *Proceedings of the 1997 International Conference on Image Processing*, volume 3, pages 78–81, October 1997.

[MDY08] C.K. Mohan, N. Dhananjaya, and B. Yegnanarayana. Video shot segmentation using late fusion technique. In *7th International Conference on Machine Learning and Applications, ICMLA '08*, pages 267–270, December 2008.

[ME07] D. Marimon and T. Ebrahimi. Efficient rotation-discriminative template matching. In *Proceedings of the Congress on pattern recognition 12th Iberoamerican conference on Progress in pattern recognition, image analysis and applications*, CIARP'07, pages 221–230, Berlin, Heidelberg, 2007. Springer-Verlag.

[MFGSMV04] J. Martinez-Fernandez, A. Garcia-Serrano, P. Martinez, and J. Villena. Automatic keyword extraction for news finder. In Andreas Nurnberger and Marcin Detyniecki, editors, *Adaptive Multimedia Retrieval*, volume 3094 of *Lecture Notes in Computer Science*, pages 405–427. Springer Berlin / Heidelberg, 2004. 10.1007/978-3-540-25981-7_7.

[MFW09] O. Medelyan, E. Frank, and I. H. Witten. Human-competitive tagging using automatic keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3*, EMNLP '09, pages 1318–1327, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[MHPG$^+$09] S. Marat, T. Ho Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guérin-Dugué. Modelling spatio-temporal saliency to predict gaze direction for short videos. *Int. J. Comput. Vision*, 82(3):231–243, 2009.

[MHS05] E.N. Mortensen, D. Hongli, and L. Shapiro. A SIFT descriptor with global context. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR '05*, volume 1, pages 184 – 190 vol. 1, june 2005.

[MHX$^+$10] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev. Semantic model vectors for complex video event recognition. *Transactions in Multimedia*, 14(1):88–101, 2010.

[MI03] Y. Matsuo and M. Ishizuka. Keyword extraction from a single document using word co-occurrence statistical information, 2003.

[Mil95] G. A. Miller. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995.

[Mis11] Agent based image segmentation method : A review. *International Journal of Computer Technology and Applications*, 2:704–708, 2011.

[MJTG98] P.J. Moreno, C. Joerg, J.-M. Van Thong, and O. Glickman. A Recursive Algorithm for the Forced Alignment of Very Long Audio Segments. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP'98)*, Sydney, Australia, 1998.

[MKBS04] V. Mezaris, I. Kompatsiaris, N.V. Boulgouris, and M.G. Strintzis. Real-time compressed-domain spatiotemporal segmentation and ontologies for video indexing and retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(5):606 – 621, May 2004.

[MKS04a] V. Mezaris, I. Kompatsiaris, and M. G. Strintzis. Still image segmentation tools for object-based multimedia applications. *IJPRAI*, 18(4):701–725, 2004.

[MKS04b] V. Mezaris, I. Kompatsiaris, and M. G. Strintzis. Still image segmentation tools for object-based multimedia applications. *International Journal of Pattern Recognition and Artificial Intelligence*, 18:701–725, 2004.

[MKS04c] V. Mezaris, I. Kompatsiaris, and M. G. Strintzis. Video object segmentation using bayes-based temporal tracking and trajectory-based region merging. *IEEE Trans. Cir. and Sys. for Video Technol.*, 14(6):782–795, June 2004.

[MLG$^+$11] R. Mertens, H. Lei, L. Gottlieb, G. Friedland, and A. Divakaran. Acoustic super models for large scale video event detection. In *Proceedings of the 2011 joint ACM workshop on Modeling and representing events*, J-MRE '11, pages 19–24, New York, NY, USA, 2011. ACM.

[MM07] K. Mikolajczyk and J. Matas. Improving descriptors for fast tree matching by optimal linear projection. In *IEEE 11th International Conference on Computer Vision, ICCV '07*, pages 1–8, October 2007.

[MOH00] T. Matsuyama, T. Ohya, and H. Habe. Background subtraction for non-stationary scenes. In *Asian Conference on Computer Vision*, 2000.

[Moh11] F. M. A. Mohsen. A new Optimization-Based Image Segmentation method By Particle Swarm Optimization. *IJACSA - International Journal of Advanced Computer Science and Applications*, (Special Issue):10–18, 2011.

[MOVY01] B.S. Manjunath, J.-R. Ohm, V.V. Vasudevan, and A. Yamada. Color and texture descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):703 –715, June 2001.

[MP04] A. Mittal and N. Paragios. Motion-based background subtraction using adaptive kernel density estimation. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR '04*, volume 2, pages 302–309, June-July 2004.

[MS98] F. Mokhtarian and R. Suomela. Robust image corner detection through curvature scale space. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(12):1376–1381, 1998.

[MS04] M.E.S. Mendes and L. Sacks. Dynamic knowledge representation for e-learning applications. In *Proceedings of the 7th ACM SIGKDD international conference on Knowledge discovery and data mining*, Enhancing the Power of the Internet, pages 255–278. Springer, 2004.

[MS05] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615 –1630, October 2005.

[MSAK09] P. Mylonas, E. Spyrou, Y. Avrithis, and S. Kollias. Using visual context and region semantics for high-level concept detection. *Trans. Multi.*, 11(2):229–243, February 2009.

[MSDK10] V. Mezaris, P. Sidiropoulos, A. Dimou, and I. Kompatsiaris. On the use of visual soft semantics for video temporal decomposition to scenes. In *IEEE 4th International Conference on Semantic Computing, ICSC '10*, pages 141–148, Septmber 2010.

[MSHvdW07] M. Marszałek, C. Schmid, H. Harzallah, and J. van de Weijer. Learning object representations for visual object class recognition, October 2007. Visual Recognition Challange workshop, in conjunction with ICCV.

[MSV$^+$11] A. Moumtzidou, P. Sidiropoulos, S. Vrochidis, N. Gkalelis, S. Nikolopoulos, V. Mezaris, I. Kompatsiaris, and I. Patras. ITI-CERTH participation to TRECVID 2011. In *TRECVID 2011 Workshop*, Gaithersburg, MD, USA, 12/2011 2011.

[MTJ06] F. Moosmann, B. Triggs, and F. Jurie. Fast discriminative visual codebooks using randomized clustering forests. In Bernhard Schlkopf, John Platt, and Thomas Hoffman, editors, *NIPS*, pages 985–992. MIT Press, 2006.

[MTT04] R. Mihalcea, P. Tarau, and P. Tarau. Textrank: Bringing order into texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 2004.

[Muh04] H. H. Muhammed. Unsupervised fuzzy clustering using weighted incremental neural networks. *International Journal of Neural Systems*, 14:355–371, 2004.

[Mun97] A. Munoz. Compound key word generation from document databases using a hierarchical clustering art model. pages 25–48, 1997.

[MY09] J.-M. Morel and G. Yu. ASIFT: A new framework for fully affine invariant image comparison. *SIAM J. Img. Sci.*, 2(2):438–469, 2009.

[MZDPAR$^+$11] M. Martínez-Zarzuela, F. J. Díaz-Pernas, M. Antón-Rodríguez, F. Perozo-Rondón, and D. González-Ortega. Bio-inspired color image segmentation on the GPU (BioSPCIS). In *Proceedings of the 4th international conference on Interplay between natural and artificial computation: new challenges on bioinspired applications - Volume Part II*, IWINAC'11, pages 353–362, Berlin, Heidelberg, 2011. Springer-Verlag.

[Nat11] P. et. al. Natarajan. BBN VISER TRECVID 2011 multimedia event detection system. In *NIST TRECVID Workshop*, Gaithersburg, MD, December 2011.

[neu] Neurotechnology. http://www.neurotechnology.com.

[Ng00] K. Ng. *Subword-based Approaches for Spoken Document Retrieval*. PhD thesis, Massachusetts Institute of Technology, February 2000.

[Ngo03] C.-W. Ngo. A robust dissolve detector by support vector machine. In *Proceedings of the 11th ACM international conference on Multimedia*, MULTIMEDIA '03, pages 283–286, New York, NY, USA, 2003. ACM.

[NJT06] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *European Conference on Computer Vision*, 2006.

[NL09] C. Niu and Y. Liu. Moving object segmentation in the H.264 compressed domain. In *ACCV '09*, pages 645–654, 2009.

[NMZ05] C.-W. Ngo, Y.-F. Ma, and H.-J. Zhang. Video summarization and scene detection by graph modeling. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(2):296–305, February 2005.

[NNS04] A. Natsev, M. R. Naphade, and J. R. Smith. Semantic representation: search and mining of multimedia content. In *Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 641–646, New York, NY, USA, 2004. ACM.

[NS03] M. R. Naphade and J. R. Smith. A hybrid framework for detecting the semantics of concepts and context. In *Proceedings of the 2nd international conference on Image and video retrieval*, CIVR'03, pages 196–205, Berlin, Heidelberg, 2003. Springer-Verlag.

[NS06] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pages 2161–2168, Washington, DC, USA, 2006. IEEE Computer Society.

[NSK00] H. Noda, M.N. Shirazi, and E. Kawaguchi. Textured image segmentation using mrf in wavelet domain. In *Proceedings of the 2000 International Conference on Image Processing*, volume 3, pages 572 –575, 2000.

[NT05] J. Nam and A.H. Tewfik. Detection of gradual transitions in video sequences using b-spline interpolation. *IEEE Transactions on Multimedia*, 7(4):667 – 679, August 2005.

[NYWC11] S.o Niu, J. Yang, S. Wang, and G. Chen. Improvement and parallel implementation of canny edge detection algorithm based on GPU. In *IEEE 9th International Conference on ASIC, ASICON '11.*, pages 641–644, October 2011.

[OGPG03] J.-M. Odobez, D. Gatica-Perez, and M. Guillemot. Spectral structuring of home videos. In *Proceedings of the 2nd international conference on Image and video retrieval*, CIVR'03, pages 310–320, Berlin, Heidelberg, 2003. Springer-Verlag.

[OLG$^+$07] J. D. Owens, D. Luebke, N. Govindaraju, M. Harris, J. Krger, A. E. Lefohn, and T. Purcell. A survey of general-purpose computation on graphics hardware, 2007.

[OPM02] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(7):971–987, July 2002.

[OPZ06] A. Opelt, A. Pinz, and A. Zisserman. Incremental learning of object detectors using a visual shape alphabet. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1*, CVPR '06, pages 3–10, Washington, DC, USA, 2006. IEEE Computer Society.

[OW05] S. Osinski and D. Weiss. A concept-driven algorithm for clustering search results. pages 48–54, 2005.

[OYBK02] H. Okamoto, Y. Yasugi, N. Babaguchi, and T. Kitahashi. Video clustering using spatio-temporal image with fixed length. In *Proceedings of the 2002 IEEE International Conference on Multimedia and Expo, ICME '02*, volume 1, pages 53 – 56, 2002.

[PAS$^+$06] K. Petridis, D. Anastasopoulos, C. Saathoff, N. Timmermann, I. Kompatsiaris, and S. Staab. M-ontomat-annotizer: Image annotation. linking ontologies and multimedia low-level features. *LNCS (LNAI)*, 4251:2006–2010, 2006.

[PC02] S.-C. Pei and Y.-Z. Chou. Effective wipe detection in MPEG compressed video using macro block type information. *IEEE Transactions on Multimedia*, 4(3):309–319, September 2002.

[PD00] N. Paragios and R. Deriche. Geodesic active contours and level sets for the detection and tracking of moving objects. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(3):266–280, March 2000.

[PDBP+09] C. Poppe, S. De Bruyne, T. Paridaens, P. Lambert, and R. Van de Walle. Moving object detection in the H.264/AVC compressed domain for video surveillance applications. *J. Vis. Comun. Image Represent.*, 20(6):428–437, August 2009.

[Pet99] N. Peterfreund. Robust tracking of position and velocity with kalman snakes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(6):564 –569, June 1999.

[Pet08] C. Petersohn. Logical unit and scene detection: a comparative survey. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 6820 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, 2008.

[PGKK05] B. Pytlik, A. Ghoshal, D. Karakos, and S. Khudanpur. Trecvid 2005 experiment at johns hopkins university: Using hidden markov models for video retrieval. In *In Proceedings of NIST TREC Video Retrieval Evaluation*, 2005.

[PKO+04] B. Popov, A. Kiryakov, D. Ognyanoff, D. Manov, and A. Kirilov. KIM a semantic platform for information extraction and retrieval. *Nat. Lang. Eng.*, 10(3-4):375–392, September 2004.

[PL03] S.-M. Park and J. Lee. Object tracking in MPEG compressed video using Mean-Shift algorithm. In *Proceedings of the 2003 Joint Conference of the Fourth International Conference on Information, Communications and Signal Processing*, volume 2, pages 748–752, December 2003.

[Por04] F. Porikli. Real-time video object segmentation for MPEG encoded video sequences, 2004.

[PP03] M. Pavan and M. Pelillo. A new graph-theoretic approach to clustering and segmentation. In *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 145–152, June 2003.

[PPL06] M.-H. Park, R.-H. Park, and S. W. Lee. Shot boundary detection using scale invariant feature matching. volume 6077, page 60771N. SPIE, 2006.

[PR09] V. A. Prisacariu and I. Reid. fastHOG- a real-time GPU implementation of HOG Technical Report No. 2310/09, 2009.

[PSE+11] G.Th. Papadopoulos, C. Saathoff, H.J. Escalante, V. Mezaris, I. Kompatsiaris, and M.G. Strintzis. A comparative study of object-level spatial context techniques for semantic image analysis. *Computer Vision and Image Understanding*, 115, 09/2011 2011.

[PSM+12] S. Papadopoulos, E. Schinas, V. Mezaris, R. Troncy, and I. Kompatsiaris. Social event detection at mediaeval 2012: Challenges, dataset and evaluation. In *Proc. MediaEval 2012 Workshop*, 2012.

[PTG12] P. Prajapati, A. Thakkar, and A. Ganatra. A survey and current research challenges in multi-label classification methods. *International Journal of Soft Computing and Engineering*, 2(1):248–252, 2012.

[QHR+07] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang. Correlative multi-label video annotation. In *Proceedings of the 15th international conference on Multimedia*, MULTIMEDIA '07, pages 17–26, New York, NY, USA, 2007. ACM.

[QMC05] A. Qamra, Y. Meng, and E. Y. Chang. Enhanced perceptual distance functions and indexing for image replica recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(3):379–391, March 2005.

[QO06] P. Quelhas and J.-M. Odobez. Natural scene image modeling using color and texture visterims. In *Proc. of CIVR*, 2006.

[RAC+03] D. A. Reynolds, W. Andrews, J. Campbell, et al. The supersid project: Exploiting high-level information for high-accuracy speaker recognition. In *In Proceedings of the 2003 International Conference on Audio, Speech, and Signal Processing, ICASSP '03*, volume 4, pages 784–787, Hong Kong, 2003.

[RDF+07] D. Rusu, L. Dali, B. Fortuna, M. Grobelnik, and D Mladenic. Triplet extraction from sentence. In *Proceedings of the 10th International Multiconference ”Information Society - IS 2007*, pages 218–222, 2007.

[RDGM10] J. Rabin, J. Delon, Y. Gousseau, and L. Moisan. MAC-RANSAC: a robust algorithm for the recognition of multiple objects. In *Proceedings of the 3DPTV 2010*, page 051, Paris, France, 2010.

[RdSVC09] Javier Ruiz-del Solar, Rodrigo Verschae, and Mauricio Correa. Recognition of faces in unconstrained environments a comparative study. *EURASIP J. Adv. Signal Process*, 2009:11–119, 2009.

[Rea08] J. Read. A Pruned Problem Transformation Method for Multi-label classification. In *Proc. 2008 New Zealand Computer Science Research Student Conference (NZCSRS 2008)*, pages 143–150, 2008.

[RHM99] Y. Rui, T. S. Huang, and S. Mehrotra. Constructing table-of-content for videos. *Multimedia Systems*, 7(5):359–368, September 1999.

[RHPA10] A. Rahman, D. Houzet, D. Pellerin, and L. Agud. GPU implementation of motion estimation for visual saliency. In *Conference on Design and Architectures for Signal and Image Processing, DASIP '10*, pages 222–227, October 2010.

[RK10] B. Rymut and B. Kwolek. GPU-supported object tracking using adaptive appearance models and particle swarm optimization. In *Proceedings of the 2010 international conference on Computer vision and graphics: Part II*, ICCVG'10, pages 227–234, Berlin, Heidelberg, 2010. Springer-Verlag.

[RLD+06] K. Rohlfing, D. Loehr, S. Duncan, A. Brown, A. Franklin, I. Kimbara, J.T. Milde, F. Parrill, T. Rose, T. Schmidt, H. Sloetjes, A. Thies, and S. Wellinghof. Comparison of multimodal annotation tools – workshop report. 7:99–123, 2006.

[RM03] B. Raytchev and H. Murase. Unsupervised face recognition by associative chaining. *Pattern Recognition*, 36(1):245 – 257, 2003.

[RMV07] N. Rasiwasia, P. J. Moreno, and N. Vasconcelos. Bridging the gap: Query by semantic example. *Transactions in Multimedia*, 9(5):923–938, 2007.

[RNKH02] M. Ramesh Naphade, I. V. Kozintsev, and T. S. Huang. Factor graph framework for semantic video indexing. *IEEE Trans. Cir. and Sys. for Video Technol.*, 12(1):40–52, January 2002.

[Roc66] J. J. Rocchio. Document retrieval systems : optimization and evaluation. 1966.

[Ros95] R. Rosenfeld. Optimizing Lexical and N-gram Coverage Via Judicious Use of Linguistic Data. In *Eurospeech-95*, pages 1763–1766, 1995.

[RQD00] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. *Digital Signal Processing*, 10:19–41, 2000.

[RS03] Z. Rasheed and M. Shah. Scene detection in hollywood movies and tv shows. In *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 343–348, June 2003.

[RS05] Z. Rasheed and M. Shah. Detection and representation of scenes in videos. *IEEE Transactions on Multimedia*, 7(6):1097–1105, December 2005.

[RSS04] M. E. S. Mendes Rodrigues, L. Sacks, and L. Sacks. A scalable hierarchical fuzzy clustering algorithm for text mining. 2004.

[RTMF08] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: A database and web-based tool for image annotation. *Int. Journal of Computer Vision*, 77(1–3):157–173, May 2008.

[RUG07] A. Ruiz, M. Ujaldón, and N. Guil. Using graphics hardware for enhancing edge and circle detection. In *Proceedings of the 3rd Iberian conference on Pattern Recognition and Image Analysis, Part II*, IbPRIA '07, pages 234–241, Berlin, Heidelberg, 2007. Springer-Verlag.

[SA06] B. Saux and G. Amato. Image classifiers for scene analysis. In K. Wojciechowski, B. Smolka, H. Palus, R.S. Kozera, W. Skarbek, and L. Noakes, editors, *Computer Vision and Graphics*, volume 32 of *Computational Imaging and Vision*, pages 39–44. Springer Netherlands, 2006.

[SA07] E. Spyrou and Y. Avrithis. High-level concept detection in video using a region thesaurus. In *Proceedings of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, pages 143–153, Amsterdam, The Netherlands, The Netherlands, 2007. IOS Press.

[SAM+12a] D. Stein, E. Apostolidis, V. Mezaris, N. de Abreu Pereira, and J. Müller. Semi-automatic video analysis for linking television to the web. In *Proc. FutureTV Workshop*, pages 1–8, Berlin, Germany, 2012.

[SAM+12b] D. Stein, E. Apostolidis, V. Mezaris, N. de Abreu Pereira, J. Müller, M. Sahuguet, B. Huet, and I. Lašek. Enrichment of news show videos with multimodal semi-automatic analysis. In *NEM-Summit*, Istanbul, Turkey, 2012.

[SAS+12] D. Stein, E. Apostolidis, M. Sahuguet, V. Mezaris, N. de Abreu Pereira, L. B. Baltussen, J. Bloom, J. Müller, I. Lašek, and B. Huet. LinkedTV Scenario Necessities for Video Enrichment with Multimodal Semi-Automatic Analysis. *Special call on Multimedia Tools and Applications: Content Analysis and Indexing for Distributed Multimedia Search & Retrieval in Broadcasting*, pages 1–30, 2012. currently under review.

[SB93] A. M. Steier and R. K. Belew. Exporting phrases: A statistical analysis of topical language. In *Second Symposium on Document Analysis and Information Retrieval*, pages 179–190, 1993.

[SB97] S. M. Smith and J. M. Brady. Susan - a new approach to low level image processing. *Int. J. Comput. Vision*, 23(1):45–78, 1997.

[SB03] J. Snchez and X. Binefa. Shot segmentation using a coupled markov chains representation of video contents. In Francisco Perales, Aurlio Campilho, Nicols de la Blanca, and Alberto Sanfeliu, editors, *Pattern Recognition and Image Analysis*, volume 2652 of *Lecture Notes in Computer Science*, pages 902–909. Springer Berlin / Heidelberg, 2003.

[SCBG+09] C. J. Solana-Cipres, L. R. Benitez, J. M. Garca, L. Jimnez, and G. Fernndez-Escribano. Real-time segmentation of moving objects in H.264 compressed domain with dynamic design of fuzzy sets. In *IFSA/EUSFLAT Conf.'09*, pages 19–24, 2009.

[SDE+09] T. Schmidt, S. Duncan, O.r Ehmer, J. Hoyt, M. Kipp, D. Loehr, M. Magnusson, T. Rose, and H. Sloetjes. An exchange format for multimodal annotations. *Multimodal Corpora, LNAI*, 5509:207–221, 2009.

[SDI06] G. Shakhnarovich, T. Darrell, and P. Indyk. *Nearest-Neighbor Methods in Learning and Vision: Theory and Practice (Neural Information Processing)*. The MIT Press, 2006.

[SFPG06] S. N. Sinha, J.-M. Frahm, M. Pollefeys, and Y. Genc. GPU-based video feature tracking and matching. Technical report, In Workshop on Edge Computing Using New Commodity Architectures, 2006.

[SFPG11] S. N. Sinha, J.-M. Frahm, M. Pollefeys, and Y. Genc. Feature tracking and matching in video using programmable graphics hardware. *Mach. Vision Appl.*, 22(1):207–217, 2011.

[SG10] K. E. A. Sande and T. Gevers. University of amsterdam at the visual concept detection and annotation tasks. In Henning Mller, Paul Clough, Thomas Deselaers, and Barbara Caputo, editors, *ImageCLEF 2010*, volume 32 of *The Information Retrieval Series*, pages 343–358. Springer Berlin Heidelberg, 2010.

[SGH98] S. Sanjay-Gopal and T.J. Hebert. Bayesian pixel classification using spatially variant finite mixtures and the generalized em algorithm. *IEEE Transactions on Image Processing*, 7(7):1014 –1028, July 1998.

[Sha00] M. Sharma. Performance evaluation of image segmentation and texture extraction methods in scene analysis. 2000.

[Sha05] G. Shakhnarovich. *Learning task-specific similarity*. PhD thesis, Cambridge, MA, USA, 2005. AAI0809132.

[SHMN04] M. Sugano, K. Hoashi, K. Matsumoto, and Y. Nakajima. Shot boundary determination on MPEG compressed domain and story segmentation experiments for trecvid 2004, in trec video retrieval evaluation forum. In *Proceedings of the TREC Video Retrieval Evaluation (TRECVID). Washington D.C.: NIST*, pages 109–120, 2004.

[SHS08] P. Sangi, J. Heikkil, and O. Silvn. Extracting motion components from image sequences using particle filters, 2008.

[SI07] C. Siagian and L. Itti. Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(2):300–312, February 2007.

[Sib11] A. Sibiryakov. Fast and high-performance template matching method. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 1417–1424, Washington, DC, USA, 2011. IEEE Computer Society.

[SJ04] M. Schultz and T. Joachims. Learning a distance metric from relative comparisons. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[SJC08] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR '08*, pages 1–8, June 2008.

[SK10] P. Strandmark and F. Kahl. Parallel and distributed graph cuts by dual decomposition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR '10*, pages 2085–2092, June 2010.

[SK11] N. Sundaram and K. Keutzer. Long term video segmentation through pixel level spectral clustering on GPUs. In *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 475–482, November 2011.

[SKK00] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In *In KDD Workshop on Text Mining*, 2000.

[SL10] Dr. V. Sankaranarayanan and S. Lakshmi. A study of edge detection techniques for segmentation computing approaches. *IJCA,Special Issue on CASCT*, (1):35–41, 2010. Published By Foundation of Computer Science.

[SLBM$^+$05] E. Spyrou, H. Le Borgne, T. Mailis, E. Cooke, Y. Avrithis, and N. O'Connor. Fusing MPEG-7 visual descriptors for image classification. In *Proceedings of the 15th international conference on Artificial neural networks: formal models and their applications - Volume Part II*, ICANN'05, pages 847–852, Berlin, Heidelberg, 2005. Springer-Verlag.

[SLL06] S. Song, C. Li, and C. Li. Improved rock for text clustering using asymmetric proximity. In *SOFSEM*, pages 501–510, 2006.

[SLT$^+$05] C.-W. Su, H.-Y.M. Liao, H.-R. Tyan, K.-C. Fan, and L.-H. Chen. A motion-tolerant dissolve detection algorithm. *IEEE Transactions on Multimedia*, 7(6):1106 – 1113, December 2005.

[SLZZ10] C. Sun, J. Li, B. Zhang, and Q. Zhang. THU-IMG at TRECVID 2010. *TRECVID workshop*, 2010.

[SM00] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, August 2000.

[SM01] N. T. Siebel and S. J. Maybank. Real-time tracking of pedestrians and vehicles. In *IEEE Workshop on PETS*, 2001.

[SMH05] F. Souvannavong, B. Merialdo, and B. Huet. Region-based video content indexing and retrieval. In *Fourth International Workshop on Content-Based Multimedia Indexing, CBMI '05*, pages 21–23, 2005.

[SMK$^+$11] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, M. Bugalho, and I. Trancoso. Temporal video segmentation to scenes using high-level audiovisual features. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(8):1163 –1177, August 2011.

[SMQS98] A. F. Smeaton, M. Morony, G. Quinn, and R. Scaife. Taiscéalaí: Information Retrieval from an Archive of Spoken Radio News. In *Proceedings of ECDL2*, pages 429–442, Crete, 1998.

[snf] SNFaceCrop, Face detection and cropping software. http://sourceforge.net/projects/snfacecrop/.

[SNN03] J.R. Smith, M. Naphade, and A. Natsev. Multimedia semantic indexing using model vectors. In *Proceedings of the 2003 International Conference on Multimedia and Expo, ICME '03*, volume 2, pages 445–448, July 2003.

[SR00] O. Sukmarg and K. R. Rao. Fast object detection and segmentation in MPEG compressed domain, 2000.

[SRP$^+$01] B. Stenger, V. Ramesh, N. Paragios, F. Coetzee, and J.M. Buhmann. Topology free hidden markov models: application to background modeling. In *Proceedings of the 8th IEEE International Conference on Computer Vision, ICCV '01*, volume 1, pages 294–301, 2001.

[SS00] R. E. Schapire and Y. Singer. BoosTexter: A Boosting-based System for Text Categorization. *Machine Learning*, 39(2/3):135–168, 2000.

[SSBT07] P. Schügerl, R. Sorschag, W. Bailer, and G. Thallinger. Object re-detection using SIFT and MPEG-7 color descriptors. In *Proceedings of the 2007 international conference on Multimedia content analysis and mining*, MCAM'07, pages 305–314, Berlin, Heidelberg, 2007. Springer-Verlag.

[SSS08] C. Saathoff, S. Schenk, and A. Scherp. Kat: the k-space annotation tool. In D.J. Duke, L. Hardman, A.G. Hauptmann, D. Paulus, and S. Staab, editors, *Third International Conference on Semantic and Digital Media Technologies (SAMT)*, Koblenz, Germany, 2008.

[SSZ04] J. Sivic, F. Schaffalitzky, and A. Zisserman. Efficient object retrieval from videos. In *European Signal Processing Conference*, 2004.

[ST00] N. Slonim and N.i Tishby. Document clustering using word clusters via the information bottleneck method. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 208–215, New York, NY, USA, 2000. ACM.

[STA08] E. Spyrou, G. Tolias, and Y. Avrithis. Large scale concept detection in video using a region thesaurus. In *Proceedings of the 15th International Multimedia Modeling Conference on Advances in Multimedia Modeling*, MMM '09, pages 197–207, Berlin, Heidelberg, 2008. Springer-Verlag.

[STMA09] E. Spyrou, G. Tolias, P. Mylonas, and Y. Avrithis. Concept detection and keyframe extraction using a visual thesaurus. *Multimedia Tools Appl.*, 41(3):337–373, February 2009.

[STV08] E. Spyromitros, G. Tsoumakas, and Ioannis Vlahavas. An empirical study of lazy multilabel classification algorithms. In *Proceedings of the 5th Hellenic conference on Artificial Intelligence: Theories, Models and Applications*, SETN '08, pages 401–406, Berlin, Heidelberg, 2008. Springer-Verlag.

[SW03] J. Stefanowski and D. Weiss. Carrot 2 and language properties in web search results clustering. 2003.

[SWFS03] M. Seki, T. Wada, H. Fujiwara, and K. Sumi. Background subtraction based on cooccurrence of image variations. In *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 65–72, June 2003.

[SWHL11] T. Schmidt, K. Wörner, H. Hedeland, and T. Lehmberg. New and future developments in exmaralda. In *Multilingual Resources and Multilingual Applications. Proceedings of the GSCL Conference*, Hamburg, Germany, 2011.

[SWKS07] C. G. M. Snoek, M. Worring, D. C. Koelma, and A. W. M. Smeulders. A learned lexicon-driven paradigm for interactive video retrieval. *Trans. Multi.*, 9(2):280–292, February 2007.

[SZ03] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. In *Proceedings of the 9th IEEE International Conference on Computer Vision*, volume 2, pages 1470–1477, October 2003.

[SZL^{+}08] Y. Song, Z. Zhuang, H. Li, Q. Zhao, J. Li, W.-C. Lee, and C. L. Giles. Real-time automatic tag recommendation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 515–522, New York, NY, USA, 2008. ACM.

[SZS06] J. Sivic, C. L. Zitnick, and R. Szeliski. Finding people in repeated shots of the same scene. In *BMVC*, volume 3, pages 909–918, 2006.

[SZZ07] H. T. Shen, X. Zhou, and A. Zhou. An adaptive and dynamic dimensionality reduction method for high-dimensional indexing. *The VLDB Journal*, 16(2):219–234, 2007.

[Tao11] R. Tao. Visual concept detection and real time object detection. *CoRR*, abs/1104.0582, 2011.

[TC88] C.-H. Teh and R.T. Chin. On image analysis by the methods of moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(4):496–513, July 1988.

[TCGP09] D.-N. Ta, W.-C. Chen, N. Gelfand, and K. Pulli. Surftrac: Efficient tracking and continuous object recognition using local feature descriptors. In *IEEE Conf. on Computer Vision and Pattern Recognition, CVPR '09*, 2009.

[TCP04] F. A. Thabtah, P. Cowling, and Y. Peng. MMAC: A New Multi-class, Multi-label Associative Classification Approach. In *Proceedings of the 4th IEEE International Conference on Data Mining, ICDM '04*, pages 217–224, 2004.

[TCYZ03] Z. Tu, X. Chen, A.L. Yuille, and S.-C. Zhu. Image parsing: unifying segmentation, detection, and recognition. In *Proceedings ot the 9th IEEE International Conference on Computer Vision*, volume 1, pages 18–25, October 2003.

[TdSL00] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319, 2000.

[TEBEH06] A. S. Tolba, A. H. El-Baz, and A. A. El-Harby. Face Recognition: A Literature Review. *International Journal of Information and Communication Engineering*, 2:88–103, 2006.

[TFH08] T. B. Terriberry, L. M. French, and J. Helmsen. GPU Accelerating Speeded-Up Robust Features, 2008.

[TFW08] A. Torralba, R. Fergus, and Y. Weiss. Small codes and large image databases for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR '08*, pages 1–8, June 2008.

[TGD$^+$11] I. Tsampoulatidis, N. Gkalelis, A. Dimou, V. Mezaris, and I. Kompatsiaris. High-level event detection system based on discriminant visual concepts. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, pages 68:1–68:2, New York, NY, USA, 2011. ACM.

[THY$^+$11] J. Tang, R. Hong, S. Yan, T.-S. Chua, G.-J. Qi, and R. Jain. Image annotation by knn-sparse graph-based label propagation over noisily tagged web images. *ACM Trans. Intell. Syst. Technol.*, 2(2):14:1–14:15, February 2011.

[TIT01] Y. Tao, T. R. Ioerger, and Y. Y. Tang. Extraction of rotation invariant signature based on fractal geometry. In *ICIP '01*, pages 1090–1093, 2001.

[TK07] G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *Int J Data Warehousing and Mining*, 2007:1–13, 2007.

[TKM$^+$09] M.A. Tahir, J. Kittler, K. Mikolajczyk, F. Yan, K.E.A. van de Sande, and T. Gevers. Visual category recognition using spectral regression and kernel discriminant analysis. In *12th IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 178 –185, October 2009.

[TKMS03] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[TKV11] G. Tsoumakas, I. Katakis, and I. Vlahavas. Random k-labelsets for multilabel classification. *IEEE Trans. on Knowl. and Data Eng.*, 23(7):1079–1089, July 2011.

[TLF08] E. Tola, V. Lepetit, and P. Fua. A fast local descriptor for dense matching. In *CVPR'08*, pages –1–1, 2008.

[TLF10] E. Tola, V. Lepetit, and P. Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):815–830, May 2010.

[TLQC10] J. Tang, H. Li, G.-J. Qi, and T.-S. Chua. Image annotation by graph-based inference with integrated multiple/single instance representations. *IEEE Transactions on Multimedia*, 12(2):131–141, February 2010.

[TM08] T. Tuytelaars and K. Mikolajczyk. *Local Invariant Feature Detectors: A Survey*. Now Publishers Inc., Hanover, MA, USA, 2008.

[TMK08] E. Tsamoura, V. Mezaris, and I. Kompatsiaris. Gradual transition detection using color coherence and other criteria in a video shot meta-segmentation framework. In *15th IEEE International Conference on Image Processing, ICIP '08*, pages 45–48, October 2008.

[TMRSSG00] L.A. Torres-Mendez, J.C. Ruiz-Suarez, L.E. Sucar, and G. Gomez. Translation, rotation, and scale-invariant object recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 30(1):125–130, February 2000.

[TP98] Y. A. Tolias and S. M. Panas. Image segmentation by a fuzzy clustering algorithm using adaptive spatially constrained membership functions. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 28(3):359–369, May 1998.

[TPG03] P. Tonella, E. Pianta, and C. Girardi. Using keyword extraction for web site clustering. In *Fifth International Workshop on Web Site Evolution (WSE-03)*, pages 41–48. IEEE Computer Society, 2003.

[Tra11] Image segmentation, available techniques, developments and open issues. *Canadian Journal on Image Processing and Computer Vision*, 2(3):20–29, March 2011.

[TRS$^+$12] P. Toharia, O. D. Robles, R. SuáRez, J. L. Bosque, and L. Pastor. Shot boundary detection using Zernike moments in multi-GPU multi-CPU architectures. *J. Parallel Distrib. Comput.*, 72(9):1127–1133, 2012.

[TSF10] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient object category recognition using classemes. In *Proceedings of the 11th European conference on Computer vision: Part I*, ECCV'10, pages 776–789, Berlin, Heidelberg, 2010. Springer-Verlag.

[TT02] D.-M. Tsai and Y.-H. Tsai. Rotation-invariant pattern matching with color ring-projection. *Pattern Recognition*, 35(1):131–141, 2002.

[TT08] J. Tao and Y.-P. Tan. Face clustering in videos using constraint propagation. In *IEEE International Symposium on Circuits and Systems, ISCAS '08*, pages 3246 –3249, May 2008.

[Tur00] P. D. Turney. Learning algorithms for keyphrase extraction. *Information Retrieval*, 2:303–336, 2000. 10.1023/A:1009976227802.

[TV07] G. Tsoumakas and I. Vlahavas. Random k-labelsets: An ensemble method for multilabel classification. In *Proceedings of the 18th European conference on Machine Learning*, ECML '07, pages 406–417, Berlin, Heidelberg, 2007. Springer-Verlag.

[TVD03] B. T. Truong, S. Venkatesh, and C. Dorai. Scene extraction in motion pictures. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(1):5–15, January 2003.

[TWY08] Y.-T. Tsai, Q. Wang, and S. You. Cdikp: A highly-compact local feature descriptor. In *19th International Conference on Pattern Recognition, ICPR '08*, pages 1–4, December 2008.

[TYB$^+$10] M. A. Tahir, F. Yan, M. Barnard, M. Awais, K. Mikolajczyk, and J. Kittler. The university of surrey visual concept detection system at imageCLEF@ICPR: working notes. In *Proceedings of the 20th International conference on Recognizing patterns in signals, speech, images, and videos*, ICPR'10, pages 162–170, Berlin, Heidelberg, 2010. Springer-Verlag.

[Uja09] M. Ujaldon. GPU acceleration of zernike moments for large-scale images. In *IEEE International Symposium on Parallel Distributed Processing, IPDPS '09*, pages 1–8, may 2009.

[UK04] F. Ullah and S. Kaneko. Using orientation codes for rotation-invariant template matching. *Pattern Recognition*, pages 201–209, 2004.

[UnLBG01] L. Ureña López, M. Buenaga, and J. Gómez. Integrating linguistic resources in tc through wsd. *Computers and the Humanities*, 35:215–230, 2001.

[USS10] J.R.R. Uijlings, A.W.M. Smeulders, and R.J.H. Scha. Real-time visual concept classification. *IEEE Transactions on Multimedia*, 12(7):665 –681, nov. 2010.

[Vap98] V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

[vdSGS08] K. E.A. van de Sande, T. Gevers, and C. G. M. Snoek. A comparison of color features for visual concept classification. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, CIVR '08, pages 141–150, New York, NY, USA, 2008. ACM.

[vdSGS10a] K. van de Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1582–1596, September 2010.

[VDSGS10b] K. E. A. Van De Sande, T. Gevers, and A. W. M. Smeulders. The University of Amsterdam's concept detection system at ImageCLEF 2009. In *Proceedings of the 10th international conference on Cross-language evaluation forum: multimedia experiments*, CLEF'09, pages 261–268, Berlin, Heidelberg, 2010. Springer-Verlag.

[vdSGS11] K. E.A. van de Sande, T. Gevers, and C. G.M. Snoek. Empowering visual categorization with the GPU. *Trans. Multi.*, 13(1):60–70, February 2011.

[vdWGB06] J. van de Weijer, T. Gevers, and A.D. Bagdanov. Boosting color saliency in image feature detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):150–156, January 2006.

[vdWS06] J. van de Weijer and C. Schmid. Coloring local feature extraction. In *Proceedings of the 9th European conference on Computer Vision - Volume Part II*, ECCV'06, pages 334–348, Berlin, Heidelberg, 2006. Springer-Verlag.

[VFSC06] C.N. Vasconcelos, B. Feijo, D. Szwarcman, and M. Costa. Shot segmentation based on the encoder signature. In *12th International Multi-Media Modelling Conference Proceedings*, 2006.

[vGVSG10] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek. Visual word ambiguity. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(7):1271–1283, July 2010.

[VHS+06] Th. Villmann, B. Hammer, F. Schleif, T. Geweniger, and W. Herrmann. Fuzzy classification by fuzzy labeled neural gas. *Neural Networks*, 19(6):772–779, July 2006.

[Vin12] E. Vincent. Advances in audio source separation and multisource audio content retrieval. In *SPIE Defense, Security, and Sensing*, Baltimore, États-Unis, 2012.

[vis] Visage Technologies. http://www.visagetechnologies.com/.

[VJ01] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR '01*, volume 1, pages 511–518, 2001.

[VN08] V. Vineet and P.J. Narayanan. CUDA cuts: Fast graph cuts on the GPU. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPRW '08*, pages 1–8, June 2008.

[VNH03] A. Velivelli, C.-W. Ngo, and T. S. Huang. Detection of documentary scene changes by audio-visual fusion. In *Proceedings of the 2nd international conference on Image and video retrieval*, CIVR'03, pages 227–238, Berlin, Heidelberg, 2003. Springer-Verlag.

[VSLB10] P. D. Z. Varcheie, M. Sills-Lavoie, and G.-A. Bilodeau. A multiscale region-based motion detection and background subtraction algorithm. *Sensors*, 10(2):1041–1061, 2010.

[VSP06] N. Vretos, V. Solachildis, and I. Pitas. A Mutual Information based Face Clustering Algorithm for Movies. In *IEEE International Conference on Multimedia and Expo, 2006*, pages 1013 –1016, July 2006.

[VW12] T. Vander Wal. Folksonomy coinage and definition, 2012.

[VZ05] M. Varma and A. Zisserman. A statistical approach to texture classification from single images. *Int. J. Comput. Vision*, 62(1-2):61–81, April 2005.

[WAC+04] J. Willamowski, D. Arregui, G. Csurka, C. R. Dance, and L. Fan. Categorizing nine visual classes using local appearance descriptors. In *In ICPR Workshop on Learning for Adaptable Visual Systems*, 2004.

[WBR+06] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. Elan: a professional framework for multimodality research. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1556–1559, 2006.

[WC02] J. Wang and T.-S. Chua. A framework for video scene boundary detection. In *Proceedings of the 10th ACM international conference on Multimedia*, MULTIMEDIA '02, pages 243–246, New York, NY, USA, 2002. ACM.

[WC08] M.-F. Weng and Y.-Y. Chuang. Multi-cue fusion for semantic video indexing. In *Proceedings of the 16th ACM international conference on Multimedia*, MM '08, pages 71–80, New York, NY, USA, 2008. ACM.

[WCK+03] J.-U. Won, Y.-S. Chung, I.-S. Kim, J.-G. Choi, and K.-H. Park. Correlation based video-dissolve detection. In *Proceedings of the International Conference on Information Technology: Research and Education, ITRE2003*, pages 104–107, August 2003.

[WFKvdM97] L. Wiskott, J.-M. Fellous, N. Kuiger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(7):775 –779, July 1997.

[WH99] L. Wixson and M. Hansen. Detecting salient motion by accumulating directionally-consistent flow. In *Proceedings of the 1999 International Conference on Computer Vision - Volume 2*, ICCV '99, pages 797–, Washington, DC, USA, 1999. IEEE Computer Society.

[WHH+09] M. Wang, X.-S. Hua, R. Hong, J. Tang, G.-J. Qi, and Y. Song. Unified video annotation via multigraph learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(5):733 –746, May 2009.

[WJN08] F. Wang, Y.-G. Jiang, and C.-W. Ngo. Video event detection using motion relativity and visual relatedness. In *Proceedings of the 16th ACM international conference on Multimedia*, MM '08, pages 239–248, New York, NY, USA, 2008. ACM.

[WMS00] P. Wu, B.S. Manjunath, and H.D. Shin. Dimensionality reduction for image retrieval. In *Proceedings of the 2000 International Conference on Image Processing*, volume 3, pages 726–729, 2000.

[WPF+99] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning. Kea: practical automatic keyphrase extraction. In *Proceedings of the 4th ACM conference on Digital libraries*, DL '99, pages 254–255, New York, NY, USA, 1999. ACM.

[WS09] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.*, 10:207–244, 2009.

[WSH09] M. Wang, Y. Song, and X.-S. Hua. Concept representation based video indexing. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 654–655, New York, NY, USA, 2009. ACM.

[WT06] S.-C. Wang and Y. Tanaka. Topic-oriented query expansion for web search. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 1029–1030, New York, NY, USA, 2006. ACM.

[WTP+09] M. Werlberger, W. Trobin, T. Pock, A. Wedel, D. Cremers, and H. Bischof. Anisotropic Huber-L1 Optical Flow. In *Proceedings of the British Machine Vision Conference (BMVC)*, London, UK, September 2009.

[WTS04] Y. Wu, B. L. Tseng, and J. R. Smith. Ontology-based multi-classification learning for video concept detection. In *IEEE International Conference on Multimedia and Expo, ICME '04*, volume 2, pages 1003 –1006, June 2004.

[WW10] P. Wang and J. Wang. Block characteristic based moving object segmentation in the H.264 compressed domain. In *International Conference on Audio Language and Image Processing, ICALIP '10*, pages 643 –647, November 2010.

[WYG08] W. Wang, J. Yang, and W. Gao. Modeling background and segmenting moving objects from compressed video. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(5):670–681, May 2008.

[WYX07] X. Wan, J. Yang, and J. Xiao. Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 2007.

[WZL+08] D. Wang, S. Zhu, T. Li, Y. Chi, and Y. Gong. Integrating clustering and multi-document summarization to improve document understanding. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 1435–1436, New York, NY, USA, 2008. ACM.

[XBF$^+$08] S. Xu, S. Bao, B. Fei, Z. Su, and Y. Yu. Exploring folksonomy for personalized search. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 155–162, New York, NY, USA, 2008. ACM.

[XC96] J. Xu and W. B. Croft. Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '96, pages 4–11, New York, NY, USA, 1996. ACM.

[XC08] D. Xu and S.-F. Chang. Video event recognition using kernel methods with multilevel temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1985–1997, November 2008.

[XGZ$^+$10] H. Xie, K. Gao, Y. Zhang, J. Li, and Y. Liu. GPU-based fast scale invariant interest point detector. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP '10*, pages 2494 –2497, March 2010.

[XMW$^+$11] E. Xie, M. McGinnity, Q. Wu, J. Cai, and R. Cai. GPU implementation of spiking neural networks for color image segmentation. In *4th International Congress on Image and Signal Processing, CISP '11*, volume 3, pages 1246–1250, October 2011.

[XNJR02] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In *Advances in Neural Information Processing Systems 15*, volume 15, pages 505–512, 2002.

[XW10] C. Xu and L. Wei. Study on shot boundary detection based on fuzzy subset-hood theory. In *International Conference on Intelligent System Design and Engineering Application (ISDEA), 2010*, volume 2, pages 476–480, October 2010.

[XXC$^+$04] L. Xie, P. Xu, S.-F. Chang, A. Divakaran, and H. Sun. Structure analysis of soccer video with domain knowledge and hidden markov models. *Pattern Recogn. Lett.*, 25(7):767–775, May 2004.

[YAR08] C. Yeo, P. Ahammad, and K. Ramchandran. Rate-efficient visual correspondences using random projections. In *15th IEEE International Conference on Image Processing, ICIP '08*, pages 217–220, October 2008.

[YCH06] R. Yan, M.-Y. Chen, and A. Hauptmann. Mining relationship between video concepts using probabilistic graphical models. In *IEEE International Conference on Multimedia and Expo*, pages 301–304, July 2006.

[YCKH07] A. Yanagawa, S.-F. Chang, L. Kennedy, and W. Hsu. Columbia University's Baseline Detectors for 374 LSCOM Semantic Visual Concepts. Technical report, Columbia University, March 2007.

[YDD05] C. Yang, R. Duraiswami, and L. Davis. Efficient mean-shift tracking via a new similarity measure. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR '05*, volume 1, pages 176 – 183, june 2005.

[YDF05] C. Yang, M. Dong, and F. Fotouhi. Region based image annotation through multiple-instance learning. In *Proceedings of the 13th annual ACM international conference on Multimedia*, MULTIMEDIA '05, pages 435–438, New York, NY, USA, 2005. ACM.

[YDT03] X.-D. Yu, L.-Y. Duan, and Q. Tian. Robust moving video object segmentation in the MPEG compressed domain. In *Proceedings of the 2003 International Conference on Image Processing, ICIP '03*, volume 3, pages 933–936, September 2003.

[YH03] X. Yin and J. Han. Cpar: Classification based on predictive association rules. In *SDM'03*, 2003.

[YH06a] R. Yan and A. G. Hauptmann. Probabilistic latent query analysis for combining multiple retrieval sources. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 324–331, New York, NY, USA, 2006. ACM.

[YH06b] J. Yang and A. G. Hauptmann. Exploring temporal consistency for video analysis and retrieval. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, MIR '06, pages 33–42, New York, NY, USA, 2006. ACM.

[YHCT12] Y. Yu, K. Huang, W. Chen, and T. Tan. A novel algorithm for view and illumination invariant image matching. *IEEE Transactions on Image Processing*, 21(1):229–240, January 2012.

[YHT10] Y. Yu, K. Huang, and T. Tan. A harris-like scale invariant feature detector. In *Proceedings of the 9th Asian conference on Computer Vision - Volume Part II*, ACCV'09, pages 586–595, Berlin, Heidelberg, 2010. Springer-Verlag.

[YJ06] L. Yang and R. Jin. Distance Metric Learning: A Comprehensive Survey. Technical report, Department of Computer Science and Engineering, Michigan State University, 2006.

[YJX09] L. Yawei, L. Jianwei, and Z. Xiaohong. A new local feature descriptor: Covariant support region. In *IEEE International Conference on Intelligent Computing and Intelligent Systems, ICIS '09*, volume 4, pages 346–351, November 2009.

[YKO$^+$00] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. The HTK book version 3.0., 2000. Cambridge, England, Cambridge University.

[YKR07] S. Yang, S.-K. Kim, and Y. M. Ro. Semantic home photo categorization. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(3):324 –335, March 2007.

[YLZ07] J. Yuan, J. Li, and B. Zhang. Exploiting spatial context constraints for automatic image region annotation. In *Proceedings of the 15th international conference on Multimedia*, MULTIMEDIA '07, pages 595–604, New York, NY, USA, 2007. ACM.

[YM09] G. Yu and J.-M. Morel. A fully affine invariant image comparison method. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '09*, pages 1597 –1600, April 2009.

[YMF07] L. Yang, P. Meer, and D.J. Foran. Multiple class segmentation using a unified framework over mean-shift patches. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR '07*, pages 1–8, June 2007.

[YMF11] R. Yumiba, M. Miyoshi, and H. Fujiyoshi. Moving object detection with background model based on spatio-temporal texture. In *IEEE Workshop on Applications of Computer Vision (WACV), 2011*, pages 352 –359, January 2011.

[YP05] M. Yokoyama and T. Poggio. A contour-based moving object detection and tracking. In *2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005*, pages 271 – 276, October 2005.

[YSK07] W. You, M. S. H.i Sabirin, and M.I Kim. Moving object tracking in H.264/AVC bitstream. In *Proceedings of the 2007 international conference on Multimedia content analysis and mining*, MCAM'07, pages 483–492, Berlin, Heidelberg, 2007. Springer-Verlag.

[YWX$^+$07] J. Yuan, H. Wang, L. Xiao, W. Zheng, J. Li, F. Lin, and B. Zhang. A formal study of shot boundary detection. *IEEE Trans. Cir. and Sys. for Video Technol.*, 17(2):168–186, February 2007.

[YXL$^+$09] Y. Yang, J. Xiao, K. Lin, G. Wu, T. Ren, and Y. Wang. Nanjing University in TRECVID 2009, 2009.

[YXZ$^+$07] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):40–51, January 2007.

[YYA10] K. Yamamoto, O Yamaguchi, and H. Aoki. Fast face clustering based on shot similarity for browsing video. In *Progress in Informatics*, volume 7, pages 56–65. March 2010.

[YYH07] J. Yang, R. Yan, and A. G. Hauptmann. Cross-domain video concept detection using adaptive SVMs. In *Proceedings of the 15th international conference on Multimedia*, MULTIMEDIA '07, pages 188–197, New York, NY, USA, 2007. ACM.

[YYL98] M. Yeung, B.-L. Yeo, and B. Liu. Segmentation of video by clustering and graph analysis. *Computer Vision Image Understanding*, 71(1):94–109, July 1998.

[ZC04] D.-Q. Zhang and S.-F. Chang. Detecting image near-duplicate by stochastic attributed relational graph matching with learning. In *Proceedings of the 12th annual ACM international conference on Multimedia*, MULTIMEDIA '04, pages 877–884, New York, NY, USA, 2004. ACM.

[ZCCY10] G. Zhao, L. Chen, G. Chen, and J. Yuan. KPB-SIFT: a compact local feature descriptor. In *Proceedings of the 2010 International Conference on Multimedia*, MM '10, pages 1175–1178, New York, NY, USA, 2010. ACM.

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

[ZD06] Y. Zheng and D. Doermann. Robust point matching for nonrigid shapes by preserving local neighborhood structures. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(4):643–649, 2006.

[ZDGH05] W. Zeng, J. Du, W. Gao, and Q. Huang. Robust moving object segmentation on H.264/AVC compressed video using the block-based MRF model. *Real-Time Imaging*, 11(4):290–299, August 2005.

[ZEX$^+$05] X. Zhu, A.K. Elmagarmid, X. Xue, L. Wu, and A.C. Catlin. Insightvideo: toward hierarchical video content organization for efficient browsing, summarization and retrieval. *IEEE Transactions on Multimedia*, 7(4):648 – 666, August 2005.

[ZG09] Z.-Q. Zhao and H. Glotin. Enhancing visual concept detection by a novel matrix modular scheme on SVM. In *Proceedings of the 9th Cross-language evaluation forum conference on Evaluating systems for multilingual and multimodal information access*, CLEF'08, pages 640–643, Berlin, Heidelberg, 2009. Springer-Verlag.

[ZGZ03] W. Zeng, W. Gao, and D. Zhao. Automatic moving object extraction in MPEG video. In *Proceedings of the 2003 International Symposium on Circuits and Systems, ISCAS '03*, volume 2, pages 524–527, May 2003.

[ZHH11] W. M. D. Wan Zaki, A. Hussain, and M. Hedayati. Moving object detection using keypoints reference model. *EURASIP J. Image and Video Processing*, pages 13–13, 2011.

[ZHLY08] J. Zhu, S. C. H. Hoi, M. R. Lyu, and S. Yan. Near-duplicate keyframe retrieval by nonrigid image matching. In *Proceedings of the 16th ACM international conference on Multimedia*, MM '08, pages 41–50, New York, NY, USA, 2008. ACM.

[ZJN06] W. Zhao, Y.-G. Jiang, and C.-W. Ngo. Keyframe retrieval by keypoints: can point-to-point matching help? In *Proceedings of the 5th international conference on Image and Video Retrieval*, CIVR'06, pages 72–81, Berlin, Heidelberg, 2006. Springer-Verlag.

[ZKF05] Y. Zhao, G. Karypis, and U. Fayyad. Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, 10:141–168, 2005. 10.1007/s10618-005-0361-3.

[ZL09] S. Zhu and Y. Liu. Video scene segmentation and semantic representation using a novel scheme. *Multimedia Tools and Applications*, 42(2):183–205, April 2009.

[ZL10] L. Zhang and Y. Liang. Motion human detection based on background subtraction. In *Second International Workshop on Education Technology and Computer Science, ETCS '10*, volume 1, pages 284–287, March 2010.

[ZLCS04] D.-Q. Zhang, C.-Y. Lin, S.-F. Chang, and J.R. Smith. Semantic video clustering across sources using bipartite spectral clustering. In *IEEE International Conference on Multimedia and Expo, ICME '04*, volume 1, pages 117 – 120, June 2004.

[ZM06] M. Zhu and A.M. Martinez. Subclass discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(8):1274–1286, August 2006.

[ZMLS07] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *Int. J. Comput. Vision*, 73(2):213–238, June 2007.

[ZMM95] R. Zabih, J. Miller, and K. Mai. A feature-based algorithm for detecting and classifying scene breaks. In *Proceedings of the 3rd ACM international conference on Multimedia*, MULTIMEDIA '95, pages 189–200, New York, NY, USA, 1995. ACM.

[ZMM99] R. Zabih, J. Miller, and K. Mai. A feature-based algorithm for detecting and classifying production effects. *Multimedia Syst.*, 7(2):119–128, March 1999.

[ZS03] J. Zhong and S. Sclaroff. Segmenting foreground objects from a dynamic textured background via a robust kalman filter. In *Proceedings of the 9th IEEE International Conference on Computer Vision*, volume 1, pages 44 –50, October 2003.

[ZS06] Y. Zhai and M. Shah. Video scene segmentation using markov chain monte carlo. *IEEE Transactions on Multimedia*, 8(4):686 –697, August 2006.

[ZS08] R. Zass and A. Shashua. Probabilistic graph and hypergraph matching. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR '08*, pages 1–8, 2008.

[ZWW$^+$07] Y. Zhao, T. Wang, P. Wang, W. Hu, Y. Du, Y. Zhang, and G. Xu. Scene segmentation and categorization using N-cuts. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR '07*, pages 1–7, June 2007.

[ZyMjWn11] C. Zhuo-yi, Z. Ming-ji, and D. Wan-ning. Efficient algorithm on video shot segmentation of the adaptive threshold. In *IEEE International Conference on Computer Science and Automation Engineering, CSAE '11*, volume 4, pages 634–637, June 2011.

[ZZ06] M.-L. Zhang and Z.-H. Zhou. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1338–1351, October 2006.

[ZZ07] M.-L. Zhang and Z.-H. Zhou. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recogn.*, 40(7):2038–2048, July 2007.

[ZZ10] C. Zhang and Z. Zhang. A Survey of Recent Advances in Face Detection. *Microsoft Research Technical Report*, 2010.