Topic Models for Semantics-preserving Video Compression

Jörn Wanke
University of Kaiserslautern
D-67663 Kaiserslautern,
Germany
j_wanke@cs.uni-kl.de

Adrian Ulges
German Research Center for
Artificial Intelligence (DFKI)
D-67663 Kaiserslautern,
Germany
adrian.ulges@dfki.de

Christoph H. Lampert
Max Planck Institute for
Biological Cybernetics
D-72076 Tübingen, Germany
chl@tuebingen.mpg.de

Thomas M. Breuel
University of Kaiserslautern
D-67663 Kaiserslautern,
Germany
tmb@cs.uni-kl.de

ABSTRACT

Most state-of-the-art systems for content-based video understanding tasks require video content to be represented as collections of many low-level descriptors, e.g. as histograms of the color, texture or motion in local image regions.

In order to preserve as much of the information contained in the original video as possible, these representations are typically high-dimensional, which conflicts with the aim for compact descriptors that would allow better efficiency and lower storage requirements.

In this paper, we address the problem of *semantic compression* of video, i.e. the reduction of low-level descriptors to a small number of dimensions while preserving most of the semantic information. For this, we adapt *topic models* – which have previously been used as compact representations of still images – to take into account the temporal structure of a video, as well as multi-modal components such as motion information.

Experiments on a large-scale collection of YouTube videos show that we can achieve a compression ratio of 20 : 1 compared to ordinary histogram representations and at least 2 : 1 compared to other dimensionality reduction techniques without significant loss of prediction accuracy. Also, improvements are demonstrated for our video-specific extensions modeling temporal structure and multiple modalities.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Retrieval and Indexing

MIR'10, March 29–31, 2010, Philadelphia, Pennsylvania, USA.

General Terms

Algorithms, Measurement, Experimentation

Keywords

Content-based Video Retrieval, Topic Models

1. INTRODUCTION

Currently, users world-wide collect more digital video data than ever before in history. These range from small home video collections to archives containing decades of TV and radio broadcast. The web also contains huge amounts of video: more than 60000 new videos are uploaded to YouTube per day, and it has been estimated that digital video will account for 91% of all internet traffic in 2013 [18].

As a manual indexing of such data quantities is infeasible, we face the question of how to extract semantic information directly from within media files. This challenge has been referred to as the *semantic gap* [29], i.e. the discrepancy between low-level content in form of raw pixel values and audio signals on the one hand and a viewer's high-level information demand on the other. Bridging this gap has been the concern of *content-based video retrieval* (CBVR) systems [11], which have been demonstrated to be effective in a variety of application scenarios: automatic concept detectors can be used to find certain locations, objects, and events [30]. Systems can also learn user preferences and recommend footage [38], or automatically detect copyright violations by finding similar or duplicate video scenes [20]. Other applications are automatic video summarization and categorization.

A typical CBVR processing pipeline is illustrated in Figure 1: the system extracts descriptive properties from the video content, so-called *low-level features.* These representations are employed by machine learning techniques, like statistical classifiers ("this clip belongs to the category 'sports videos'"), a clustering of videos into similar groups for later browsing, or a nearest neighbor search to detect visually similar (or even copied) material.

One key aspect with video content is that the amount of data is magnitudes higher than it is for images or text

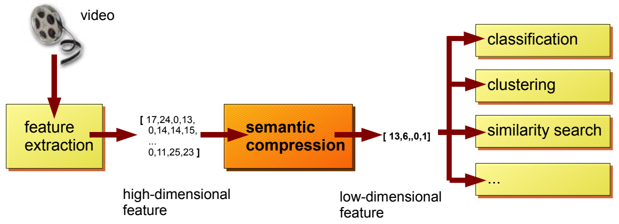

Figure 1: A typical content-based video retrieval (CBVR) processing pipeline: features are extracted from the raw video content and reduced to a few descriptive dimensions using semantic compression (which is the core concern of this paper). The resulting low-dimensional features are fed to machine learning techniques such as classification or clustering.

(for example, a 60-second video consists of 1500 images). This leads to several problems for CBVR methods: training and model application become inefficient, features do not fit into storage, and high-dimensional representations lead to overfitting.

One strategy to resolve these problems is to employ dimensionality reduction techniques, which reduce the amount of data to describe a video. We would like these methods to adapt to different domains of video content, to scale and generalize well, and to be reasonably fast while achieving good compression results. Most importantly, however, the resulting low-dimensional descriptions should preserve the semantic information required for tasks such as classification and clustering. Accordingly, we will refer to dimensionality reduction in the context of CBVR as *semantic compression* in the following.

One solution to achieve such semantic compression are topic models, which will be the main focus of this paper. Topic models have originally been developed in the text domain, and have already been successfully applied to still images [1, 9, 26, 28]. Their underlying idea is to decompose a document (or image) collection such that each item is described by a mixture of latent associated topics. The mixture coefficients form a feature vector, and as the number of topics is usually low, a strong dimensionality reduction is achieved. It should also be kept in mind that topic models work unsupervised, i.e. they estimate topics from plain video data without additional information. This is a key requirement, as manually annotated training data is difficult to acquire at a large scale.

Obviously, this approach also seems reasonable in the video domain: for example, a soccer video might generally consist of topics such as *'game play', 'interview', 'audience'* and so on (c.f. Figure 2). Therefore, we will investigate topic models for the semantic compression of video data in the following. Thereby, we will cover three key aspects:

1. We evaluate topic models in a video categorization scenario, and show that a performance comparable to high-dimensional standard descriptors can be reached at a compression rate of 1/20. Also, topic models outperform other dimensionality reduction techniques such as Principal Component Analysis (PCA, [7]) or a direct reduction of visual codebooks.

2. While topic models are usually applied to a single modality (for example, patch-based features [26]), video content offers multiple feature modalities, including frame content but also motion and audio. We show how topic models for video can benefit from an **integration of multi-modal features**.

3. Finally, we employ the fact that video streams come with a **temporal structure**, which is composed of units (or shots) separated by cuts (or other transitions). We argue that shots are not only temporal but also semantic units, such that switches in the semantics of a video – and correspondingly in the aspects of the topic model – should coincide with shot transitions. Based on this observation, we adapt a combination of topic models with Hidden Markov Models [10] for the video domain, and show that this leads to more stable and meaningful topics.

2. RELATED WORK

Topic models like Probabilistic Latent Semantic Analysis (PLSA) [12] and Latent Dirichlet Allocation (LDA) [3] have originally been developed in the text domain as strategies for decomposing document collections into latent aspects. The concept has been transferred to the image domain successfully, where a variety of approaches have employed topic models for scene classification [21], object recognition [8], and image retrieval [16]. An overview is given in the following.

The standard approach of adapting topic models to the image domain is to draw an analogy between the well-known 'bag of words' representation in the textual domain and 'bag of visual words' in the image domain [28]. Thereby, an image is described as a collection of local patches, and topic models are used to discover latent groups of correlated image parts. Barnard et al. [1] followed this approach for automatic image annotation using a hierarchical topic model. Similarly, Quelhas et al. [26] investigated the use of PLSA for scene classification and compared the dimensionality reduced features to the original visual-words representation on a dataset of 9500 images.

Topic models for dimensionality reduction on larger image data sets have been investigated by Hörster and Lienhart [15], who compared PLSA and LDA with alternative dimensionality reduction techniques like Restricted Boltzmann Machines (RBMs). While all three approaches perform similarly, PLSA gives slight improvements over both LDA and RBMs. In this paper, we will confirm similar results for video data, but beyond this address other video-specific issues like temporal structure and multiple feature modalities.

Several extensions to topic models have been proposed to tailor them more specifically to the image domain. Monay and Gatica-Perez [24] argued that, in word-image associations, the semantic information gained by words is much higher than for images. They proposed an asymmetrical model, which first trains PLSA on an images' captions and then extends topics to its visual features. Hörster et al. [17] explored an extension to PLSA where they model visual words as continuous distributions rather than quantized high-dimensional descriptors. They argue that this is more suited for the image domain and show a performance gain over the discretized standard PLSA model.

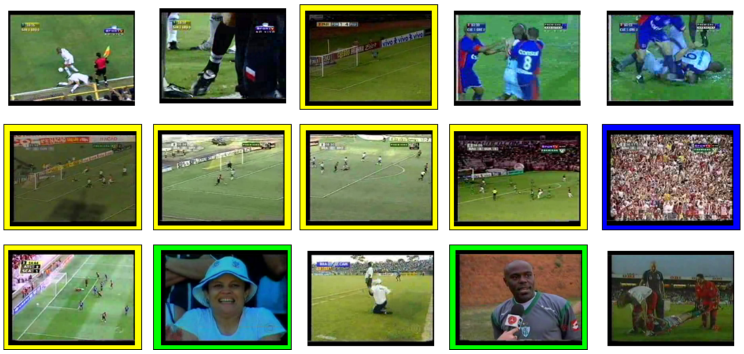

Figure 2: A web video from the category "soccer". Each shot of the video can belong to a particular category, like gameplay (yellow), close-ups and interviews (green), or shots of the crowd (blue). ***Topic models*** **can capture these categories by associating visual content with specific latent aspects.**

Other work has been targeted at integrating the spatial position of patches within images. For example, Tirilly et al. [35] proposed to capture patch centers by projecting them to the main axis of patch positions (computed using PCA). This is supposed to introduce structural order, shaping the analogy of "visual sentences". Liu and Chen [22] also used spatial information by integrating correspondence (shape and location of patches). They give "rewards" R for corresponding patches and learn their distribution depending on the topics. Fergus et al. [8] learned object categories by retrieving images from Google Image Search, using PLSA to filter noise content. To better capture the position of objects, they introduced a latent variable into the PLSA model that describes the object bounding box, resulting in invariance to translation and scaling. Hohl et al. [13] enhanced Latent Semantic Analysis (LSA) to capture spatial information by modifying the vocabulary of visual features: instead of treating features C_i individually, tuples (C_i, C_j) are used to capture feature co-occurrences.

There are only few prior contributions that utilize topic models in the video domain. Souvannavong et al. [31, 32, 33] explored LSA with region-based descriptors for tasks like object retrieval and scene classification. Niebles et al. [25] used PLSA to model human action categories like walking, running etc. They replace static image patches with spatio-temporal visual words, and thus employ a video-specific feature representation. In contrast to this work, we will follow a different approach of multi-modal fusion – namely a late fusion of single-modality topic models.

3. TOPIC MODELS

This section provides a compact review of topic models. While a variety of variations exist (e.g. LDA [3], Hierarchical Dirichlet Processes [34]), we will focus on two approaches here: first, we will introduce Probabilistic Latent Semantic Analysis (PLSA), which is a frequently used approach (e.g. [26, 28]) and has been shown to perform comparably to other topic models [15]. After this, we will review Hidden Topic Markov Models (HTMMs) [10]. This approach has been introduced by Gruber et al. [10] and adapts topic models such that the sentence structure of text documents is taken into account.

We start with some notation: a document (or video) collection will be denoted with $\mathcal{D}$. Each document $D_i \in \mathcal{D}$ contains N_i words out of an alphabet W. To capture the semantic relations between documents and words, intermediate latent variables (or *topics*) $Z_1, .., Z_K$ are introduced. A document is described as a mixture of topics, which in turn are mixtures of words. For example, a news report might be related to topics like '*financial world crisis*', '*soccer*' and '*weather*'. In turn, the topic '*soccer*' would be strongly associated with words like '*goal*', '*ball*', and '*referee*'.

3.1 Probabilistic Latent Semantic Analysis

PLSA was originally introduced by Hofmann [12]. It defines a generative model for sampling the words $W_1, .., W_{N_i}$ of a document D_i given a multinomial topic mixture $P(Z|D_i)$ and topics $P(W|Z)$ as follows (c.f. Figure 3):

1. For $j = 1, .., N_i$:
   (a) Sample a topic $Z_j \sim P(Z|D_i)$
   (b) Sample a word $W_j \sim P(W|Z = Z_j)$

Thus, the distribution of words within document D_i is approximated by a mixture of latent aspects:

$$P(W_j|D_i) = \sum_{k=1}^{K} P(W_j|Z_k)P(Z_k|D_i) \qquad (1)$$

Model Fitting.

The fitting of a PLSA model to a collection of training documents $\mathcal{D}$ is based on the idea that the observed joint probability of documents and words should be approximated well by the model. This is done by maximizing the log likelihood over all documents:

$$L = \sum_{i=1}^{N} \sum_{j=1}^{N_i} log\ P(D_i, W_j) \qquad (2)$$

For optimization, the Expectation Maximization (EM) algorithm is used [6], which consists of two alternating steps: in the *E-Step*, posteriors for the latent variables (here, the topics Z from which words are generated) are calculated. In the *M-Step*, system parameters (here, the topics $P(W|Z)$ and mixture coefficients $P(Z|D)$) are updated. Both steps are alternated until either the algorithm converges or an alternative termination condition like a maximum number of

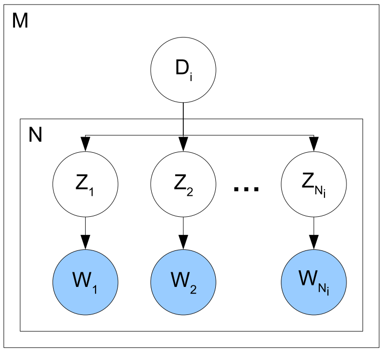

Figure 3: Generative model of PLSA: given a topic mixture of a document D_i, repeatedly a topic Z_k is picked and a word W_j is sampled from the corresponding topic distribution.

iterations is met. To increase the chance of escaping local maxima, EM is sometimes extended with an annealing procedure (for more details, please refer to Hofmann [12]).

Inference.

A priori, PLSA defines topic mixtures only on the training documents $\mathcal{D}$. To apply the model to a previously unseen test document D^*, its topic mixture is computed using a so-called *fold-in* heuristic: given $P(W|D^*)$, $P(Z|D^*)$ is estimated using the EM algorithm, whereas the topics $P(W|Z)$ learned previously in training are kept fixed (for more information, please refer to [12]).

Properties of PLSA.

One interesting property of PLSA should be kept in mind, namely that its topics help capture synonymy and polysemy within documents. Synonymy means that different words imply the same meaning, like the words '*buy*' and '*purchase*'. These words would have a high probability of being associated with the same topic, as they tend to be used in the same context. Similarly, polysemy implies that a word might have several meanings, like the word '*football*' implying either American football or European soccer. Polysemious words have a good chance of being associated with multiple topics. '*Football*' could have a high probability for a topic which also includes words like '*quarterback*', '*receiver*' and '*touchdown*' (american football), as well as be made up of words such as '*goalkeeper*', '*striker*' and '*corner kick*' (soccer).

3.2 Hidden Topic Markov Model

One important drawback of conventional topic models such as PLSA is that they reduce documents to a "bag-of-words" representation. This means that syntactical information is discarded, and it is merely counted how often a word occurs within a document. Correspondingly, when applying topic models to images the spatial arrangement of patches is typically neglected, and so is the temporal order of frames in a video sequence.

The Hidden Topic Markov Model (HTMM) [10] described in the following overcomes this problem by modeling documents as sequences of sentences. Words within a sentence are assumed to be derived from the same topic, while topic transitions occur only between sentences.

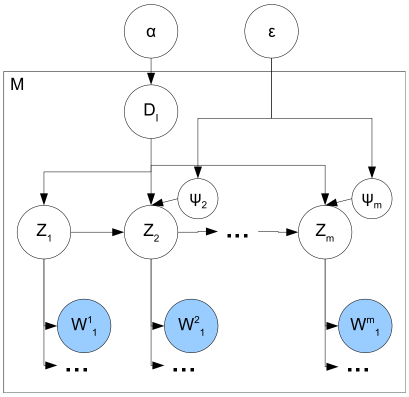

Figure 4: Generative model of the HTMM: given a topic mixture for a document D_i (drawn from a Dirichlet prior α), consider each word in the document: if the word does not start a new sentence ($\Psi_i = 0$, depending on the probability ϵ), the topic equals the topic of the previous sentence. Otherwise, a new topic is chosen according to the documents' topic mixture.

In the following, we give a quick overview of the HTMM model (for more information, please refer to [10]): each document $D \in \mathcal{D}$ consists of sentences $S_1, .., S_m$, and each sentence S_j again consists of words $W_1^j, .., W_{|s_j|}^j$. A topic distribution $P(Z|D)$ is assumed to be given (which is drawn from a Dirichlet prior with parameter α as in LDA [3]). At the beginning of each sentence, it is decided whether a topic transition occurs (which happens with probability ε). Either the topic from the previous sentence is used or a new one is chosen according to $P(Z|D)$. Each word within the sentence is then sampled from the chosen topic (see Figure 4 for a graphical illustration):

```
1. Draw $P(Z|D) \sim Dirichlet(\alpha)$
2. For $j = 1, .., m$: // sample sentences
    (a) Sample $\Psi_j \sim Binom(\varepsilon)$
    (b) If $(\Psi_j == 0)$ set $Z_j = Z_{j-1}$
        else $Z_j \sim P(Z|D)$
    (c) For $i = 1, .., |s_j|$: // sample words in a sentence
        i. Draw $W_i^j \sim P(W|Z_j)$
```

The idea of HTMMs is that words within a sentence are semantically connected, such that enforcing a single topic per sentence leads to a more stable document structure (which has been demonstrated by Gruber et al. [10]). The connection of the sentences is modeled using a Hidden Markov Model. Training of the HTMM is done by Expectation Maximization (c.f. Section 3.1) and the forward-backward algorithm [27]. EM is used to estimate two kinds of latent variables: aside from the topics Z_k, the decision Ψ_j whether a topic transition occurs at a sentence transition is to be determined. Thus, Expectation Maximization is slightly modified:

- In the E-step, $P(Z_m, \Psi_m | D, W_1, \ldots, W_M; \theta, \epsilon)$ is calculated for each sentence using the forward-backward algorithm.

- In the M-Step, the parameters $P(Z|D)$ and $P(W|Z)$ are calculated. Additionally, the parameter ϵ is estimated, which serves as a prior for the probability of topic transitions.

4. TOPIC MODELS FOR VIDEO

Previously, topic models have been applied to still pictures using a "bag-of-visual-words" representation: images are viewed as collections of local patches, which are discretized into clusters (so-called *visual words* [26]), and the number of occurrences of each visual word within the image is counted, resulting in the so-called *bag-of-visual-words* histogram. Thereby, the codebook of visual words draws an analogy to the vocabulary W in the textual domain, and the latent aspects discovered by the topic model can be viewed as groups of correlated patches occurring in the training images (for example, sampled from similar object regions).

In this section, we extend this approach and introduce a system that employs topic models for the video domain. Particularly, we take two additional video-specific key aspects into account: first, videos come with additional feature modalities – such as motion and audio – which offer valuable additional information sources that should be taken into account by CBVR systems (and correspondingly by a topic model generating low-level descriptions). We will outline this aspect of our system in Section 4.1. Second, video streams come with a temporal structure, i.e. their frames are not independent but heavily correlated, forming certain patterns over time. Accordingly, the semantic aspects given by a topic model should adapt to this structure, and we will take this into account in our system (Section 4.2).

4.1 Multi-modal Features

Video offers much richer information than images, as it comes with additional feature modalities like motion and audio. This opens the question of how to integrate multimodal features with topic models. A straightforward approach would be an *early fusion*, in which features are concatenated before feeding them to the topic model. However, the problem with this approach is that the resulting feature dimensionalities will become larger the more features are added (this curse of dimensionality is a well-known problem to machine learning techniques in general). Instead, we will follow a *late fusion* approach: given a video D_i that is described by features from M different modalities, we train a specific topic model on each feature modality. This results in M dimensionality-reduced topic vectors $P^m(Z_k|D_i)_{k=1}^{K_m}$ (where $m = 1,..,M$). These low-dimensional features are concatenated to a joint multi-modal topic descriptor of target dimensionality $\sum_{m=1}^{M} K_m$. Hörster and Lienhart [14] have reported previously for static images that such a fusion of different features gives improved results. In the case of multiple modalities, a similar approach of dimensionality reduction followed by concatenation has been pursued in [23]. Another alternative would be to apply modality-specific classifiers and fuse their results using classifier combination [37].

4.2 Shot Structure

A key problem with topic models like PLSA and LDA is that they simplify input documents to a "bag-of-words" representation, and syntactical information is widely neglected. For video content, this means that topic models do not offer a rigorous way to model the temporal structure of content. We could represent the video stream as a set of keyframes, but neglect the fact that content within the same scene is semantically related. On the other hand, when viewing a whole video (say, a 10-minute YouTube clip) as a single global document, temporal structure is lost and no information about specific events in the video is maintained.

Finally, we could employ shot boundary information to some extent and view *shots* within the video as a PLSA document. While this approach uses the structure of video to some extent, it still make only weak use of temporal information, as the order of shots is neglected, and patches within shots may be related to very different semantic aspects. This is illustrated in Figure 5 (top), which shows a result of a shot-level PLSA analysis of a sample YouTube video: from a "good" topic model, we would expect that patches from the same object share the same topic. In the PLSA result, however, even patches within the same image region (for example, the grass or the soccer players) are highlighted in different colors, i.e. linked with different topics. Obviously, PLSA topics are not accurately related to real-world semantics.

This raises the question whether the temporal structure of video can be used to infer more stable topics. Therefore, we propose an alternative that puts stronger emphasis on this aspect: we understand shots as the semantic units of video. This is motivated by the assumption that the content within a shot is usually heavily semantically related, while switches of semantics occur at shot boundaries (which is true at least for professionally produced videos).

Based on this observation, we adopt the HTMM model (Section 3.2) for the video domain by drawing an analogy between sentences in a text and shots in a video, which both form semantic units in a sense that a sentence (shot) should be related to a single topic only. Previous to the topic model, shot boundary detection is applied, and patches from within the same shot are viewed as words sampled from a single coherent topic. This is illustrated in Figure 5: while PLSA leads to unstable results, HTMM topics (bottom) are enforced to be consistent within shots and change only at shot boundaries. This way, more stable (and – as will be demonstrated later – meaningful) topics are obtained.

5. EXPERIMENTS

In a series of quantitative experiment on a large-scale dataset of real-world web video content, we demonstrate that the proposed approach leads to a highly efficient semantics-preserving compression of video material. We show that topic models for video achieve a performance comparable to high-dimensional standard descriptors and outperform other dimensionality reduction techniques. Also, we demonstrate that the proposed video-specific extensions – namely, the integration of multi-modal features and the use of temporal shot structure – lead to more meaningful topics and to improvements of CBVR systems.

We start with an outline of the experimental setup (Section 5.1). After this, three key issues are addressed: first, we compare topic models with other dimensionality reduction techniques (Section 5.2). Second, we evaluate the integration of multimodal features (here: patch-based features and motion information, Section 5.3). Finally, we test the Hidden Topic Markov Model for making better use of shot structure (Section 5.4).

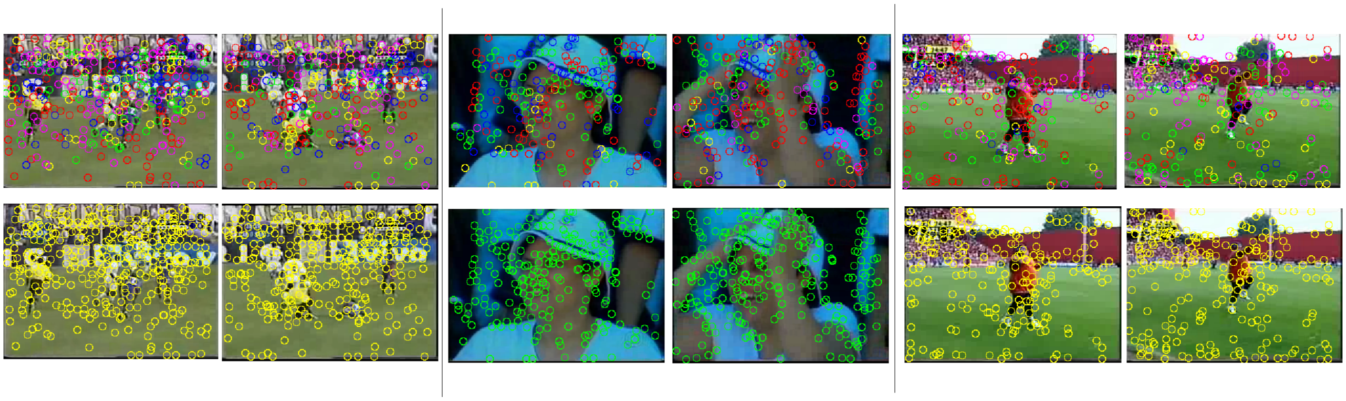

Figure 5: Three shots from a soccer video, which was analyzed using PLSA (top) and an HTMM (bottom). Patches are colored according to the topic that generated them. The topic mixtures resulting from PLSA are unstable: even patches from the same object region are assigned to different latent aspects, i.e. PLSA topics are not accurately related to real-world semantics. In contrast, the HTMM enforces changes of semantics to coincide with shot transitions and leads to topics that are well-associated with semantic categories ("crowd shots", "gameplay").

5.1 Experimental Setup

We evaluate the proposed approach for the semantic compression of video content in a CBVR application on web video clips. Standard bag-of-visual words features [28] are extracted, reduced using topic models, and finally fed to machine learning techniques like video categorization and clustering.

For evaluation purposes, we use a dataset of 195 hours of web video content. Ten categories were chosen, related to scene types, object categories, and sports (*basketball, cats, desert, eiffeltower, helicopter, riot, sailing, soccer, swimming, tank*). For each category, clips were downloaded from YouTube by searching for the respective category and similar terms. For example, soccer videos were obtained by querying YouTube with words like '*soccer*' or '*soccer world championship*'. To increase the diversity of the dataset, only one video per YouTube user was retrieved. The resulting content shows enormous intra-class variation: for example, clips within the category *hiking* range from outdoor trips to presentations about hiking boots. In total, the dataset is comprised of 3618 videos. 250 videos of each category are in the training set (amounting to a total of 2500 clips) and the rest in the test set (1118). Note that the amount of test samples for each category varies: for example, there are only 57 videos in the *eiffeltower* category, whereas the categories *riot* and *sailing* have more than twice as many.

Since a core concern of our experiments is the use of shot information, we extracted shot boundaries using a standard method that comes with a publicly available reference implementation by Lienhart[1]). Keyframes were extracted using the method by Borth et al. [4].

To evaluate the use of shot information, we also collected shot-level ground truth annotations with respect to the 10 target concepts: for each keyframe, a human annotator decided whether or not the associated concept was actually present in the video. Shots were only accepted as containing a concept if all keyframes within that shot were marked accordingly. All shots that did not show any of the 10 test concepts were discarded, obtaining a refined dataset of 12900 manually annotated shots (out of a total of 90854 shots) and a test set size of 5830 shots. This dataset was used in all following experiments.

For feature representation, a standard bag-of-visual-words approach was used: from each video, SURF features with 128 dimensions were extracted on each 10th frame [2]. These features were then aggregated into a frame-level bag-of-visual-words histogram with 2000 entries. The visual codebook was trained previously on a bigger, generic dataset of Youtube videos using a K-Means clustering. Thus, each video was described by $\frac{|Frames|}{10} * 2000$ values. These frame-wise histograms were aggregated to shot-level.

5.2 Comparing Semantic Compression Techniques

In a first experiment, we compare topic models with other dimensionality reduction techniques. We reduced the original 2000-dimensional bag-of-visual-words histograms using PLSA and Principal Component Analysis (PCA), and also tested a reduction of the number of visual words itself (the latter was only performed for 2 data points for efficiency reasons). Finally, the MPEG-7 Color Layout Descriptor (CLD) [19] – a manually designed 12-dimensional feature based on the distribution of color in a frame – was included in the evaluation.

Feature vectors were reduced to $10, 20, ..., 200$ dimensions. The resulting video descriptors were then used as input for an SVM classifier (the libsvm standard implementation [5] was used with an RBF kernel), which performed concept detection on the shot test set after learning on the training set of 12900 shots. As a performance measure, mean average precision (MAP) was used.

Results are illustrated in Figure 6, where system performance is plotted against the feature dimensionality. It can be seen that PLSA comes close to the performance of the full descriptor (58.9% for 170 topics as opposed to 62.1% for the full 2000-dimensional feature). Even for as few as 100 topics, performance is comparable to the full bag-of-visual-

[1] http://www.informatik.uni-mannheim.de/pi4/projects/MoCA/downloads.html

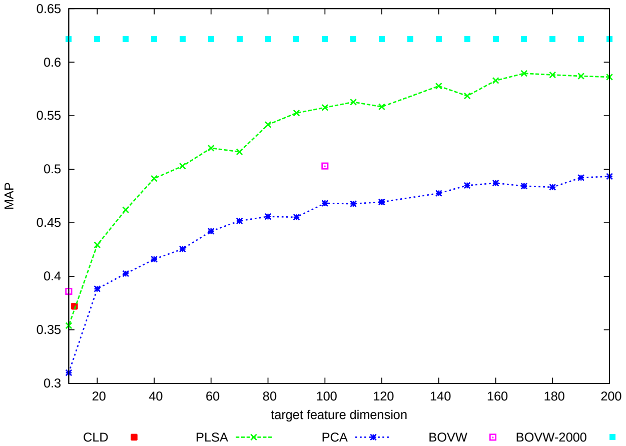

Figure 6: The mean average precision (MAP) plotted against the feature dimensionality when using PLSA, PCA, a reduction of the codebook size (BOVW), and the MPEG-7 color layout descriptor (CLD).

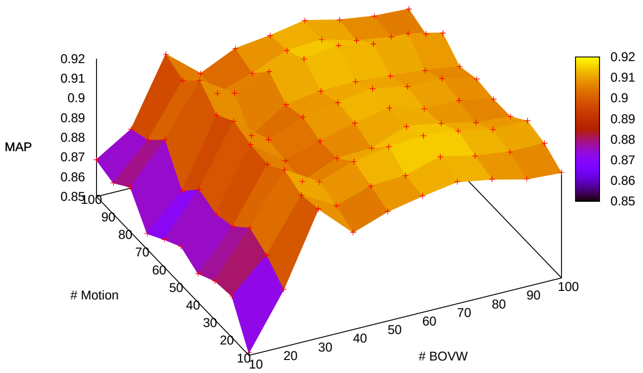

Figure 7: The mean average precision plotted against the number of topics used for fusing bag-of-visual-words (BOVW) and motion (MHW) features. The feature axes indicate the number of topics used for a single feature.

words representation (a relative performance loss of 10.3% occurs at a compression rate of 1/20). With this, PLSA outperforms PCA by up to 10.0%, which means that we can achieve the same performance at a much higher compression rate – for example, PLSA with 40 topics is on par with PCA when using 200 topics, i.e. a compression of 5 : 1 is achieved. While PLSA performs similarly to a codebook with only 10 entries, it seems to scale better, i.e. the performance of a codebook with 100 entries is reached given only 50 topics.

Finally, PLSA for 12 topics performs comparably to the manually designed MPEG-7 color layout descriptor of equal size. This result is encouraging, as PLSA can be adapted by varying the number of topics if a higher precision is required. Also, our current approach could be improved by using color-based patch descriptors (for which strong improvements have been reported by van de Sande et al. [36]). Overall, this experiment confirms that topic models are indeed a promising approach to effectively compress semantic information of videos. This confirms earlier results for the image domain reported by Hörster and Lienhart [15].

5.3 Multiple Modalities

This experiment investigates the combination of several feature descriptors in the context of topic models. Using more and different features and combining them is known to be beneficial in many CBVR applications. In this experiment, we demonstrate that topic models can be improved by a combination of multimodal features as well: we follow a late fusion approach, i.e. specific topic models are trained for each modality, and the resulting topic scores are concatenated afterwards.

Aside from bag-of-visual-word features, we use motion as a second video-specific modality. For this purpose, motion window histograms (MWHs) over MPEG-4 motion vectors[2] are extracted: each frame is split into 4 × 3 rectangular subregions, and for each subregion all block motion vectors within are quantized into one of seven bins in both X- and Y-direction (i.e., each motion block is assigned to one of 49 bins). This information is aggregated into a histogram describing the motion in each subregion. To describe the whole video, all subregion histograms are then concatenated, obtaining a descriptor of 588 dimensions. Both visual word histograms and MWHs are reduced to 10, .., 200 topics using PLSA. The resulting descriptors are concatenated and fed to an SVM classifier, which performed a categorization on video level.

Results of the different fusions can be found in Figure 7, where the categorization accuracy on video level is plotted against the number of topics used for each modality. It can be seen that pure PLSA-MWH (using only motion information) performs poorly. However, when fusing motion information with patch information, the resulting models outperform the pure PLSA on visual words (albeit only by a small margin). For example, using 15 motion topics and 30 patch topics gives an improvement of 3.3% (from 88.3% to 91.6%) over the single-modality case (45 patch topics), and using 15 motion and 75 patch topics results in an improvement of 2.7% (from 89.7% to 92.4%). This is particularly remarkable considering that the motion-based PLSA is not a very strong feature on its own. These results indicate that assigning different topic sizes to different features can reach a certain target dimensionality and achieve an improved performance (though it seems preferable to use more topics for better feature modalities).

5.4 Temporal Structure

While the previous experiments have been conducted with PLSA as a representative topic model, we have also proposed HTMMs as an alternative for making better use of the temporal structure of video (Section 3.2). We have argued that by enforcing changes of semantics to coincide with shot boundaries, we can obtain more stable and meaningful topics.

This last experiment compares the topics obtained by PLSA and by the proposed HTMM approach. Both PLSA and HTMM were applied to the training set of 12900 shots,

[2] motion vectors were extracted using the XViD codec: www.xvid.org

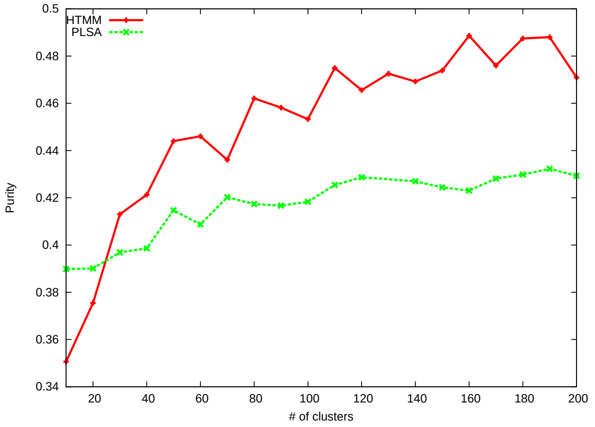

Figure 8: The purity for PLSA and HTMM topics, plotted against the number of topics.

and the reduced topic features were used to assign shots to clusters: each topic corresponds to one cluster, and each shot D is assigned to the topic with maximum posterior.

$$C(D) = \underset{k=1,..,K}{\arg\max}\ P(Z_k|D)$$

From a good topic model, we would expect that shots from the same semantic category are assigned to the same topic. Therefore, we measure the correlation of clusters and ground truth categories using purity [39] as a simple evaluation criterion. Purity counts the number of samples per category k in each topic C_i. The category $C_{i,k}$ with the maximum samples is divided by the total number of samples in a topic and (with adjusted weights) summed up over all topics C_i:

$$purity(C_i) = \frac{1}{|C_i|}\ \underset{k}{max}\ (|C_{i,k}|)$$

$$purity(C) = \sum_j \frac{|C_j|}{|D|} purity(C_j)$$

Note that purity increases monotonically with the number of clusters, and ultimately peaks at 100% in case every shot is assigned to its own singleton cluster. Yet, purity is a fair approach for comparing different methods on the same number of clusters.

The performance of both HTMM and PLSA is plotted against the number of topics in Figure 8. It can be seen that the HTMM gives a higher purity than PLSA, which indicates that the temporal structure enforced by the HTMM indeed leads to a better correspondence of topics and semantic categories.

6. CONCLUSIONS

In this paper, we have adapted topic models – which have previously been applied successfully for still images – for the domain of video. Our key contributions are extensions of topic models such that two video-specific aspects are taken into account: first, we make use of additional information such as audio and motion by combining multi-modal topics. Second, we draw an analogy between sentences in a text and shots in a video, and thus take the temporal structure of video content into account.

In quantitative experiments, we have shown that the proposed approach leads to a highly effective semantic compression of video content:

- Topic models were demonstrated to preserve most discriminative information at a compression rate at $20 : 1$ compared to full bag-of-visual-word descriptors.
- The late fusion of visual and motion-related features was found to be beneficial, giving improvements over single-modality image-based descriptions.
- It was found that the proposed approach for modeling the temporal structure of video content leads to more meaningful topics that are much stronger related with semantic categories.

Several interesting aspects might be pursued further. For instance, the Hidden Topic Markov Model could be adapted to model each sentence as a topic mixture instead of a single topic, as this would further enhance its usability for tasks like classification or similarity search. Another extension might be to learn statistics of specific topic transitions (for example, capturing the fact that “weather reports” never directly follow an “interview”, but frequently an “anchorman”).

Finally, online learning in the scope of topic models is a task which seems worth investigating: services like YouTube with roughly 60000 new videos per day are not capable of training completely new topic models with an ever increasing amount of data. Thus, adaption of existing models to incorporate new data is definitely an interesting research direction.

7. REFERENCES

[1] K. Barnard, P. Duygulu, D. Forsyth, N. D. Freitas, D. M. Blei, J. K, T. Hofmann, T. Poggio, and J. Shawe-taylor. Matching Words and Pictures. *J. Machine Learning Research*, 3:1107–1135, 2003.

[2] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-Up Robust Features. *CVIU*, 110(3):346–359, 2008.

[3] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[4] D. Borth, A. Ulges, C. Schulze, and T. Breuel. Keyframe Extraction for Video Tagging & Summarization. In *GI-Informatiktage*, pages 45–48, 2008.

[5] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[6] A. Dempster, N. Laird, and D. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[7] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification.* Wiley-Interscience, 2000.

[8] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning Object Categories from Google”s Image Search. In *ICCV*, pages 1816–1823, 2005.

[9] R. Fergus, P. Perona, and A. Zisserman. Object Class Recognition by Unsupervised Scale-invariant Learning. In *CVPR*, pages 264–271, 2003.

[10] A. Gruber, M. Rosen-Zvi, and Y. Weiss. Hidden Topic Markov Models. In *AISTATS*, 2007.

[11] A. Hanjalic, R. Lienhart, W. Ma, and J. Smith. The Holy Grail of Multimedia Information Retrieval: So Close or Yet So Far Away? *Proc. IEEE*, 96(4):541–547, 2008.

[12] T. Hofmann. Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, 42:177–196, 2001.

[13] L. Hohl, F. Souvannavong, B. Mérialdo, and B. Huet. Enhancing Latent Semantic Analysis Video Object Retrieval with Structural Information. In *ICIP*, pages 1609–1612, 2004.

[14] E. Hörster and R. Lienhart. Fusing Local Image Descriptors for Large-Scale Image Retrieval. *CVPR*, pages 1–8, 2007.

[15] E. Hörster and R. Lienhart. Deep Networks for Image Retrieval on Large-scale Databases. In *ACM MM*, pages 643–646, 2008.

[16] E. Hörster, R. Lienhart, and M. Slaney. Image Retrieval on Large-scale Image Databases. In *CIVR*, pages 17–24, 2007.

[17] E. Hörster, R. Lienhart, and M. Slaney. Continuous Visual Vocabulary Models for PLSA-based Scene Recognition. In *CIVR*, pages 319–328, 2008.

[18] C. S. Inc. Cisco Visual Networking Index: Forecast and Methodology, 2008-2013. available from `http://www.cisco.com` (retrieved: June'09).

[19] E. Kasutani and A. Yamada. The MPEG-7 Color Layout Descriptor: A Compact Image Feature Description for High-speed Image/Video Segment Retrieval. *Image Processing, 2001*, 1:674–677, 2001.

[20] W. Kraaij. TRECVID-2008 Content-based Copy Detection Task: Overview. In *TRECVID Workshop (available from: http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html)*, 2008.

[21] F.-F. Li and P. Perona. A Bayesian Hierarchical Model for Learning Natural Scene Categories. In *CVPR*, pages 524–531, 2005.

[22] D. Liu and T. Chen. Unsupervised Image Categorization and Object Localization using Topic Models and Correspondences between Images. *ICCV*, pages 1–7, 2007.

[23] J. Magalh aes and S. Rüger. Information-theoretic Semantic Multimedia Indexing. In *CIVR*, pages 619–626, 2007.

[24] F. Monay and D. Gatica-Perez. PLSA-based Image Auto-annotation: Constraining the Latent Space. In *ACM MM*, pages 348–351, 2004.

[25] J. C. Niebles, H. Wang, and L. Fei-fei. Unsupervised Learning of Human Action Categories using Spatial-temporal Words. In *BMVC*, 2006.

[26] P. Quelhas, F. Monay, J. Odobez, D. Gatica-Perez, and T. Tuytelaars. A Thousand Words in a Scene. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(9):1575–1589, 2007.

[27] L. R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In *Proc. IEEE*, pages 257–286, 1989.

[28] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering Object Categories in Image Collections. In *ICCV*, 2005.

[29] A. Smeulders, M. Worring, S. Santini, and A. G. R. Jain. Content-Based Image Retrieval at the End of the Early Years. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[30] C. Snoek and M. Worring. Concept-based Video Retrieval. *Foundations and Trends in Information Retrieval*, 4(2):215–322, 2009.

[31] F. Souvannavong, L. Hohl, B. Merialdo, and B. Huet. Structurally Enhanced Latent Semantic Analysis for Video Object Retrieval. *IEEE Proc. Vision, Image and Signal Processing, Volume 152, No. 6*, 2005.

[32] F. Souvannavong, B. Mérialdo, and B. Huet. Latent Semantic Analysis for an Effective Region-based Video Shot Retrieval System. In *MIR*, pages 243–250, 2004.

[33] F. Souvannavong, B. Mérialdo, and B. Huet. Latent Semantic Indexing for Semantic Content Detection of Video Shots. In *ICME*, pages 1783–1786, 2004.

[34] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *J. American Statistical Association*, 101(476):1566–1581, 2006.

[35] P. Tirilly, V. Claveau, and P. Gros. Language Modeling for Bag-of-visual Words Image Categorization. In *CIVR*, pages 249–258, 2008.

[36] K. E. van de Sande, T. Gevers, and C. G. Snoek. A Comparison of Color Features for Visual Concept Classification. In *CIVR*, pages 141–150, 2008.

[37] Y. Wu, E. Chang, K. Chang, and J. Smith. Optimal Multimodal Fusion for Multimedia Data Analysis. In *ACM Multimedia*, pages 572–579, New York, NY, USA, 2004.

[38] B. Yang, T. Mei, X. Hua, L. Yang, S. Yang, and M. Li. Online Video Recommendation Based on Multimodal Fusion and Relevance Feedback. In *CIVR*, pages 73–80, 2007.

[39] Y. Zhao and G. Karypis. Criterion functions for document clustering: Experiments and analysis, 2001.