Video Corpus Annotation using Active Learning

Stéphane Ayache and Georges Quénot

Laboratoire d'Informatique de Grenoble (LIG)
385 rue de la Bibliothèque - BP 53
38041 Grenoble - Cedex 9 - France

Abstract. Concept indexing in multimedia libraries is very useful for users searching and browsing but it is a very challenging research problem as well. Beyond the systems' implementations issues, semantic indexing is strongly dependent upon the size and quality of the training examples. In this paper, we describe the collaborative annotation system used to annotate the High Level Features (HLF) in the development set of TRECVID 2007. This system is web-based and takes advantage of *Active Learning* approach. We show that *Active Learning* allows simultaneously getting the most useful information from the partial annotation and significantly reducing the annotation effort per participant relatively to previous collaborative annotations.

1 Introduction

Semantic content-based access to image and video documents is a strong need for many industrial applications. Indexing concepts in images and in video segments is the main key to enable it and it is still a research challenge. Due to the so called *semantic gap* between the raw image or video contents and the elements that makes sense to human beings, indexing concepts in image or video documents is a very hard task. It is most often carried out using classifiers or networks of classifiers [10, 14, 3] trained using supervised learning. Systems' performance depends a lot upon the implementation choices and details but it also strongly depends upon the size and quality of the training examples. While it is quite easy and cheap to get large amounts of raw data, it is usually very costly to have them annotated because it involves human intervention for the judging of the "ground truth".

Many research works on content-based image and video indexing are conducted in the context of the TRECVID campaigns [13]. These campaigns provide to the participants a complete framework with data collections, well defined tasks, ground truth and metrics for the evaluation of indexing and/or retrieval systems. Additionally, annotated data are provided for some tasks like the "High Level Feature (HLF) extraction task" which is actually a concept indexing task. Large annotation efforts were organized in 2003 [8] and 2005 [15, 9] in order to produce a complete annotation of the development set for a series of target concepts. These initiatives produced very valuable resources but at a very high cost.

While the volume of data that can be manually annotated is limited due to the cost of manual intervention, there remains the possibility to select the data

samples that will be annotated so that their annotation is "as useful as possible" [1]. Deciding which samples will be the most useful is not trivial. *Active learning* is an approach in which an existing system is used to predict the usefulness of new samples. This approach is a particular case of *incremental learning* in which a system is trained several times with a growing set of samples. The objective is to select as few samples as possible to be manually indexed and to get from then the best possible classification performance.

In this paper, we describe the use of active learning technique for annotation of unlabeled video corpus. In order to provide manually annotation on the TRECVID 2007 development set at cheapest cost, we organized a web-based collaborative annotation tool in the spirit of what was done in the 2003 and 2005 [15]. Active learning has been used in order to simultaneously get the most useful information from the partial annotation and significantly reduce the annotation effort per participant relatively to previous collaborative annotations. In the following of this paper, we first describe previous active learning experiments and then present the principles and the organization of this project and the lessons learnt from it.

2 Simulated active learning

In a previous work, [2] simulated an active learning process using the TRECVID 2005 fully annotated development set and the TRECVID 2006 test set and metrics. By progressively including annotations in the training set, various active learning strategies have been evaluated in a variety of conditions. Results have been obtained using a particular corpus (TRECVID 2005/2006), a particular type of concepts (LSCOM-lite) and using a particular learning system (network of SVM classifiers). They might not transpose directly to other types of contents, target concepts or learning system though we expect the observed general trends to still be valid.

Three strategies were compared: "relevance sampling", "uncertainty sampling", and "random sampling". The two first strategies respectively select the most probable and the most uncertain samples [7]. The third one is a random choice. Here are the main conclusions:

- For easy concepts, the "relevance sampling" strategy is the best one when less than 15% of the dataset is annotated and the "uncertainty sampling" strategy is the best one when 15% or more of the dataset is annotated.
- The "relevance sampling" and "uncertainty sampling" strategies are roughly equivalent for moderately difficult and difficult concepts. In all cases, the maximum performance is reached when 12 to 15% of the whole dataset is annotated.
- The previous results depend upon the step size and the training set size. $1/40^{\text{th}}$ of the training set size is a good value for the step size.
- The size of the subset of the training set that has to be annotated in order to reach the maximum achievable performance varies with the square root of the training set size.

– The “relevance sampling” strategy is more “recall oriented” while the “uncertainty sampling”’ strategy is more “precision oriented”.

Figure 1 shows the evolution of the system Mean Average Precision (MAP, actually inferred average precision as it was introduced in TRECVID 2006) with the number of annotated samples for the three strategies and with an active learning step size of $1/40^{\text{th}}$ of the training set size. The active learning process was initialized with a set of 10 positive samples and 20 negative samples randomly chosen (the assumption is that the user has at least a few positive examples of what he is looking for and that negative examples are easy to find). What is remarkable is that the maximum system performance is reached when only a small fraction of the development set is annotated if this fraction is carefully chosen. Here the fraction is of about 12-15% for a development set size of 36014 samples. Other experiments (not shown here) indicate that this is also the case for different development set sizes and that the optimal fraction varies with the square root of the development set (it is of about 25-30% of the development set if its size is reduced to 9003 samples).

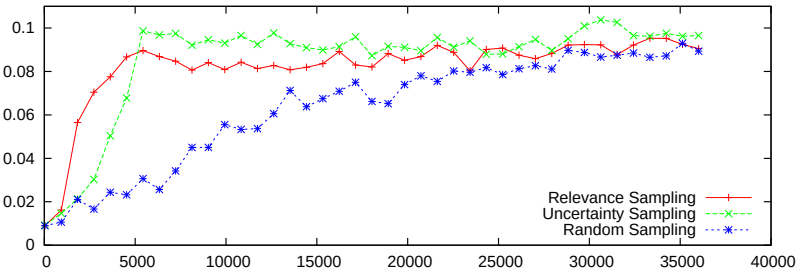

Fig. 1. Evolution of system MAP with the number of annotated samples for the three strategies, all concepts.

Figure 2 shows the evolution of the number of positive samples found (average on all concepts) as a function of the number of annotated samples for the three strategies. The rate of finding of positive samples near the beginning are of about 2.4:1 and 4.5:1 for “uncertainty sampling” and “relevance sampling”’ strategies respectively relatively to the “random” choice.

3 Collaborative annotation system

For the TRECVID 2003 annotation effort, [8] provided a tool to facilitate multimedia annotation tasks for general users. This tool generated MPEG-7 compatible outputs and included various features from video shot segmentation to ontology editing and region based annotation. However, Videoannex was a standalone system, thus each user needs to get possession of the entire collection

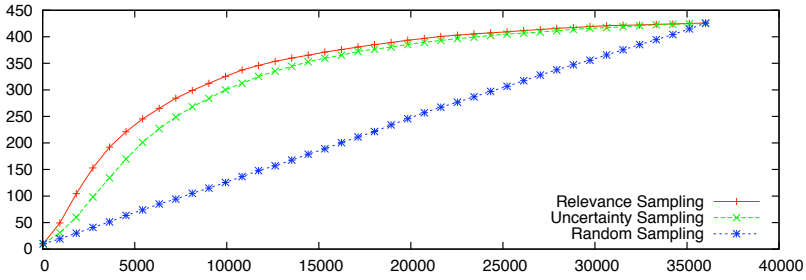

Fig. 2. Evolution of the number of positive samples found with the number of annotated samples for the three strategies, all concepts.

and the annotation data must be collected afterwards. Moreover, this tool was not user centered as it forced to annotate all available concepts from the ontology simultaneously. The TRECVID 2005 collaborative annotation system was a web-based application that allowed users to annotate using a web browser [15]. Thanks to the centralized architecture, the system was able to display a set of overall statistics during an annotation session.

Our system is web-based and relies on an active learning approach. Similar approaches have already been considered for image and video indexing or retrieval [5, 11] but not yet in the context of a web based collaborative annotation. As this was done in the previous collaborative annotation, we produced samples at the subshot level since these are much more likely to have a homogeneous and non ambiguous content. In order to ease the annotation process, annotation is consists to judge one key frame per subshot. We finally extracted 21532 key frames using the video segmentation tool described in [12]. The following subsections describe the interface and organization of the collaborative annotation system.

3.1 Web interface

The TRECVID 2007 Collaborative Annotation system has been designed to be efficient and easy to use. Like the TRECVID 2005 collaborative annotation system [15], it operates through the Web and requires no local software installation. Participation is restricted to groups that are registered TRECVID participants and that have signed a license agreement to access the video data.

The system has two modes of operation: a sequential mode in which the images to annotate are displayed one by one and a parallel mode (Figure 3) in which the images are displayed by groups in a two-dimensional array. In the parallel mode, users can define the dimensions of the array in order and adapt visualization to his screen size.

Users were required to annotate only one concept at a time. The system gave priority to the concept which had the less annotated samples. For the current concept to annotate, images are displayed, either one by one or by group depending upon the mode chosen, and for each image the user has three choices

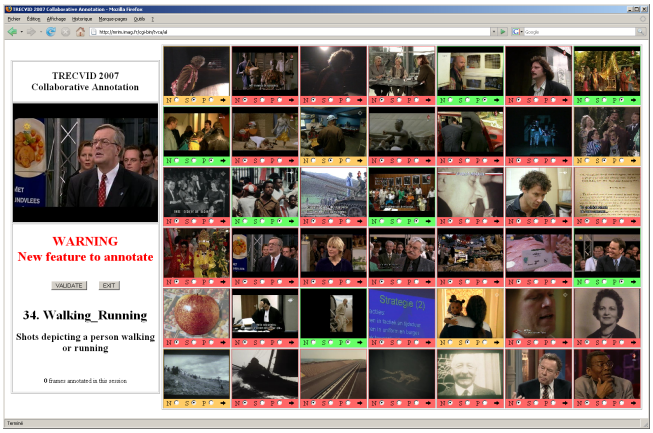

Fig. 3. Parallel interface of the annotation system.

for its annotation: POSITIVE (the concept is clearly there), NEGATIVE (the concept is clearly not there), SKIPPED (any other case, whatever the cause of uncertainty).

In the parallel mode, users see by default an image at a smaller resolution than the video one (160 × 120 instead of 352 × 288). By passing the mouse over one of the small images they can get an enlarged view of it in a corner of the screen. In both modes, users also have the possibility to play the whole video shot if they feel that this can help them to make a better decision. This is often the case for "dynamic" concepts like "Walking_Running".

3.2 Organization

TRECVID participants register as teams and each team may have several users doing the annotation. In order to encourage participation to the collaborative annotation, the resulting annotation is available only to the teams that have completed a minimum amount of annotations, as this was also the case in previous TRECVID collaborative annotations. The minimum annotation effort was set to 3% of the total number of annotations that should be done in order to annotate each key frame/subshot for each concept once. This amounts to 23255 annotations per team and can be completed in about 13 hours considering an average annotation time of 2 seconds per key frame × concept.

4 Active learning system

We implemented the same system described in [2]: an iterative process which use samples score from previous iteration to sort samples depending of the strategies. The active learning process was running permanently during the whole annotation period (over two months). It has been optimize in order to run with a

parallel implementation on 10 bi-processor (3 GHz P4) servers. The process continuously computed (training/prediction) one concept at once. Hence, in order to have similar annotation progress for the concepts, the system continuously chooses the concept which received the largest number of annotations since its last training. Consequently, there is not any step size as iterations occur when a concept has been selected by the system.

The collaborative annotation system also runs permanently and independently of the active learning process. The Collaborative annotation process uses the last version of the classification system produced by the active learning system in order to select the samples for annotation. Similarly, the active learning system uses the last available set of annotations to re-train the classification systems.

4.1 Classification system

The classification system used for the active learning process is derived from the one used for our participation the TRECVID 2006 high level feature extraction task. Since the language used in both collections is different and since the English machine translation was not available yet, we used two variants, one using the text input and the other not using it.

The system is detailed in [3]. It uses visual and text features when available. Visual features include local and global features and both include color, texture and motion low-level features. The system uses network of SVM classifiers [4] and implements a mix of early and late fusion schemes. Its performance on the TRECVID 2006 HLF extraction task was slightly above the median with an Inferred Average Precision of 0.088.

4.2 Cold start and strategies

Since the concepts to annotate are the same in 2005/2006 and 2007, we can use a system trained only on 2005 data for starting the selection of the samples on the 2007 data. This is a challenge since the 2005 and 2007 corpora are quite different on visual, sound and text modalities. The "cold start" strategy was finally to begin the training with only 2005 samples and then to progressively replace as many as possible of them by 2007 samples. This was done until enough 2007 positive and negative samples were found. This was quite hard to judge but we finally decided to remove the last 2005 samples and therefore switch to "2007 only" training when 25% of the development set was annotated.

During the mixed training phase, using both 2005 and 2007 samples, it was not possible to use the text features in the classification system since no common representation was possible (English vs. Dutch language). This phase was therefore completed using only the visual content. The text was finally added as an additional feature for classification after the switch to 2007 only. It was actually introduced when about 40% of the development set was annotated both because we wanted to observe and distinguish both effects.

We started with the "relevance sampling" strategy as it was identified as the most efficient for the beginning of the process. Switching to the "most uncertain" strategy was considered at a time but we finally did not activate it as the expected gain was low and because we still wanted to observe other effects that might have interacted with it.

We implemented an additional strategy in order to boost annotation of positive samples, we call "neighborhood sampling". It consists in looking for new positive samples in the temporal neighborhood of already found positive samples. Each time a positive sample has been found, the preceding and following samples (previous and next subshots in the same video file) are selected with the highest priority for annotation. This additional strategy was used jointly with the "relevance sampling" strategy and it was activated early, when about 1.5% of the development set was annotated.

5 Quality

From the TRECVID 2005 collaborative annotation study [15], it was observed that disagreement among annotators occurred for about 3% of the annotated key frame $\times$ concepts. These are due sometimes to obvious mistakes, to misunderstanding of the concept or to subjective interpretation of the key frame/subshot contents. We had an additional source of inconsistency that is that some users apparently failed sometimes to notice the change of the concept to annotate despite the displayed warning. Such changes occur quite frequently since they are required by the active learning framework. Those various wrong annotations introduced some false positive and negatives which could affect the active learning process.

Since we wanted to keep the annotation effort reasonable, we did not want to have most of the concept being annotated several times. We decided to have a multiple check of only the most suspect annotations. We used for that the active learning approach by re-proposing the samples that have been predicted as most misclassified (i.e. positive annotated samples that were most probably predicted as negative and vice versa). All samples marked as skipped were also proposed for a second annotation. In case of disagreement between the first and second annotation of a key frame $\times$ concept, this one was proposed for a third judgment and a majority voting was used for making the final decision. As indicated in the following section, only a small fraction of the samples have been annotated twice, an even smaller fraction was annotated three times and so on while these were done as cleverly as possible to clean up as much as possible the collaborative annotation.

6 Analysis

32 teams participated to the 2007 TRECVID collaborative annotation effort and produced a total of 711566 annotations. Table 1 gives some statistics on

these annotations. “Pass 1”, “Pass 2”, “Pass 3” and “Pass 4” corresponds to the number of annotations that were done respectively at least once, at least twice, at least three times and at least four times for a given key frame × concept. The “Synthesis” correspond to the global annotation when a “majority” rule is applied if there is more than one annotation for a key frame × concept.

	Annotated	% Annotated	Negative	Skipped	Positive	% Positive
Pass 1	641223	82.7	578299	13163	49761	7.76
Pass 2	46864	6.05	11904	7478	27482	58.6
Pass 3	21987	2.84	9383	4040	8564	39.0
Pass 4	1492	0.19	324	940	228	15.3
Synthesis	641223	82.7	578683	15348	47192	7.36

Table 1. Annotation statistics by pass, average on all concepts.

Table 2 indicates the frequency of the concepts in the collection. These figures come from incomplete data and this may cause a bias. Thanks to the active learning approach and to the fact that 75-90% of the corpus has been annotated, the bias is expected to be negligible except for the most frequent concepts like “Face” or “Person”.

Flag-US	0.06	Maps	0.60	Military	2.31	Crowd	8.56
Prisoner	0.15	Mountain	0.65	TV-screen	2.99	Walking_Run.	9.69
Weather	0.18	Truck	0.67	Car	3.68	Urban	9.70
Explosion_Fire	0.24	Court	0.73	Studio	4.22	Building	12.1
Natural-Disaster	0.25	Snow	0.75	Meeting	4.42	Vegetation	14.3
Airplane	0.30	Police_Security	1.40	Animal	4.63	Sky	17.4
Bus	0.30	People-Marching	1.43	Waterscape	5.07	Outdoor	39.3
Desert	0.35	Sports	1.50	Road	5.92	Face	56.3
Charts	0.60	Boat_Ship	1.58	Office	7.25	Person	72.4
		Median	1.95	Average	7.36		

Table 2. Frequency of concepts (in percent).

The annotation finally reached a level of about 82% in average varying from about 75 to 90% depending upon the concepts, some having been more often multiply annotated than others. Figure 4 shows how the collaborative effort was spread over time. Horizontal units correspond to the days of May 2007 between 1 and 31 included and extrapolated outside. The effort started slowly with only the organizers (LIG) participating in order to control the size of the first active learning steps and to keep them small for an efficient start. Other users were asked to participate after a few days and to do their main effort during the following 15 days. Additional teams joined from time to time afterwards and contributed with a small but sustained effort which was mainly used for cleaning

up the collaborative annotation with double and triple checks of suspect or inconsistent annotations.

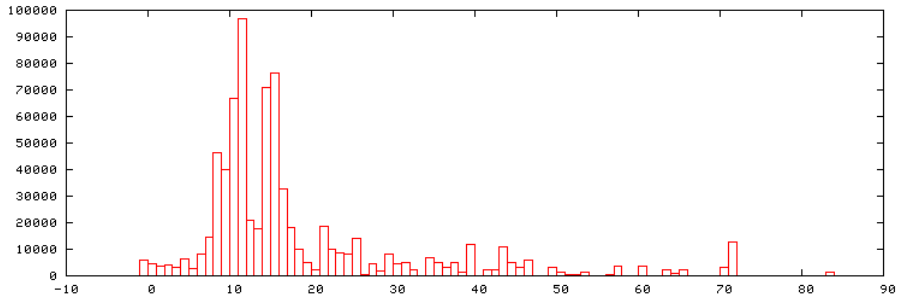

Fig. 4. Daily annotations in the collaborative annotation project (GMT time, May 2007 days).

Evolution of the number of positive samples found with the fraction of annotated samples gives idea of the reduction of effort provided by active learning method. Figure 5 shows this evolution (average for all concepts) for the TRECVID 2007 collaborative annotation. The prediction of what would have been the case for a random or sequential scan is shown as the diagonal. The shape is similar and the scale of the active learning effect is comparable. Three particular behaviors can be observed though the effects are small:

- Near the origin, at about 0.015, an increase in the finding rate is probably due to the activation of the "neighborhood sampling" strategy.
- After 0.25, an increase in the finding rate is probably due to the closing of the "cold start". Before this point, active learning uses a mix of 2005 and 2007 data; after this point, it uses only 2007 data.
- After 0.40 an increase in the finding rate is probably due to the inclusion of text feature in the classification system.

Though all these events have small effects of the overall finding rate, they may have larger effects for individual concepts. This is the case for example for the "Prisoner" concept when text features are included.

Figure 5 only shows the general trend of the evolution of the positive annotations with the total of annotations but this evolution is highly variable according to the considered concept. Figure 6 shows a superposition of the same curve for each of the 36 target concepts. The active learning effect is visible everywhere but it is more important for some concepts and sometimes more important in different regions.

Some effect linked to the fact that the cold start was done using a different collection can be observed. For instance, in the "Court", "Charts" and "Studio" concepts, the visual aspect is quite different in both collections and the active

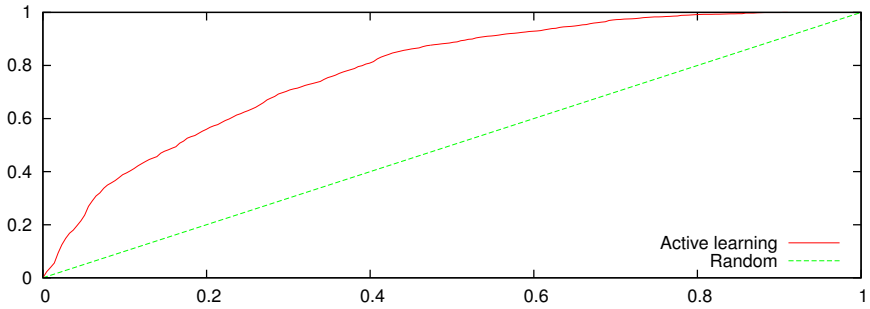

Fig. 5. Evolution of the fraction of positive samples found with the fraction of annotated samples; comparison between active learning and random annotation, all concepts.

learning has first a negative effect (less positive samples are found than what a random choice would provide) and then, when a few are finally encountered (possibly by chance) the effect becomes positive and quite strong. In figure 6, the first and second curves close to the upper left corner have these behavior and correspond respectively to "Court" and "Studio" concepts. Furthermore, we observe some "step" shapes for some concepts, this effect typically happens for some visually heterogeneous concepts. When a positive sample is found, the system possibly finds many others positives in his temporal neighborhood. In figure 6, the lower curve corresponds to the "Prisoner" concept.

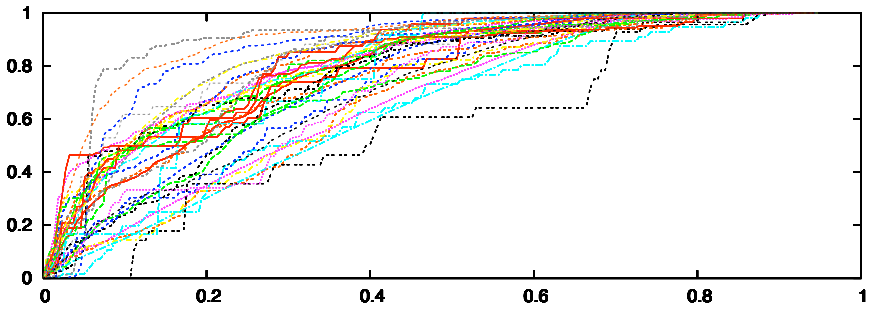

Fig. 6. Evolution of the fraction of positive samples found with the fraction of annotated samples for the 36 concepts individually.

In order to study the benefit provided the quality and diversity of the samples selected by the active learning process, we computed classification of the test set with several fraction of the learning set from 5% to 90%. Figure 7 shows the evolution of the Inferred Average Precision (IAP) (average for the 20 con-

cepts selected by TRECVID2007 for evaluation) with the number of annotated samples. The experiment has been conducted with two different systems: one from LIG wich is close to the one used for active learning during the annotation process and another from Helsinki University [6]. For the LIG system, it appears that the most useful samples are quickly selected: classification based on the 15% first annotated samples gives satisfying performance, while the classification based on the 35% first annotated samples gives the best performance. For the Helsinki University system, the best performance is reached slightly afterwards when about 50% of the samples have been annotated.

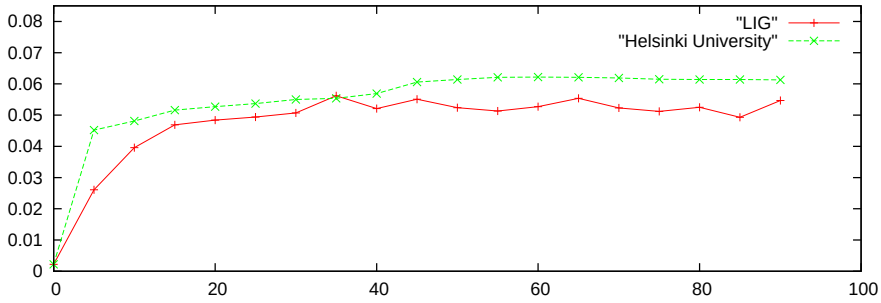

Fig. 7. Evolution of the mean of IAP of the 20 evaluated concepts with the fraction of annotated samples.

7 Conclusion

We organized the collaborative annotation of the High Level Features (HLF) in the development set of TRECVID 2007. These annotations have been used by the TRECVID 2007 participants to train their systems for the HLF extraction task. The annotation system is web-based and takes benefits of the *Active Learning* approach. This system allows participants to simultaneously get the most useful information from the partial annotation and significantly reduce the annotation effort relatively to previous collaborative annotations. We described the principles and the organization of this project and the lessons learnt from it. Previous experiments indicated that annotating only 20 to 30% of the development set would not hurt the systems' performances if these are carefully chosen. A similar behavior in the finding rate of positive samples was observed in the TRECVID 2007 collaborative annotation. While the development collection of TRECVID 2007 was quite small compared to the TRECVID 2003 and 2005 development collections, the benefits of the active learning approach for corpus annotation would be even more visible on a larger corpus to be annotated. Such an annotation system would be valuable in other machine-learning based areas, but not necessarily take benefit of the neighborhood sampling.

Bibliography

[1] D. Angluin. Queries and concept learning. *Machine Learning*, 2:319–342, 1988.

[2] S. Ayache and G. Quénot. Evaluation of active learning strategies for video indexing. *Signal Processing: Image Communication*, 2007.

[3] S. Ayache, G. Quénot, and J. Gensel. CLIPS-LSR Experiments at TRECVID 2006. In *Proceedings of the TRECVID 2006*, 2006.

[4] 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.

[5] P. H. Gosselin and M. Cord. A comparison of active classification methods for content-based image retrieval. In *CVDB '04: Proceedings of the 1st international workshop on Computer vision meets databases*, pages 51–58, New York, NY, USA, 2004. ACM Press.

[6] M. Koskela, M. Sjberg, V. Viitaniemi, J. Laaksonen, and P. Prentis. PicSOM Experiments in TRECVID 2007. In *Proceedings of the TRECVID 2007*, 5-6 Nov. 2007.

[7] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In W. B. Croft and C. J. van Rijsbergen, editors, *Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval*, pages 3–12, Dublin, IE, 1994. Springer Verlag, Heidelberg, DE.

[8] C.-Y. Lin, B. L. Tseng, and J. R. Smith. Video collaborative annotation forum: Establishing ground-truth labels on large multimedia datasets. In *Proceedings of the TRECVID 2003*, 17-18 Nov. 2003.

[9] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 13(3):86–91, 2006.

[10] M. R. Naphade and J. R. Smith. On the detection of semantic concepts at trecvid. In *MULTIMEDIA'04: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 660–667, New York, NY, USA, 2004. ACM Press.

[11] G.-J. Qi, Y. Song, X.-S. Hua, H.-J. Zhang, and L.-R. Dai. Video annotation by active learning and cluster tuning. In *CVPRW '06: Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop*, page 114, Washington, DC, USA, 2006. IEEE Computer Society.

[12] G. Quénot. Computation of Optical Flow Using Dynamic Programming. In *IAPR Workshop on Machine Vision Applications*, pages 249–52, 12-14 Nov. 1996.

[13] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[14] C. G. M. Snoek, M. Worring, and A. G. Hauptmann. Learning rich semantics from news video archives by style analysis. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2(2):91–108, 2006.

[15] T. Volkmer, J. R. Smith, and A. P. Natsev. A web-based system for collaborative annotation of large image and video collections: an evaluation and user study. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 892–901, New York, NY, USA, 2005. ACM Press.