Identifying Relevant Frames in Weakly Labeled Videos for Training Concept Detectors

Adrian Ulges
Department of Computer Science, Technical University of Kaiserslautern
D-67663 Kaiserslautern, Germany
a_ulges@informatik.uni-kl.de

Christian Schulze
German Research Center for Artificial Intelligence (DFKI)
D-67663 Kaiserslautern, Germany
schulze@iupr.dfki.de

Daniel Keysers
German Research Center for Artificial Intelligence (DFKI)
D-67663 Kaiserslautern, Germany
keysers@iupr.dfki.de

Thomas M. Breuel
DFKI and Technical University Kaiserslautern
D-67663 Kaiserslautern, Germany
tmb@iupr.dfki.de

ABSTRACT

A key problem with the automatic detection of semantic concepts (like 'interview' or 'soccer') in video streams is the manual acquisition of adequate training sets. Recently, we have proposed to use online videos downloaded from portals like *youtube.com* for this purpose, whereas tags provided by users during video upload serve as ground truth annotations.

The problem with such training data is that it is *weakly* labeled: Annotations are only provided on video level, and many shots of a video may be "non-relevant", i.e. not visually related to a tag. In this paper, we present a probabilistic framework for learning from such weakly annotated training videos in the presence of irrelevant content. Thereby, the relevance of keyframes is modeled as a latent random variable that is estimated during training.

In quantitative experiments on real-world online videos and TV news data, we demonstrate that the proposed model leads to a significantly increased robustness with respect to irrelevant content, and to a better generalization of the resulting concept detectors.

Categories and Subject Descriptors

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

General Terms

Algorithms, Measurement, Experimentation

Keywords

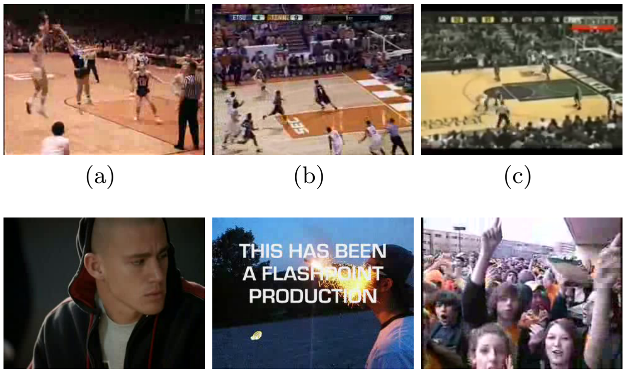
Content-based Video Retrieval, Video Tagging, Online Videos

1. INTRODUCTION

Content-based video retrieval is drawing more and more attention as the amount of digital video being stored and published is growing rapidly. Online video distributed via portals like *youtube.com* or *blinkx.com* has emerged as a rapidly growing market and as a serious competitor for traditional TV stations, which again maintain archives containing decades of video broadcast. To grant efficient access to such vast databases, most commercial systems rely on user-generated meta-data that are time-consuming to acquire, subjective, and can be incomplete or just not at hand. This is why *content-based retrieval* tries to complement traditional meta-data search with statistical models of the content of a video. In this context, a key task is the construction of systems that automatically annotate videos with high-level semantic concepts like objects, locations, or actions. Such *concept detectors* can support users with tagging their videos. They also allow a text-based search in video databases by mining them for concepts like 'interview' or 'US flag'.

A key problem is that concept detectors need to be trained on annotated video data, i.e. shots manually labeled with semantic concepts (or "tags", respectively). Since the distribution of such tags in low-level feature space can be arbitrarily complex, training sets must be large-scale and are thus difficult to acquire in practice. This is a reason why concept detection – though drawing strong attention from research and industry – has not been widely applied in practice.

In previous work [18, 19], we have proposed *online videos* as a data source that is both large-scale and easy to acquire. Such videos can be downloaded in large quantities from portals like *youtube.com*, and the tags users provide when uploading content can serve as ground truth annotations. Consequently, systems trained on such data can learn to tag videos autonomously by automatically downloading their annotated training sets.



(a) (b) (c)

(d) (e) (f)

Figure 1: Frames sampled from a video downloaded from youtube (the video is tagged 'basketball'). While some frames do show basketball (a,b,c), other "non-relevant" frames are not visually related to the concept (d,e,f).

On the downside, what makes learning from online videos difficult is that they are *weakly* labeled, i.e. tags are only given on video level. Knowing that a video is tagged with a concept, however, does not tell us *where* in the video the concept appears. Figure 1 illustrates this problem by showing frames from an online video tagged with 'basketball'. While the concept is visually present in some of its frames (a,b,c), others are not visually related to the tag at all, like (d,e,f). We will refer to such frames as "non-relevant" in the following. It is obvious that including them in the training set can have a negative impact on the performance of the resulting concept detectors.

Our previous work [19] did not explicitly model non-relevant frames and relied on a voting strategy to reduce their influence instead. Our results have demonstrated that visual learning from online videos is possible in general[1], and that fair tagging results can be obtained when using state-of-the-art visual features and fusion strategies. In this paper, we try to improve concept detection further using a rigorous model for the relevance of video content. More precisely, we describe the relevance of video frames as a latent random variable in a probabilistic framework of kernel densities. During training, these latent variables are estimated in an EM fashion.

Our key contributions are: (1) a rigorous statistical formulation of the problem of learning from weakly labeled videos, (2) a probabilistic approach for training concept detectors from weakly labeled videos that explicitly models the relevance of frames as a latent random variable. To the best of our knowledge, our framework is the first specifically targeted at training concept detectors on weakly annotated videos. (3) We present quantitative experiments on real-world online and TV video data in which we demonstrate the general applicability of the proposed approach. Further, we show that concept detectors can be built which rely less on non-relevant content and generalize better to different data sources.

[1] A demo of our prototype can be found at `http://demo.iupr.org/videotagging/`.

2. RELATED WORK

To the best of our knowledge, detectors of high-level semantic concepts (or "taggers", respectively) have been trained on strongly annotated video data until now, i.e. training usually takes place on datasets that have been annotated on shot or keyframe level. Using such data, researchers have tackled the problem in the annual TRECVID video retrieval contest[2], which hosts over 50 participating research groups world-wide and includes evaluations on an large corpus of news video. In TRECVID's "High-Level Features" task, the tagging problem has been addressed, whereas recent trends are headed towards patch-based descriptions [16] that have been developed in the computer vision community [10, 11], towards incorporating multiple features, discriminative classifiers, and fusion strategies [2], and towards specialized features like face detectors [13].

The problem of non-relevant content in training videos also draws an analogy to visual object recognition in still images, where modern systems are trained on unsegmented data labeled with the object name. While segmented training images are called *strongly labeled*, unsegmented ones are referred to as *weakly labeled* [14]. Like in the case of weakly labeled videos, the reason why an expert labels a visual document (here, an image instead of a video) with a certain tag is not associated with the whole document, but only with a part of it. For object recognition, this is the object region, and research has been targeted at "learning away" the irrelevant background (or *clutter*, respectively). For weakly labeled videos, we draw an analogy to the temporal domain, where this object region turns into "relevant parts of a video" that the training of a concept detector should focus on.

Object recognition is an intensively researched problem that has recently made a step forward due to the introduction of *patch-based methods* that view images as collections of local image parts. Like for video annotation, researchers have organized an annual contest – the PASCAL Visual Object Challenge[3] – which hosts object category recognition on very difficult datasets of weakly labeled images. To learn away clutter, discriminative strategies over local patches have been proposed, like boosting [12] or maximum entropy [4]. As generative models, topic models like Probabilistic Latent Semantic Analysis (PLSA) [8] have been adapted for the image domain to discover visual topics associated with object presence [6, 15]. Closest to our work is another approach by Rosenberg [14] that models the relevance of a pixel as a Boolean random variable which is estimated in an EM fashion. Finally, our setup is also related to multiple instance learning (MIL) [20], which views visual documents as bags of local entities (or "instances") and tries to detect a single instance that is discriminative for each category of bags. In previous experiments, however, we found the assumption of a single discriminative entity too restrictive given the potentially multimodal distribution of concepts in low-level feature space.

Other object recognition approaches model the location of features in an image, like graphical models [7] or discretizations of the image domain into tiles [9]. For videos, however, the temporal structure of their content seems less significant than the image position of the parts of an object. Thus, we do not model the temporal order of frames here.

[2] http://www-nlpir.nist.gov/projects/trecvid/

[3] http://www.pascal-network.org/challenges/VOC/

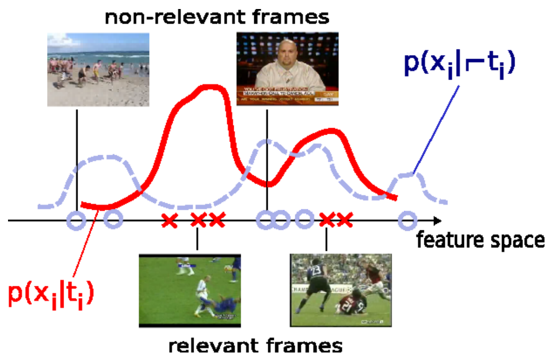


Figure 2: For a concept like soccer, both relevant and irrelevant frames are represented in a feature space, in which distributions of relevant and irrelevant content, $p(x_i|t_i)$ and $p(x_i|\neg t_i)$, are modeled.

3. APPROACH

In the following, a probabilistic framework is described for learning statistical models of the appearance of semantic concepts. The model achieves robustness to "irrelevant" training content by explicitly identifying it and "learning it away". Once the model has been trained, it can be used as a concept detector to tag a previously unseen video.

We first introduce some basic notation (which is also subsumed in Table 1). After this, it is described how the distribution of relevant and non-relevant content in feature space is modeled. Training – which involves estimating the relevance of training content – is outlined afterwards, and finally it is described how tagging works, i.e. how the trained model is used to score a video.

3.1 Basic Notation

We assume a semantic concept (or "tag") is given together with a video X consisting of n keyframes associated with feature vectors $x_1, .., x_n$. The Boolean random variable $\mathbf{T}$ is true whenever the tag is visible in X ("$\mathbf{T} = true$" will be abbreviated with "T" and "$\mathbf{T} = false$" with "$\neg T$"). Tagging the video is viewed as a binary classification problem of estimating the value of $\mathbf{T}$, giving the *score* $P(T|X)$.

The concept detector is trained on a set of videos $Y_1, .., Y_m$. A training video Y_j has keyframes associated with features $y_{j1}, .., y_{jm_j}$. Like for test videos, we use a random variable $\mathbf{T_j}$ for the presence of the tag in Y_j ("$\mathbf{T_j} = true$" is abbreviated with "T_j" and "$\mathbf{T_j} = false$" with "$\neg T_j$"). Since each training video Y_j is only weakly annotated, it is known whether the concept is present *somewhere* in it (T_j) or it is not ($\neg T_j$), but it is not clear in *which* of its frames y_{jk} the concept is visible. We will denote the presence of the concept in y_{jk} with another random variable $\mathbf{t_{jk}}$ with values t_{jk} ($\mathbf{t_{jk}} = true$) and $\neg t_{jk}$ ($\mathbf{t_{jk}} = false$). $\mathbf{t_{jk}}$ is a *latent* variable, i.e. its value is unknown.

3.2 Distributions of Non-Relevant and Relevant Frames

Our general approach is to represent video frames as vectors in a feature space, in which the distribution of relevant and non-relevant content is modeled (see Figure 2 for an illustration). As a model, we use non-parametric *kernel densities* [5]. Our motivation for this is that the distribution of concepts in feature space may be arbitrarily complex and thus difficult to model with parametric methods. Also, previous experiments with non-parametric methods have given positive results [19].

Table 1: Notation used in Section 3.

X	test video (to be scored)
$\mathbf{T}$	presence of a tag in X. Possible values: T, $\neg T$
$P(T\|X)$	tag score (to be estimated)
$x_1, .., x_n$	features for keyframes of X
$Y_1, .., Y_m$	training videos
$\mathbf{T_j}$	presence of concept in Y_j (given). Possible values: T_j, $\neg T_j$
$y_{j1}, .., y_{jm_j}$	features for keyframes of Y_j
$\mathbf{t_{jk}}$	presence of concept in y_{jk} (unknown). Possible values: t_{jk}, $\neg t_{jk}$
$p(x_i\|t_i)$	distribution of relevant frames
$p(x_i\|\neg t_i)$	distribution of non-relevant frames
$P(t_{jk}\|y_{jk}, \mathbf{T_j})$	"relevance score": the probability that a training frame is relevant (unknown)
$P(t_{jk}\|T_j)$	"relevance prior": fraction of relevant frames in relevant videos (given)

If a keyframe x_i is relevant – i.e., it shows the concept –, its features are assumed to be drawn from the following density:

$$p(x_i|t_i) \propto \sum_{j=1}^{m} \sum_{k=1}^{m_j} P(t_{jk}|y_{jk}, \mathbf{T_j}) \cdot K_h(x_i; y_{jk}), \quad (1)$$

and an equivalent density is defined for non-relevant frames *not* showing the concept:

$$p(x_i|\neg t_i) \propto \sum_{j=1}^{m} \sum_{k=1}^{m_j} [1 - P(t_{jk}|y_{jk}, \mathbf{T_j})] \cdot K_h(x_i; y_{jk}). \quad (2)$$

$p(x_i|t_i)$ and $p(x_i|\neg t_i)$ are kernel densities over all keyframes from all training videos, whereas each training sample is weighted by its probability of showing the concept or of *not* showing the concept ($P(t_{jk}|y_{jk}, \mathbf{T_j})$ or $1 - P(t_{jk}|y_{jk}, \mathbf{T_j})$). As a kernel function K_h, the well-known Epanechnikov kernel [5] is used with an empirically determined bandwidth h:

$$K_h(x; y) \propto max\,(0, 1 - \frac{d(x,y)^2}{h^2}),$$

where $d(.,.)$ denotes some distance function in feature space.

To estimate the densities $p(x_i|t_i)$ and $p(x_i|\neg t_i)$, for each training frame y_{jk} the probability of relevance $P(t_{jk}|y_{jk}, \mathbf{T_j})$ must be known. The estimation of these *relevance scores* will be outlined in the following section.

3.3 Identifying Relevant Frames

We view concept presence in training frames y_{jk} as a hidden variable $\mathbf{t_{jk}}$, for which the posterior $P(t_{jk}|y_{jk}, \mathbf{T_j})$ needs to be estimated.

We distinguish between two cases: if the concept is *not* present in a training video, we assume that it does not appear in any of its keyframes, i.e $P(t_{jk}|y_{jk}, \neg T_j) = 0\ \forall k$. On the other hand, if the video *is* tagged with the concept, $\mathbf{t_{jk}}$ is unknown, and $P(t_{jk}|y_{jk}, T_j)$ needs to be inferred. For this purpose, the following iterative procedure is used:

1. set the iteration $u = 0$ and

$$P^0(t_{jk}|y_{jk}, \mathbf{T_j}) = \begin{cases} P(t_{jk}|T_j), & \mathbf{T_j} = \textit{true} \\ 0, & \textit{else} \end{cases}$$

2. set $P^{u+1}(t_{jk}|y_{jk}, T_j) =$

$$\frac{p^u(y_{jk}|t_{jk}) \cdot P(t_{jk}|T_j)}{p^u(y_{jk}|t_{jk}) \cdot P(t_{jk}|T_j) + p^u(y_{jk}|\neg t_{jk}) \cdot (1 - P(t_{jk}|T_j))}$$

3. if convergence: exit.

   else: set $u = u + 1$ and goto (2)

The densities $p^u(y_{jk}|t_{jk})$ and $p^u(y_{jk}|\neg t_{jk})$ are computed by plugging $P^u(t_{jk}|y_{jk}, \mathbf{T_j})$ into Equations (1) and (2).

The term $P(t_{jk}|T_j)$ models the fraction of keyframes from videos tagged with the concept that actually *do show* the concept. It is referred to as the *relevance prior* in the following and is considered prior expert knowledge that is given or can be estimated via cross-validation. If we set $P(t_{jk}|T_j) = 1$, all training frames in relevant videos are considered relevant, i.e. $P(t_{jk}|y_{jk}, T_j) = 1 \ \forall k$. In this case, the model degenerates to a standard kernel density model.

The proposed iterative procedure resembles the well-known "Expectation Maximization" (EM) scheme [3], which optimizes a bound on the data likelihood by alternating so-called "E" steps (in which posteriors for hidden variables are estimated) and "M" steps (in which the parameters of a distribution are updated from this knowledge). The situation, however, is more simple for kernel densities: Since the underlying distributions are non-parametric and the kernel bandwidth is fixed, the system is parameter-free and no "M" step is required.

3.4 Concept Detection

Once the relevance scores of all training keyframes have been estimated, the densities of relevant and non-relevant frames (Equations (1) and (2)) can be computed for the keyframes $x_1, .., x_n$ of a test video X. Using this information, a result score $P(T|X)$ is estimated as described in the following:

The sum rule is used for fusing weak pieces of evidence from the keyframes $x_1, .., x_n$, an approach which has been successful in previous experiments [19]:

$$P(T|X) = P(T|x_1, .., x_n) \approx \frac{1}{n} \sum_{i=1}^{n} P(T|x_i)$$

By applying Bayes' rule, we obtain:

$$P(T|x_i) = \frac{p(x_i|T) \cdot P(T)}{p(x_i|T) \cdot P(T) + p(x_i|\neg T) \cdot (1 - P(T))} \quad (3)$$

We further assume that the prior $P(T)$ of a video being relevant is known. In this way, the problem of tagging a video is reduced to estimating $p(x_i|T)$ and $p(x_i|\neg T)$ for all of its keyframes.

Two cases can be distinguished: first, for an irrelevant video all its keyframes are drawn from the distribution of irrelevant frames: $p(x_i|\neg T) = p(x_i|\neg t_i)$ (Equation 2). The second case is that a video *does* show the concept (T). We then marginalize over the latent variable of relevance t_i:

$$\begin{aligned} p(x_i|T) &= p(x_i, t_i|T) + p(x_i, \neg t_i|T) \\ &\approx p(x_i|t_i) \cdot P(t_i|T) + p(x_i|\neg t_i) \cdot [1 - P(t_i|T)], \end{aligned}$$

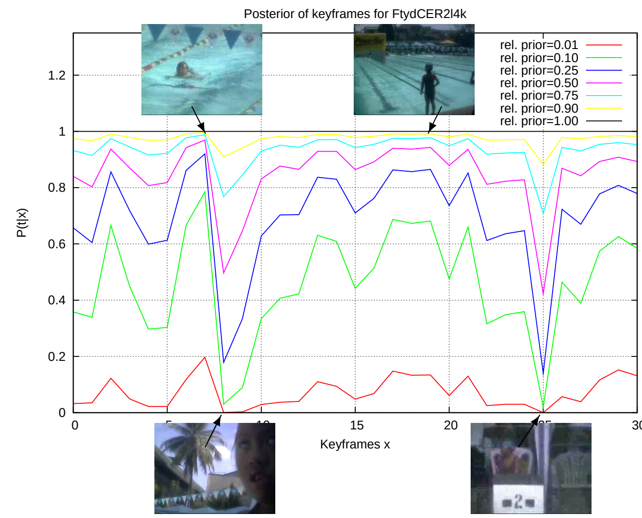


Figure 3: Posterior probabilities for each keyframe of a 'swimming' video using different values of the relevance prior $P(t_{jk}|T_j)$. For sample frames that do not show the concept, the system estimates a low relevance score.

where the relevance prior $P(t_i|T)$ is assumed to be given. By plugging $p(x_i|t_i)$ (Equation (1)) and $p(x_i|\neg t_i)$ (Equation (2)) into the aforementioned terms, we can thus compute the final score $P(T|X)$.

4. EXPERIMENTS

In quantitative experiments on online videos and TV news data, we demonstrate that learning from weakly labeled videos is possible, and that irrelevant content can be successfully identified. We also show that such explicit modeling of relevance makes concept detection more robust to irrelevant content in the training set.

4.1 Experiment 1: Youtube Data

We test our framework on a large-scale dataset of real-world online videos from the portal *youtube.com*. The dataset consists of 2600 videos (about 230 hrs.). It was downloaded by simulating queries for 22 representative tags including locations (e.g., 'desert' and 'beach'), actions (e.g., 'hiking', 'interview'), objects (e.g., 'cat','eiffeltower'), and sports (e.g., 'swimming', 'soccer'). The full list of tags can be found in [18].

Our framework was tested for the 4 sports concepts 'basketball', 'golf', 'soccer', and 'swimming'. Since fair tagging results were achieved with simple features in previous experiments [19], these concepts were considered good candidates to study the influence of relevance modeling.

The dataset was split into a training set (1500 videos, 150 of which were labeled with each of of the 4 sports concept) and a test set (1100 videos, 50 for each concept).

We use a keyframe extraction based on a shot boundary detection and intra-shot clustering of frames [1], which gives about 97.000 keyframes for the whole dataset. For each keyframe, color histograms and Tamura texture features [17] were combined in an early fusion (i.e., concatenated) and used as a feature vector. A concept detector was trained on this data for each of the 4 sports concepts using the approach

Table 2: Averaged results when learning the relevance prior using 10-fold cross-validation (bandwidth 0.00075). It can be seen that the estimated relevance prior is always close to the optimal one (see the plots in Figure 7). The tagging performance is improved compared to a system that does not model relevance, to a baseline using NN matching, and to our previous system .

concept	estim. rel. prior	AP with relevance [%]	impr. over no rel. [%]	impr. over NN [%]	impr. over [19] [%]
basketball	0.25	79.5	8.9	30.7	7.8
golf	0.75	56.3	3.1	9.5	-1.6
soccer	0.10	83.6	5.0	8.2	3.1
swimming	0.25	82.6	4.5	11.0	3.0

described in Section 3. The kernel bandwidth was varied between 0.0005 and 0.001. As a distance function, the χ^2 distance was used, which is a standard choice for histogram features [21].

Figure 3 gives a first impression of whether the system identifies the correct content as relevant. For a training video tagged 'swimming', the estimated relevance of all keyframes is plotted against the keyframe number. It can be seen how relevance learning behaves for several relevance priors: with decreasing relevance prior $P(t_{jk}|T_j)$, the overall relevance of training keyframes decreases, while the order of the keyframes is mostly preserved. Also, it can be seen from sample frames that content with a high relevance value is in fact visually related to the concept.

Instead of studying a single video, a similar check can be done for the whole dataset. For each concept, the frames from the training set that were assigned the highest relevance scores $P(t_{jk}|y_{jk}, T_j)$ are visualized, as well as the frames for which the relevance score is the lowest. The results are illustrated in Figure 7: it can be seen that the system correctly identifies relevant content (for example, the frames with high relevance scores from basketball videos do in fact show basketball scenes), and also classifies the majority of irrelevant content correctly (for basketball, there are some frames showing soccer action. These appear in videos labeled basketball, but have been identified as irrelevant by our model). Only few false negatives can be found that are visually related to the topic, like an untypical swimming or golf scenes that the system labels irrelevant.

While these results demonstrate that the proposed approach seems to identify non-relevant content well, the key question is whether the robustness of tagging is improved by modeling the relevance of frames. Therefore, for each concept a plot is given in Figure 7, showing the performance of the system (measured by the average precision) for the bandwidths 0.0005, 0.00075, and 0.001, plotted against the relevance prior $P(t_{jk}|T_j)$. This parameter models the fraction of relevant frames in relevant videos. If it is set to 1, the framework degenerates to a standard kernel density case in which all frames from positive training videos are assumed to be relevant. This setup can thus be used as a baseline for a system that models the relevance of frames. It corresponds to the rightmost data point in all plots, and is also indicated by dotted horizontal lines.

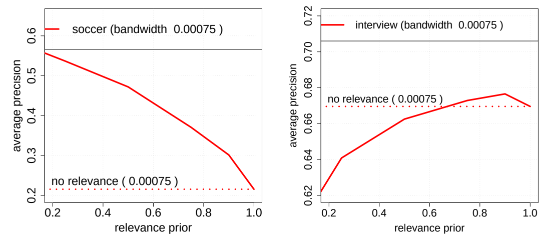


Figure 4: The average precision plotted against the relevance prior when generalizing taggers trained on youtube to a dataset of TV news. Relevance detection can improve tagging performance by 36 % for the concept 'soccer' (left). For 'interview', relevance modelling does not improve tagging significantly, which corresponds to the observations in Figure 6.

All 4 plots indicate that the tagging performance can be improved compared to the baseline system that assumes all training frames to be relevant. This improvement varies between concepts and bandwidths: it is strongest (up to 20 %) when using a low bandwidth of 0.0005 together with a low relevance prior. For the highest bandwidth of 0.001, the improvements are not as strong (for the concept 'golf', modelling relevance even has a negative impact).

The best result for all concepts is achieved with the moderate bandwidth of 0.00075. For all concepts, improvements between 4 % ('golf') and 8 % ('basketball') are achieved by modelling the relevance of frames. Thereby, the optimal choice of the relevance prior differs between the concepts: for basketball, the best choice is about 0.25, while for golf a higher relevance prior is optimal. Obviously, an interesting question is how to set this parameter automatically.

We demonstrate in another experiment that cross-validation is suitable for this purpose. 10-fold cross-validation is used on the test set (the bandwidth of 0.00075 was used, which gave the best overall results). Table 2 gives the estimated values of the relevance prior together with the achieved average precision. It can be seen that the relevance prior determined from cross-validation is close to the one giving the peak performance (see Figure 7). Correspondingly, performance improvements between 3.1 % and 8.9 % can be observed relative to the system that does *not* model relevance.

We also compared our system to a simple baseline that does nearest neighbor (1-NN) matching in the same feature space (color and Tamura histograms), revealing strong improvements between 8.2 % ('soccer') and 30.7 % ('basketball'). Finally, the proposed approach is also compared to the system we introduced in [19]. This prototype makes use of several additional features, including motion information and a discriminative patch-based approach. Our results demonstrate that by modelling the relevance of content, a system that uses much simpler features can achieve comparable ('golf') or even better results (all other concepts).

4.2 Experiment 2: Generalization to TV Data

In a second experiment, we evaluate the generalization properties of concept detectors when applied to a different

(a)

(b)

Figure 5: The 10 shots with the highest 'soccer' score on the TV news dataset: when modelling relevance (a), significantly better results can be achieved than if not (b).

data source (here: TV news video). Again, the focus is on how tagging is influenced by relevance modeling.

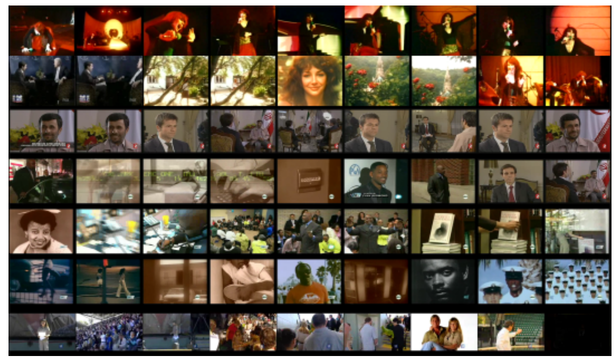
Our framework is trained on the weakly annotated youtube database that has been used in Experiment 1 (Section 4.1). For testing, a dataset of 5.5 hrs. of German news TV is used (which corresponds to 7200 shots). Tests are run for the two concepts 'soccer' and 'interview', which were expected to appear frequently in the TV dataset. The test set was manually assessed in a bootstrap manner by labeling top results given by several visual features and fusion strategies. For 'soccer', 17 % of the dataset were assessed, whereas 150 'soccer' shots were found. For the concept 'interview', significantly more positive shots were found (1300) by assessing 36 % of the dataset.

For the 'soccer' detector, the best setup from Experiment 1 was used (color and Tamura features, χ^2 distance, bandwidth 0.00075). For the 'interview' tagger a motion-based descriptor of tiled histograms was used that showed a superior performance compared to color and Tamura features in previous experiments [19] (χ^2 distance, bandwidth 0.00075).

We obtain very different results when testing the proposed relevance modelling for both concepts. For 'soccer', Figure 4 (left) shows a significant improvement by 36 % when modelling relevance. Figure 5 illustrates this improvement: the 10 shots with the highest 'soccer' score on the TV news dataset are plotted if using relevance (a) and if not (b). While the baseline system (relevance prior 1) gives 6 false positives, the proposed approach (relevance prior 0.1) yields a perfect result.

The situation is different for the 'interview' concept, where relevance modelling has a negative impact on tagging performance. An in-depth analysis of the filtered content (as shown in Figure 6) reveals that relevant content is not properly separated from non-relevant one. Obviously, the concept 'interview' – which belongs to the difficult tags in our youtube dataset – is not captured well by the motion features used, and relevance modelling "throws away" a significant fraction of valuable information. The conclusion we draw from this is that for relevance modelling to work, the underlying features must be discriminative for the concept.

(a)



(b)

Figure 6: Frames classified as relevant (a) and irrelevant (b) for the concept 'interview'. A high fraction of frames judged as non-relevant shows false negatives.

5. DISCUSSION

In this paper, the challenge of training concept detectors on videos downloaded from online portals was addressed. The benefit of this data source is that training can take place without human supervision by learning from automatically downloaded online videos, involving no manual extra annotation work.

A key problem when learning from online videos is that they are weakly labeled: Videos are not tagged on shot level and may contain a significant fraction of *non-relevant* material, i.e. shots not visually related to the semantic concept. We have presented a rigorous formulation of the problem as well as a probabilistic approach for training concept detectors in the presence of non-relevant content, whereas the relevance of a training keyframe is modelled as a latent random variable.

In quantitative experiments with sports tags on online videos and TV news data, we have demonstrated that learning such relevance is possible, and that by doing so the performance of concept detectors can be improved significantly. An experiment with the visually complex tag 'interview' revealed further that the performance of relevance detection is bound to the suitability of the underlying features used. Otherwise, information can be lost which is valuable for tagging.

While this paper gives a proof of concept with experiments on a limited number of tags, we plan to investigate the applicability of the proposed approach further in the future. This includes evaluations on a broader range of tags. Of particular interest for this might be standard datasets like the ones used in the TRECVID video evaluation campaign. These provide manually annotated ground truth of high quality and at a large scale. Long-term, we hope this work will arouse more research interest in learning from weakly labeled video data.

6. ACKNOWLEDGEMENTS

This work was supported in part by the Stiftung Rheinland-Pfalz für Innovation, project InViRe (961-386261/791).

7. REFERENCES

[1] D. Borth, A. Ulges, C. Schulze, and T. Breuel. Keyframe Extraction for Video Tagging and Summarization. In *GI-Informatiktage*, 2008.

[2] M. Campbell, A. Haubold, M. Liu, A. Natsev, J. Smith, and J. Tesic. IBM Research TRECVID-2007 Video Retrieval System. In *TRECVID Workshop*, Gaithersburg, USA, November 2007.

[3] A. Dempster, N. Laird, and D. Rubin. Maximum Likelihood from Incomplete Data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[4] T. Deselaers, D. Keysers, and H. Ney. Discriminative Training for Object Recognition Using Image Patches. In *CVPR*, pages 157–162, 2005.

[5] R. Duda, P. Hart, and D. Stork. *Pattern Classification (2nd Edition)*. Wiley Interscience Publications, 2000.

[6] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning Object Categories from Google's Image Search. *Computer Vision*, 2:1816–1823, 2005.

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

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

[9] S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *CVPR*, pages 2169–2178, 2006.

[10] K. Mikolajczyk, R. Mohr, and C. Bauckhage. Evaluation of Interest Point Detectors. *Intern. J. Compt. Vis.*, 37(2):1–38, 2007.

[11] K. Mikolajczyk and C. Schmid. A Performance Evaluation of Local Descriptors. In *CVPR*, pages 257–263, 2007.

[12] A. Opelt, M. Fussenegger, and P. Auer. Generic Object Recognition with Boosting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(3):416–431, 2006.

[13] J. Philbin, O. Chum, J. Sivic, V. Ferrari, M. Marin, A. Bosch, N. Apostolof, and A. Zisserman. Oxford TRECVID 2007, Notebook paper. In *TRECVID Workshop*, 2007.

[14] C. Rosenberg and M. Hebert. Training Object Detection Models with Weakly Labeled Data. In *BMVC*, 2002.

[15] J. Sivic, B. Russell, A. Efros, and A. Zisserman. Discovering Objects and their Locations in Images. In *ICCV*, pages 370–377, 2005.

[16] C. G. M. Snoek, I. Everts, J. C. van Gemert, J.-M. Geusebroek, B. Huurnink, D. C. Koelma, M. van Liempt, O. de Rooij, K. E. A. van de Sande, A. W. M. Smeulders, J. R. R. Uijlings, and M. Worring. The MediaMill TRECVID 2007 Semantic Video Search Engine. In *TRECVID Workshop*, November 2007.

[17] H. Tamura, S. Mori, and T. Yamawaki. Textural Features Corresponding to Visual Perception. *IEEE Trans. on Sys., Man, Cybern.*, 6(8):460–472, 1978.

[18] A. Ulges, C. Schulze, D. Keysers, and T. M. Breuel. Content-Based Video Tagging for Online Video Portals. In *MUSCLE/Image-CLEF Workshop, Budapest*, 2007.

[19] A. Ulges, C. Schulze, D. Keysers, and T. M. Breuel. A System that Learns to Tag Videos by Watching Youtube. In *ICVS (accepted for publication)*, 2008.

[20] C. Yang and T. Lozano-Perez. Image Database Retrieval with Multiple-Instance Learning Techniques. In *Int. Conf. on Data Eng.*, 2000.

[21] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study. *International Journal of Computer Vision*, 73(2):213–238, June 2007.

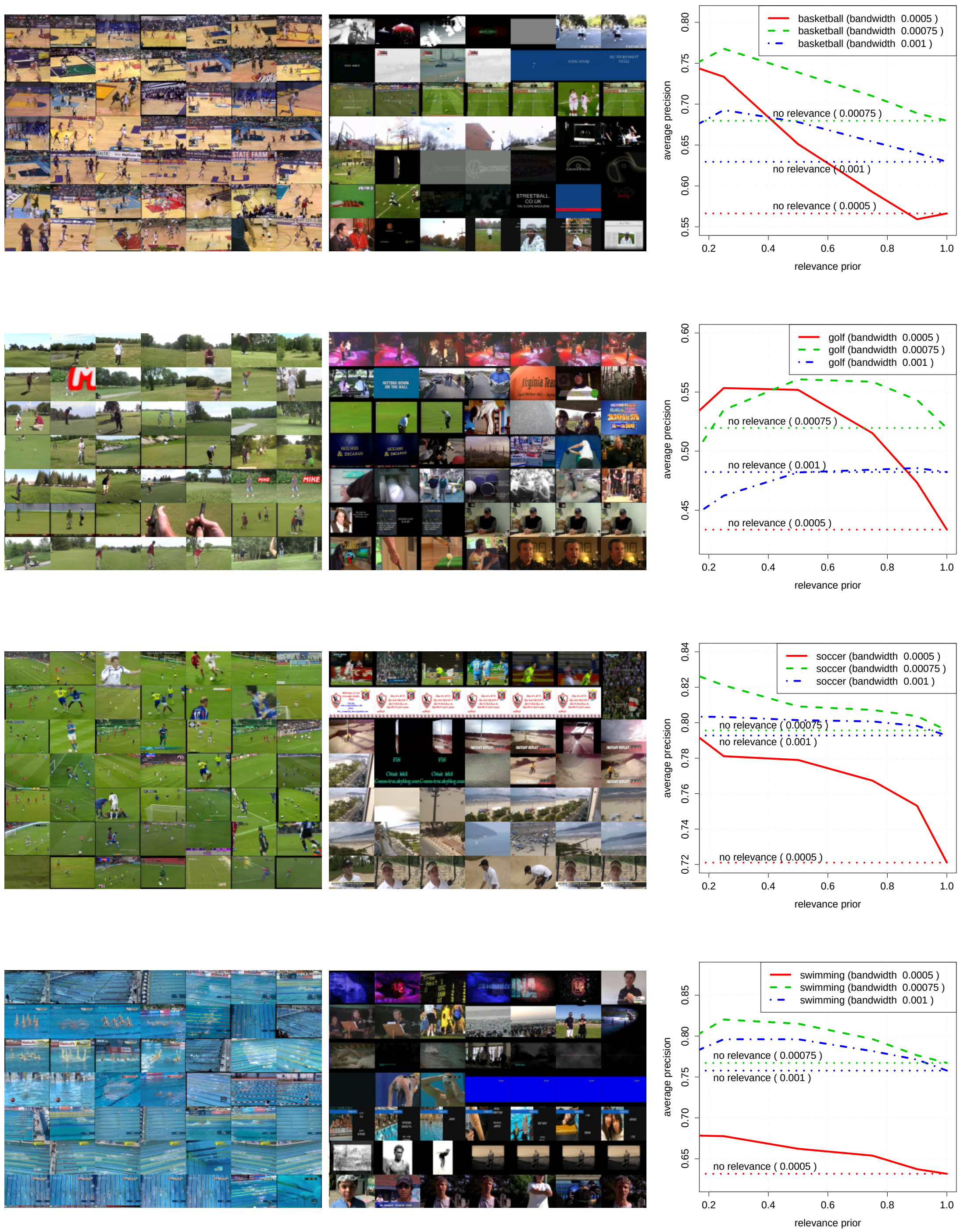


Figure 7: Results of relevance modelling for the 4 concepts 'basketball','golf','soccer', and 'swimming' (each concept corresponds to a row). Left column: the frames our approach estimates to be most relevant for the concept. Center column: frames our approach classifies as irrelevant. Right column: the tagging performance (average precision) for several kernel bandwidths, plotted against the relevance prior (the fraction of relevant frames in relevant videos). A relevance prior of 1 corresponds to a baseline system that assumes all training frames to be relevant (dotted horizontal lines).