PARTITION SAMPLING FOR ACTIVE VIDEO DATABASE ANNOTATION

Fabrice Souvannavong, Bernard Merialdo and Benoît Huet

Département Communications Multimédias
Institut Eurécom
2229, route des crêtes
06904 Sophia-Antipolis - France
(Fabrice.Souvannavong, Bernard.Merialdo, Benoit.Huet)@eurecom.fr

ABSTRACT

Annotating a video-database requires an intensive human effort that is time consuming and error prone. However this task is mandatory to bridge the gap between low-level video features and the semantic content. We propose a partition sampling active learning method to minimize human effort in labeling. Formally, active learning is a process where new unlabeled samples are iteratively selected and presented to teachers. The major problem is then to find the best selection function that maximizes the knowledge gain acquired from new samples. In contrast with existing active learning approaches, we focus on the selection of multiple samples. We propose to select samples such that their contribution to the knowledge gain is complementary and optimal. Hence, at each iteration we ensure to maximize the knowledge gain. Our method offers many advantages; among them the possibility to share the annotation effort among several teachers.

1. INTRODUCTION

Because of the growth of numerical storage facilities, many documents are now archived in huge databases or extensively shared over the Internet. The advantage of such mass storage is undeniable. However the challenging tasks of multimedia content indexing and retrieval remain unsolved without the expensive human intervention to archive and annotate contents. Many researchers are currently investigating methods to automatically analyze, organize, index and retrieve video information [2, 6]. This effort is further stressed by the emerging Mpeg-7 standard that provides a rich and common description tool of multimedia contents. It is also encouraged by Video-TREC which aims at developing video content analysis and retrieval.

Currently, one of the main challenges in the field is to bridge the gap from low-level video features to the semantic content. Classical approaches build statistical models from training data. Unfortunately, given the complexity and diversity of semantic contents, a great amount of labeled data is necessary to build efficient models. In June 2003, Video-TREC has launched a collaborative effort to annotate video sequences in order to build a labeled reference database. It is composed of about 63 hours of news videos that are segmented into shot. These shots were annotated with items in a list of 133 labels which root concepts are the event taking place, the context of the scene and objects involved. Twenty one worldwide institutes participated to this huge collaborative annotation effort. We noticed that the database is composed of many redundant shots like news anchor person, weather maps, ... In that case, it is very interesting to limit the annotation effort by removing the redundant information. We propose an active learning approach to achieve this task.

Active learning aims at training an efficient statistical model with the smallest training set. To achieve this goal, it iteratively selects new samples to be labeled by teachers. Samples are selected to optimize the knowledge gain at each iteration. Existing active learning approaches concentrate on the selection or creation of a single element to be annotated by a teacher at each round. We propose a partition sampling approach to select a set of ambiguous samples that contain complementary information. This selection strategy allows to gain time during the annotation effort but also to share query samples among several teachers.

In the following, we first introduce active learning and related work in the literature. Then we present a common mathematical approach to uncertainty sampling to set up our mathematical framework for partition sampling. We detail an algorithm that allows to efficiently annotate many samples in a round. Finally we conclude with a brief summary including future work.

2. RELATED WORK

Annotating content is time consuming and subject to errors. However it is necessary and compulsory in many applications to build statistical models that require training data. Limiting this effort has raised the interest of the machine learning community. Two approaches were proposed to this problem, semi-supervised and active learning. On one hand, a semi-supervised learner combines a small set of labeled samples with a large set of unlabeled samples [7]. The latter set does not provide any direct information but their distribution can be used to boost the performance of the classifier. On the other hand, an active learner starts from a very small number of labeled samples and then it iteratively asks for new samples to be labeled by a teacher, in order to optimally update the statistical model and increase its performance and accuracy with few samples.

The major task in active learning is to determine the optimal sample selection strategy. New samples can either be created by the system, but they can lack coherence. Typically a digit recognition system could create and ask to be labeled a non existing digit. They can also be selected from a unlabeled set. This approach, called selective sampling, is the most common, and many researchers proposed selection methods, such as query by committee [4] or uncertainty sampling [3]. Applications of active learning techniques are now emerging in the field of multimedia database

annotation [1, 8]. In the following section we propose a new uncertainty sampling strategy, called partition sampling, that allows to select multiple samples as opposed to classical approaches.

3. PROPOSED APPROACH

3.1. Notation and Terminology

We have a database of video sequences, denoted D, whose shots have to be annotated. A shot is represented by a vector x taking values in X. Formally, the learning algorithm takes a set of training examples $L = \{(x_1, y_1), ..., (x_N, y_N)\}$ as input where y_i is the label assigned to x_i. It produces an hypothesis $f_L : X \mapsto \Re$ that minimizes the generalization expected error:

$$E_L = \int_X E_{Y|X}[C(f_L(x), y)]P(x)dx \quad (1)$$

Where P(x) is the marginal distribution of x and $C : X, Y \mapsto \Re^+$ a predefined loss function.

Active learning starts from an initial annotated set and lets the learner iteratively update its training set while learning at each step from the new knowledge gain, i.e. knowledge provided by new samples. There are two main components involved in selective sampling: the classifier $f_L(.)$ trained on the labeled samples L; the selection function $s_f(P)$. The goal of $s_f(P)$ is to select the most appropriate samples S of a unlabeled pool P given the knowledge already acquired by the trainer.

3.2. Active Learning

An active learner has to efficiently select a set of samples S in P to be labeled by teachers. The optimal set, $L^+ = L \cup S$, is the one that will result in the maximal error reduction, denoted R_S.

$$R_S = \int_X (E_{Y|X}[C(f_L(x), y)] - E_{Y|X}[C(f_{L^+}(x), y)])P(X)dx \quad (2)$$

$$S = s_f(P) = arg\max_S R_S \quad (3)$$

There are two difficulties in the task. First it is intractable to compute all possible combinations for S. The common approach is, then, to select one query sample at each round. We call this method greedy-like sampling. Secondly, we can not exactly determine the error because the target distributions $P(X)$ and $P(Y|X)$ are not known. Several assumptions have to be made leading to different selection strategies.

A classical approach consists in approximating the integral in equation (1) with a sum over the pool. P is build from a large number of unlabeled samples. We can thus assume that its size is big enough to approximate the true distribution. Hence, the expected error reduction can be expressed as:

$$\hat{R}_S = \sum_P E_{Y|X}[C(f_L(x), y)] - E_{Y|X}[C(f_{L^+}(x), y)] \quad (4)$$

To learn the hypothesis $f_{L^+}(.)$ of equation (2) for each possible query sample S is now the major problem to compute the estimated error reduction. In [8], the authors first assume that all losses for any $x \in P \setminus L$ have an equal influence. Hence, the sum over P is reduced over S. Then they can neglect $C(f_{L^+}(x), y)$ over $C(f_L(x), y)$ since the new learner is expected to have a very small loss error over S, if not null, compared to the current learner. A worst case model is, then, used to approximate $E_{Y|X}[C(f_L(x), y)]$. Let $\hat{y}$ be the estimated label of x, the best approximated error reduction is finally obtained for:

$$s_f(P) = arg\max_{x \in S} C(f_L(x), \hat{y}) \quad (5)$$

The idea behind this formulation is to select the most ambiguous sample at each iteration.

3.3. Our Approach

We propose a selection strategy to select a set of samples at each round. Learning algorithms make the assumption that close elements are similar. Thus the knowledge of one sample should induce the knowledge of its neighbors. This is implicitly used in the greedy-like active learning and it is emphasized in [1], where they proposed to weight the selection function value of a sample with an estimation of its probability density function to increase learning speed. However most ambiguous points are likely to be neighbors. Thus a strategy that would select the n most ambiguous samples would mostly ask the teacher to annotate similar content; resulting in sub-optimal selection.

It is therefore important to select ambiguous points spread over the distribution of X. We have to ensure that most of selected points are far from each other and also as ambiguous as possible. Let assume that we constructed a partition of P, i.e. $P = \cup U_i$ and $U_i \cap U_j = \emptyset$ for $i \neq j$, such that U_i are connex and that given $\varepsilon \in \Re$ then:

$$\forall (x_1, x_2) \in U_i \times U_i$$
$$\|x_1 - x_2\| < \varepsilon$$

Consider a representative element of each set selected with a selection function $m_i = \dot{s}_f(U_i)$, for example mean element, maximum ambiguity, maximum density. Let $M = \{m_i\}$, then we approximate equation (4) with:

$$\hat{R}_S = \sum_M (E_{Y|X}[C(f_L(x_i), y_i)] - E_{Y|X}[C(f_{L^+}(x_i), y_i)])N_i \quad (6)$$

Where N_i is the cardinal of U_i. This approximation relies on the assumption that neighbors have the same behavior with respect to learners, i.e. similar loss value for a given learner. Let

$$\Delta_{L,L^+}(x_i, y_i) = E_{Y|X}[C(f_L(x_i), y_i)] - E_{Y|X}[C(f_{L^+}(x_i), y_i)]$$

We are looking for S such that:

$$s_f(P) = arg\max_S [\sum_S \Delta_{L,L^+}(x_i, y_i))N_i + \sum_{M \setminus S} \Delta_{L,L^+}(x_i, y_i)N_i] \quad (7)$$

We now further assume that for $x \in M \setminus S$, $\Delta_{L,L^+}(x_i, y_i)$ is small. Indeed, given the partition we do not expect L^+ to improve classification of elements of $M \setminus S$. Hence,

$$s_f(P) = arg\max_S \sum_S \Delta_{L,L^+}(x_i, y_i)N_i \quad (8)$$

Moreover the new learner is expected to have a very small loss error on S,

$$\forall S, \sum_S \Delta_{L,L^+}(x_i, y_i))N_i \approx \sum_S E_{Y|X}[C(f_L(x_i)]N_i \quad (9)$$

Finally,

$$s_f(P) = arg\max_{S \subset M} \sum_S E_{Y|X}[C(f_L(x_i)]N_i \quad (10)$$

The idea behind this formulation is to select the most ambiguous samples spread over the distribution of x.

In practice, we propose to create a partition of the pool thanks to clustering techniques. In our experiments we use the well-known k-means algorithm. Then we choose equation (5) to select representative elements of each set of the partition. In fact we select the most ambiguous element per cluster. Finally S is composed of the n most relevant representatives.

4. EXPERIMENTS

We evaluate our method on synthetic data and on the Video-TREC 2003 annotated database. First, we present the learner we used, then the experimental framework and finally results on both data sets.

Since we do not have any knowledge about the distribution of features in the semantic space, we opt for the k-nearest neighbors classifier in the experiments presented here. Let N_s be the neighborhood of a shot s in L, i.e. k-nearest neighbors in the training set, and $y_i \in \{0,1\}$ the semantic value of the neighbor i. The hypothesis is defined as:

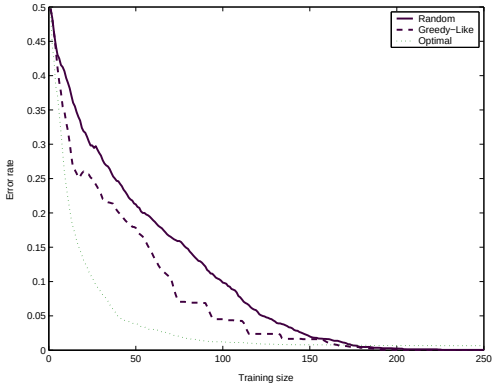
$$f_L(s) = \frac{\sum_{N_s} sim(s,n_i) * y_{n_i}}{\sum_{N_s} sim(s,n_i)}$$
$$\text{where } sim(s,n_i) = \cos(s,n_i)$$

and the estimated label of s as:

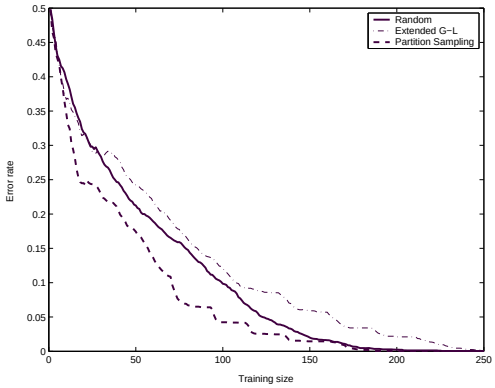
$$\hat{y}_s = arg\min_y \|y - f_L(s)\|$$

The evaluation consists in comparing five approaches. Two reference experiments that do not rely on the selection strategy are presented. The first one gives the error rate when samples are randomly selected in the pool. This should be the worst case. The second one is an approximation of the optimal selection sequence that is obtained thanks to a greedy maximization of the error reduction R_S, see equation (2), knowing labels of the database. Then we draw the error rate curve for the greedy-like sampling strategy, see equation (5). And for its direct extension to the selection of n best samples, i.e. extended greedy-like. Finally we compare the error rate evolutions with respect to the number of samples with our partition sampling strategy.

Figure (1) compares the different approaches on synthetic data. The dataset is composed of 2,000 points associated with a label in {0,1}. Then around each original points are added 20 additional points with the same label. Figure (1(a)) shows performance when selecting one sample at each iteration. We notice that the random selection already performs very well. With 25,000 samples, i.e. half of the database, the error rate is null. The optimal sampling achieve an error rate of 0.01 with a training size of 10,000 that represents 5 elements per cluster. At last, the greedy-like sampling performs well compared to random sampling. Figure (1(b)) shows performance when selecting 100 samples at each iteration. As expected, extended greedy-like sampling has it performance drastically decreased while partition sampling outperforms the random approach. To summarize, we notice that partition sampling performs as well as the greedy-like approach (dashed curves in figures (1(a)) and (1(b))). Indeed the greedy-like approach is the optimal



(a) unique selection at each round



(b) pool selection at each round

Figure 1: Comparison of Active Learning strategies on synthetic data. Random (independent of the number of selected samples), Greedy-Like (unique selection), Optimal (Greedy selection of the best samples knowing all labels), Extended Greedy-Like (selection of n best samples) and Partition Sampling (selection of n best elements spread over a partition).

solution of our framework when selecting one sample. However partition sampling provides more advantages with similar performance. First of all, teachers are involved in 100 times less loops. The annotation can also be shared among many teachers. Finally, we can reserve more computational power between rounds to find optimal elements since we do expect teachers to have a rest between rounds.

Figure (2) presents results on a real database. It is composed of 40,000 annotated shots from news sequences [5]. Shots are described by HS color histograms and Gabor's energies of their key frame. We use Latent Semantic Analysis, as described in [9], to capture local information, remove noise and emphasize information occurrence in frames. We draw the same conclusions. However the benefit over the extended greedy-like sampling is not as important as previously. Finally, the gain of active learning is undeniable since annotating half of the data set gives an error rate of only 0.005.

5. FUTURE WORK

We proposed a new selection strategy that allows to build a set of optimal query samples to be annotated. The set may then be

shared among teachers in a collaborative work to efficiently annotate complex contents, which require many examples. In the context of a single teacher, it simply reduces the time spend by the annotator. An initial mathematical framework was set up to justify our approach to the problem. Then we presented experimental results on synthetic data and on the real problem of video database annotation. The partition sampling approach outperforms random sampling and reaches its optimal learning sequence in the actual framework.

Future work will involve the improvement of selection strategies to achieve better performances, closer to the optimal selection strategy. Then we will thoroughly study the partitioning problem that we introduced. For example a clustering driven by the density would be more appropriate than k-means algorithm. We will investigate ways to select the partition size and look at whether there is a need to partition the pool at each iteration or all the database at the beginning. Finally we should extend our framework to take into account the annotation over multiple labels.

6. REFERENCES

[1] Tsuhan Chen Cha Zhang. An active learning framework for content-based information retrieval. In *IEEE Transactions on Multimedia*, volume 4, pages 260–268, 2002.

[2] Shih-Fu Chang, W. Chen, H.J. Meng, H. Sundaram, and Di Zhong. A fully automated content-based video search engine supporting spatiotemporal queries. In *IEEE Transactions on Circuits and Systems for Video Technology*, volume 8, pages 602– 615, 1998.

[3] Micahel I. Jordan David A. Cohn, Zoubin Ghahramani. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

[4] Yoav Freund, H. Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28:133–168, 1997.

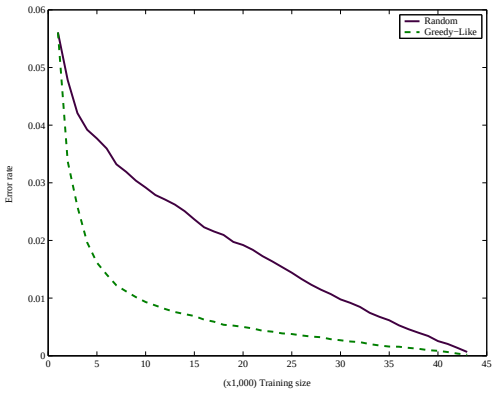
[5] Ching-Yung Lin, Belle L. Tseng, and John R. Smith. Video collaborative annotation forum: Establishing ground-truth labels on large multimedia datasets. In *Proceedings of the TRECVID 2003 Workshop*, 2003.

[6] M.R. Naphade, T. Kristjansson, B. Frey, and T.S. Huang. Probabilistic multimedia objects (multijects): a novel approach to video indexing and retrieval. In *IEEE International Conference on Image Processing*, volume 3, pages 536–540, 1998.

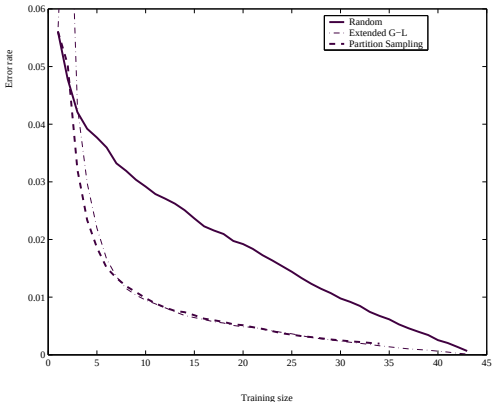
[7] Kamal Nigam, Andrew K. McCallum, Sebastian Thrun, and Tom M. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134, 2000.

[8] Alexander Hauptmann Rong Yan, Jie Yang. Automatically labeling video data using multi-class active learning. In *IEEE International Conference on Computer Vision*.

[9] Fabrice Souvannavong, Bernard Merialdo, and Benoˆ Huet. Video content modeling with latent semantic analysis. In *Third International Workshop on Content-Based Multimedia Indexing*, 2003.



(a) unique selection at each round



(b) pool selection at each round

Figure 2: Comparison of Active Learning strategies on real data. Random (independent of the number of selected samples), Greedy-Like (unique selection), Extended Greedy-Like (selection of n best samples) and Partition Sampling (selection of n best elements spread over a partition).