EVALUATION OF ACTIVE LEARNING STRATEGIES FOR VIDEO INDEXING

Stéphane Ayache and Georges Quénot

Laboratoire d'informatique de Grenoble
BP 53, Grenoble Cedex 9, France
Georges.Quenot@imag.fr

ABSTRACT

In this paper, we compare active learning strategies for indexing concepts in video shots. Active learning is simulated using subsets of a fully annotated dataset instead of actually calling for user intervention. Training is done using the collaborative annotation of 39 concepts of the TRECVID 2005 campaign. Performance is measured on the 20 concepts selected for the TRECVID 2006 concept detection task. The simulation allows exploring the effect of several parameters: the strategy, the annotated fraction of the dataset, the number of iterations and the relative difficulty of concepts.

Three strategies were compared. The first two respectively select the most probable and the most uncertain samples. The third one is a random choice. For easy concepts, the "most probable" strategy is the best one when less than 15% of the dataset is annotated and the "most uncertain" strategy is the best one when 15% or more of the dataset is annotated. The "most probable" and "most uncertain" 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.

1. INTRODUCTION

Image and video databases become more and more common and large. They are found in a variety of places including home, companies and institutions and for a variety of applications. In order to keep them manageable, powerful tools are needed for searching and browsing. These tools need other tools for contents indexing. This indexing can be done at the signal level (color, texture, motion ...) or at the semantic level (concepts). From both indexing types, the latter is by far the most useful for the users but it is also by far the most difficult one to extract from the contents. 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. This task is most often carried out using classifiers or networks of classifiers [1, 2].

Georges Quénot is with CNRS

Supervised learning consists in training a system from sets of positive and negative examples. The learning system may be composed of various types of feature extractors, classifiers and fusion modules. The 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". In the context of LSCOM (Large Scale Concept Ontology for Multimedia) [3], 449 concepts have been annotated on 61901 samples. If we consider the performance obtained by the state of the art systems using this annotated data during the TRECVID 2006 evaluation campaign (a Mean Average Precision of only 0.1 to 0.2), even this seems far from being sufficient.

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 sample that will be annotated so that their annotation is as useful as possible. Deciding which samples will be the more 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. Our objective is to select as few sample shots as possible to be manually indexed but still have a good performance classifier (in the sense of Mean Average Precision).

Several strategies or heuristics can be considered to predict the samples' usefulness. Most of them operate by *selective sampling* which consists in progressively adding to the training set the samples expected to be the most informative ones. The most popular ones include:

- When several systems are available, choose the samples which maximize the disagreement amongst them ("query by committee" [4]). This strategy cannot be used if a single system is available or inefficient if the used systems are too close to each other.
- Choose the most uncertain samples (*uncertainty sampling* [5]). This strategy tries to increase the sample

density in the neighborhood of the frontier between positives and negatives and therefore improve the system's precision.

- Choose the most probable positive samples. This strategy tries to maximize the size of the set of positive samples (the positive samples are most often sparse within the whole set and finding negative samples is easy and these generally come as numerous enough whatever the selection strategy).

- Choose the farthest samples from already evaluated ones. This strategy tries to maximize the variety of the evaluated samples and therefore to improve the system's recall. It is based on a distance between sample that has to be appropriately defined and it does not require the availability of a system (the considered distance, however, may be related to some system outputs).

More complex strategies can be used including combinations of these. For instance, the system may choose the samples for annotation amongst the most probable ones *and* amongst the farthest from the already evaluated ones. Another possibility is to select samples by groups in which maximize the expected global knowledge gain [6].

Quite often, indexing systems classify samples for several concepts. Many strategies are based on the assumption that when a sample is selected for annotation, all the concepts are evaluated by the annotator. This is what was done in the context of the collaborative annotation effort of TRECVID 2003 [7] where 133 concepts were annotated at once by an operator for each proposed shot. The resulting annotation was of poor quality because of the cognitive load applied to the annotator. For instance, many false negative were observed because the operator was not able to keep in mind all the concepts to be annotated. Therefore, for the TRECVID 2005 [8] collaborative annotation effort and for the LSCOM annotations as well [3], the choice was made to give to the operators a single concept to annotate on a series of proposed shots. From the active learning point of view, the type of annotation has some implications. The second one excludes methods that require that each selected sample is annotated against all concepts. On the other hand, it permits a finer grain application of the various strategies because sample selection can be done independently for each concept.

A particular case is the use of active learning within relevance feedback. In this case, there is only one concept considered and it corresponds to the user's need. The system learns it from the growing set of user's judgments across feedback cycles. This approach has been successfully used in content-based image retrieval [9, 10].

In this work we investigate how efficient active learning strategies are for the indexing of concepts in video shots. We consider the annotation type used in TRECVID 2005 and LSCOM collaborative annotation efforts: samples are presented to the annotator for a single concept at once. We use a single classification system which is one of the variant that we used for our participation to the "high level feature extraction" task of TRECVID 2006 [11]. One originality in the evaluation approach is that we use the full TRECVID 2005 collaborative annotation to simulate the incremental annotations required by the various active learning strategies. Our assumption is that the annotations made by an annotator do not depend upon the order in which the annotations are proposed to him or that, if this is the case, this does not significantly affect our conclusions.

2. SIMULATED ACTIVE LEARNING

Actual experiments for comparing strategies in active learning are difficult and costly to organize because of the involvement of humans in the process. Active learning methods are especially developed for contexts in which it is possible to annotate only a small fraction of a large data set. It sometimes happens however that large datasets are fully annotated even if the corresponding cost if very high. Active learning is not relevant in these cases since nothing remains to be annotated but, on the opposite, such large scale full annotations constitute opportunities to simulate, evaluate and compare strategies in active learning without the need to involve again a user. In *simulated active learning*, methods are executed as if no annotation is available in the beginning. Then, each time a human annotation is needed, the corresponding subset of the full annotation is made available.

The performance of the system during the various stages of the simulated active learning process will normally not exceed the performance it achieves using the whole annotated dataset. What simulated active learning brings is the possibility to compare the performance that the system would have reached with the various strategies if only a given fraction of the whole dataset could have been annotated.

From this, it is possible to make predictions on what would be the better strategy if an even larger dataset becomes available but only a small fraction of it can be annotated. It is also possible to make predictions on what would be the better strategy for annotating a large number of new concepts on a small fraction of the existing dataset. A large number of strategies can be compared under equivalent and repeatable conditions.

In order to evaluate the efficiency of the different strategies, it is necessary to measure the system performance on a test dataset which is different from the training dataset.

The TRECVID 2005 and 2006 test datasets with their annotations and performance metrics can be used for this purpose in conjunction with the above mentioned training data. The annotation of the TRECVID 2005 development set was done in conditions similar to those of active learning.

3. SYSTEM DESCRIPTION

For the evaluation of active learning strategy, the exact type of system used has probably a strong influence. It can be expected that a better the system will globally increase the efficiency of most strategies. Also, the optimal strategy may vary with systems' characteristics. We did not evaluate here the influence of the classifying system implementation. We selected one of the variants of the classification system we used for the TRECVID 2006 concept detection task [11]. This variant has the advantages of having a relatively short training time and of working independently on the different concepts. It also has a global performance (when trained on the whole development set) which is close to the median performance of the participating systems.

Concept detection is performed using networks of SVM classifiers arranged in order to take into account a variety of low level descriptors combining text, local and global visual information as well as conceptual context. The 20 assessed concepts of the TRECVID 2006 campaign are derived from "intermediate" concepts, themselves derived from low level descriptors and not necessarily related to the target final concepts. This approach is linked to the idea that it may be better to bridge the semantic gap in several steps within which the complexity remains low and the correlation between the inputs and the outputs is kept high.

3.1. "Visual" intermediate concepts

In the visual modality, intermediate concepts are computed on image patches. There are 260 (20×13) half-overlapping 32×32 pixels patches. We have built a set of 15 intermediate concepts (ANIMAL, BUILDING, CAR, CARTOON, CROWD, FIRE, FLAG-US, GREENERY, MAPS, ROAD, SEA, SKIN, SKY, SPORTS and STUDIO BACKGROUND), which have been learned from the collaborative corpus annotation of TRECVID 2003 and 2005 that we cleaned up and enriched (there is no overlap between the TRECVID 2005 concepts used as intermediate concepts for training and those used as target concepts). We have trained 15 intermediate concepts with a single classifier that takes as inputs:

- 9 color components (RGB means, variances and covariances)
- 24 texture components (8 orientations $\times$ 3 scales Gabor transforms)
- 7 motion components (the central velocity components plus the mean, variance and co-variance of the velocity components within the patch; a velocity vector is computed for every image pixel using an optical flow tool [12] on the whole image).

The 15×260 outputs of those intermediate concepts are inputs for the higher level classifiers (the 20 classifiers corresponding to the TRECVID 2006 assessed concepts). In practice, not all of the 15 intermediate concepts are used for all of the 20 concepts but only a subset of them. This subset is manually chosen for each of the concepts and typically contains 5 or 6 intermediate concepts. Therefore, we consider those intermediate concepts as "local" descriptor which typically contains about 1500 components. The vector components corresponding to the intermediate concepts are real values between 0 and 1 corresponding to the estimated probability of the patch of containing the concept as they are computed by the libsvm package [13].

3.2. Visual global features

The intermediate concepts are completed by low-level visual features at the global level. The two descriptors are simply concatenated, as an "early fusion" scheme. The global low-level image descriptors include:

- 64 color components ($4 \times 4 \times 4$ color histogram),
- 24 texture components (8 orientations $\times$ 5 scales Gabor transform),
- 5 motion components (the mean, variance and covariance of the velocity components within the image).

Adding global descriptors aim to add context into the concept detection process. Exploiting local and global descriptors can help to overcome the ambiguity often faced in concept detection.

3.3. Textual intermediate concepts

We also computed "Textual" intermediate concepts on each audio segment of the ASR-MT transcription. A list of 2500 terms optimized for each of the 20 concepts is built considering the most frequent terms and excluding stop words, which co-occur with a considered concept. The text descriptor is a Boolean vector whose components are 0 or 1 if the term is absent or present in the audio segment. Again the vectors built at the level of the audio segments are projected on the key frames and in the same way.

3.4. Normalized Early Fusion

The number of extracted features depends upon the modalities and the type of the features. Hence, an early fusion

scheme based on simple vector concatenation is much affected by the vector which has the highest number of inputs. Such fusion should ha an impact on the classification, especially with a RBF kernel which is based on Euclidian distance between each training sample.

In traditional SVM implementation, a normalization process is integrated and aims to transform each input in the same range (e.g. $[0..1]$, $[-1..1]$) in order to unbiased the Euclidian distance. But, for the scope of merging features, this normalization does not take into account the number of input from individual features. Hence, we used a normalized early fusion scheme to avoid the problem of imbalanced features input by reprocessing each feature vectors before concatenation. We normalized each individual vector so that its average norm is about the same. The normalization formula becomes:

$$x_i\prime = \frac{x_i - min_i}{(max_i - min_i) \times \sqrt{Card(x_i)}}$$

where x_i is an input of the feature vector x, min_i and max_i are respectively the minimum and maximum value of the i^{th} input among the training samples and $Card(x_i)$ is the number of dimensions of the source vector of x_i.

4. EXPERIMENTATIONS

Evaluations were done using the TRECVID 2005 collaborative annotation for training and the TRECVID 2006 concept detection task for testing. In this task, systems are required to provide, for 20 concepts, a ranked list of the 2000 shots most likely to contain it (amongst 146328 candidate shots). System performance is measured using the Mean Average Precision of the system computed on the returned list (in practice, a variant called Inferred Average Precision is used). This metric is the one that we use for all our evaluations and for comparing the efficiency of the different strategies.

4.1. Strategies

Three strategies have been evaluated each with four different step sizes. The first strategy is to always select the most probable positive samples. The second strategy is to always select the most uncertain samples. The third strategy in rather an absence of strategy and corresponds to a random choice, it is used as a baseline for the other two. The system always outputs a score for each test sample. This score is used for ordering the results list and it correspond to the probability of presence of the target concept in the sample computed as such by the last SVM stage of the system. In the "most probable" strategy, samples with the probability closest to 1.0 are selected. In the "most uncertain" strategy, samples with the probability closest to 0.5 are selected. The three strategies are implemented using a step size (fraction of the whole set that is proposed for annotation to a human) of 1/5th, 1/10th, 1/20th and 1/40th respectively.

We did not investigate yet the effect of the initial conditions. We started all experiments with an initial set of ten positives and twenty negative samples per concept randomly selected in the set of all positive and negative samples. That is: for the "cold start", we assume that in actual experiments, users have at least ten examples of what they are looking for. We also consider that "negatives come for free" because it is much easier to find negative samples than positive ones. For system training, we use twice as many negative samples as positive samples because this appears to be an optimum value for the kind of classifier we use. In all our experiments, whatever the strategy, the target concept, the step size and the step number, we always got at least this ratio in the annotations we asked for. Since there are always much more negative samples than necessary, the selected ones are randomly chosen.

4.2. Global trends

Figure 1 displays the evolution of system MAP (actually Inferred Average Precision as measured in TRECVID 2006) with the number of annotated samples for the three strategies, for all concepts and with the smallest step size (2.5% of the sample set). We first notice a significant level of "noise": the performance has some fluctuations even considering an average on 20 concepts and large numbers so annotated samples. The performance is not always increasing with the number of annotations. The fluctuations are probably due to the fact that added concepts, though positives, are not always very representative. We also noticed that they depend upon the random choice of negative samples and upon the initial choice of the positive samples for the cold start (not shown). This is the case even at the last iteration when all positive samples are taken into account because the negative samples have been selected in a different way. General trends can be observed anyway by making abstraction of the fluctuations:

- The "most probable" strategy is the best one when a small fraction (less than 15%) of the dataset is annotated. It gets very close to the best "random" performance with the annotation of only about 12.5% of the whole sample set. The performance increases then very slowly.

- The "most uncertain" strategy is the best one when a medium to large fraction (15% or more) of the dataset is annotated. It gets slightly over the best "most probable" and "random"performances with the annotation of only about 15% of the whole sample set. The performance does not increase after.

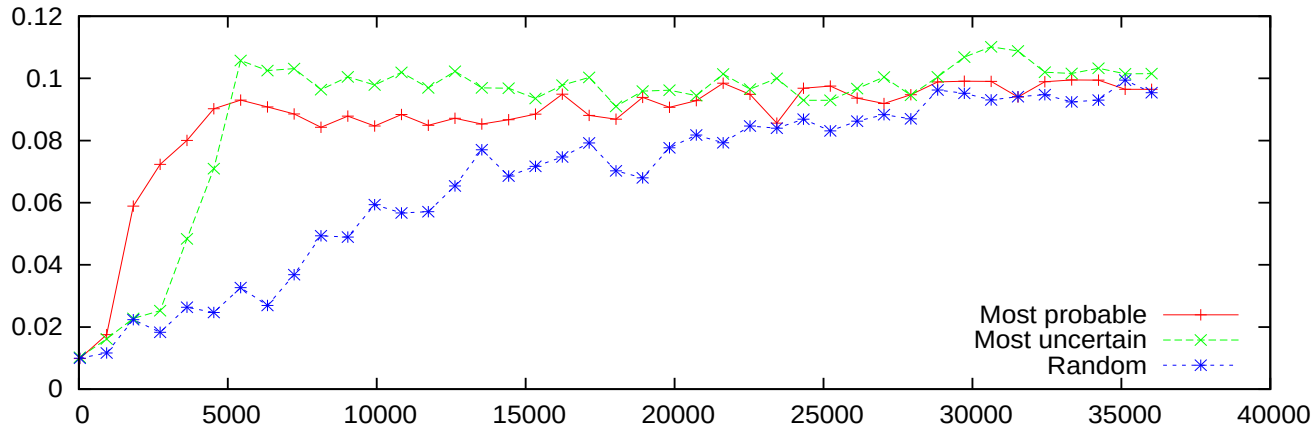

Fig. 1. Evolution of system MAP with the number of annotated samples for the three strategies, all concepts

- The "random" strategy shows a continuous increase in performance with the size of the sample set with a higher rate near the beginning. The maximum performance is reached only when 100% of the sample set is annotated.

4.3. Relation with concept difficulty

Figures 2, 3 and 4 display the same information for easy, moderately difficult and difficult concepts respectively. Easy concepts are: WEATHER (0.4539), SPORT (0.3006) and MAPS (0.2171). Moderate concepts are: MILITARY (0.0985) and CAR (0.0771) WATERSCAPE-WATERFRONT (0.0755), CHARTS (0.0708), MEETING (0.0671), FLAG-US (0.0634), DESERT (0.0557) and EXPLOSION-FIRE (0.0548). Difficult concepts are: COMPUTER-TV-SCREEN (0.0411), TRUCK (0.0355), MOUNTAIN (0.0329), PEOPLE-MARCHING (0.0284), POLICE-SECURITY (0.0257), AIR-PLANE (0.0206), ANIMAL (0.0058), OFFICE (0.0027) and CORPORATE-LEADER (0.0000). The number between parentheses is the MAP for this concept when the system is trained on the whole dataset.

The general trend is conserved for each group of concepts even if the MAP absolute values are on different scales. There are more fluctuations in the results for difficult concepts. These are also less frequent and each new positive sample added has a significant impact which may be negative if the sample, though positive, is not very representative.

The trends are more clear for easy concepts: the "most probable" strategy is the best one if the annotated fraction is 12.5% or less. It reaches there the maximum performance there which is the performance of the "random" strategy with an annotated fraction of 100%. The "most uncertain" strategy is the best one if the annotated fraction is 15% or more. It reaches there the maximum performance there which is consistently greater thant the performance of both "most probable" and "random" strategy with an annotated fraction of 100%.

For moderately difficult and difficult concepts, the difference between the "most probable" and "most uncertain" strategies is hard to distinguish from the fluctuations. The general trend is more difficult to analyse too. Both strategies seem to get close to their maximum values when 15 to 20% of the dataset is annotated.

For moderately difficult concepts, The maximum value reached by these two strategies seem to be consistently lower than the one reached by the "random" strategy with an annotated fraction of 100% (that might not be due to fluctuations). The "random" strategy seems to be better when 80% or over of the dataset is annotated. This can be explained by a different selection of the negative samples (at the end, all positive samples are taken into account).

For difficult concepts, The maximum value reached by these two strategies seem to be consistently higher than the one reached by the "random" strategy with an annotated fraction of 100%. This can again be explained by a different selection of the negative samples.

4.4. Finding of positive and negative samples

Figure 5 displays the evolution of the number of positive samples found with the number of annotated samples for the three strategies, for all concepts and with the smallest step size. As expected, the "random" strategy finds positive samples in a quasi-linear way; the "most uncertain" strategy finds them faster and the "most probable" strategy even faster. The finding rates relatively to random near the beginning are of about 2.4:1 and 4.5:1 for "most uncertain" and "most probable" strategies respectively. These ratios proba-

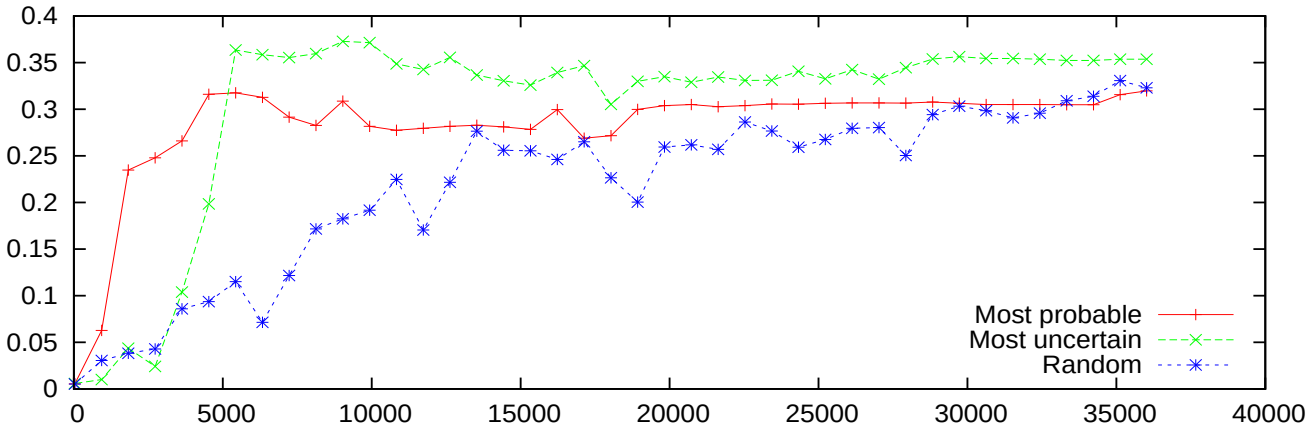

Fig. 2. Evolution of system MAP with the number of annotated samples for the three strategies, easy concepts

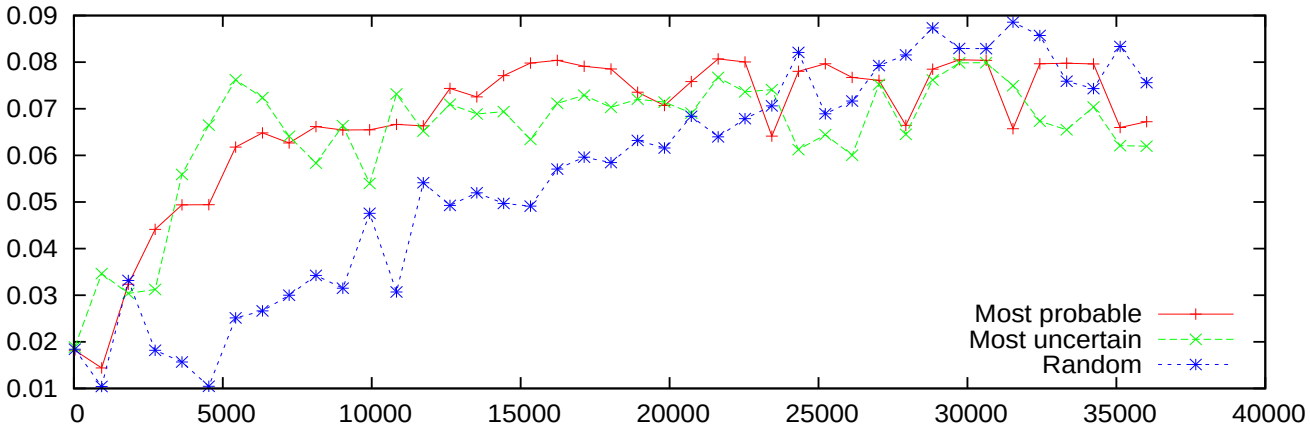

Fig. 3. Evolution of system MAP with the number of annotated samples for the three strategies, moderate concepts

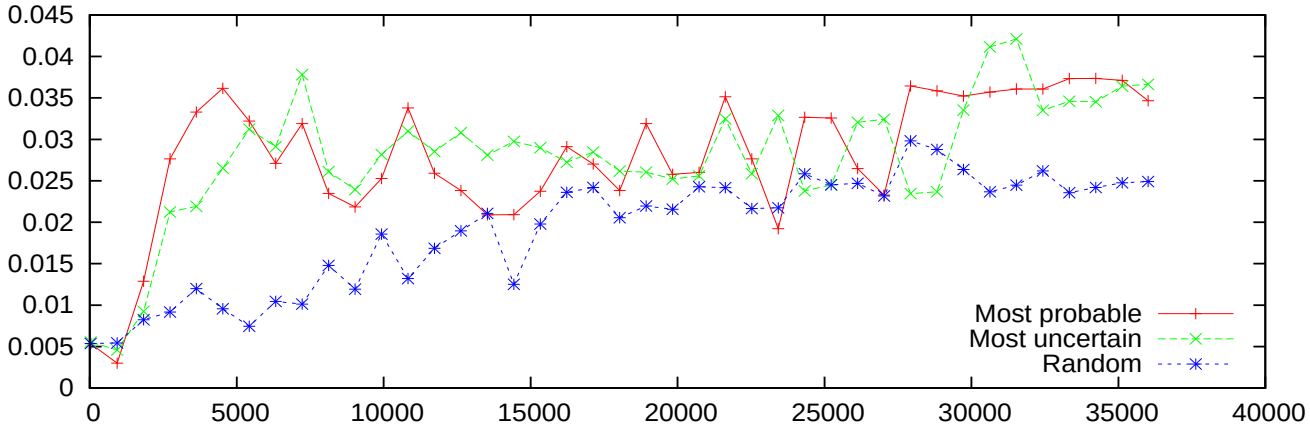

Fig. 4. Evolution of system MAP with the number of annotated samples for the three strategies, difficult concepts

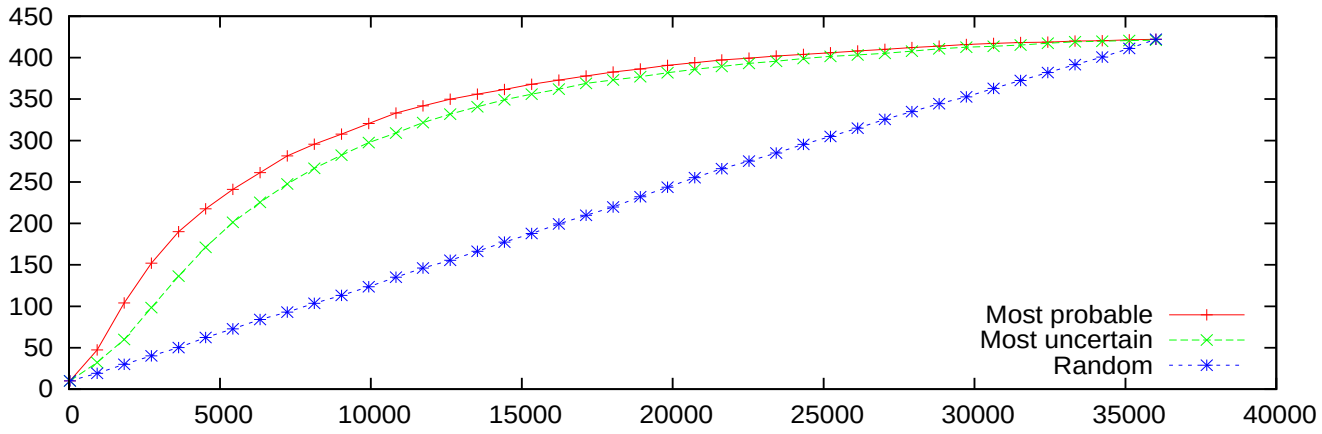

Fig. 5. Evolution of the number of positive samples with the number of annotated samples for the three strategies

bly depend significantly of the system performance but the relative performance of the strategies is clear either from the number of positive samples found or from the Mean Average Precision and we expect it to be representative.

The negative samples selected by the "most probable" strategy are more useful than those selected by the "most uncertain" strategy because they look more like positive samples. Negative samples that are far from the positive ones are usually not considered by the SVM classifiers and do not help. Therefore, the "most probable" strategy does not only select more positive samples, it also selects more helpful negative samples.

4.5. Effect of the step size

We tried the three strategies with four step sizes corresponding to 1/5th, 1/10th 1/20th and 1/40th of the dataset size. We presented the previous results with the smallest step size since it is the most realistic. Figure 6 shows the evolution of system MAP with the number of annotated samples on the first half of the dataset for the "most incertain" strategy for the four step sizes and for all concepts. As expected, smaller step sizes leads to a faster and stronger active learning effect. Similar results are obtained with the "most probable" strategy (not shown).

The effect of the step size seems to be larger within the first few iterations. Also, the training associated to the last iterations becomes longer and longer due to the increasing number of positive and negative samples. This suggests that an increasing step size strategy would optimize the gain relatively to both the number of annotation available and the cumulated training time.

5. CONCLUSION

We have compared active learning strategies for indexing concepts in video shots. Our objective was to select as few sample shots as possible to be manually indexed but still have a good performance classifier (in the sense of Mean Average Precision). Active learning was simulated using subsets of a fully annotated dataset instead of actually calling for user intervention. Training was done using the collaborative annotation of 39 concepts of the TRECVID 2005 campaign. Performance was measured on the 20 concepts selected for the TRECVID 2006 concept detection task. The simulation allowed exploring the effect of several parameters: the strategy, the annotated fraction of the dataset, the number of iterations and the relative difficulty of concepts.

Three strategies were compared. The first two respectively select the most probable and the most uncertain samples. The third one is a random choice. For easy concepts, the "most probable" strategy is the best one when less than 15% of the dataset is annotated and the "most uncertain" strategy is the best one when 15% or more of the dataset is annotated. The "most probable" and "most uncertain" 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.

Current annotated corpora, TRECVID 2003 and 2005, and LSCOM, constitute very good resources but they are still insufficient for efficient system training. Their exhaustive annotation has been very useful for the current study and for the design and evaluation of many concept indexing systems. These annotated corpora will have to be further enriched and active learning techniques should be considered in order to get as much useful information as possible from every paid annotation. These corpora can themselves be used for the cold start and the set of already developed sys-

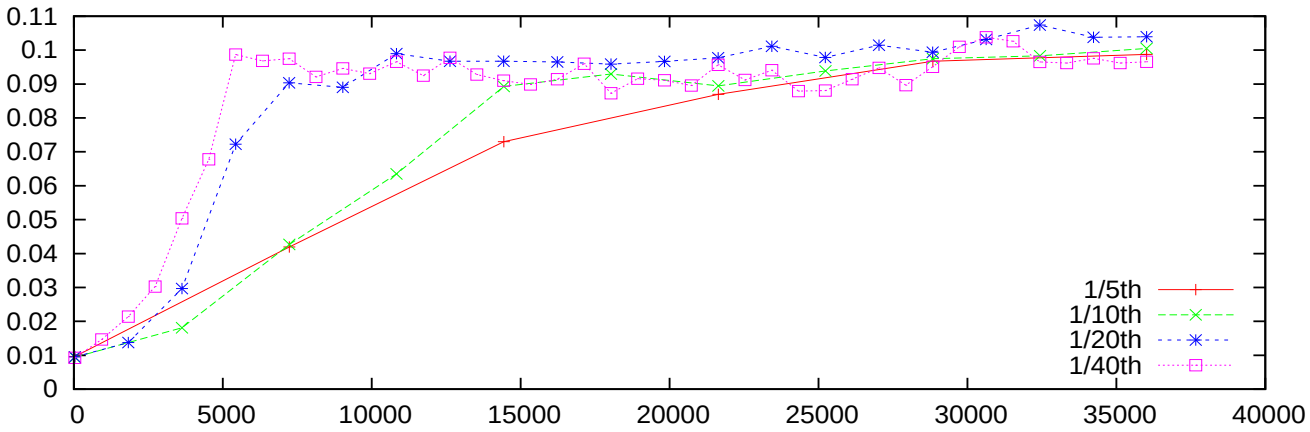

Fig. 6. Evolution of system MAP with the number of annotated samples for the "most uncertain" strategy for the four step sizes, all concepts

tems could be used in a collaborative way for the implementation of the most efficient active learning strategies.

Future work will be conducted to better characterize the optimal strategy. Other strategies can be investigated like "query by committee" if several system are accessible. A selection of negative samples based on a maximum variability could be considered instead of a random selection. Combination of several strategies should also be investigated.

6. REFERENCES

[1] Milind R. Naphade and John 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.

[2] Stéphane Ayache, Georges Quénot, and Shin'ichi Satoh. Context-based conceptual image indexing. In *ICASSP'06: IEEE International Conference on Acoustics, Speech and Signal Processing*, 15-19 May 2006.

[3] Milind Naphade, John R. Smith, Jelena Tesic, Shih-Fu Chang, Winston Hsu, Lyndon Kennedy, Alexander Hauptmann, and Jon Curtis. Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 13(3):86–91, 2006.

[4] H. S. Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Computational Learning Theory*, pages 287–294, 1992.

[5] David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In W. Bruce Croft and Cornelis 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.

[6] Fabrice Souvannavong, Bernard Mérialdo, and Benoit Huet. Partition sampling for active video database annotation. In *WIAMIS'04, 5th International Workshop on Image Analysis for Multimedia Interactive Services, April 21-23, 2004, Instituto Superior Técnico, Lisboa, Portugal.*, 21-23 April 2004.

[7] C.-Y. Lin, B. L. Tseng, and J. R. Smith. Video collaborative annotation forum: Establishing ground-truth labels on large multimedia datasets. In *NIST TREC-2003 Video Retrieval Evaluation Conference*, 17-18 November 2003.

[8] Timo Volkmer, John R. Smith, Apostol (Paul) Natsev, Murray Campbell, and Milind Naphade. A web-based system for collaborative annotation of large image and video collections. In *13th ACM international Conference on Multimedia,*, 6-11 November 2005.

[9] Simon Tong and Edward Chang. Support vector machine active learning for image retrieval. In *MULTIMEDIA '01: Proceedings of the ninth ACM international conference on Multimedia*, pages 107–118, New York, NY, USA, 2001. ACM Press.

[10] Philippe H. Gosselin and Matthieu 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.

[11] Stéphane Ayache, Georges Quénot, and Jérôme Gensel. CLIPS-LSR Experiments at TRECVID 2006. In *NIST TREC-2006 Video Retrieval Evaluation Conference*, 13-14 November 2006.

[12] G.M. Quénot. Computation of optical flow using dynamic programming. In *IAPR Workshop on Machine Vision Applications*, pages 249–252, 1996.

[13] Chih-Chung Chang and Chih-Jen LIn proceedings of. *LIBSVM: A Library for Support Vector Machines*, 2001. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.