Content-based Video Tagging for Online Video Portals*

Adrian Ulges[1], Christian Schulze[2], Daniel Keysers[2], Thomas M. Breuel[1]

[1]University of Kaiserslautern, Germany

[2]German Research Center for Artificial Intelligence (DFKI), Kaiserslautern

{a_ulges,tmb}@informatik.uni-kl.de,

{christian.schulze,daniel.keysers}@dfki.de

Abstract

Despite the increasing economic impact of the online video market, search in commercial video databases is still mostly based on user-generated meta-data. To complement this manual labeling, recent research efforts have investigated the interpretation of the visual content of a video to automatically annotate it. A key problem with such methods is the costly acquisition of a manually annotated training set.

In this paper, we study whether content-based tagging can be learned from user-tagged online video, a vast, public data source. We present an extensive benchmark using a database of real-world videos from the video portal *youtube.com*. We show that a combination of several visual features improves performance over our baseline system by about 30%.

1 Introduction

Due to the rapid spread of the web and growth of its bandwidth, millions of users have discovered online video as a source of information and entertainment. A market of significant economic impact has evolved that is often seen as a serious competitor for traditional TV broadcast. However, accessing the desired pieces of information in an efficient manner is a difficult problem due to the enormous quantity and diversity of video material published. Most commercial systems organize video access and search via meta-data like the video title or user-generated tags (e.g., *youtube*, *myspace*, *clipfish*) – an indexing method that requires manual work and is time-consuming, incomplete, and subjective.

While commercial systems neglect another valuable source of information, namely the content of a video, research in *content-based video retrieval* strives to automatically annotate (or 'tag') videos. Such systems learn connections between low-level visual features and high-level semantic concepts from a training set of annotated videos. Acquiring such a training set manually is costly and poses a key limitation to these content-based systems.

*Work supported partially by the Stiftung Rheinland-Pfalz für Innovation, project InViRe (961-386261/791)

(a) (b) (c) (d) (e) (f)

Figure 1: Some sample keyframes extracted from a video with the tag 'sailing'. Tagging such material is aggravated by the complexity of concepts (a,b), varying appearance (b,c), shots not directly visually linked to sailing (d,e), and low production quality (f).

In this paper, we study a different kind of training set, namely videos downloaded from online video portals (a similar idea has been published for images before, learning from Google Image Search [3]). Online videos are publicly available and come in a quantity that is unmatched by any dataset annotated for research purposes, providing a rich amount of tagged video content for a large number of concepts.

On the backside, online video content is extraordinarily difficult to interpret automatically due to several of its characteristics. First, its diversity is enormous: online video is produced world-wide and under various conditions, ranging from private holiday snapshots to commercial TV shows. Second, semantic concepts are often only linked indirectly to the actual visual content. These phenomena are illustrated in Figure 1, which shows keyframes from online videos tagged with the concept 'sailing'. The visual appearance of frames varies due to several reasons: first of all, the concept itself is so complex that it can be linked to multiple views, like shots of the boat or ocean views (1(a), 1(b)). Second, the appearance varies greatly among shots of the same kind (1(b), 1(c)). Third, there are shots not directly linked to sailing in a visual sense (1(d),1(e)) and garbage frames that occur frequently in home video (1(e)).

In this paper, we study whether - despite these difficulties - automatic tagging can be learned from online video. Our key contributions are: first, the description of an extensive, publicly available database of real-world online videos associated with 22 user-defined tags. Second, a benchmark of online video tagging that presents quantitative results for several combinations of visual features and strategies of fusing evidence over the shots of a video.

2 Related Work

Though to our knowledge there exists no prior work on video tagging with the focus on online material, we briefly discuss related work in content-based image and video retrieval that forms the basis for our study.

Content-based image retrieval [16] provides a useful pool of techniques and is strongly related to video annotation via the use of keyframes. Particularly, image annotation has been dealt with by modelling latent visual concepts in images [4, 14] or joint occurrences of local descriptors and tags [11]. Also, multiple instance learning methods have been used to detect local features associated with a concept [21]. However, only few prototypes perform this task online and at large scale, e.g. the (ALIPR) server [8].

When it comes to video content, annotation often follows a keyframe extraction for efficiency reasons, leading to an image retrieval problem [11, 20]. However, some approaches employ the video-specific concept of *motion*, for example in form of activity descriptors [19] or spatio-temporal features [2]. Generally, significantly less work can be found for video. One fundamental reason for this are copyright issues that cause a lack of large, publicly available datasets and hence aggravate quantitative comparisons.

A remarkable exception is TRECVID[1], an annually video retrieval contest with the goal of creating a stock of best practice for video retrieval. The contest includes quantitative evaluations on an extensive corpus of news video. In its "high-level features" task, the automatic annotation of shots is addressed. To boost standardization and comparability of results, groups share large sets of manual annotations , low-level features, and baseline results [17].

3 Database

Our evaluation is done on a database of real-world online videos we downloaded from the video portal *youtube.com*, and is therefore publicly available via the video URLs. These videos are tagged and grouped into semantic categories (like *sports* or *travel & places*) by users during upload. We selected 22 frequent tags and associate a canonic category with each of them, which helps to remove ambiguities that may be present if using tags only: e.g., a search for 'beach' returns beach scenes as well as music videos by the Beach Boys. Categories offer a straightforward way to avoid such problems in our analysis.

The youtube API was used to download 75 videos for each tag+category combination, obtaining a database of 1650 videos (total: 145 hrs.). The whole set was separated into a training set (50 videos per tag) and a test set (25 videos per tag). You can find more details, including the complete list of tags and the database itself (all URLs) at *demo.iupr.org/videotagging*.

3.1 Duplicate Removal

Duplicate videos uploaded multiple times by different users pose a problem for our evaluation. We identify them by extracting a signature from each downloaded video. This signature consists of the combined information of the change of color and the amount of motion of the bright and dark centroids between adjacent frames of a video [6]. In contrast to the $YCbCr$-converted frames used in [6], we represent the change of color by the MPEG-7 Color Layout Descriptor (CLD) [10], which proved to be more robust to color gradation. If the edit distance between the signatures of two videos from the same category is below a certain threshold, one of them is considered a duplicate and removed. Most of the duplicate videos could be eliminated using the combined signatures. Videos that were modified by adding additional content to the beginning or end of the video, however, could not be reliably identified as duplicate. Those near-duplicate videos caused suspiciously good results in the tagging experiments, which we used to detect and eliminate them manually.

[1] http://www-nlpir.nist.gov/projects/t01v/

3.2 Keyframe Extraction

To cope with the large amount of 145 hrs. of video data, we first reduce each video to a set of representative keyframes (though we enrich our representations with shot-level motion-based descriptors as well). In practice, often the first frame or center frame of a shot is chosen, which causes information loss in case of long shots containing considerable zooming and panning. This is why unsupervised approaches have been suggested that provide multiple keyframes per shot [5, 12]. Since for online video the structure varies strongly, we use a two-step approach that delivers multiple keyframes per shot in an efficient way by following a divide-and-conquer strategy: shot boundary detection – for which reliable standard techniques exist – is used to divide keyframe extraction into shot-level subproblems that are solved separately.

shot boundary detection: shot boundaries are detected using an adaptive thresholding [9] of differences of MPEG-7 color layout features [10].
intra-shot clustering: to determine the keyframes for a shot, we follow an unsupervised approach similar to [12]. We fit a Gaussian mixture model to the descriptors of all frames within a shot using K-Means, and for each mixture component the frame nearest to the center is extracted as a keyframe. The number of components is determined using the Bayesian Information criterion (BIC), which balances the number of keyframes explaining the shot versus the fitting error.

The tradeoff of this simplification is the loss of inter-shot reasoning: for example, in a dialog scene that alternates between two actors, the shot produces many similar keyframes. An additional grouping step might overcome this problem, but we omit it here. The keyframe extraction gives us a set of 75.000 frames for the whole database.

4 Tagging System

The purpose of our system is – given a video X – to return a *score* for each tag t that corresponds to the posterior $P(t|X)$. Figure 2 gives an overview of our setup: in a preprocessing stage, visual features are extracted (see Section 4.1). These features can be *global* frame descriptors based on color, texture, or motion, or based on *local* patches. For both classes, different statistical models are used (Section 4.2).

Since a video may contain many shots, each of them with several keyframes, the evidence gathered from all shots and keyframes must be fused to a global decision, a problem that we tackle using *direct voting* [13]. Optionally, the results for different features can be combined in a late fusion step.

4.1 Visual Features

In the feature extraction step, we scale videos to the same format (320×240) and extract several visual features. The first category contains descriptors for the whole frame and are thus referred to as *global* features:

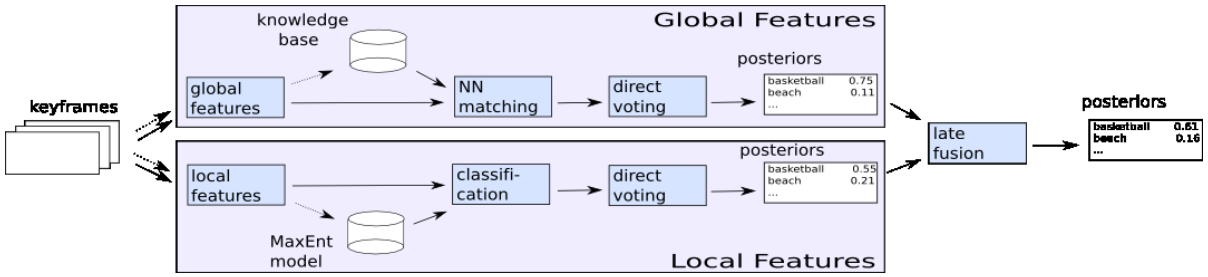
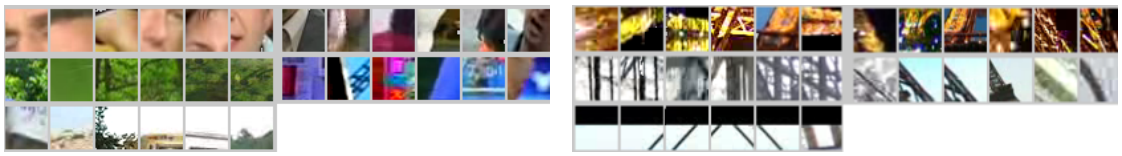


Figure 2: An overview of our tagging system: during offline processing (dashed line) features extracted from a training set are used to train tag models. When tagging a new video (solid line), its features are classified, and evidence from all keyframes is combined using direct voting. Results for local and global features can be combined in a late fusion.

color: RGB color histograms with $8 \times 8 \times 8$ bins
texture: Tamura texture features [18]
motion: some semantic features can be characterized better by their motion pattern than by color or texture. Here, we use a simple compressed domain feature of motion vectors extracted by the MPEG-4 codec XViD[2] to describe *what* motion occurs as well as *where* it occurs. For this, the spatial domain is divided into 4×3 regular tiles, and for each tile a regular 7×7 histogram of all motion vectors in the associated shot is stored (clipped to $[-20, 20] \times [-20, 20]$). The resulting 588-dimensional descriptor is the same for all keyframes in a shot.

While these global features are frequently used in practice, modern recognition systems are often based on collections of local image regions to make them more robust against partial occlusion, deformations, and clutter. Many of these *patch-based* methods have been published recently, among them the 'bag-of-visual-words' model [1, 3, 15, 17]. Here, visual features are clustered according to their appearance, and histograms over the clusters indicate what 'kinds' of features appear with what frequency, an analogy to the 'bag-of-words' model from textual information retrieval. This approach is referred to as *local* in the following:
visual words: a visual vocabulary is learned by extracting and clustering features of overlapping 32×32 patches regularly sampled in steps of 16. To extract the features, patches are transformed to YUV color space, and the discrete cosine transform (DCT) is applied to each channel. From the resulting DCT coefficients, 36 low-frequency components are extracted in a zigzag-pattern for the intensity, and 21 coefficients for each chroma component, obtaining a 78-dimensional descriptor. K-Means clustering in Euclidean space is performed with 500 clusters, five of which are illustrated by a few sample patches in Figure 3(a). Interestingly, it can be seen that some of those visual words can be related to certain semantics, like 'plants' or 'face parts'. During recognition, patches are extracted and assigned to the nearest cluster. A histogram of cluster frequencies is stored as a 500-dimensional feature.

[2] www.xvid.org



(a) (b)

Figure 3: Left: samples from our visual codebook: 6 patches belong to the same cluster (or 'visual word', respectively). Right: sample patches from 5 of the most discriminative visual words for the tag 'eiffeltower' (patches sampled from eiffeltower images only).

4.2 Statistical Modelling

Given a new video X associated with features $x_1, .., x_n$ each extracted from one of its keyframes, our tagging relies on the estimation of the posterior $P(t|x_1, .., x_n)$. This is done in two steps: first, for each frame, a posterior $P(t|x_i)$ is estimated for each feature x_i, and second, the single estimates are fused to a global posterior $P(t|X)$. For the first step, we use separate strategies for global and local features.

Global Features - NN Matching: for each global feature x_i, a nearest neighbor x' is found among all keyframes in the knowledge base (a kd-tree with Euclidean distance is used for fast matching [13]), giving an estimate $P(t|x_i) \approx \delta(t, t(x'))$.

Local Features - Maximum Entropy: For histograms over visual words, we adapt a discriminative approach based on maximum-entropy that has successfully been applied to object recognition before [1]. The posterior is modeled in a log-linear fashion:

$$P(t|x_i) \propto \exp\left(\alpha_t + \sum_{c=1}^{C} \lambda_{tc} x_i^c\right), \tag{1}$$

where x_i^c is the cth entry in the visual word histogram for frame x_i. The parameters $\{\alpha_t, \lambda_{tc}\}$ are estimated from our training set using an iterative scaling algorithm [1].

Fusing Key Frames - Direct Voting: to tag the whole video, the evidence from all keyframes must be fused to a global posterior $P(t|x_1, .., x_n)$. For this purpose, we propose a simple *direct voting* strategy:

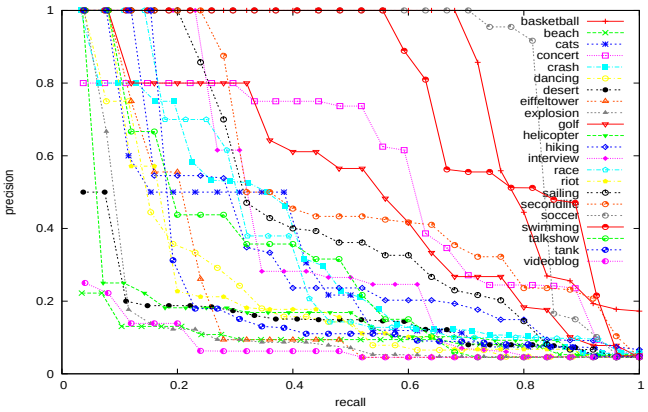
$$P(t|X) = \frac{1}{n} \sum_{1}^{n} P(t|x_i) \tag{2}$$

Direct voting can be seen as an equivalence of the sum rule in classifier combination. Statistical motivations for the approach can be found in [7, 13]. Particularly, in [7] theoretical reasons for the excellent robustness against errors in weak local estimates are described and validated in experimental studies – a property that is important in our context, since many keyframes in $\{x_1, .., x_n\}$ may not be visually related to the true tag and thus give erroneous matches.

Late Fusion - Sum Rule: To combine posteriors obtained from local and global features, we use a *late fusion* by applying the sum rule.

method	MAP
(1) color+texture, EF	0.267
(2) motion	0.170
(3) col+tex+motion, EF	0.290
(4) (col+tex, EF) + mot, LF	0.274
(5) visual words	0.281
(6) (col+tex+mot, EF) + vis. words, LF	0.345
(7) (col+tex, EF) + mot + vis. words, LF	0.345

(a)



(b)

Figure 4: Left: Experimental results in terms of Mean Average Precision (MAP). EF=early fusion, LF=late fusion. Right: the recall-precision curves for our best system (7). The average precision per concept varies between 0.840 (soccer) and 0.117 (beach)

5 Experiments

In our benchmark, we apply several combinations of features and fusion strategies to our test set of 550 videos. We obtain a posterior (*score*) for each combination of test video and tag. By sorting all test videos according to this score, we obtain a ranked list for each tag in which relevant videos should come first. We measure quality in terms of the *mean average precision* (MAP). Our results are subsumed in Figure 4(a) and outlined in the following.

(1) Baseline - Color and Texture: Our baseline system uses color histograms and Tamura features in an *early fusion* (EF), i.e. NN matching is performed on concatenated feature vectors. The nearest neighbors for some frames are illustrated in Figure 5.

(2) Motion: While the baseline gives good results for concepts with a fixed global composition of frames (e.g., soccer, basketball), it performs poorly when the action that takes place is critical. Such action can be captured well using the motion descriptor as outlined in Section 4.1. Though we obtain a worse overall result, the motion descriptor gives a strong improvement for some concepts that come with typical motion patterns like 'interview' (40%) or 'riot' (44%).

(3)+(4) Fusing Baseline and Motion: We combine the baseline features and motion features using early fusion (EF) or late fusion (LF). First improvements relative to the baseline can be achieved (MAP=0.290 for early fusion, 0.274 for late fusion).

(5) Visual Words: We train a discriminative maximum entropy model on all visual word histograms from the training set and use the posterior from Equation (1) as the score. Our results give a slight improvement compared to the baseline (MAP=0.281).

To visualize which patches the model uses to discriminate between classes, we use the fact that coefficients from the model are related to the discriminative power of a visual word: if $\lambda_{tc} \gg \lambda_{t'c} \quad \forall t' \neq t$, the feature c is a strong discriminator for class t. Figure 3(b) illustrates 5

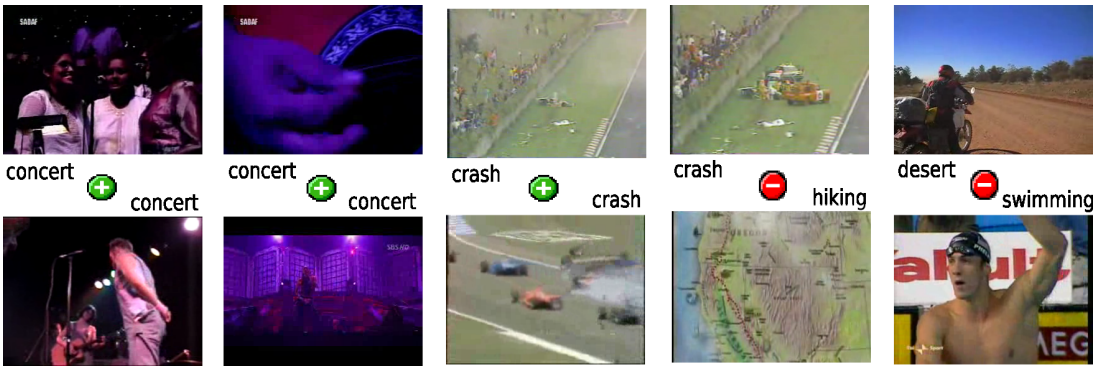


Figure 5: Some sample votes given by our baseline system (1). The upper row shows frames from the test set, below each of them its nearest neighbor in the training set. The three left columns show correct votes, erroneous ones are on the right.

of the 20 visual words that maximize the discriminative power

$$disc_t(c) = \lambda_{tc} - \max_{c'} \lambda_{tc'} \tag{3}$$

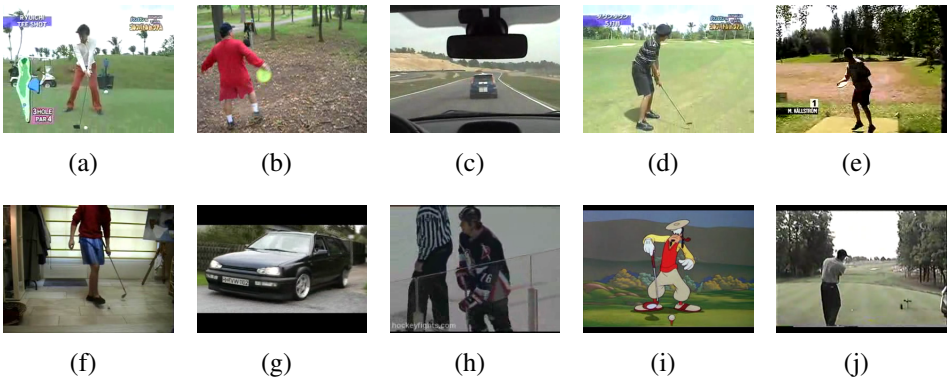
for 'eiffeltower', the tag with the strongest improvement (148%) relative to the baseline. These patches indicate how the system learns the visual structure of the tower.

(6)+(7) Best System - All Features: finally, we integrate global features with visual words using a late fusion. We obtain the best results with an MAP of 0.345 (an improvement of about 30% relative to the baseline). The associated recall-precision curves for all concepts are illustrated in Figure 4(b). It can be seen that the MAP varies strongly between concepts: the best results are achieved for sports videos (soccer - MAP=0.83, basketball - 0.80, swimming - 0.77), which often come with a well-defined color layout of the screen. Rather difficult concepts for our approach are associated ill-defined classes of actions (like 'dancing' - 0.21) or widely varying appearance (like 'beach' - 0.10).

Figure 6 illustrates some of the results for the category 'golf' (MAP=0.50), with representative key frames from the top 5 videos of the ranked retrieval list (first row) and from the golf videos with lowest posterior are plotted. While the top 5 show 4 correct samples (2 $\times$ golf, 2 $\times$ frisbee golf) and one false positive (6(c)), the 5 false negatives show only one typical golf video (6(j)). The others - though tagged with 'golf' - turn out to be difficult special cases: an indoor golf scene (6(f)), a VW Golf (6(g)), a hockey match (6(h)), and a comic (6(i)). It is obvious that finding the correct tag from such content is very difficult using visual features only.

6 Discussion

We have proposed online videos as a source for training models for content-based video retrieval. Our experimental results on a publicly available database of online videos suggest that such training is possible, though a difficult problem, and that the performance varies strongly between concepts.



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Figure 6: Sample frames from the 5 videos with the highest score for the tag ‘golf’ (first row) and from the 5 golf videos with lowest score (second row).

Another problem that should be mentioned is a certain redundancy of visual content in the database whose influence is difficult to quantify. Though we remove duplicate videos, still different levels of near-duplicates exist: first, popular scenes are reused and recomposed by different users (e.g., the ‘best trick plays in soccer’). Second, *series* of videos often play at the same location or share a similar production style. Our system makes use of such redundancy, but quantifying its influence on tagging performance is difficult, and we have not tackled it yet.

References

[1] Deselaers T. and Keysers D. and Ney H., ‘Discriminative Training for Object Recognition Using Image Patches’, *CVPR*, pp.157-162, Washington, DC, 2005.

[2] DeMenthon D. and Doermann D., ‘Video Retrieval using Spatio-Temporal Descriptors’, *ACM Intern. Conf. on Multimedia*, pp.508-517, Berkeley, CA, 2003.

[3] Fergus R. and Fei-Fei L. and Perona P. and Zisserman, A., ‘Learning Object Categories from Google’s Image Search’, *Computer Vision*, Vol. 2, pp.1816-1823, 2005.

[4] Barnard K. and Duygulu P. and Forsyth D. and de Freitas N. and Bleib D. and Jordan M., ‘Matching Words and Pictures’, *J. Mach. Learn. Res.*, Vol. 3, pp.1107-1135, 2003.

[5] Hanjalic A. and Zhang H., ‘An Integrated Scheme for Automated Video Abstraction based on Unsupervised Cluster-Validity Analysis’, *IEEE Trans. Circuits Syst. for Video Tech.*, Vol. 9, No. 8, pp.1280-1289, 1999.

[6] Hoad T.C. and Zobel J., ‘Detection of Video Sequences using Compact Signatures’, *ACM Trans. Inf. Systems*, Vol. 24, No. 1, 2006.

[7] Kittler J. and Hatef M. and Duin R. and Matas J., ‘On Combining Classifiers’, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 20, No. 3, pp.226-239, 1998.

[8] Li J. and Wang J., 'Real-time Computerized Annotation of Pictures', *Intern. Conf. on Multimedia*, pp.911-920, Santa Barbara, CA, 2006.

[9] Lienhart R., 'Reliable Transition Detection in Videos: A Survey and Practitioner's Guide', *International Journal of Image and Graphics*, Vol. 1, No. 3, pp.469-286, 2001.

[10] Manjunath B.S. and Ohm J.-R. and Vasudevan V.V. and Yamada A., 'Color and Texture Descriptors', *IEEE Trans. on Circuits Syst. for Video Techn.*, Vol. 11, No. 6, 2001.

[11] Feng S.L. and Manmatha R. and Lavrenko V., 'Multiple Bernoulli Relevance Models for Image and Video Annotation', *CVPR*, pp.1002-1009, Washington, DC, 2004.

[12] Hammoud R. and Mohr R., 'A Probabilistic Framework of Selecting Effective Key Frames for Video Browsing and Indexing', *Intern. Worksh. on Real-Time Img. Seq. Anal.*, pp.79-88, Oulu, Finland, 2000.

[13] Paredes R. and Perez-Cortes A., 'Local Representations and a Direct Voting Scheme for Face Recognition', *Workshop on Pattern Rec. and Inf. Systems*, pp.71-79, 2001.

[14] Fei-Fei L. and Perona P., 'A Bayesian Hierarchical Model for Learning Natural Scene Categories', *CVPR*, pp.524-531, San Diego, CA, 2005.

[15] Sivic J. and Zisserman A., 'Video Google: A Text Retrieval Approach to Object Matching in Videos', *ICCV*, pp.1470-1477, Washington, DC, 2003.

[16] Smeulders A. and Worring M. and Santini S. and Gupta A. and Jain R., 'Content-Based Image Retrieval at the End of the Early Years', *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 22, No. 12, pp.1349-1380, 2000.

[17] Snoek C. et al., 'The MediaMill TRECVID 2006 Semantic Video Search Engine', *TRECVID Workshop* (unreviewed workshop paper), Gaithersburg, MD, 2006.

[18] Tamura H. and Mori S. and Yamawaki T., 'Textural Features Corresponding to Visual Perception', *IEEE Trans. on Systems, Man, and Cybern.*, No. 6, Vol. 8, pp.460-472, 1978.

[19] Vasconcelos N. and Lippman A., 'Statistical Models of Video Structure for Content Analysis and Characterization', *IEEE Trans. Image Process.*, Vol. 9, No. 1, pp.3-19, 2000.

[20] Snoek C. and Worring M. and van Gemert J. and Geusebroek J.-M. and Smeulders A., 'The Challenge Problem for Automated Detection of 101 Semantic Concepts in Multimedia', *Intern. Conf. on Multimedia*, pp.421-430, Santa Barbara, CA, 2006.

[21] Yang C. and Lozano-Perez T., 'Image Database Retrieval with Multiple-Instance Learning Techniques', *Int. Conf. on Data Eng.*, pp.233-243, San Diego, CA, 2000.