TKK Reports in Information and Computer Science
Espoo 2010 TKK-ICS-R39

CONCEPT-BASED VIDEO SEARCH WITH THE PICSOM MULTIMEDIA RETRIEVAL SYSTEM

Ville Viitaniemi, Mats Sjöberg, Markus Koskela and Jorma Laaksonen

Aalto University
School of Science
and Technology

TKK Reports in Information and Computer Science
Espoo 2010 TKK-ICS-R39

CONCEPT-BASED VIDEO SEARCH WITH THE PICSOM MULTIMEDIA RETRIEVAL SYSTEM

Ville Viitaniemi, Mats Sjöberg, Markus Koskela and Jorma Laaksonen

Aalto University School of Science and Technology
Faculty of Information and Natural Sciences
Department of Information and Computer Science

Aalto-yliopiston teknillinen korkeakoulu
Informaatio- ja luonnontieteiden tiedekunta
Tietojenkäsittelytieteen laitos

Distribution:
Aalto University School of Science and Technology
Faculty of Information and Natural Sciences
Department of Information and Computer Science
PO Box 15400
FI-00076 AALTO
FINLAND
URL: http://ics.tkk.fi
Tel. +358 9 470 01
Fax +358 9 470 23369
E-mail: series@ics.tkk.fi

ISBN 978-952-60-3545-1 (Print)
ISBN 978-952-60-3546-8 (Online)
ISSN 1797-5034 (Print)
ISSN 1797-5042 (Online)
URL: http://lib.tkk.fi/Reports/2010/isbn9789526035468.pdf

AALTO ICS
Espoo 2010

ABSTRACT: In this report we describe the structure of the PicSOM multimedia retrieval system and elaborate on its automatic concept detection and video search subsystems. We evaluate several alternative techniques for implementing these two components of the PicSOM system in a comprehensive series of experiments employing the large-scale setups of the TRECVID video retrieval evaluation campaigns of 2008 and 2009. Based on the results of the experiments, we conclude that fusion-based shot-wise visual analysis together with N-gram temporal concept-wise post-processing make an efficient combination of techniques for automatic semantic concept detection from video material. It has previously been shown that semantic concepts are very beneficial for video search. Here we investigate the lexical and visual-example-based selection of concepts for search queries, concluding that both selection methods are successful in matching useful concepts. Finally, we show that the performance of the PicSOM system has improved since the TRECVID 2008 evaluation and now compares very well with the state-of-the-art in concept detection and video search.

KEYWORDS: multimedia databases, content-based video retrieval, semantic concept detection, concept-based retrieval, temporal multimedia analysis, self-organising map, support-vector machine

ACKNOWLEDGEMENT: This work has been supported by the Centre of Excellence in Adaptive Informatics Research, funded by the Academy of Finland.

CONTENTS

1 INTRODUCTION

Digital video has become commonplace, both in professional use and in various consumer products such as camcorders, webcams, mobile phones, digital TV recorders, video sharing websites, CCTV surveillance, and virtual and augmented reality applications. The growing amount of data available requires more effective solutions for organising and searching the video content. While the capturing, storing, and transmitting of digital video has steadily become easier and more cost-effective, the current methods for the automatic analysis and semantic representation of the video content itself are considerably less mature.

Content-based visual information retrieval addresses the problem of finding visual data relevant to the users' information needs from multimedia databases. In early content-based image and video retrieval systems, the retrieval was usually based solely on querying by examples and measuring the similarity of the database objects (images, video shots) with *low-level features* automatically extracted from the objects. Generic low-level features are often, however, insufficient to discriminate content well on a conceptual level. This "semantic gap" is the fundamental problem in content-based multimedia retrieval.

Extracting semantic concepts from multimedia data has attracted a lot of research attention in the recent years [37, 26, 51, 16]. The main aim of automatic concept extraction is to facilitate semantic indexing and concept-based retrieval of multimedia content. The leading principle has been to build semantic representations by extracting intermediate semantic levels (events, objects, locations, people, etc.) from low-level visual and aural features by using machine learning techniques. Recently, the introduction of large-scale multimedia ontologies, such as LSCOM [36] and ImageNet [11] and large manually annotated data sets (e.g. [2]) have enabled generic analysis of multimedia content as well as an increase in multimedia lexicon sizes by orders of magnitude.

The modelling of *mid-level semantic concepts* can be seen as an attempt to fill, or at least reduce, the semantic gap. Indeed, in recent studies it has been observed that, despite the fact that the accuracy of the concept detectors is far from perfect, they can be useful in supporting *high-level indexing and querying* on multimedia data [16]. This is mainly because such semantic concept detectors can be trained off-line with computationally more demanding supervised learning algorithms and with considerably more positive and negative training examples than what are typically available at query time. In addition, this automatic approach is scalable to large numbers of multimedia objects and features.

One of the tasks in the annual TRECVID video retrieval evaluation [47] is to detect the presence of predefined *high-level features* (HLFs)—such as *sports*, *meeting* or *urban*—in broadcast videos that are already partitioned into shots. These kinds of HLFs can be seen as practical examples of mid-level semantic concepts one might find both possible to be extracted and useful as building blocks for media search on higher semantic levels. Evaluation of semantic video search is also an important constituent task of the TRECVID evaluation.

In order to be able to exploit the idea of concept-based multimedia retrieval, one needs to implement the idea as a system. In this report we give a detailed description of our implementation, the PicSOM multimedia analysis and retrieval framework. This is the first time the framework is described in its entirety. The framework includes subsystems for video corpus preparation, semantic concept detection and concept-based automatic video search. The subsystems divide further into numerous components. In the following sections we describe several alternative techniques for implementing these components.

As one part of the description, we propose a set of post-processing techniques to take advantage of correlations that semantic concepts occurring in video material typically exhibit, both in temporal dimension between shots and across different concepts. In particular, the post-processing techniques combine an N-gram intra-concept inter-shot temporal modelling technique with a simple clustering approach that takes advantage of temporal and instantaneous inter-concept co-occurrences. Many of the current state-of-the-art multimedia retrieval systems do not include inter-shot temporal analysis.

Having implemented several alternative techniques for different components of the video retrieval system, the techniques can be compared in experiments. In this report, we use the large-scale experimental setups of the TRECVID video retrieval evaluation campaign, in particular the setups of its HLF extraction and automatic video search tasks. Performing a large number of such experiments enables us to deduce which of the available techniques one should use to implement the components of a multimedia retrieval system. We are also able to compare different overall operational modes in which the PicSOM framework can be used.

As the experiments are performed with the commonly used TRECVID standard setups, we can relate the PicSOM system's performance to the other systems of the video retrieval community. The experiments give insight to the question, what makes some systems in the literature work and what not. Some of the techniques applied in them might be essential, whereas some other, attractive-looking ones might just be parts of otherwise well-functioning systems, without being particularly effective themselves. Our experiments enable us also to assess whether the multimedia search community as a whole has made progress in concept detection and video search between TRECVIDs of 2008 and 2009. This is because we have performed the experiments with an identical system setup for both years' tasks. Thus we can use the performance of the PicSOM system as a yardstick to relate the general performance level of TRECVID 2008 submissions to that of TRECVID 2009.

The succeeding sections of this report are organised as follows. In Section 2 we describe the parts of a generic video retrieval system and provide examples of the ways in which they have been implemented by others and in our PicSOM system. Sections 3 and 4 contain the essential theoretical and methodological novelty of this report. They describe, respectively, the concept detection techniques and how their outputs are utilised for video search in the PicSOM system. Section 5 presents empirical evaluations of the proposed concept detection and video search algorithms in the TRECVID context. In Section 6 we give our final conclusions from the experiments.

2 PARTS OF A VIDEO RETRIEVAL SYSTEM

The operation of a video retrieval system generally consists of two phases. In the first phase, the system is *prepared* for a video corpus. The corpus is divided into an annotated training part and an unannotated testing part, on which video retrieval is going to be performed in the second *search* phase.

In the preparing phase the whole video corpus is first segmented into shots and the available annotations are associated with the shots. A number of low-level visual, audio and textual feature descriptors are extracted from each shot and content-based indices prepared based on the features. In systems that rely on automatic detection of concepts, the annotated part can then be used to train shot-wise detectors for the concepts existent in the annotations. The detectors apply supervised learning techniques to learn the mapping between low-level shot features and the annotation concepts, also often referred to as high-level features. The preparing phase is allowed to be time-consuming as it is intended to be performed off-line prior to the actual on-line use of the retrieval system.

After the preparation phase, the retrieval system is ready to be used for video retrieval in the search phase. In this phase, the system is queried with a combination of textual phrase, and image and video examples of the desired query topic. The result of a query is a list of video shots, ranked in the order of decreasing predicted likelihood to match the query. The system operation in the search phase is intended to be sufficiently fast to enable the retrieval needs of a real user to be satisfied while the user is waiting, typically in a couple of seconds. The example images and video shots will require pre-processing, feature extraction and classification that cannot be performed during the preparing phase but will inevitably need to be done while the user is waiting for the output.

The rest of this section addresses the components of the outlined video retrieval system architecture in more detail with bibliographic references and descriptions of our implementations in the PicSOM system.

2.1 Shot segmentation and keyframe selection

The first task of the preparing phase for a comprehensive video retrieval system is to segment the video corpus temporally into sequential basic units. Depending on the video material, such a segmentation can be performed on various levels, such as stories, events, scenes, groups, sequences, and shots. For scripted content, the basic semantic unit is a shot, as shots are intrinsically linked to the production of the video. Shot boundaries can be coarsely divided into *abrupt cuts* and *gradual transitions* (wipes, dissolves, fades) based on their duration. In cut transitions there is an instantaneous change from one shot to another without any special effects, whereas in gradual transitions the shift has a nonzero duration, and there are transitional frames that do not belong exclusively to either one of the two shots. Furthermore, as shots can usually be identified by automatic methods with a reasonable accuracy and they provide a suitable representation level for the higher-level video analysis tasks, the contemporary video retrieval systems customarily treat the shot as the basic unit of retrieval.

There has been a multitude of different methods for shot boundary detection proposed in the literature, including pixel-based, feature-based, transform-based, and motion-based approaches. See [8] for a recent review. In the TRECVID evaluations, a master definition of shots is provided in [41]. Their method uses a set of detectors for different types of shot boundaries, based on pixel, edge, and histogram difference statistics. The resulting shots then serve as the common units of retrieval needed for commensurable evaluation of all submissions.

In cases where a master definition of shots has not been available, we have applied one of two shot segmentation algorithms available in the PicSOM system. First, a shot segmentation method based on the Self-Organising Map (SOM) artificial neural network [23] was published in [35]. The main idea is to spot discontinuities in the visual stream by monitoring video frame trajectories of the best-matching units of frame-wise feature vectors on the Self-Organising Maps trained with video corpus. The SOM mapping compensates for the probability density differences in the feature space, and consequently distances between SOM coordinates are more informative than, e.g. Euclidean distances between raw feature vectors. The method compares two sliding best-matching unit windows instead of just measuring distances between two trajectory points, which increases the robustness of the detector. Furthermore, the robustness is increased by using a committee machine of multiple SOM-based detectors, each created with different low-level features.

For video material consisting of only hard cuts as shot boundaries, we have used a method based on thresholding the proportion of successfully tracked interest points between video frames [28]. A shot boundary corresponds to the case when the proportion of interest points that are lost is greater than a heuristically set threshold. This simple approach works reasonably well if there are no drastic changes in the video stream during the shots and the shot boundaries are mostly cuts.

In typical video retrieval systems, one or more keyframes are extracted from each video shot segment. The keyframes are needed both for extracting visual features to describe the content of the shot and for presenting them to the users of the system as still replacements for the dynamic video content. The most straightforward keyframe selection method is to use the centremost frame of each shot. Better results can be obtained by selecting the keyframe on the basis of the content of the shot. In the PicSOM system, the frames of a shot are given a keyframe selection score that is awarded for closeness to the temporal centre of the shot, and penalised for distance from the calculated average image and for having big changes as compared to neighbouring frames. The keyframe is selected as the frame with the highest score, the idea being that it should be close to the centre, but at the same time be "typical" and not contain rapid movement, which could cause, e.g. motion blurring.

2.2 Low-level features

Automatic extraction of low-level features is the foundation of large-scale content-based multimedia processing. Using the video or image data directly in search and retrieval is typically not feasible because of the high dimensionality of the data. Extracted low-level features should thus ideally be of

reasonable dimensionality and discriminative of semantic differences in the data, i.e. the feature extractors should be sensitive to those characteristics of the raw data that are somehow relevant to the human perception of the media contents. Effective features combined with an appropriate distance or similarity measure facilitates the use of the statistical vector space model approach, which is the basis of most current multimedia analysis methods.

For video analysis in particular there is an opportunity to combine several data modalities, such as keyframe images, video motion, audio and text from speech recognition or subtitling. From these modalities diverse feature representations can be extracted to represent different relevant and complementary aspects of the underlying data.

Image features

Video analysis has its roots in image processing, and many techniques developed for content-based image retrieval can be used for video data as well. If a video shot is short and contains only one visually homogeneous scene, a single well-chosen keyframe can compactly express the most central visual characteristics of that shot. Therefore, one can use still-image features, often originally developed for image-only retrieval systems, as a way to compare video shots.

Many image features are global, i.e. calculated from all pixels of the image, thus representing global characteristics of the image. An increasingly popular alternative has been to calculate features separately for smaller image segments, for example for each block in a grid structure placed over the image. It is also possible to use automatic segmentation, where the image is split into visually homogeneous segments, for which features are calculated separately [4, 44].

Gaining increasing prominence today is the approach of characterising images by describing the statistics of their local feature descriptors. The local descriptors can be calculated for visually salient *interest points* [1]. For instance, the points can be edge or corner points where the the image content changes substantially. Another strategy is to sample image area evenly and calculate local descriptors for the sample of image locations. Histograms of robust, scale-invariant local descriptors—such as Scale-Invariant Feature Transform (SIFT) [33] and the Speeded Up Robust Features (SURF) [5]—provide the state-of-the-art image descriptors of the day. Histograms of localised features are also called "bag of visual words" (BoV) in analogy to the traditional bag-of-words approach in textual information retrieval. In this interpretation each histogram bin—representing a specific local pattern—is seen as a "visual word" in the vocabulary of all the histogram bins. The BoV features can be enhanced by calculating the histograms for different subdivisions of the image, in addition to the entire image [31, 60]. Another recent improvement to the BoV methodology is to use soft-assignment in histogram generation as demonstrated e.g. in [54, 59, 55].

The PicSOM system uses a wide range of image features that have been added over the years of its development. Many of our global image features are based on the standardised MPEG-7 descriptors [34]. We have used both the implementations of the MPEG-7 XM reference software and our own more efficient implementations of the following MPEG-7 features: *Color*

Layout, *Color Structure*, *Dominant Color*, *Scalable Color*, *Edge Histogram*, and *Region Shape*. Furthermore, PicSOM implements some non-standard image features developed in-house: *Average Colour*, *Colour Moments*, *Texture Neighbourhood*, *Edge Histogram*, *Edge Co-occurrence* [6] and *Edge Fourier*. These have been calculated either globally or for five spatial zones (centre, top, bottom, left, right) of the image. In the case of zoning, the final image-wise feature vector has been obtained as a concatenation of the zone-wise features. These features are explained in more detail in [45].

The BoV features of the PicSOM system are based on the *SIFT* local descriptors and the opponent colour space version of the *Color SIFT* [53] descriptor. We have employed two different strategies for selecting the points from which the local descriptors are extracted: the Harris-Laplace interest point detector and dense sampling of images. The codebooks have been generated with k-means and self-organizing map (SOM) clustering algorithms. In the experiments of this report, we have used codebooks with 1000 and 2000 codebook vectors. The BoV features represent the images as histograms of local descriptors over these codebooks.

Video features

In many cases the static visual properties of a video keyframe are not enough to describe the salient features of the full scene. In some situations the motion of objects or the camera might be semantically significant, for example in distinguishing between a ball that is rolling from one that is still. Also, the dynamic properties may in some cases make the computational learning problem easier. For example it may be easy for a human to recognise a running person even from a still keyframe image, but such videos are surely easier to distinguish based on the temporal properties of the person moving across the scene. It has been reported in various recent publications that using video features beyond the single keyframe approach can improve the results [45, 50, 18].

The set of video features used in PicSOM include the MPEG-7 *Motion Activity* descriptor, our own motion feature *KLT Histogram* [28], and temporal extensions of some of the still-image features described in the previous section. The temporal versions of originally still-image features are calculated as follows. The video shot is first divided into five non-overlapping parts with equal lengths. The resulting short video clips are called slices. All the frames of the five slices are then extracted, and each frame is divided into five spatial zones as described earlier in the image feature section. A feature vector is calculated separately for each zone, and then the zone-wise feature vectors are concatenated to form a vector depicting the whole frame. All the frame feature vectors of a video slice are then averaged to form the slice-wise feature vector. Finally, the feature vectors of the five slices are concatenated to form one shot-wise feature vector.

Audio features

Most video shots include a sound track, containing for example human speech, music or different environment sounds. Speech can often be automatically recognised and thus handled as text as will be described in the next section. Depending on the video analysis and retrieval task at hand,

analysing music and environment sounds may or may not be beneficial.

A certain distinctive musical tune may perhaps indicate the beginning of a news broadcast, or indicate for example the occurrence of an action scene in a movie. The music might also just be more or less random background music which does not correlate strongly with the visual contents and semantic interpretation of the scene. The same observation goes for environment sounds. For example, a crowd cheering in a football game is a strong cue of an important event such as a goal being scored, but the sound of a car passing by might be totally irrelevant. Hence, the usefulness of aural features is highly task-dependent and should be carefully considered and evaluated for each video corpus and search problem separately.

If the video shots are short and mostly relatively homogeneous, it is reasonable to use a global aural feature calculated for the entire video shot. A popular approach is to use the mel-scaled cepstral coefficients feature (MFCC), which is the discrete cosine transform (DCT) applied to the logarithm of the mel-scaled filter bank energies [9]. This feature is quite generic and commonly used in speech recognition, musical genre classification, and classification of the type of sound, e.g. whether it consists of speech, music or background noise. Instead of a single global feature the sound can also be segmented into shorter homogeneous parts, which can then be classified separately using features calculated locally [22].

In PicSOM we have used two implementations of MFCC. The first implementation produces 12 coefficients organised as a statistical vector in which the total power of the signal is appended, resulting in a feature vector of length 13. The second implementation is by the MUVIS group at Tampere University of Technology [21] and produces a 24-dimensional MFCC vector.

2.3 Textual search

Often, the video material includes textual data or meta-data that can facilitate text-based indexing and retrieval. Textual data for video shots may originate e.g. from speech recognition, closed captions, subtitles, or video OCR. As text-based information retrieval methodology is very mature and text indices can provide fast and accurate results [43, 3], an effective video retrieval system will definitely benefit from a text search component.

Text-based information retrieval is beyond the scope of this report, as the textual search module can easily be implemented as a separate component whose output is fused afterwards with the other modalities. The module can then utilise all common text processing methods, such as stemming and part-of-speech tagging, and existing tools for text-based indexing, such as the Lemur toolkit [40] or Apache Lucene [15]. In the experiments of this report, we use the Lucene search engine for indexing text obtained with automatic speech recognition and machine translation.

2.4 Concept detection

After having extracted low-level video features from each shot, supervised learning techniques can be applied in order to learn the associations between

the low-level features and the concepts in the annotations of the video corpus.

The standard practice in concept detection has been that the concept detectors are created with a procedure in which each detector is trained and each shot is classified separately. This can be seen as a manifestation of an implicit assumption that (i) the concepts in the shots and mutually independent, and (ii) there are no temporal dependencies between the adjacent shots. However, it is almost self-evident that the concepts or in videos have temporal structure, for example the concept distributions of subsequent shots are likely to be similar. Another nearly obvious characteristic of videos is that the concepts exhibited by a video shot are mutually dependent. For instance, the concept *snow* almost always implies *outdoor*, whereas concepts like *sports* and *weather forecast* are practically mutually exclusive. There might also be dependencies between different concepts that extend temporally over the time span of multiple shots. In a certain type of programming, for example, a shot of *weather forecast* could be a strong cue that some of the immediately subsequent shots would be likely to exhibit the concept *news anchor*.

In the PicSOM system, the shot-wise concept detection is followed by a temporal post-processing stage that can leverage on the inter-shot and inter-concept temporal structures. The details of the implementation of concept detection are addressed in Section 3.

2.5 Video search

The ultimate goal of video retrieval is to find relevant video content for a specific information need of the user. The conventional approach has been to rely on textual descriptions, keywords, and other meta-data to achieve this functionality, but this requires manual annotation and does not usually scale well to large and dynamic video collections. In some applications, such as YouTube, the text-based approach works reasonably well, but it fails when there is no meta-data available or when the meta-data cannot adequately capture the essential content of the video material.

Content-based video retrieval, on the other hand, utilises techniques from related research fields, such as image and audio processing, computer vision, and machine learning, to automatically index the video material with low-level features (see Section 2.2). Content-based queries are typically based on a small number of provided examples (i.e. *query-by-example*). The material of a video collection is ranked based on its similarity to the examples according to material's low-level features [49, 12, 48].

In recent works, the content-based techniques are commonly combined with separately pre-trained detectors for various semantic concepts (*query-by-concepts*) [51, 16]. It has been empirically observed that visual concept lexicons or ontologies are an integral part of effective content-based video retrieval systems. However, the use of concept detectors brings out a number of important research questions, including (i) how to select the concepts to be detected, (ii) which methods to use when training the detectors, (iii) how to deal with the mixed performance of the detectors, (iv) how to combine and weight multiple concept detectors, and (v) how to select the concepts used for a particular query instance.

In *automatic concept-based video retrieval*, the fundamental problem is

how to map the user's information need into the space of available concepts in the used concept ontology [38]. The basic approach is to select a small number of concept detectors as active and weight them based either on the performance of the detectors or their estimated suitability for the current query. Negative or complementary concepts are not typically used. In [38] the methods for automatic selection of concepts were divided into three categories: *text-based*, *visual-example-based*, and *results-based methods*. Text-based methods use lexical analysis of the textual query and resources such as WordNet [13] to map query words into concepts. Methods based on visual examples measure the similarity between the provided example objects and the concept detectors to identify suitable concepts. Results-based methods perform an initial retrieval step and analyse the results to determine the concepts that are then incorporated into the actual retrieval algorithm.

The second problem is how to fuse the output of the concept detectors with the other modalities such as text search and content-based retrieval. It has been observed that the relative performances of the modalities significantly depend on the types of queries [20, 38]. For this reason, a common approach is to use *query-dependent fusion* where the queries are classified into one of a set of pre-determined query classes (e.g. *named entity*, *scene query*, *event query*, *sports query*, etc.) and the weights for the modalities are set accordingly.

In addition to automatic retrieval, *interactive concept-based retrieval* constitutes a parallel paradigm. Interactive video retrieval systems include the user in the loop at all stages of the retrieval session and therefore call for sophisticated and flexible user interfaces. A global database visualisation tool providing an overview of the database as well as a localised point-of-interest with increased level of detail are typically needed. Relevance feedback can also be used to steer the interactive query toward video material the user considers relevant [25].

In recent works, semantic concept detection has been recognised as an important component also in interactive video retrieval [16], and current state-of-the-art interactive video retrieval systems (e.g. [10]) typically use concept detectors as a starting point for the interactive search functionality. A specific problem in concept-based interactive retrieval is how to present to a non-expert user the list of available concepts from a large and unfamiliar concept ontology.

The details of the implementation of video search in the PicSOM system are addressed in Section 4.

3 CONCEPT DETECTION IN PICSOM

The PicSOM multimedia retrieval system includes a supervised concept detection subsystem trained in the preparing phase of the video corpus. Figure 1 illustrates the overall architecture of this system. All the K concepts are first detected from each shot, based on the shot's low-level features, K being the number on concepts that have been annotated in the training part of the video corpus. This step results in a K-dimensional vector of detection scores. After the shot-level concept detection, the scores are re-adjusted in a

post-processing step according to the score vectors of temporally neighbouring shots, based on the estimated likelihood of observing particular temporal concept patterns.

The remaining part of this section describes the techniques that are applied in the concept detection subsystem. Section 3.1 details the shot-wise concept detection module. The temporal post-processing techniques are described in Section 3.2.

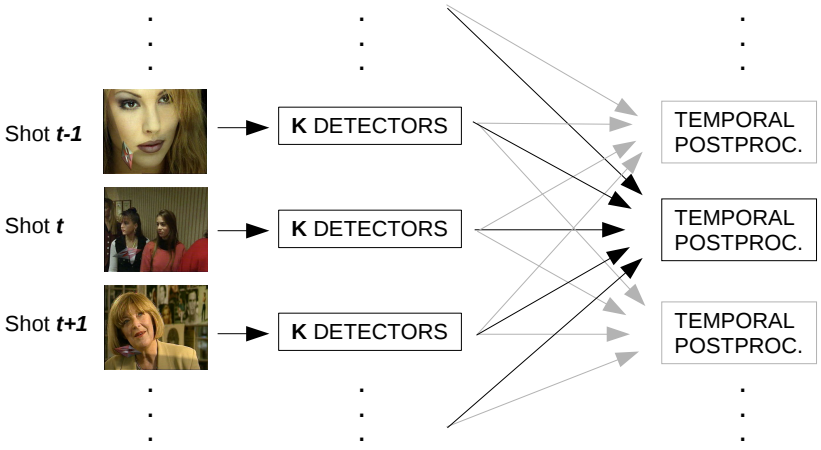

Figure 1: General architecture of the PicSOM concept detection subsystem. Here K denotes the number of concepts that is to be detected.

3.1 Shot-level concept detection

Figure 2 illustrates the shot-wise concept detection submodule. The task is addressed with a well-established fusion-based architecture. The fusion-based approach has been used in the PicSOM system for various image content analysis tasks since late 1990s (e.g. [29, 30, 56, 58]) and is common also in other well-performing state-of-the-art image and video analysis systems (e.g. [50, 39]). In our approach, dozens of supervised probabilistic detectors are first trained for each concept, based on the different shot-wise low-level features, detailed in Section 2.2, and their early-fusion combinations. The feature-wise detector outcomes are then fused in a post-classifier fusion (also called late fusion) step. The outlined shot-level detection architecture contains a number of components that can be implemented in a number of alternative ways. In the following we describe the techniques we have implemented in the PicSOM system.

Supervised detector algorithms

Given the extracted shot-wise features, the first stage in our fusion-based concept detection approach is the feature-wise supervised detection. There each concept and feature is treated symmetrically, i.e. every concept is detected with the same algorithm using each of the features. As some of the subsequent stages in the concept-detection pipeline are supervised, the feature-wise concept detection of this stage needs to be performed also for the shots in the training part of the video corpus, not only for the test part. For this purpose, we employ 10-fold cross-validation. As the supervised detection al-

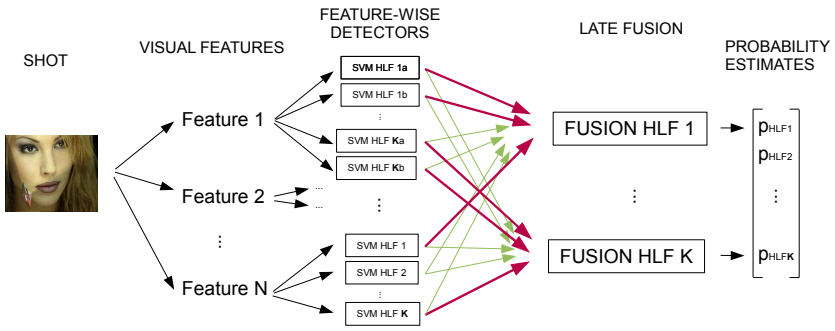

Figure 2: Fusion-based shot-wise concept detection module in PicSOM system. K denotes the number of concepts that is to be detected. The solid red lines between the feature-wise detector and fusion stages are intra-concept connections, the dashed green lines represent cross-concept links (cf. Section 3.1).

gorithm we have used both the Self-organising Map (SOM) based classifiers as well as Support Vector Machines (SVMs). The advantage of SVM-based detectors is their much greater accuracy, whereas SOM-based classifiers are much faster to train. In particular, a new SOM-based classifier is very fast to generate to discriminate any pair of two semantic classes once the data set specific preparation step has been performed. This makes the SOM-based approach scalable to extremely large concept ontologies. For example, it has enabled us to model a total of 294 concepts from the LSCOM ontology [36] without excessive computational requirements [45].

The SVM implementation used in the PicSOM system is an adaptation of the C-SVC classifier of the LIBSVM software library [7]. We use the radial basis function (RBF) SVM kernel

$$g_{\mathrm{RBF}}(\mathbf{x}, \mathbf{x}') = \exp\left(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2\right) \tag{1}$$

for all the shot-wise features and also have the option to use the χ^2 kernel

$$g_{\chi^2}(\mathbf{x}, \mathbf{x}') = \exp\left(-\gamma \sum_{i=1}^{d} \frac{(x_i - x_i')^2}{x_i + x_i'}\right) \tag{2}$$

for histogram-like visual features.

The free parameters of the SVMs are selected with an approximate 10-fold cross-validation search procedure that consists of a heuristic line search to identify a promising parameter region, followed by a grid search in that region. To speed up the computation, the data set is radically downsampled for the parameter search phase. Further speed-up is gained by optimising the C-SVC cost function only very approximately during the search. For the final detectors we also downsample the data set, but less radically than in the parameter search phase. Usually there are much fewer annotated example shots of a concept (positive examples) than there are example shots not exhibiting that concept (negative examples). Consequently, for most of the concepts, the sampling is able to retain all the positive examples and just limit the number of negative examples. The exact amount of applied sampling varies according to the computation resources available and the required accuracy

of the outputs. Generally we have observed the downsampling always to degrade detection accuracy.

The construction of the SOM-based classifiers begins with quantising the feature spaces using the TS-SOM [24] algorithm, a tree-structured variant of the SOM. Typically we have used TS-SOMs with four stacked levels, the bottom levels measuring 256×256 map units. This preparation step needs to be performed only once for each feature type in a video corpus and after it, generating a classifier for any binary partitioning of the training data is very fast. Any partitioning is characterised by the division of the training data into positive and negative examples. The classifier for the partitioning is created by subtracting the proportion of negative examples that fall into each SOM unit or quantisation bin from the corresponding proportion of positive examples, which assigns a classification score to each quantisation bin. After this, the scores are low-pass filtered on the two-dimensional TS-SOM grid surface, taking advantage of the topology-preserving characteristic of the SOM clustering and efficiently emphasising the differences between the feature space regions where positive and negative examples are well separated, or occur mixed with each other. After preparing a classifier, a detection score is associated with each quantisation cell of a feature space. Assigning a feature-wise detection score to an unannotated video shot is then simple: the extracted feature vector of the shot is quantised using the same quantisation scheme and the shot receives the detection score of the quantisation cell into which its feature vector is mapped. [26] provide more details on the SOM-based concept detection.

Selection of detectors to be fused

The supervised training step produces a large number of feature- and concept-specific detectors. The question is then, which of these detectors should be combined in a post-classifier fusion step for the detection of each concept. As an answer to this question, we consider alternative strategies for the selection of the set of feature-wise detectors that are fused for detecting each concept. Our basic alternative is to fuse the outcomes of the detectors that were trained for detecting this particular concept. We also implement an attempt to exploit cross-concept correlations by including detectors trained for all the other concepts in the fusion as well. This idea is realised as a two-stage fusion scheme where the detectors for each concept are first fused separately. In the second stage, the concept-wise fusion is repeated otherwise in the same way, but the set of detector outcomes to be fused is augmented with fused detection outcomes for all the other concepts and their temporally smoothed versions.

Fusion algorithms

After having selected which set of detectors to combine in the fusion step, one has to decide which algorithm to use for the fusion. Once again, the PicSOM system includes several alternative algorithms. As a baseline approach we form the geometric mean of all the detector outcomes. Besides this unsupervised fusion approach, we also implement several supervised fusion methods that make use of the detector outcomes for the training set.

One of the supervised techniques is SVM-based fusion employing RBF

kernels, another one the Bayesian Binary Regression (BBR) [14]. The further implemented alternatives are variations of the scheme where the basic fusion mechanism is the geometric mean, but the mean is calculated only of a subset of the detector outcomes, selected by a sequential forward-backward search (SFBS).

In addition to basic SFBS, we implement the idea of partitioning the training set into multiple folds. In our implementation we have used a fixed number of six folds. The SFBS algorithm is run several times, each time leaving one fold outside the training set. The final fusion outcome is the geometric mean of the fold-wise geometric means. For later reference, we denote this fusion algorithm multifold-SFBS.

We also consider reserving a part of the training set for validation and early-stopping the search based on the performance in this validation set. This early-stopping can be combined with both the basic SFBS and multifold-SFBS algorithms. For the basic SFBS, one sixth of the training data is used as a validation set. In case of multifold-SFBS, the left-out fold for each fold-wise run is re-used as the validation set.

3.2 Temporal post-processing

For temporal post-processing of the fusion outcomes, the PicSOM system implements techniques published in [61]. The techniques operate on a stream of K-tuples corresponding the concept detector outputs for the sequential video shots, where K is the number of the detected concepts. The methods thus ignore the absolute timing and duration of the video shots, preserving only their ordering. Methodologically, two types of techniques are combined in the temporal post-processing step: concept-wise N-gram modelling and clustering of temporal neighbourhoods in the concept space.

N-gram models

The N-gram model is applied to each concept individually. In the following, $c_n \in \{0, 1\}$ is an indicator variable of the occurrence of the concept to be detected at time instant n and $s_n \in \mathbf{R}$ is the output of the corresponding concept detector. H_n denotes the recursive prediction history known at time instant n, extending $N-1$ steps backwards in time:

$$H_n = \{\hat{p}(c_{n-i}|s_{n-i}, H_{n-i})\}_{i=1}^{N-1} \ . \tag{3}$$

Using this notation, we can write the recursive N-gram model as

$$\hat{p}\,(c_n|s_n, H_n) \propto \hat{p}(s_n|c_n)\hat{p}\,(c_n|H_n) \tag{4}$$

if we assume the conditional independence of s_n and H_n given c_n, i.e.

$$\hat{p}(s_n|c_n, H_n) = \hat{p}(s_n|c_n) \ . \tag{5}$$

In the recursive model

$$\hat{p}\,(c_n|H_n) \quad = \quad \sum_{c_{n-1}} \cdots \sum_{c_{n-N+1}} p_0(c_n|c_{n-1}, \ldots, c_{n-N+1}) \prod_{i=1}^{N-1} \hat{p}\,(c_{n-i}|s_{n-i}, H_{n-i}) \ . \tag{6}$$

Here p_0 is the marginalised N-gram probability that is estimated from the training data. The N-gram model is initialised in the beginning of each video by using models of lower order, e.g. bigram model is used on the second time instant. The conditional distributions of detector outputs $\hat{p}(s_n|c_n)$ are modelled as exponential distributions

$$\hat{p}(s_n|c_n) = \frac{1}{\lambda} e^{-s_n/\lambda}. \tag{7}$$

For concept-wise parameters λ we use the maximum likelihood estimates

$$\hat{\lambda} = \frac{1}{M} \sum_n s_n \ , \tag{8}$$

where the summation is over the M shots of the training set.

In addition to this causal model, we also form the corresponding anticausal model that is obtained by reversing the time flow. The causal and anticausal models are then combined by logarithmic averaging of the model outcomes.

Clusterings of temporal neighbourhoods

The N-gram model is augmented with information C_n that is obtained by clustering the shot-wise detector outputs within temporal neighbourhoods around the prediction time instant n. The clustering is based simultaneously on all the detected K concepts. For clustering, the shot-wise detector outputs are converted to K-dimensional probability estimate vectors. The sum of the vectors is aggregated within a temporal window, and the sum vectors are clustered with the LBG algorithm [32]. Based on some earlier experimentation, the number of clusters has been chosen to be 16 in the PicSOM implementation.

The cluster information is combined with the N-gram model by estimating the N-gram model separately for each cluster. This results in cluster-specific models $p_0(c_n|C_n, c_{n-1}, \ldots, c_{n-N+1})$. The cluster-specific detector outcome distribution $\hat{p}(s_n|C_n, c_n)$ is modelled as a linear interpolation between the global logistic model and a logistic model estimated for each cluster separately. In our implementation, several different clusterings are combined by taking logarithmic averages of the detection probability estimates based on each clustering. The different clusterings result from neighbourhoods of different time spans.

4 VIDEO SEARCH IN PICSOM

The PicSOM multimedia retrieval system supports both automatic and interactive content-based search modes. In this report, however, our emphasis is on automatic video retrieval.

4.1 Automatic search

Figure 3 gives an overview of the automatic video search process within PicSOM, with a detailed view of the concept-based submodule. In the top part

of the figure a search query is presented, typically containing a *text query* and possibly also *visual examples*. The visual examples may consist of videos and/or images, demonstrating the visual properties of the desired retrieval response. Either or both of these two modalities of the search query are then used as input to the three parallel submodules of the search system: *text search*, *concept-based search* and *content-based search*. Based on its input, each module produces an estimate of the relevance of each database video to the given query. These scores are finally fused to produce the final search result which is a list of video shots ordered with decreasing estimated relevance to the query.

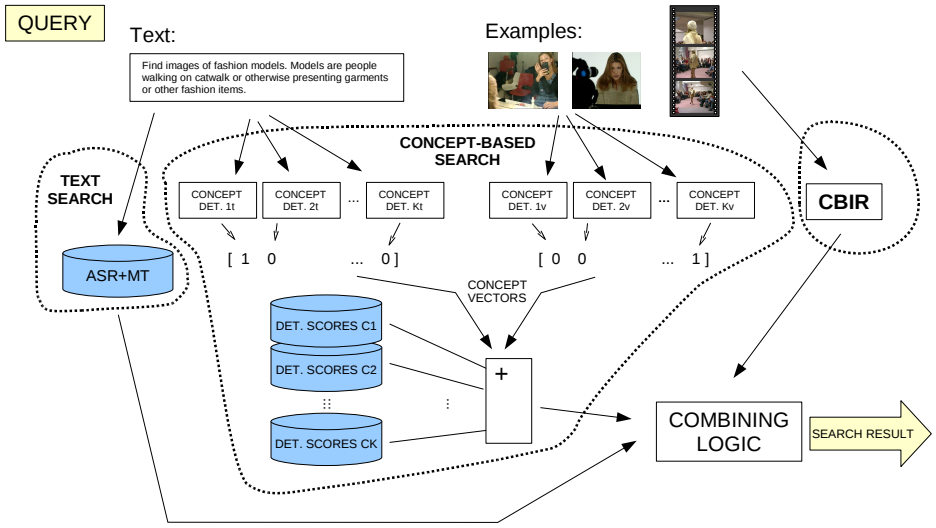

Figure 3: General architecture of the PicSOM search module.

Text-based search

The text search submodule uses only the text query and performs a traditional information retrieval role for the videos of the database that have associated textual data, e.g. automatically recognised speech or closed-captioning text. We use the Apache Lucene as the text search engine, along with the included Snowball stemmer with its default stop word list. The textual queries are analysed using the Stanford part-of-speech tagger [52], stemmed, and expanded with synonyms using WordNet [13].

Concept-based search

The concept-based module utilises the high-level semantic concept detectors discussed in Section 3. This functionality is highly important as using pre-trained concept detectors have been observed to consistently surpass the retrieval performance of direct content-based retrieval [27]. Here both the text query and the visual examples can be used for the selection of which concept detectors to enable. The textual query is compared to words associated with the concepts, for example the concept *animal* may be triggered for names of common animals such as dog, cat, horse etc. These trigger words are generated by taking the concept name itself as the initial word or words and expanding with WordNet synonyms. These lists are then cleaned up

by hand (without knowledge of the particular search topics). The explanation for including this manual step is that similar automatic process would be straightforward to implement, if only one would be willing to invest the required programming effort in details of language processing. In addition to triggering based on individual words, we also consider more complex text-matching triggers via Perl regular expressions.

The visual examples may also be used for selection of concepts. The examples are of the same modality as the database objects, i.e videos or keyframe images, and can thus be scored using the concept detectors that were already trained in the preparing phase (excepting, of course, the temporal correlations). A particular concept is then included if the sum of scores over all examples exceeds a heuristically determined threshold.

In some setups, it is profitable to use other specialised concept detectors, e.g. to detect faces, empty shots, test screens, presence of motion, or speech. The PicSOM system is also able to make use of outputs of external concept detectors, such as those contributed by the other research groups taking part in the TRECVID evaluations. Such detector scores can simply replace the output of one of the internal detector scores $c_1, \ldots, c_K$ shown in Figure 3.

Content-based search

The content-based retrieval part of the PicSOM system makes use of the supervised learning method based Self-Organising Maps (SOM), described in Section 3.1. Content-based search starts with extracting shot-level features from the image and video examples of a search query. The features are the same shot-level features that are extracted from the shots of the video corpus in the preparation phase of the system (Section 2.2). Several probabilistic SOM classifiers—each based on different shot-level feature—are then used in parallel for estimating the feature-wise similarity of the video shots in the video corpus with the query examples. A global similarity score is formed by combining the feature-wise similarities. It should be remembered the SOM mappings—once formed during the video corpus preparation phase—are static during the search phase and therefore the search time application of SOM-based classifiers is computationally extremely light.

Combining logic

In the final stage the scores from all the submodules of the search system are combined. This is accomplished by linearly combining the subsystems' outputs. The combination weights vary depending on the quality and availability of textual and concept annotations, video-corpus-specific relative importance of input modalities, and other particularities of the corpus. For some queries one or two of the submodules will produce a zero output, for example if there are no visual examples, or if no concepts can be matched to the query. In such cases the system relies on the other available submodules. We can adjust the combining logic by detecting special query types, for example if a named entity (i.e. a famous person or location) is detected, we can increase the weighting of the textual search for that name. Finally, after combining the scores from all the three submodules, an ordered list of video shots is returned, with the most relevant shot to the query at the top.

4.2 Interactive search

In addition to automatic video search, the PicSOM system can also be used for interactive retrieval via an intuitive web-based user interface. The retrieval process then initially follows the scheme described above, but the search result is presented to the user who can provide relevance feedback by marking a number of the received video clips as relevant or non-relevant. These will then give an updated set of positive and negative examples which will initialise the next retrieval round. Naturally, such relevance feedback cannot be used in the automatic retrieval mode. However, the feedback can be simulated using *pseudo relevance feedback*, where an initial retrieval round is performed and the best-scoring shots are marked as additional positive examples and used to re-rank the retrieval results.

5 EXPERIMENTS

In this section, we describe the experiments we have performed in high-level concept detection and automatic video search using the TRECVID data of 2008–2009, and present an analysis of the results. Our experiments are based on our submissions to corresponding TRECVID evaluations [28, 46] but we augment the submitted results with additional experiments based on retrospective analysis of the annual results.

5.1 TRECVID evaluation campaign

The video material and the search topics used in these experiments are from the TRECVID evaluations [47] in 2008–2009. TRECVID is an annual workshop series organised by the National Institute of Standards and Technology (NIST) and arguably the leading venue for evaluating research on content-based video analysis and retrieval. It provides the participating organisations large test collections, uniform scoring procedures, and a forum for comparing the results. Each year the TRECVID evaluation contains a set of video analysis tasks, such as high-level feature (i.e. concept) extraction, video search, video summarisation, and content-based copy detection. For video search, TRECVID specifies three modes of operation: fully-automatic, manual, and interactive search. Manual search refers to the situation where the user specifies the query and optionally sets some retrieval parameters based on the search topic before submitting the query to the retrieval system. In the experiments of this report, we focus on the settings of high-level feature extraction (HLFE) and automatic video search tasks of TRECVID 2008 and TRECVID 2009.

In 2008–2009 the type of video material used in TRECVID consisted of documentaries, news reports, and educational programming from Dutch TV. The video data is always divided into separate development and test sets. The same development set of approximately 100 hours in length is used both in 2008 and 2009. The amount of test data was approximately 100 and 280 hours in 2008 and 2009, respectively. To obtain training data for the high-level feature extraction or concept detection task, a collaborative annotation

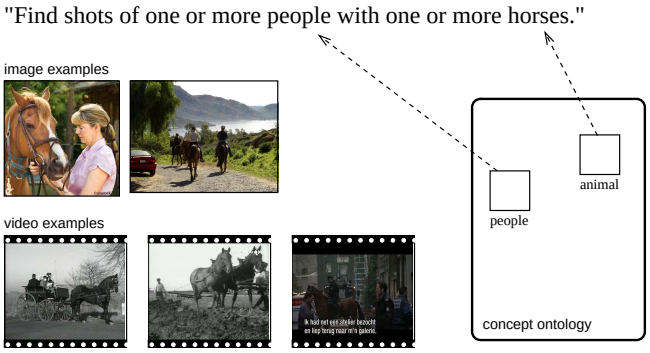

Figure 4: An example TRECVID search topic, with one possible lexical mapping from a concept ontology with two matching concepts.

effort [2] was organised in both years.

NIST also defines sets of standard search topics for the video search tasks and then evaluates the results submitted by the participants. The search topics contain a textual description along with a small number of both image and video examples of an information need. Figure 4 shows an example of a search topic, including a possible mapping of concept detectors from a concept ontology based on the textual description. The number of topics evaluated for automatic search was 48 in the year 2008 and 24 in 2009. The search topics are available in the TRECVID guidelines documents at *http://www-nlpir.nist.gov/projects/trecvid/*.

The video material used in the search tasks is divided into shots in advance and these reference shots are used as the unit of retrieval [41]. The shot segmentation step resulted in 36 000 shots for the training set and 97 000 shots for the full 2009 test set. The output from an automatic speech recognition (ASR) software is provided to all participants. In addition, the ASR result from all non-English material is translated into English by using automatic machine translation. It is therefore quite unsurprising that the quality of the textual data is remarkably poor and pure text queries can only obtain a very modest performance.

Due to the size of the test corpora, it is infeasible within the resources of the TRECVID initiative to perform an exhaustive examination in order to determine the topic-wise ground truth. Therefore, the following pooling technique is used instead. First, a pool of possibly relevant shots is obtained by gathering the sets of shots returned by the participating teams. These sets are then merged, duplicate shots are removed, and the relevance of only this subset of shots is assessed manually. It should be noted that the pooling technique can result in the underestimation of the performance of new algorithms and, to a lesser degree, new runs, which were not part of the official evaluation, as all unique relevant shots found in them will be missing from the ground truth.

The basic performance measure in TRECVID is the *average precision* (AP):

$$\mathrm{AP} = \frac{\sum_{r=1}^{N} P(r) \cdot R(r)}{N_{\mathrm{rel}}} \tag{9}$$

where r is the rank, N is the number of retrieved shots, $R(r)$ is a binary function stating the relevance of the shot retrieved with rank r, $P(r)$ is the

precision at the rank r, and N_{rel} is the total number of relevant shots in the test set. In TRECVID search tasks, N is set to 1000. The mean of the average precision values over a set of queries, *mean average precision* (MAP) has been the standard evaluation measure in TRECVID. In recent years, however, average precision has been occasionally replaced by *inferred average precision* (IAP) [62], which approximates the AP measure very closely, but requires only a subset of the pooled results to be evaluated manually. The query-wise IAP values are similarly combined to form the performance measure *mean inferred average precision* (MIAP).

5.2 Experiments with concept detection techniques

For measuring the accuracy of various concept detection techniques, we use the experimental setup and the same MIAP performance metric as in the official high-level feature extraction task of the TRECVID evaluations of 2008 and 2009. In most of the results, the fluctuation of IAP results for any individual concept was observed to be so strong that we felt we could not reliably distinguish different algorithmic techniques. Therefore, we settle for drawing conclusions on average only, i.e. on basis of MIAP which we consider more reliable. However, there probably are genuine differences between the concepts—one technique might really be the most suitable for detecting a particular concept, while other techniques may work better for other concepts. Such concept-specific differences will be missed by our average case analysis. As an example of the types of concepts we are dealing with, Table 1 lists all concepts detected in TRECVID 2009.

In the experiments reported in the following subsections we have used Support Vector Machine (SVM) as the feature-wise supervised concept detector component of our system. For reference, we also evaluated the SOM-based detectors baseline in the same setting, but the resulting detector accuracy was considerably worse than that of the SVMs'. The shot-wise feature sets that we have used as a starting point for the concept detection consist solely of various combinations of visual features, i.e. keyframe and video features. Audio and text have not been used in these experiments.

Table 1: The 20 concepts detected in TRECVID 2009 high-level feature extraction task

Classroom	Person playing a musical instrument	Hand
Chair	Person playing soccer	People dancing
Infant	Cityscape	Nighttime
Traffic	Person riding a bicycle	Boat or ship
Doorway	Female human face closeup	Telephone
Airplane flying	Person eating	Singing
Bus	Demonstration or protest	

Table 2: Concept detection accuracy (MIAP) based on various BoV image features.

Feature	sampling	histograms	spatial partitioning	MIAP
Color SIFT	dense	soft histograms	spatial pyramid	0.1166
Color SIFT	dense	soft histograms	global	0.1031
Color SIFT	interest points	soft histograms	spatial pyramid	0.1014
Color SIFT	interest points	soft histograms	global	0.0961
Color SIFT	sense	hard histograms	global	0.0988
SIFT	interest points	hard histograms	global	0.0832

Shot-wise features

As a preparation for the post-classifier fusion, we trained a number of individual SVM detectors, each based on a single shot-level feature. This lets us compare different shot-level features in terms of their detection accuracies, although the individual detectors are only used as components of the final fusion-based detection subsystem.

The best individual feature performances we observed resulted from histograms of local image features collected according to the bag-of-visual-words (BoV) paradigm, i.e. variants of SIFT and Color SIFT features. Table 2 compares different BoV feature variants in terms of MIAP. As expected, Color SIFT outperforms normal SIFT. Dense sampling is a more effective approach than interest point detection. The soft histogram technique and spatial pyramids improve the performance of the BoV features as well. These results hold on average, but concept-wise differences are large. It does not seem likely that all the differences would result from statistical fluctuations. Table 3 lists the most accurate non-BoV features.

We tried two different degrees of downsampling of the training data for SVM training. In the first, more radical sampling alternative we retained at most 5000 training examples. The other alternative was to retain at most 12,000, resulting in more elaborate and time-consuming SVM training. On average, the more elaborate SVM training produced somewhat more accurate detectors than the faster training for 80% of the visual features. For histogram-type features, the χ^2 SVM kernels produced clearly more accurate results than the RBF kernel.

In these experiments, augmenting the feature set with early fusion did not improve the overall performance of the system even though in some earlier studies early fusion has been beneficial [57]. On the level of single SVM detectors, combined features surely resulted in more accurate detectors than their individual constituent features, the best early fusion combination of image features had MIAP 0.0601. One explanation is that the best individual features (i.e. BoV and video features) were not involved in the early fusion.

Usefulness of cross-concept correlations

On average, we were not able to exploit cross-concept correlations for our advantage in this experimental setup when selecting the set of SVM detectors to be fused. That is, we did not gain any advantage of the two-stage fusion algorithm. However, there exists empirical evidence that one can

Table 3: Selection of feature-wise concept detection accuracies (MIAP).

Feature	type	MIAP
Edge Histogram	video	0.0625
Color Moments	image	0.0438
MPEG-7 Edge Histogram	image	0.0417
Edge Histogram	image	0.0403
Color Layout	video	0.0340
Color Layout	image	0.0309
Scalable Color	image	0.0330
Edge Fourier	image	0.0290
MPEG-7 Color Structure	image	0.0263

often benefit somewhat by taking the correlations between concepts into account [42, 61, 39].

Still, it can be questioned whether one should even expect to benefit from instantaneous cross-concept co-occurrences. In the limiting case with enough training data and flexible enough learning algorithms for accurately approximating the conditional probabilities involved, the answer is namely negative as long as the training data is fully annotated and all the concepts are detected based on the same visual features. The situation is different if the detections for different concepts come from (partially) independent sources. In the practical case with limited training data and learning algorithms, there probably is no general guarantee to one direction or another. For non-instantaneous cross-concept correlations the situation is somewhat different if the dependency occurs outside the temporal window whose shot-wise features are being used for the concept detection.

There exist several potential explanations for our negative experimental results that seem to contradict the experience of others. It is possible that there simply are no such non-instantaneous cross-concept dependencies in the TRECVID 2008–2009 HLF annotations that could be exploited. On the other hand, we might not have been able to learn such dependencies reliably from the training data due to limited amount of independent training examples. It is also questionable whether the temporal dependency patterns are similar enough in the training and test portions of the video corpus for being exploited in our current straightforward fashion.

Fusion algorithms

We performed a preliminary evaluation of the various post-classifier fusion algorithms in a setting where the annotated part of the video corpus was further partitioned to a training and validation part in 2:1 proportions. In this preliminary experiment SVM and BBR based fusion algorithms were significantly and consistently outperformed by geometric mean based fusion algorithms, both by the unsupervised basic version and by the supervised SFBS variants. Moreover, the SVM and BBR fusion mechanisms are computationally much more costly. Consequently, the remaining evaluation with the full data set was constrained to the variants of geometric mean fusion.

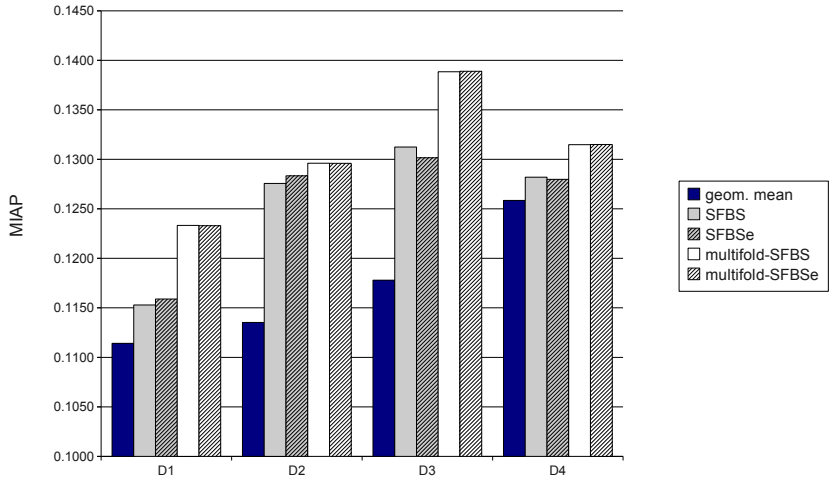

Figure 5: Comparison of algorithms for selecting detectors for geometric mean fusion for four different sets of detectors D1–D4. The SFBSe and multifold-SFBSe bars with diagonal hatching correspond to algorithms with early stopping.

Figure 5 compares different geometric mean based fusion algorithms with the whole video corpus and four different sets D1–D4 of detectors to be fused. These sets result from different sets of shot-wise features, different SVM training parameters and different cross-concept strategies. The number of fused detectors ranges between 77 (D1) and 26 (D4). We can see that the geometric mean of all detectors (the leftmost bar) is always inferior to methods where the set of detectors is selected with sequential forward-backward search (SFBS). This has not always been the case in our earlier experiments as SFBS easily overfits to the training data. The figure also shows that multifold-SFBS performs better than the basic SFBS. Early stopping has no essential effect on the average performance. It, however, seems to increase the variance of the results. These experiments thus confirm that early stopping is not a suitable way of regularising SFBS.

The results of this section—when compared with the MIAP values of the best individual features in Section 5.2—can be used to confirm the observation that fusion of features usually outperforms individual features, even if the individual features are rather good compared with other of the fused features. With a good fusion algorithm, benefit can be obtained from individually rather badly-performing features. In one experiment we picked approximately 75% of the best features for fusion, thus leaving just the worst performing 25% of the features outside. Still, with the multifold-SFBS fusion algorithm the fusion accuracy improved when the worst 25% were returned to the feature set. With a less-developed fusion algorithm, the saturation point is reached earlier where further addition of features no longer improves the fusion result. An example of this behaviour can be seen in Figure 5 when comparing sets of detectors D3 and D4. Here set D3 is a superset of D4 having almost three times as many detectors. When the geometric mean fusion is used, better performance is obtained by using the smaller set D4, whereas with the more advanced SFBS fusion algorithms the situation reverses: benefit can be obtained from the extra detectors in D3.

Temporal post-processing

Figure 6 shows the effect of temporal post-processing for a selection of shot-wise fusion-based detectors F1–F4. The detectors employ different sets of shot-wise features and fusion algorithms. From the figure we can observe that the N-gram post-processing (bars with diagonal hatching) improves MIAP markedly over the baseline with no post-processing (white bars). We evaluated two strategies for choosing the order of N-gram models. In one strategy, the N-gram order was selected for each concept separately based on a validation experiment performed with 2:1 split of the training data. The other strategy was to choose the order globally, i.e. select the order of N-grams that resulted in the best mean performance over all the concepts in the validation experiment. As the results show, the global order-selection approach works somewhat better. This may indicate that our validation setup was inadequate: a better use of the training data could have been made with a more elaborate cross-validation setup. On the other hand, there might simply be too few independent temporal patterns in the training data, just as argued before. Similarly, the explanation of different temporal patterns in training and test data might be valid also here. This is supported by the anecdotal piece of evidence that in many cases the validation experiment failed to indicate the usefulness of N-grams for concept *people dancing*, although in the test data N-grams improved the detection accuracy hugely, by almost 200%.

When selecting the order of N-grams globally, the selection resulted in almost all cases in the selection of order eight, the maximum value that was considered. Generally, in the beginning the mean performance seems to increase rapidly with increasing N-gram order. Gradually the performance starts to saturate and eventually begins to degrade slowly when the order is further increased.

The post-processing methods marked with identifier "any" (solid dark bars) refer to the concept-wise selection of the post-processing method from a larger pool of methods according to the best performance in the 2:1 validation experiment. The pool includes the clustering-based inter-concept techniques of Section 3.2 in addition to the N-gram methods. From the

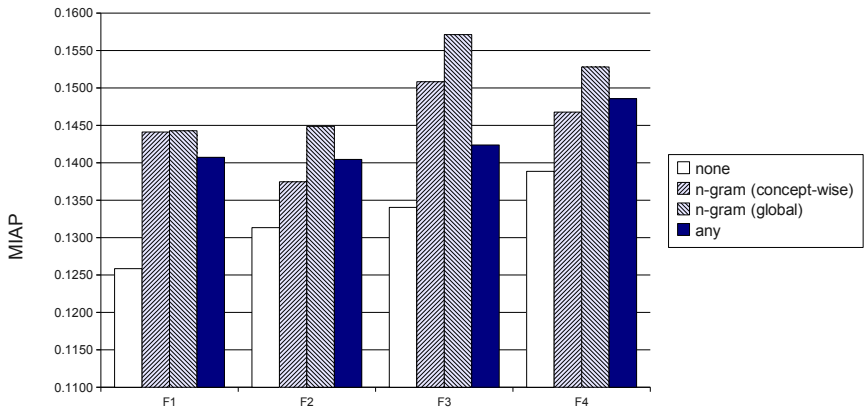

Figure 6: The effect of applying temporal post-processing on four different shot-wise fusion based detectors. The bars with diagonal hatching correspond to the N-gram technique with two different strategies for order selection.

figure we observe that the inter-concept methods do not bring any improvement over N-grams in this concept detection task. This has not always been the case. For example, the inter-concept techniques turned out to be useful in an experiment with data sets of TRECVID 2005–2007 [61], although in that case the baseline detectors were based on less powerful SOM detectors instead of SVMs.

5.3 Video search experiments

In this section we describe our automatic search experiments that have been performed using the setups of TRECVID 2008 and 2009 automatic search tasks, i.e their query topics and test corpora. In addition, we present experiments where interactive search has been simulated. As described in Section 5.1, in TRECVID search tasks the system is given a set of search topics, each containing a textual description and image and video examples of the desired result. The number of visual examples is usually quite small, on average six videos and two images per query topic.

Due to the importance of the mid-level concepts in obtaining state-of-the-art retrieval performance, the main challenge in the TRECVID automatic search setting is selecting the right concepts for a particular query topic. Content-based retrieval is used only for topics to which no concepts could be matched.

Despite the low quality of the textual data, the inclusion of the text search component generally improves the overall results. We thus include text search in all search experiments in this report. The machine-translated ASR data is used on the shot level, i.e. all snippets of text are mapped to the corresponding shots and the shot-wise collections of text are used as the set of retrieved documents. The shot-wise text search scores are then spread to the temporally neighbouring shots using a triangular kernel of five shots in width. For more details, see [28, 46].

The mean results over all query topics of TRECVID 2008 and 2009 for different concept lexicons and concept selection strategies are summarised in Table 4. To set a baseline, the first row shows the results of the purely content-based search, i.e. search without utilising concept detector outputs.

Table 4: Search results with various concept lexicons

	2008 (MIAP)	2009 (MAP)
no concepts	0.0110	0.0095
own (text)	0.0312	0.0593
own (visual)	0.0341	0.0397
own (visual+text)	0.0381	0.0599
donated (text)	0.0678	0.1021
donated + own (text)	0.0735	0.1051
donated + own (visual+text)	0.0719	0.1048

Comparison of concept lexicons

We used two concept lexicons, with differing detectors, in the experiments described in this report. The first is a set of 30 of our *own* concept detectors, consisting of 20 detectors submitted to the HLFE task of TRECVID 2009, and 10 detectors for concepts that were evaluated in 2008 but not in 2009 [28, 46]. Rows 2–4 in Table 4 show the results for these runs with different concept selection strategies to be explained in detail in the next section.

The second, *donated* set consists primarily of the 64 concept detectors shared by MediaMill [50]. Secondarily, we included concepts from the contributed CU-VIREO374 [19] detector set for words in the textual queries for which no match was found among the MediaMill concepts. The semantic concepts of this set are a superset of those in our own set, however with different detectors for the concepts. The results for this set are given on row 5 of Table 4, and a union of both detector sets are used on the last two rows.

The results clearly show a marked improvement when using the larger donated concept lexicon. This is probably best explained by the fact that from a wider set of concepts, it is more probable to find a concept matching a given query. This is illustrated by the fact that while using only our own concepts, 50% of the topics could not be matched to any concept by lexical selection, while with the donated set this was reduced to 7%.

Furthermore, we can conclude that the union of both sets of detectors (donated + own) provides a small additional improvement in retrieval accuracy. Since our own concept set is a subset of the donated set, the difference must lie in the concept detectors themselves. Even for the same semantic concept the two different detectors provide somewhat complementary information.

Matching query topics to concepts

Given a concept lexicon, the next task is to select which concepts to activate for a particular query topic. We have tried both text-based (lexical) and visual-example-based matching as explained in Section 4.1. For the donated concepts, we had only access to the final detection scores for the TRECVID development set, and thus visual matching could not be used for them. In Table 4 the different concept selection schemes are given in parentheses after the concept lexicon name.

Looking first at our own concepts only, the picture is somewhat mixed. In both datasets using the union of the lexical and visual-example-based selection is the best approach, however the improvement over using purely text-based matching is not substantial. In the 2009 dataset, using only visual-based matching performs considerably worse, while in 2008 it is even slightly better than the lexical matching. The donated set uses only lexical matching, but including the visual-example-based matches from our set slightly degrades the mean result, which is somewhat surprising. We also experimented with selecting the same concepts in the donated set as were selected in our own set by visual examples. This caused a further decrease of the score (not shown here).

For the visual-example-based matching with our own concepts, there is a threshold or sensitivity level for how many concepts should be included. For the TRECVID submissions we activated a given concept for those search topics for which the matching score was more than six times larger than the

median score for that concept. This limit was chosen heuristically based on our previous experience (for more details, see [46]). After the TRECVID evaluations we performed some after-the-fact tests for the optimal value of this sensitivity limit. It turns out that the limit chosen heuristically was very close to the optimal one for the 2008 set, but for 2009 some improvement could be made by including more concepts (accepting scores twice the median was optimal). In general it seems that on average 2–3 concepts per topic should be picked this way when using only visual-based selection. However, surprisingly, when combined with text-based selection, using more concepts from the visual selection degrades the result. This "oracle" parameter selection optimised on the test set has not been used in the results of this report.

Simulated interactive search

We have not performed interactive search experiments with the 2008 and 2009 TRECVID data. However, inspired by [17], we made a simulated interactive run assuming the following setup. A single automated search query is performed first returning an ordered list of potentially matching video shots to the user. In the 10 minutes allowed in the TRECVID guidelines the user is then able to examine N first shots of the list, by visually inspecting the keyframes and possibly playing a few of them if he or she is unsure of their contents. In this way the user can now discard all incorrect ones from the displayed results.

Since this experiment can be done automatically, we tried this on the TRECVID 2009 topics as a simulated interactive run. In this way, we achieved an improvement in MAP from 0.1051 to about 0.14 with $N = 100$ and to 0.22 with $N = 500$. These correspond to increases of about 30% and 110% over our best automatic run MAP. The latter result is already comparable to the best real interactive systems used in TRECVID 2009 (the best result was 0.2463 by the MediaMill group). If one would be able to check 1000 shots per topic, the resulting MAP would be 0.28, surpassing the real interactive runs. In [17] it was reported that users participating in an experiment were always able to reach 1500 shots in 15 minutes, and sometimes even reaching 5000, albeit with trained users and a sophisticated user interface.

This kind of an experiment setup is naturally inconclusive and optimistic as the simulated human user is assumed to make no errors. Still, it would suggest that automatic search followed by manual re-ranking is capable of rivaling the interactive retrieval systems in the TRECVID video search experimental setting.

5.4 The best PicSOM system and its performance

In the above sections we have investigated many alternative techniques and components of the video retrieval system. In this section we collect the results together and describe the best-performing video retrieval system that we can assemble from the discussed components. We compare the performance of such a system with that of the state-of-the-art systems participated in the TRECVID high-level feature extraction (HLFE) and automatic search tasks in years 2008 and 2009.

In concept detection, our experiments have shown that with our fusion algorithms, the PicSOM system can benefit from all the shot-wise visual features we have extracted. For concept detection, we thus train one or two SVM detectors based on each shot-wise feature. In SVM training, we have to make a compromise between accuracy and training time. The detectors from all the features are fused together with the multifold-SFBS post-classifier fusion algorithm. We perform the detection separately for each concept as the cross-concept links have turned out not to be useful in our case. The concept detection is finalised with an N-gram temporal post-processing stage where we use the same N-gram order (eight) for all the concepts.

The search subsystem of our choice uses the donated set of concept detectors (64 concepts by MediaMill plus a few from CU-VIREO374) in addition to our own concept detectors. Concept selection is done by lexical analysis of each topic query by regular expressions, which can match more complex structures than simple single-word matching, reducing the number of false matches. Visual-example-based matching is not used since it did not improve the results for the full set of concept detectors. Text search on the shot-level is included, since it consistently improves the results slightly, despite the poor quality of the available textual data.

Figure 7 shows the MIAP concept detection performance of the PicSOM system in the TRECVID HLFE tasks of years 2008 and 2009 in comparison with the best performing systems of the respective years. The preparation of PicSOM A is computationally somewhat costly, thus for the 2008 HLFE task we have only evaluated the system PicSOM B that is closely similar, but computationally lighter. The system PicSOM B is evaluated for both the years' 2008 and 2009 HLFE tasks. By taking the performance of the PicSOM B system as yardstick, we can see that the concept detection community as a whole has improved from 2008 to 2009—not dramatically, but still clearly. The figures also show that the concept detection performance of the PicSOM system has improved dramatically since the year 2008 submission. In 2009, while not being absolutely the best, PicSOM's HLFE performance compares well with the state-of-the-art systems.

In Figure 8 the automatic search performance of the PicSOM system is compared with the top automatic search systems submitted to TRECVID 2008 and 2009. We see that also in the automatic video search task the performance of the PicSOM system has been enhanced and now compares very well with the other top systems. This result is strengthened by the fact that the top system (BUPT) of 2009 exploited manual annotation of the training data according to 2008 search topics and training detectors for these very specific concepts. This was not forbidden by the evaluation rules, but was unfortunate as over half of the search topics were re-used in 2009 in either unchanged form or with only small modifications. This makes it difficult to interpret the significance of the BUPT results and techniques. Nevertheless, if we ignore this one system, the community seems to have made progress also in automatic video search from 2008 to 2009. This can be seen, once again, if we compare the general performance level with that of the current PicSOM system. In search, not very much can be said about the usefulness of various choices for system architecture since all the top systems are rather similar in this respect.

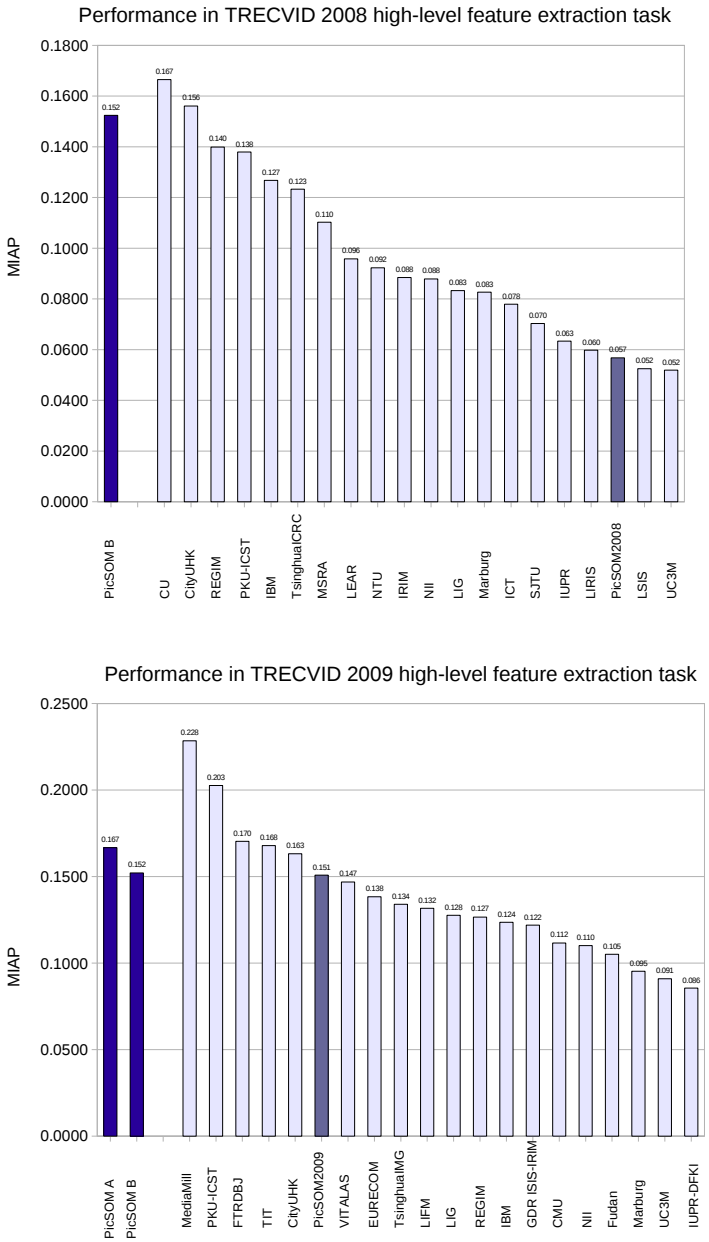

Figure 7: The MIAP performance in TRECVID 2008 and TRECVID 2009 high-level feature extraction task compared with the systems submitted by the best groups to the evaluation in the respective years. The comparison is made within the most popular class of systems: the systems that use the training data provided by the TRECVID organisers. The dark bars correspond to the PicSOM system discussed here, not any submitted system. Note that the figures show only the best-performing end of the distribution, all the systems are significantly more accurate than median of the submissions (0.048 and 0.049 in 2008 and 2009, respectively).

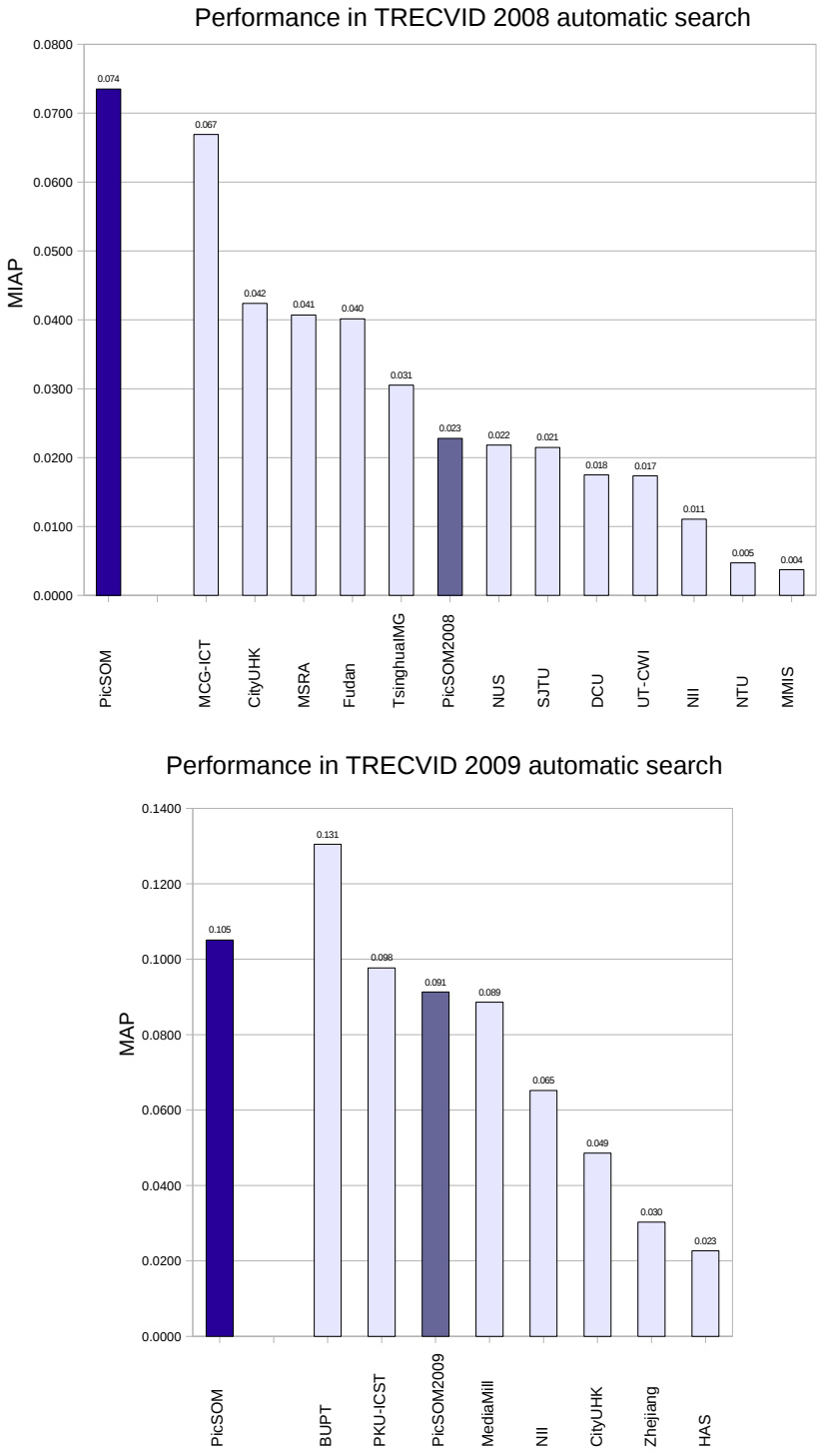

Figure 8: The MIAP performance in automatic search task of TRECVID 2008 and MAP of TRECVID 2009. The PicSOM system is compared with the best systems of the groups that submitted their runs of the evaluation. The comparison is made within the most popular class of systems: the systems that use the training data provided by the TRECVID organisers.

6 CONCLUSIONS

In this report we have described the PicSOM multimedia retrieval system, especially its subsystems for automatic semantic concept detection and concept-based video search. In the experiments with the TRECVID 2008 and 2009 high-level feature extraction setups, we have shown that the performance of the PicSOM concept detection subsystem compares well with other state-of-the-art systems. A closer look at the descriptions of the other systems reveals that most of the top-performing systems use more sophisticated and richer set of low-level features than ours. Audio, motion and camera work features as well as multi-keyframe analysis are examples of beneficial techniques employed by the others. However, the fusion architecture and temporal analysis in the PicSOM system is able to compete with any other system.

The results of our experiments confirm that fusion of a large set of features results in much better concept detection accuracy than the use of just a single feature. Of our fusion algorithms, the geometric mean of multifold-SFBS selected set of feature-wise detectors gives the best performance. SVM-based fusion and Bayesian Binary Regression are clearly inferior. Of the post-processing techniques, the temporal intra-concept re-ranking with N-gram models is clearly beneficial, whereas no further gain is obtained from cross-concept techniques on average.

The main conclusion from our search experiments with the TRECVID 2008 and 2009 video material is that the ability to detect and utilise a wide set of good semantic concepts is crucial. The decisive factor for search performance is finding at least one matching concept for each query topic. In our studies with lexical and visual-example-based concept selection, both methods were successful in matching useful concepts. Purely text-based selection of concepts, however, seemed to result in best overall performance, especially with large concept lexicons. Only when the concept set was small, using both selection methods in combination showed an improvement. Still, considering the semantic gap and the small number of visual examples, it is surprising that visual-example-based matching is even comparable with text-based matching, which has a semantic level much closer to the search query.

As already mentioned, we have confirmed concept-based search to be an effective approach for content-based multimedia retrieval. Concept-based search can be interpreted as a form of transfer learning: the training annotations for one learning problem—concept detection—are utilised in synthesising a solution to another problem. Transfer learning has attracted intense research attention in the statistical machine learning community in recent times. A promising future direction in enhancing multimedia search techniques could thus be to systematically explore the advanced transfer learning methodology that has been developed recently.

Overall, our experiments support the conclusion that collectively the multimedia search community has made progress in both concept detection and automatic video search between TRECVIDs 2008 and 2009.

REFERENCES

[1] Laurent Amsaleg and Patrick Gros. Content-based retrieval using local descriptors: Problems and issues from a database perspective. *Pattern Analysis & Applications*, 4(2+3):108–124, June 2001.

[2] Stéphane Ayache and Georges Quénot. Video corpus annotation using active learning. In *Proceedings of 30th European Conference on Information Retrieval (ECIR'08)*, pages 187–198, Glasgow, UK, March-April 2008.

[3] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.

[4] Kobus Barnard and Nikhil V. Shirahatti. A method for comparing content based image retrieval methods. In *Proceedings of SPIE Internet Imaging IV*, volume 5018, pages 1–8, Santa Clara, CA, USA, January 2003.

[5] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded up robust features. In *Proc. ECCV 2006*, May 2006.

[6] Sami Brandt, Jorma Laaksonen, and Erkki Oja. Statistical shape features for content-based image retrieval. *Journal of Mathematical Imaging and Vision*, 17(2):187–198, September 2002.

[7] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[8] Costas Cotsaces, Nikos Nikolaidis, and Ioannis Pitas. Video shot boundary detection and condensed representation: A review. *IEEE signal processing magazine*, 23(2):28–38, 2006.

[9] Steven B. Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In A. Waibel and K. Lee, editors, *Readings in speech recognition*, pages 65–74. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990.

[10] Ork de Rooij, Cees G. M. Snoek, and Marcel Worring. Balancing thread based navigation for targeted video search. In *Proceedings of the International Conference on Image and Video Retrieval (CIVR 2008)*, pages 485–494, Niagara Falls, Canada, 2008.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[12] N. Dimitrova, Hong-Jiang Zhang, B. Shahraray, I. Sezan, T. Huang, and A. Zakhor. Applications of video-content analysis and retrieval. *IEEE MultiMedia*, 9(3):42–55, Jul-Sep 2002.

[13] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. The MIT Press, 1998.

[14] A. Genkin, D. D. Lewis, and D. Madigan. BBR: Bayesian logistic regression software, 2005. Software available at http://www.stat.rutgers.edu/ madigan/BBR/.

[15] Erik Hatcher and Otis Gospodnetic. *Lucene in Action*. Manning Publications Co., 2004.

[16] Alexander G. Hauptmann, Michael G. Christel, and Rong Yan. Video retrieval based on semantic concepts. *Proceedings of the IEEE*, 96(4):602–622, April 2008.

[17] Alexander G. Hauptmann, Wei-Hao Lin, Rong Yan, Jun Yang, and Ming-Yu Chen. Extreme video retrieval: joint maximization of human and computer performance. In *Proceedings of ACM Multimedia '06*, Santa Barbara, USA, October 2006.

[18] N. Inoue, S. Hao, T. Saito, K. Shinoda, I. Kim, and C. Lee. TITGT at TRECVID 2009 workshop. In *Proceedings of the TRECVID 2009 Workshop*, Gaithersburg, MD, USA, November 2009.

[19] Yu-Gang Jiang, Akira Yanagawa, Shih-Fu Chang, and Chong-Wah Ngo. CU-VIREO374: Fusing columbia374 and VIREO374 for large scale semantic concept detection. In *Columbia University ADVENT Technical Report #223-2008-1*, August 2008.

[20] Lyndon S. Kennedy, Apostol (Paul) Natsev, and Shih-Fu Chang. Automatic discovery of query-class-dependent models for multimodal search. In *Proceedings of ACM Multimedia (ACM MM'05)*, pages 882–891, Singapore, November 2005.

[21] S. Kiranyaz and M. Gabbouj. Generic content-based audio indexing and retrieval framework. *Vision, Image and Signal Processing, IEE Proceedings*, 153(3):285–297, June 2006.

[22] S. Kiranyaz, A. F. Qureshi, and M. Gabbouj. A generic audio classification and segmentation approach for multimedia indexing and retrieval. *IEEE Transactions on Speech and Audio Processing*, 14(3):1062–1081, May 2006.

[23] Teuvo Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer-Verlag, Berlin, third edition, 2001.

[24] Pasi Koikkalainen and Erkki Oja. Self-organizing hierarchical feature maps. In *Proceedings of International Joint Conference on Neural Networks*, volume II, pages 279–284, San Diego, CA, USA, 1990.

[25] Markus Koskela. *Interactive Image Retrieval using Self-Organizing Maps*. PhD thesis, Laboratory of Computer and Information Science, Helsinki University of Technology, November 2003. Available online at: http://lib.hut.fi/Diss/2003/isbn9512267659/.

[26] Markus Koskela and Jorma Laaksonen. Semantic concept detection from news videos with self-organizing maps. In Ilias Maglogiannis, Kostas Karpouzis, and Max Bramer, editors, *Proceedings of 3rd IFIP Conference on Artificial Intelligence Applications and Innovations*, pages 591–599, Athens, Greece, June 2006. IFIP, Springer.

[27] Markus Koskela, Mats Sjöberg, and Jorma Laaksonen. Improving automatic video retrieval with semantic concept detection. In *Proceedings of 16th Scandinavian Conference on Image Analysis (SCIA 2009)*, volume 5575 of *Lecture Notes in Computer Science*, pages 480–489, Oslo, Norway, 2009. Springer Verlag.

[28] Markus Koskela, Mats Sjöberg, Ville Viitaniemi, and Jorma Laaksonen. PicSOM experiments in TRECVID 2008. In *Proceedings of the TRECVID 2008 Workshop*, Gaithersburg, MD, USA, November 2008. Available online at http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html.

[29] Jorma Laaksonen, Markus Koskela, and Erkki Oja. Content-based image retrieval using self-organizing maps. In *Third International Conference on Visual Information Systems (VISual'99)*, pages 541–548, Amsterdam, The Netherlands, June 1999. Springer-Verlag.

[30] Jorma Laaksonen, Markus Koskela, and Erkki Oja. PicSOM—Self-organizing image retrieval with MPEG-7 content descriptions. *IEEE Transactions on Neural Networks, Special Issue on Intelligent Multimedia Processing*, 13(4):841–853, July 2002.

[31] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. of IEEE CVPR*, volume 2, pages 2169–2178, 2006.

[32] Y. Linde, A. Buzo, and R. M. Gray. An algorithm for vector quantizer design. *IEEE Transactions on Communications*, 28(1):84–95, January 1980.

[33] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.

[34] B. S. Manjunath, Philippe Salembier, and Thomas Sikora, editors. *Introduction to MPEG-7: Multimedia Content Description Interface*. John Wiley & Sons Ltd., 2002.

[35] Hannes Muurinen and Jorma Laaksonen. Video segmentation and shot boundary detection using self-organizing maps. In *Proceedings of 15th Scandinavian Conference on Image Analysis (SCIA 2007)*, pages 770–779, Aalborg, Denmark, June 2007.

[36] Milind Naphade, John R. Smith, Jelena Tešić, Shih-Fu Chang, Winston Hsu, Lyndon Kennedy, Alexander Hauptmann, and Jon Curtis. Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 13(3):86–91, 2006.

[37] Milind R. Naphade and Thomas S. Huang. Extracting semantics from audiovisual content: The final frontier in multimedia retrieval. *IEEE Transactions on Neural Networks*, 13(4):793–810, July 2002.

[38] Apostol (Paul) Natsev, Alexander Haubold, Jelena Tešić, Lexing Xie, and Rong Yan. Semantic concept-based query expansion and re-ranking for multimedia retrieval. In *Proceedings of ACM Multimedia (ACM MM'07)*, pages 991–1000, Augsburg, Germany, September 2007.

[39] Chong-Wah Ngo, Yu-Gang Jiang, Xiao-Yong Wei, Wanlei Zhao, Yang Liu, Jun Wang, Shiai Zhu, and Shih-Fu Chang. VIREO/DVMM at TRECVID 2009: High-level feature extraction, automatic video search, and content-based copy detection. In *Proceedings of the TRECVID Workshop*, pages 415–432, 2009.

[40] Paul Ogilvie and Jamie Callan. Experiments using the Lemur toolkit. In *Proceedings of the 2001 Text REtrieval Conference (TREC 2001)*. National Institute of Standards and Technology, special publication 500-250, 2002.

[41] C. Petersohn. Fraunhofer HHI at TRECVID 2004: Shot boundary detection system. In *TREC Video Retrieval Evaluation Online Proceedings*. TRECVID, 2004.

[42] Slav Petrov, Arlo Faria, Pascal Michaillat, Andreas Stolcke, Dan Klein, and Jitendra Malik. Detecting categories in news video using acoustic, speech and image features. In *TRECVID Online Proceedings*. TRECVID, November 2006.

[43] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. Computer Science Series. McGraw-Hill, New York, 1983.

[44] Mats Sjöberg, Jorma Laaksonen, and Ville Viitaniemi. Using image segments in PicSOM CBIR system. In *Proceedings of 13th Scandinavian Conference on Image Analysis (SCIA 2003)*, volume 2749, pages 1106–1113, Halmstad, Sweden, June/July 2003. Springer Verlag. Available online at http://www.springerlink.com/content/65xgpgqc9cjquqcl/.

[45] Mats Sjöberg, Hannes Muurinen, Jorma Laaksonen, and Markus Koskela. PicSOM experiments in TRECVID 2006. In *Proc. of the TRECVID 2006 Workshop*, Gaithersburg, MD, USA, November 2006.

[46] Mats Sjöberg, Ville Viitaniemi, Markus Koskela, and Jorma Laaksonen. PicSOM experiments in TRECVID 2009. In *Proceedings of the TRECVID 2009 Workshop*, Gaithersburg, MD, USA, November 2009. Available online at http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html.

[47] Alan F. Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and TRECVid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[48] Alan F. Smeaton, Peter Wilkins, Marcel Worring, Ork de Rooij, Tat-Seng Chua, and Huanbo Lua. Content-based video retrieval: Three example systems from TRECVid. *International Journal of Imaging Systems and Technology*, 18(2-3):195–201, 2008.

[49] Stephen W. Smoliar and HongJiang Zhang. Content-based video indexing and retrieval. *IEEE MultiMedia*, 1(2):62–72, 1994.

[50] C. G. M. Snoek, K. E. A. van de Sande, O. de Rooij, B. Huurnink, J. C. van Gemert, J. R. R. Uijlings, and et al. The MediaMill TRECVID 2008 semantic video search engine. In *Proceedings of the TRECVID Workshop*, 2008.

[51] Cees G. M. Snoek and Marcel Worring. Are concept detector lexicons effective for video search? In *Proceedings of the IEEE International Conference on Multimedia & Expo (ICME 2007)*, pages 1966–1969, Beijing, China, July 2007.

[52] Kristina Toutanova and Christopher D. Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC 2000)*, pages 63–70, Hong Kong, October 2000.

[53] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.

[54] J. C. van Gemert, J. M. Geusebroek, C. J. Veenman, and A. W. M. Smeulders. Kernel codebooks for scene categorization. In *Proceedings of ECCV 2008*, pages 696–709, 2008.

[55] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek. Visual word ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[56] Ville Viitaniemi and Jorma Laaksonen. Evaluating the performance in automatic image annotation: example case by adaptive fusion of global image features. *Signal Processing: Image Communications*, 22(6):557–568, July 2007.

[57] Ville Viitaniemi and Jorma Laaksonen. Improving the accuracy of global feature fusion based image categorisation. In Bianca Falcidieno, Michela Spagnuolo, Yannis S. Avrithis, Ioannis Kompatsiaris, and Paul Buitelaar, editors, *Proceedings of the 2nd International Conference on Semantic and Digital Media Technologies (SAMT 2007)*, volume 4669 of *Lecture Notes in Computer Science*, pages 1–14, Genova, Italy, December 2007. Springer.

[58] Ville Viitaniemi and Jorma Laaksonen. Techniques for image classification, object detection and object segmentation. In Monica Sebillo, Giuliana Vitiello, and Gerald Schaefer, editors, *Proceedings of*

the 10th International Conference on Visual Information Systems (VISUAL 2008), volume 5188 of *Lecture Notes in Computer Science*, pages 231–234, Salerno, Italy, September 2008. Springer.

[59] Ville Viitaniemi and Jorma Laaksonen. Combining local feature histograms of different granularities. In *Proceedings of 16th Scandinavian Conference on Image Analysis (SCIA 2009)*, volume 5575 of *Lecture Notes in Computer Science*, pages 636–645, Oslo, Norway, 2009. Springer Verlag.

[60] Ville Viitaniemi and Jorma Laaksonen. Spatial extensions to bag of visual words. In *Proceedings of ACM International Conference on Image and Video Retrieval (CIVR 2009)*, Fira, Greece, July 2009.

[61] Ville Viitaniemi, Mats Sjöberg, Markus Koskela, and Jorma Laaksonen. Exploiting temporal and inter-concept co-occurrence structure to detect high-level features in broadcast videos. In *Proceedings of 9th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2008)*, pages 12–15, Klagenfurt, Austria, May 2008.

[62] Emine Yilmaz and Javed A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of 15th International Conference on Information and Knowledge Management (CIKM 2006)*, Arlington, VA, USA, November 2006.

TKK REPORTS IN INFORMATION AND COMPUTER SCIENCE

TKK-ICS-R29 Miki Sirola, Jaakko Talonen, Jukka Parviainen, Golan Lampi
Decision Support with Data-Analysis Methods in a Nuclear Power Plant. March 2010.

TKK-ICS-R30 Teuvo Kohonen
Contextually Self-Organized Maps of Chinese Words. April 2010.

TKK-ICS-R31 Jefrey Lijffijt, Panagiotis Papapetrou, Niko Vuokko, Kai Puolamäki
The smallest set of constraints that explains the data: a randomization approach. May 2010.

TKK-ICS-R32 Tero Laitinen
Extending SAT Solver With Parity Constraints. June 2010.

TKK-ICS-R33 Antti Sorjamaa, Amaury Lendasse
Fast Missing Value Imputation using Ensemble of SOMs. June 2010.

TKK-ICS-R34 Yoan Miche, Patrick Bas, Amaury Lendasse
Using Multiple Re-embeddings for Quantitative Steganalysis and Image Reliability Estimation. June 2010.

TKK-ICS-R35 Teuvo Kohonen
Contextually Self-Organized Maps of Chinese Words, Part II. August 2010.

TKK-ICS-R36 Antti Ukkonen
Approximate Top-k Retrieval from Hidden Relations. August 2010.

TKK-ICS-R37 Mikko Kurimo, Sami Virpioja, Ville T. Turunen (Editors)
Proceedings of the Morpho Challenge 2010 Workshop. September 2010.

TKK-ICS-R38 Jaakko Peltonen, Samuel Kaski
Generative Modeling for Maximizing Precision and Recall in Information Visualization. November 2010.

ISBN 978-952-60-3545-1 (Print)
ISBN 978-952-60-3546-8 (Online)
ISSN 1797-5034 (Print)
ISSN 1797-5042 (Online)