Multiple Instance Learning from Weakly Labeled Videos

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

[1] Department of Computer Science, Technical University of Kaiserslautern
[2] Image Understanding and Pattern Recognition Group
German Research Center for Artificial Intelligence (DFKI), Kaiserslautern
{ulges,schulze,tmb}@iupr.dfki.de

Abstract. Automatic video tagging systems are targeted at assigning semantic concepts ("tags") to videos by linking textual descriptions with the audio-visual video content. To train such systems, we investigate online video from portals such as YouTube^{TM}as a large-scale, freely available knowledge source. Tags provided by video owners serve as weak annotations indicating *that* a target concept appears in a video, but not *when* it appears. This situation resembles the *multiple instance learning* (MIL) scenario, in which classifiers are trained on labeled bags (videos) of unlabeled samples (the frames of a video).

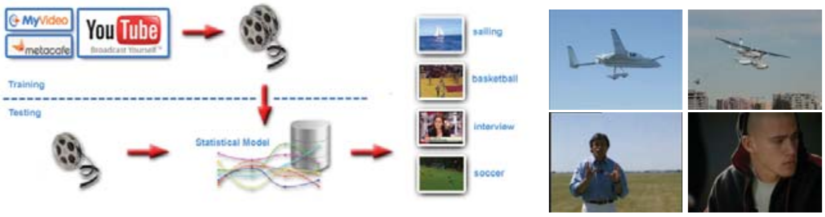
We study MIL in quantitative experiments on real-world online videos. Our key findings are: (1) conventional MIL tends to neglect valuable information in the training data and thus performs poorly. (2) By relaxing the MIL assumption, a tagging system can be built that performs comparable or better than its supervised counterpart. (3) Improvements by MIL are minor compared to a kernel-based model we proposed recently [13].

1 Introduction

Content-based video retrieval has recently attracted growing interest by research and industry, which is triggered by the enormous amount of digital video being generated and published world-wide. While efficient methods for data acquisition and storage exist, finding useful information within the sheer plethora of content remains a challenge.

The most successful and comfortable search strategy from a user perspective remains the text-based search allowing users to "google" inside a database. Conventionally, this approach demands a previous manual annotation of content, which is time-consuming, expensive and simply not possible in many situations. To overcome this problem, *automatic video tagging* has been investigated (the task is alternatively referred to as "concept detection" or "high-level feature extraction"), which aims at building computer systems that detect semantic concepts in video data by analyzing its audio-visual content. This way, a text-based index can be created automatically.

To realize such a computer-based tagging, machine learning techniques are used that analyze correlations between the audio-visual content of a video and textual keywords associated with it. This requires training sets of annotated videos, which must be large-scale due to the high number and visual complexity of concepts. The cost associated with the manual acquisition of such large training sets poses a key burden for the practical use of concept detection systems.



(a) (b)

Fig. 1. (a) A video tagger that learns its underlying statistical models autonomously from video sharing portals. (b) Frames sampled from youtube videos tagged with "airplane". While some *relevant* frames do show the target concept (top), other *non-relevant* frames (bottom) are not (or only indirectly) related to it.

2 Learning from Weakly Labeled Online Videos

To overcome the lack of annotated training data, we have recently proposed to train tagging systems on online video [12], which is freely available at a large scale from portals like YouTube^{TM}. In this setup, the tags provided by users during upload serve as labels in a machine learning framework, which allows for a totally autonomous training free of manual annotation effort as is illustrated in Figure 1: the tagging system acquires its training set from online video portals and trains itself on these web videos.

On the downside, online video as training data is of a significantly lower quality than conventional datasets, which are labeled according to precise visual criteria (assessors are given instructions like "find shots that take place outdoors at night, but no sporting events under lights" [TRECVID 2008]). In contrast to this, tags at portals like YouTube^{TM}are often given with an intention that links the tag only indirectly to the visual content. Further, online videos come with tags for the whole video that do not tell us *when* in a video a target concept appears. Overall, online video clips contain *relevant* content as well as *irrelevant* parts. This phenomenon is illustrated in Figure 1, which shows key frames from clips tagged with the concept "airplane": some *relevant* frames do in fact show airplanes, while other *irrelevant* ones are not visually related to the target concept at all.

Obviously, training a video tagger on all (including non-relevant) content can lead to significant performance loss. This issue is addressed here by using *multiple instance learning* (MIL). In contrast to standard supervised learning – which is targeted at classifying single samples – MIL aims at classifying *bags* of samples (which correspond to videos in our case), whereas latent positive samples in the bag (here: relevant frames) cause it to be labeled positive regardless of the presence of negative samples (irrelevant frames). Training is performed on labeled bags of unlabeled samples.

We study two MIL models of varying strictness in quantitative experiments on real-world online videos. Our contributions are: (1) to our knowledge the first study applying MIL to video tagging, (2) a modification of the MIL-BPNET [14] algorithm based on a relaxed formulation of MIL, and (3) experiments demonstrating moderate performance gains for the modified multiple instance classifier.

3 Related Work

To the best of our knowledge, video annotation on weakly labeled videos has not been studied except for prior work we presented in [13], where a probabilistic framework for modelling content relevance has been introduced. Here, we compare this framework to multiple instance learning.

Research on *video tagging* in general is concentrated in the TRECVID campaign[3], an open metrics-based evaluation that provides researchers with common datasets, evaluations, and a forum for discussion [10]. In TRECVID's "high-level features" task, the automatic annotation of videos is addressed. Our work differs from the one in TRECVID in the sense that it is targeted at *weakly* labeled training content, whereas in TRECVID strong but cost-intensive high-quality annotations on shot level are used.

For learning from *weakly labeled training data*, several models have been studied in the context of *image* annotation. Here, ground truth labels are usually given on image level and no region-wise annotations are provided. Kück et al. [7] proposed a constrained semi-supervised model for this setup that is solved using an MCMC sampler. Other models build joint co-occurrences of tags and regions [6] – an approach that does not explicitly identify relevant parts but takes *context* into account.

Research on *multiple instance learning* has been started by Dietterich et al. [4]. The majority of approaches is based on the assumption of a single critical positive sample per bag [1,14] or a critical region in feature space [4,8]. We will later refer to such formulations as "strict". Other, "relaxed" MIL approaches make the assumption that a bag label is not caused by a single instance, but by an arbitrary non-empty subset. Examples are the mi-SVM [1] or the model by Kück et al. [7]. While we are not aware of any work relating MIL with video content, it has been applied to the domain of still images, where images are viewed as bags of regions causing bag-level tags. Unfortunately, in this scenario MIL techniques have not been demonstrated to give performance improvements over standard supervised algorithms [9]. Results presented in this paper indicate why and suggest more relaxed MIL assumptions.

4 Approach

Given an input video X and a target concept, a tagging system estimates $P(T|X)$, where T is a Boolean random variable with $T = 1$ iff. the target concept is present in X. X is viewed as a "bag" of representative "key frames" associated with feature vectors $x_1, .., x_n \in \mathbb{R}^d$. To infer $P(T|X)$, a classifier $\Phi : 2^{\mathbb{R}^d} \rightarrow [0,1]$ is constructed such that $\Phi(X) \approx P(T|X)$. $2^{\mathbb{R}^d}$ denotes the space of all "bags", i.e. all finite sets of feature vectors.

To construct Φ, the system is trained on a set of videos $X_1, .., X_N$. Again, each X_j is a bag of key frame-related features $x_{j1}, .., x_{jn_j}$. We assume that training videos are *weakly labeled*, i.e. the associated concept presence variables T_j are observed, but we do not know *which* key frames of a video are "critical", i.e. visually related to the target concept.

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

In the following, we discuss multiple instance learning in the context of weakly labeled videos. Two formulations corresponding to different assumptions on the relevance of training data are given: (1) the multiple-instance neural network MIL-BPNET (which has been presented in [14]). (2) a relaxed formulation assuming that not a *single* instance is relevant, but a certain *fraction* of samples.

4.1 Multiple Instance Learning

This section briefly discusses the MIL-BPNET approach from [14]. The model works similar to a plain backpropagation Multi-Layer Perceptron (MLP) [5], only that it is used as a classifier for *bags* of samples $X = \{x_1, .., x_n\}$ instead of single samples.

Classification works by feeding all samples x_i to the network, obtaining scores $\phi(x_i) \in [0, 1]$. The sample with maximum output is assumed to be critical and is used for classification, such that the bag's posterior is approximated by

$$P(T|X) \approx \Phi(X) := \max_{i=1,..,n} \phi(x_i)$$

Correspondingly, training minimizes classification error on the training set of bags $X_1, .., X_N$ with relevance labels $T_1, .., T_N \in \{0, 1\}$, whereas only the critical sample in each bag is taken into account:

$$E = \sum_{j=1}^{N} \left(\max_{k=1,..,n_j} \phi(x_{jk}) - T_j \right)^2$$

This energy function is minimized in a procedure similar to backpropagation: samples are iteratively fed to the network, and for each bag $X_j = \{x_1, .., x_{n_j}\}$ the "critical sample" is determined as $\arg\max_{k=1,..,n_j} \phi(x_{jk})$. Then, the network weights are updated with respect to the critical samples previously identified.

4.2 Relaxed Multiple Instance Learning

The strict multiple instance learning setup outlined in Section 4.1 is very restrictive in a sense that it assumes a single critical sample per bag. This neglects valuable information in the training data if multiple relevant samples per bag exist. In our video-related setup it is fair to assume that videos contain more than a single key frame related to the target concept, which is why we relax the multiple instance learning condition.

The basic idea is to assume a known fraction $\alpha \in [0, 1]$ of critical instances for each bag. This set of critical samples for bag X is:

$$S_\alpha(X) = \{ \ x_i \mid \phi(x_i) \geq Q_{1-\alpha}[\phi(X)] \ \}$$

where $Q_{1-\alpha}[\phi(X)]$ is the empirical $(1-\alpha)$-quantile of the MLP scores $\phi(x_1), .., \phi(x_n)$. For example, we might assume that $\alpha = 70$ % of key frames sampled from each "basketball" video are in fact visually related to the concept, and assume that these are the ones with the highest MLP scores $\phi(.)$.

The parameter α is unknown and can vary between videos and between concepts. In the current setting, we assume that it is learnable using a validation set, and that no variation between videos of a concept takes place (a more sophisticated way would be to introduce a prior for α similar to topic priors in topic models [2]).

Classification is again based on the assumption that a fraction α of samples in the bag is positive. Their scores are averaged:

$$P(T|X) \approx \Phi(X) := \frac{1}{|S_\alpha(X)|} \sum_{x \in S_\alpha(X)} \phi(x)$$

Training is also similar to the multiple instance version. Iteratively, critical samples are estimated and used for updating the network. Again, instead of a single critical sample per bag, a fraction of samples $S_\alpha(X)$ is estimated and used for updating the network weights in order to minimize the following energy function:

$$E = \sum_j \sum_{x \in S_\alpha(X_j)} (\phi(x) - T_j)^2$$

Let us briefly discuss the behavior of the model for extreme choices of the "critical fraction" α: if $\alpha \to 0$, only the single sample with the highest score is considered critical, i.e. the model corresponds to the MIL model outlined in Section 4.1. For $\alpha \to 1$, the model becomes fully supervised: *all* samples are considered relevant, and bag classification is done using a sum rule fusion over all samples in the bag.

5 Experiments

In the following experiments on a dataset of real-world online videos it is studied whether multiple instance learning can make video taggers more robust to irrelevant training content. First, the dataset and experimental protocol are outlined. After this, we study whether strict MIL (Section 4.1) leads to performance improvements compared to a conventional supervised setup. Third, the relaxed MIL model introduced in Section 4.2 is tested. Finally, the results are compared to the generative weakly supervised approach introduced in [13].

5.1 Dataset and Protocol

We test our framework on the *youtube-22concepts* dataset of real-world online videos downloaded from the portal YouTube$^{\text{TM}}$. The dataset consists of 2200 videos (about 194 hrs.) and has been made available for research purposes[4]. It contains 100 clips for each of 22 semantic concepts that include locations (e.g., "desert", "beach"), actions (e.g., "hiking", "interview"), objects (e.g., "cat", "eiffeltower"), and sports (e.g., "swimming", "soccer"). Tests were run for 5 concepts: the sports "basketball", "golf", "soccer", and "swimming", as well as the tag "desert". Our motivation for the sports

[4] http://sites.google.com/a/iupr.com/iupr-image-and-video-computing/youtube-22-concepts-dataset

candidates is that good tagging performance has been achieved even with simple features in previous experiments, which makes them good candidates for studying the influence of multiple instance learning. “desert” is a more challenging tag, for which weakly supervised learning similar to MIL has previously been demonstrated to give good results [13].

The dataset was split into a training set of 66 % and a test set of 33 % using stratified sampling. Key frames were extracted using a shot boundary detection and intra-shot clustering of frames [3]. For each key frame, color histograms and Tamura texture features [11] were concatenated to a joint 1024-dimensional feature vector.

The parameters for neural network training were set to 5 hidden units, a fixed backpropagation learning rate (0.5) and 100 training epochs. In previous test, results were found rather insensitive with respect to these parameters. Since results vary with the initialization of backpropagation, the median results over multiple runs are presented.

To measure performance for each concept, test videos are sorted by their score $P(T|X)$, obtaining a ranked retrieval list. By thresholding this list at different positions, a recall-precision curve is obtained, and the average precision AP (which corresponds to the area under the curve) is used as a performance measure.

5.2 Multiple Instance Learning

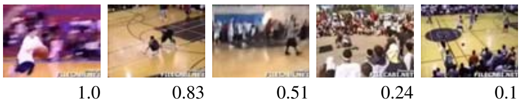
We first evaluate multiple instance learning as discussed in Section 4.1. As a baseline serves the fully supervised MLP, which uses all key frames for training (including the ones not visually related to the target concept). Note that another interesting benchmark would be training on relevant key frames, but this is thus omitted here due to the time and cost associated with manual ground truth annotation.

Figure 2 gives the median results over 7 backpropagation restarts for both methods. For all concepts tested, multiple instance learning is significantly outperformed and gives poor results. An explanation for this is also given in Figure 2: for a “basketball” training clip, the 5 samples are plotted for which the MIL classifier gives the highest scores at the end of training. Under each image, the corresponding score can be found. Obviously, the system has overfitted to a few examples it assumes to be critical. The rest of the bag – though it contains a significant amount of basketball content – is neglected in training and thus given incorrect low scores. Similar observations have been made by Ray and Craven [9] in the context of image annotation.

Fig. 2. left: The average precision (AP) (%) for the MIL-BPNET and a supervised MLP (median over 7 runs). **right:** the five most critical frames of a “basketball” training video with their associated MIL-BPNET scores. The algorithm overfits to a few sample, and other relevant content is given incorrect low scores.

	basketb.	golf	soccer	swim.	desert
MIL	42.9	17.3	46.7	34.8	15.2
sup.	70.1	51.9	83.1	76.1	22.2

(a)



(b)

5.3 Relaxed Multiple Instance Learning

While the multiple instance learning assumption seems too restrictive, the next experiment studies the relaxed MIL setup discussed in Section 4.2. Thereby, the fraction of relevant material α was tested for $\alpha \in \{0.1, 0.25, 0.5, 0.75, 0.9, 1.0\}$, and the choice of α was picked that performs best on the test set. We assume that α can be determined reliably via cross-validation, as has been demonstrated for a similar parameter in [13].

Table 1 gives the median performance for both methods. Relaxed MIL outperforms strict MIL significantly and gives performance improvements over the supervised case for three of five concepts (4.9 % for "basketball", 2.8 % for "desert", and 3.5 for "golf"). For "soccer" and "swimming", the standard supervised setup performs comparably.

Table 1. The average precision (AP) (%) for the MLP framework presented here and the kernel density framework presented in [13]. Both methods are compared with their fully supervised counterparts.

	concept	basketball	golf	soccer	swimming	desert
MLP	rel. MIL	75.0	55.4	83.5	76.1	25.3
	sup.	70.1	51.9	83.1	76.1	22.5
Kernel Density Estimation	Kernel RM	**85.6**	**57.6**	**84.4**	**80.5**	**29.0**
	sup.	71.2	51.6	78.5	76.5	11.8

5.4 Comparison with Probabilistic Relevance Estimation

We also compare MIL to another system following the idea of modelling irrelevant training video content. This "kernel-based relevance modelling" (KRM) approach was introduced in [13]. It models densities of relevant and irrelevant material in a probabilistic manner, whereas – similar to the MIL approach – the relevance of training frames is estimated using Expectation Maximization. The differences to the MLP framework presented here are: (1) the decision whether a training frame is relevant is absolute in MIL, whereas KRM estimates smooth probabilities of relevance, (2) both systems deal differently with non-relevant content: while it is simply ignored in the MIL framework, it influences the corresponding density in KRM (3) both relaxed MIL and KRM assume a fraction of relevant material in relevant videos. However, while the first one assumes this fraction on video level, the other defines it over the whole collection.

It can be seen in Table 1 that KRM tends to benefit more from modelling irrelevant content. For all concepts, an improvement compared to the fully supervised case is achieved that can reach up to 17.2 % (desert). Due to this fact, KRM outperforms both MIL and the supervised system on all concepts.

6 Discussion

We have addressed the challenge of visual learning from weakly labeled videos with the practical motivation of training automatic video tagging systems on online video content. To achieve such weakly supervised learning, a relaxed formulation of MIL was suggested as a modification of the MIL-BPNET algorithm [14].

In quantitative experiments using on a dataset of real-world online videos, the following key observations have been made: (1) While strict MIL was significantly outperformed, relaxed MIL gave moderate performance improvements compared to a supervised system. (2) The benefits of multiple instance learning are lower than for a kernel-based model we have introduced recently [13]. However, the MIL approach followed here offers a better scalability, which is interesting in real-world large-scale setups.

Finally, we point to the annotations of web videos as another information pool of potential interest. In a video search framework, this information might help to link queries with concepts or concepts with detectors. While we have focused on the visual aspect of concept detection in this paper, we believe that video concept detection and search could benefit from a joint learning on textual descriptions and content[5].

References

1. S. Andrews, I. Tsochantaridis, and T. Hofmann. Support Vector Machines for Multiple-Instance Learning. In *NIPS '03*, pages 561–568, 2003.
2. D. Blei, A. Ng, and M. Jordan. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.
3. D. Borth, A. Ulges, C. Schulze, and T. M. Breuel. Keyframe Extraction for Video Tagging and Summarization. In *Proc. Informatiktage 2008*, pages 45–48, 2008.
4. T. Dietterich, R. Lathrop, and T. Lozano-Perez. Solving the Multiple Instance Problem with Axis-parallel Rectangles. *Artif. Intell.*, 89(1-2):31–71, 1997.
5. R. Duda, P. Hart, and D. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.
6. S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple Bernoulli Relevance Models for Image and Video Annotation. In *Proc. Int. Conf. Computer Vision Pattern Recognition*, pages 1002–1009, June 2004.
7. H. Kück, P. Carbonetto, and N. de Freitas. A Constrained Semi-Supervised Learning Approach to Data Association. In *Proc. Europ. Conf. Computer Vision*, pages 1–12, 2004.
8. O. Maron and T. Lozano-Perez. A Framework for Multiple Instance Learning. In *NIPS '97*, pages 570–576, 1998.
9. S. Ray and M. Craven. Supervised versus Multiple Instance Learning: an Empirical Comparison. In *ICML '05*, pages 697–704, 2005.
10. A. F. Smeaton. Large Scale Evaluations of Multimedia Information Retrieval: The TRECVid Experience. In *Proc. Int. Conf. Image and Video Retrieval*, pages 11–17, July 2005.
11. H. Tamura, S. Mori, and T. Yamawaki. Textural Features Corresponding to Visual Perception. *IEEE Trans. System, Man, Cybernetics*, 8(6):460–472, 1978.
12. A. Ulges, C. Schulze, D. Keysers, and T. M. Breuel. A System that Learns to Tag Videos by Watching Youtube. In *Proc. Int. Conf. on Vision Systems*, pages 415–424, May 2008.
13. A. Ulges, C. Schulze, D. Keysers, and T. M. Breuel. Identifying Relevant Frames in Weakly Labeled Videos for Training Concept Detectors. In *Proc. Int. Conf. Image and Video Retrieval*, pages 9–16, Jul 2008.
14. M.-L. Zhang and Z.-H. Zhou. Improve Multi-Instance Neural Networks through Feature Selection. *Neural Process. Lett.*, 19(1):1–10, 2004.

[5] This work was supported in part by the Deutsche Forschungsgemeinschaft (DFG), project MOONVID (BR 2517/1-1).