

Active Cleaning for Video Corpus Annotation

Bahjat Safadi, Stéphane Ayache, Georges Quénot

▶ **To cite this version:**

Bahjat Safadi, Stéphane Ayache, Georges Quénot. Active Cleaning for Video Corpus Annotation. MMM 2012 - International MultiMedia Modeling Conference, Jan 2012, Klagenfurt, Austria. pp.518-528, 10.1007/978-3-642-27355-1_48 . hal-00767345

HAL Id: hal-00767345
https://hal.science/hal-00767345

Submitted on 19 Dec 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Active Cleaning for Video Corpus Annotation

Bahjat Safadi[1]*, Stéphane Ayache*[2] *and Georges Quénot*[1]

[1]UJF-Grenoble 1 / UPMF-Grenoble 2 / Grenoble INP / CNRS, LIG UMR 5217,
Grenoble, F-38041, France
{Bahjat.Safadi, Georges.Quenot}@imag.fr
[2]LIF UMR 6166, CNRS / Université de la Méditerranée / Université de Provence,
F-13288 Marseille Cedex 9, France
Stephane.Ayache@univmed.fr

Abstract. In this paper, we have described the active cleaning approach that was used to complement the active learning approach in the TRECVID collaborative annotation. It consists in using a classification system in order to select the most informative samples for multiple annotations, in order to improve the quality and the reliability of the annotations. We have evaluated the actual impact of the active cleaning approach on TRECVID 2007 collection. The evaluations were conducted using complete annotations that were collected from different resources, including the TRECVID collaborative annotations and the MCG-ICT-CAS annotations.

From our experiments, a significant improvement of the annotation quality was observed when applying the cleaning by cross-validation strategy, which selects the samples to be re-annotated. Experiments show that higher performance can be reached with a double annotations of 10% of negative samples or 5% of all the annotated samples selected by the proposed cleaning strategy using cross-validation. It has been shown that, with an appropriate strategy, using a small fraction of the annotations for cleaning improves much more the system's performance than using the same fraction for adding more annotations.

Keywords: Corpus annotation, active learning, annotation cleaning.

1 Introduction

Concept indexing in image and video documents is very important for content-based retrieval. It is a fundamental image/video retrieval problem: given a data set of images and a query (visual concept), which images do present the given visual concept? Generally, classical keyword based search is not possible due to the frequent absence of appropriate text annotation. Signal-level descriptions (e.g. color and texture) are also known to be inappropriate for the task since they do not represent the semantic content well, and are not easy to handle for users. Automatic concept indexing has been one of the main focus of the TRECVID campaigns (evaluation of video retrieval systems, [12]) since 2002.

Most concept indexing systems use a supervised learning approaches [7, 13], in which concepts are learned from sets of positive and negative samples. The models and training algorithms are important for systems' performance, but the training data also play

an important role. While it is quite easy and cheap to get large amounts of raw data, it is usually very costly to have them annotated, due to the involvement of human intervention for judging the "ground truth".

While the volume of data that can be manually annotated is limited due to the cost of manual intervention, it still possible to select the data samples that will be annotated, so 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 the system is trained several times with a growing set of labeled samples. The objective is to select as few samples as possible to be manually annotated so that these annotations lead to better classification performance.

The quantity of the annotated samples is important for system's performance, Their quality is also very important since most advanced classification methods are sensitive to mislabeled training examples. Using crowd-sourcing [3, 14] methods leads to quickly changing the landscape for the quantity and the quality of labeled data available to supervised learning. While such data can now be obtained more quickly and cheaply than ever before, the generated labels also tend to be far noisier due to limitations of quality control mechanisms. The quality of the labels obtained from annotators varies. Some annotators provide random or bad quality labels in the hope that they will go unnoticed and still be paid, and yet others may have good intentions but completely misunderstand the task at hand or they become distracted or tired over time. The standard solution to the problem of "noisy" labels is to assign the same labeling task to annotators, in the hope that at least a few of them will provide high quality labels or that a consensus emerges from a great number of labels. In either case, a large number of labels is necessary, and although a single label is cheap, the costs can accumulate quickly. It can be observed, that if one is aiming to produce a quality labels within minimum time and money, it makes more sense to dynamically decide on the number of labelers needed. For instance, if an expert annotator provides a label, we can probably rely on it being of high quality, and we may not need more labels for that particular task. On the other hand, if an unreliable annotator provides a label, we should probably ask for more labels until we find an expert or until we have enough labels on which we can apply the majority vote to decide the final label.

Given the substantial human effort required to gather good training sets -as well as the expectation that more data is almost always advantageous-, researchers have begun to explore new ways to collect labeled data. Both active learning and crowd-sourced labeling are promising ways to efficiently build up training sets for concept indexing and retrieval. The active learning techniques aim to minimize human effort by focusing label requests on those that appear to be the most informative samples to the classifier [8, 4, 15, 10, 2], whereas crowd-sourcing work explores how to package annotation tasks in such a way that they can be dispersed effectively [15, 5, 11]. The interesting questions raised in these areas - such as dealing with noisy labels, measuring reliability, mixing strong and weak annotations - make it clear that data collection is no longer an ordinary necessity, but a thriving research area in itself.

Recent years have seen significant growth in label aggregation researches. For example, Vijayanarasimhan et al. presented an approach for live learning of object detectors

[15], in which the system autonomously refines its models by actively requesting crowd-sourced annotations on images crawled from the worldwide web. Kumar et al. showed that generating additional labels for labeled examples reduces the potential label noise [5], and faster learning can be achieved by incorporating knowledge of worker accuracies into consensus labeling. Sheng et al. in [11] presented repeated-labeling strategies of increasing complexity, and their results show clearly that when labeling is not perfect, selective acquisition of multiple labels is a strategy that data miners should have in their repertoire; and for certain label-quality/cost regimes, the benefit is substantial.

Using multiple annotations to reduce labeling noise have also been used in the context of crowd-sourcing; although a full double or triple annotation is even more costly than a simple full one; and it is not in the spirit of data annotation based active learning approaches, in which we do not need to annotate all the samples in the data set. In this paper, we propose to use an active learning approach for selecting samples for second or third annotations. We call this approach *Active Cleaning*. Using the simulated active learning approach and all the available annotations on TRECVID 2007 development set, we have designed different experiments in order to evaluate the benefits of the active cleaning approach, as well as the relative efficiency of the associated strategies.

The outline of the paper continues as follows: the annotation type is presented in section 2; the active cleaning approach is discussed in section 3; section 4 describes the experimental results, while Section 5 presents concluding remarks.

2 Annotation type

We consider the binary annotations, which are often used for image/video classification, such as "Does the video-shot contain an instance of the given visual concept C or not?". Let t_x the target value for the sample x and y_{xk} the k^{th} label for the sample x given by an annotator. The set of target values T and the set of labels Y are binary scalars, hence $y_{xk}, t_x \in \{-1, 1\}$. T values are decided by applying the majority vote on Y values. In the collaborative annotation we have a third case that we call *skipped*: the user already saw the shot but he/she was confused of its label. Three possible annotations were considered: *Positive, Skipped* and *Negative* we name them *pos, skip* and *neg* respectively.

3 Active cleaning

Active cleaning is the method of using an existing classification system for selecting samples for re-annotation, in order to improve the quality of an annotated corpus. It may be implemented in an incremental way, in conjunction with an active learning based annotation algorithm. In this case, the annotations may be cleaner and more correct, which makes the active learning more effective and efficient. Active cleaning may also be used for cleaning an already existing annotation, which can be either complete or partial. In this case, the benefits are only at the level of the resulting annotation.

Cleaning during active learning is the approach that was used in TRECVID collaborative annotation system. The active cleaning algorithm based concept annotation is

detailed in Algorithm 1, which applies the classical active learning algorithm in which we added the cleaning process. Let *D* be the data set which needs to be labeled as containing a target concept (e.g. Airplane, Person..); L, U the labeled and unlabeled subsets respectively, thus $L \cup U = D$ and $L \cap U = \phi$. N a set of the possible choices of the user to label sample x as containing a given concept or not. Three possible choices are allowed by the annotation system: *Positive, Skipped* and *Negative*, (see section 2). We denote Q_{al} and Q_{cl} to be the selection strategies of the active learning and cleaning respectively (see section 3.1). Before explaining the algorithm let us introduce some definitions in order to facilitate the understanding of our algorithm:

1. The set of available annotations: $Y = \{y_{xk} \in N : x \in L; k \in \{1, 2, \ldots, t\}\}$, where y_{xk} defines the k^{th} label of sample x given from an annotator. Hence we ask, orderly, for up to three annotations for each sample, we set $t = 3$.
2. The subset of conflicting samples: $ConfANN = \{x \in L : y_{x1}, y_{x2} \in Y \wedge y_{x1} \neq y_{x2}\}$, a subset of L that have two different annotations for each sample.
3. The subset of second-annotations: $SANN_{Q_{cl}} = \{x \in L : y_{x1} \in Y \wedge y_{x2} \notin Y\}$, a subset of L that have only one annotation for each sample, selected according to the cleaning strategy Q_{cl}.
4. The subset of primary-annotations: $PANN_{Q_{al}} = \{x \in U\}$ samples have no available annotations, selected according to the active learning strategy Q_{al}.

Algorithm 1 Active Cleaning Algorithm Based Concept Annotations

D: all data samples.
L_i, U_i: labeled and unlabeled subsets of S, ($L_i \cup U_i = D$).
A=(train, predict): the elementary learning algorithm.
Q_{al}, Q_{cl}: the selection strategies, respectively, for the active learning and cleaning.
Y_i: available annotations for L_i.
Initialize L_0 and Y_0.
while $D \setminus L_i \neq \emptyset$ **do**
  $m_i \leftarrow$ Train(A, L_i, Y_i)
  $P_u \leftarrow$ Predict(U_i,m_i)
  $P_l \leftarrow$ Predict(L_i,m_i)
  (*) Select the subset $ConfANN \subset L_i$
  (**) Apply Q_{cl} on P_l in order to select the subset $SANN \subset L_i$.
  (***) Apply Q_{al} on P_u in order to select subset $PANN \subset U_i$.
  $\tilde{Y}$ = (Label ($ConfANN$)) $\cup$ (Label ($SANN$)) $\cup$ (Label ($PANN$))
  $Y_{i+1} \leftarrow Y_i \cup \tilde{Y}$
  $L_{i+1} \leftarrow L_i \cup PANN$
  $U_{i+1} \leftarrow U_i \setminus PANN$
end while

The algorithm is iterative, for implementation purposes, the elementary learning algorithm A is split into two parts: train and predict. The algorithm starts by initializing the L_0 set, which can be done by collecting initial labels Y_0 for some samples of D, through the annotators. Iteratively, the development set D is split into two parts: labeled samples L_i, and unlabeled samples U_i. Then classifier A is trained using L_i with its associated labels Y_i and obtains the model m_i, which is then used to predict the scores

- likeliness to contain the target concept - P_l and P_u of the samples in L_i and U_i sets respectively. These predicted scores are used to select the samples to be labeled in the next iteration. However, the selection is done in three steps: first the algorithm chooses the samples with conflicting labels *ConfANN* (*); then it apply the cleaning strategy Q_{cl} on the predicted scores P_l of the samples in L_i, and selects the samples of the *SANN* set to be re-annotated by different users (**). Finally, the predicted scores P_u of unlabeled samples in U_i are passed to the Q_{al} strategy, which selects the *PANN* set (***). The annotators are asked to annotate all the samples in these three sets, taking into account that a data sample x can be examined maximum once by the same annotator, and annotators cannot access the judgments of other annotators. When the new annotations set $\tilde{Y}$ is completed, it will be added to the global annotations set Y. The set *PANN* is added to the L_i set to produce the set L_{i+1}, and it is also removed from the U_i set to produce the U_{i+1} set. Thus a new iteration is started.

3.1 Active learning and Cleaning strategies, Q_{al} and Q_{cl}

In this paper, the selection strategy of the active learning, Q_{al} has been chosen to implement the relevance sampling, which selects the most probable positive samples regarding to their classification scores (samples with high prediction scores). It was observed that this is a good strategy for sparse concepts [2, 10] where the objective is to find as many positive samples as possible from the unlabeled set U to be annotated.

For the active cleaning, several strategies Q_{cl} can be used for the selection of samples to be re-annotated. They may depend upon the type of annotation (number of possible judgments for instance) and the problem of highly imbalanced dataset, which is a very frequent case in video indexing. Furthermore, these strategies can depend on whether the first annotations were done incrementally or at once. We propose here a cleaning strategy, denoted *Cross-Val*. It is based on re-annotating the wrongly labeled samples due to an error of the annotator (for instance if the annotator missed the change of the concept to annotate). Detecting the wrongly labeled samples is done by training classifiers on these labeled samples and using the trained models to predict the correctness of these labeled samples. Thus, through the predicted score of each sample we can expect if the sample has a correct label or not. The wrongly labeled samples are then those having positive labels with low scores, or negative labels with high scores. Basically, this strategy selects fractions of the labeled samples. These fractions denoted as $P\%, N\%$ and $S\%$ and refer to annotated samples as positive, negative and skipped respectively, (see section 2). Furthermore, the selected samples are then proposed to annotators for a second annotation round.

In *Cross-Val* strategy, the *N%, P% and S%* correspond to the percentage of the labeled samples as *Negative, Positive* and *Skipped*. This includes the baseline (no second annotations), when *N=P=S=0*, re-annotating all skipped and positive samples (*Skip-Pos*) by *P=S=100 and N=0*, and the extreme fully cleaning *N=P=S=100*. In this paper, we evaluated the Cross-Val strategy with different fractions and several ways of re-annotations as in table 1. Our goal is to study the system performance with the Cross-Val strategy for cleaning annotations, furthermore to find the best fraction values for this process.

3.2 A posteriori cleaning

In the case of a posteriori cleaning, we assume that first annotations have been done, thus we have one annotation for each sample, and they will be cleaned globally with a single iteration. A system is trained using the available annotations and the samples are ranked according to their probability of being positive by the system. The given fractions P%, S% and N% of samples annotated as positive, skipped and negative will be used respectively to select the samples for second annotation round. For the positive samples, the system chooses the P% of positive samples with false prediction (have lowest predicted scores). For the negative samples, it chooses the first N% of negatives samples with the highest predicted scores annotated. For the skipped samples we chose the S% of the skipped samples that have uncertainty scores (predicted score is close to the classifier boundaries). In all cases, a third annotation is required from the annotators when conflicting is detected, between the first and second annotations.

4 Experiments

We have evaluated the active cleaning approach based on the *Cross-Val (Q_{cl})* strategy in a variety of contexts. It has been applied with a classification system using four types of image descriptors, which are taken from IRIM GDR-ISIS partners [9], including *the combination of Histogram and Gabor, Global-Tlep, Global-Qwm and Bow-Sift*. The multiple-SVM classifiers with RBF kernel was applied as the classification algorithm, which was implemented as in [10]. The evaluations were conducted using the TRECVID 2007 collection metrics and protocol. The TRECVID 2007 collection contains two main sets: the development set consists 21532 sub-shots with 36 concepts (or "high level features") selected from the LSCOM-lite [6] set for annotation, and the test set which consists of 22084 sub-shots. In TRECVID 2007, the evaluation was done on the test set using only 20 concepts which were chosen by the National Institute of Standards and Technology (NIST). In order to carry out the experiments on the simulated active cleaning, three annotations are needed for each concept (c)$\times$ sub-shot (x) in this dataset. We have collected and completed all the annotations, which were produced by the collaborative annotation on the considered database, that we get at least two labels for each $c \times x$. In addition, we used a complete set of annotations: one label for each video shot, produced independently by a group from the Multimedia Content Group, Institute of Computing Technology, Chinese Academy of Sciences (MCG-ICT-CAS).

Since our goal, in this work, is to study the system performance with the *Cross-Val (Q_{cl})* strategy for cleaning annotations, we present the different fractions that were used in our experiments in table 1. In which $E1$ is the baseline, $E8$ refers to the cleaning of all skipped and positive samples, and $(E2, E3, \ldots, E7)$ indicates the cross-validation strategy with different *(N%, P%, S%)* fractions.

Q_{cl}	E1	E2	E3	E4	E5	E6	E7	E8
pos %	0	10	0	0	5	10	20	100
neg %	0	0	0	10	5	10	20	0
skip %	0	0	10	0	5	10	20	100

Table 1. The *(P%,N%,S%)* fraction values that were used in our experiments with our active cleaning strategy.

4.1 The active learning steps and the cold-start

In calculating the number of the required annotations at each active learning iteration (including the third, second and first annotations), a variable step size function can be used. In practice we used 30 steps in total, considering the geometric scale function with the following formula: $s_k = s_0 \times (n/s_0)^{k/K}$, where n is the total size of the development set, s_0 is the size of the training set at the cold-start, K is the total number of steps and k is the current step. At each step (or iteration) the algorithm calculates the s_k to be the size of the new training set and it chooses the number of samples that needs to be cleaned cl_k, and the new samples to be labeled with size equal to $new_k = s_k - s_{k-1} - cl_k$.

In this evaluation, the harmonic mean has been applied as a fusion function for the multiple-SVM results (scores). The cold start problem was solved by using another TRECVID collection, the 2005 one. We trained SVM classifiers on the TRECVID 2005 collection and predicted the usefulness on the development set of TRECVID 2007; we have started with annotating the first 100 samples at the top of the ranked list (samples having high scores), then the Active learning and cleaning system was run to label all the shots within the development set.

4.2 Available annotations

In the following we present the two resources of the considered annotations:

1- Collaborative annotations (*CA*): annotations were done in collaboration with 32 groups of participants at TRECVID, each group contributed with several annotators. The annotation system used is based on the active learning approach. For each concept (c) $\times$ sub-shot (x) in the data set, the annotators have left the choice to label x as containing an instance of concept t or not, *pos* and *neg* respectively, they also can skip annotating it in the case of confusing on its label. This can be considered as crowd-sourcing, since each shot could be proposed to several annotators to judge whether it contains c or not. Since we were limited in time of the annotating phase of TRECVID, this data set was not fully annotated. Furthermore, there are multiple annotations for the annotated samples L for each concept c, and they are still available and can be used as multiple judgments for the experiments on simulated active cleaning approach. For our experiments, these judgments have been completed to have at least two judgments for each sample.

2- MCG-ICT-CAS annotations (*MCG*): The MCG-ICT-CAS team has produced, on its own, complete and independent annotations of all the concepts (c)$\times$ sub-shots (x). The annotations were made by a pool of students. Each student could annotate shots to contain only a specific concept, and the annotations were done on all the data set (active learning was not considered). Each $c \times x$ has only one label, since only one annotator (student) could examine and label it, which means that it does not contain multiple annotations. This annotations set has the advantage of being complete, and since it was made using a smaller number of annotators, one can say it is more consistent.

These annotations were taken by different annotators and two different systems, and they have some noise in annotations. These noises came from the annotation systems

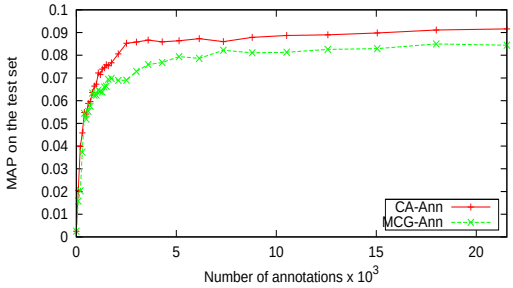


Fig. 1. The MAP calculated on 20 concepts of the TRECVID 2007 test set, with two different annotation sources.

used and the annotators themselves. For instance, given concept *Sports*: we got 482 positive samples from the CA annotations, while from MCS annotations we got only 226 positives; furthermore, the two sources were agreed on only 168 positive samples.

The performance of our baseline system, by using only single annotations from the two annotation resources (CA and MCG), is shown in figure 1. This figure shows the effectiveness in performance, of the classification system, with the number of the annotated sub-shots from the development set. Thus, it presents the MAP, of the 20 concepts, calculated on the test set. For both curves, we consider a better curve to be: the fastest in growing, and the highest MAP value, it reaches, especially in the beginning. As we can see, the system performance using the annotations produced by the CA is much higher than using the MCG annotations. This can be due to the annotation strategy, which is different in the two cases as described above, and it may also be related to the annotators themselves.

From this result, we assume that for each concept (c) $\times$ sub-shot (x), the annotations taken from CA are cleaner than the MCG, and we planned two main experiments to study the effectiveness of the active cleaning strategies:

1. (MCG-CA): the first annotation, for each $c \times x$, is taken from low-quality annotators, (MCG), and the second annotation was taken from better-quality annotators (CA).
2. (CA-MCG): the first annotation, for each $c \times x$, is taken from good-quality annotators, (CA), and the second annotation was taken from lower-quality annotators (MCG).

In both experiments, we have used the second annotation produced by CA as the third annotation, and it was used when the two annotations (CA and MCG) are conflicting.

4.3 Active cleaning effectiveness

We have studied the performance of the annotation system using the cleaning strategy, *Cros-Val* with different P%, N% and S% fractions as set in table 1. Thus, we report the obtained results from our two main experiments MCG-CA and CA-MCG. For simplicity, we report the results of the last iteration of the active cleaning, in table 2. Furthermore, in figure 2 we present the full iterative results of the cleaning performance, for some experiments.

Table 2 presents the evaluation results of the two main combinations MCG-CA and CA-MCG, using the cleaning strategy, *Cros-Val* with different P%, N% and S% fractions as set in table 1. Moreover, it presents the number of cleaning annotations required for each experiment in the two considered combinations. As we can see from this table,

	MCG-CA	#Annotations	CA-MCG	#Annotations
E1=N0P0S0	0.084	21532	0.091	21532
E2=N0P10S0	0.084 +0%	+65	0.091 +0%	+46
E3=N0P0S10	0.086 +2%	+50	0.092 +1%	+11
E4=N10P0S0	0.095 +14%	+2100	0.096 +5%	+2150
E5=N5P5S5	0.096 +14%	+1100	0.095 +4%	+1100
E6=N10P10S10	0.097 +15%	+2200	0.090 -1%	+2215
E7=N20P20S20	0.097 +15%	+4400	0.095 +4%	+4420
E8=N0P100S100	0.086 +2%	+1150	0.093 +2%	+580

Table 2. The result of the cleaning strategies with the eight experiments described in table 1.

some experiments do not have a real effect on the system performance, especially when the cleaning system does not include the negative samples, as in E2, E3 and E8. This is due to the fact, that the number of re-annotated samples is very small, since there are few positive and skipped samples in the data set. However, the performance is higher when the negative samples were included in the cleaning system; moreover it goes up to 15% in the case of MCG-CA and 5% in CA-MCG. This is expected since, as shown in figure 1, we consider that annotations from MCG have lower-quality than CA.

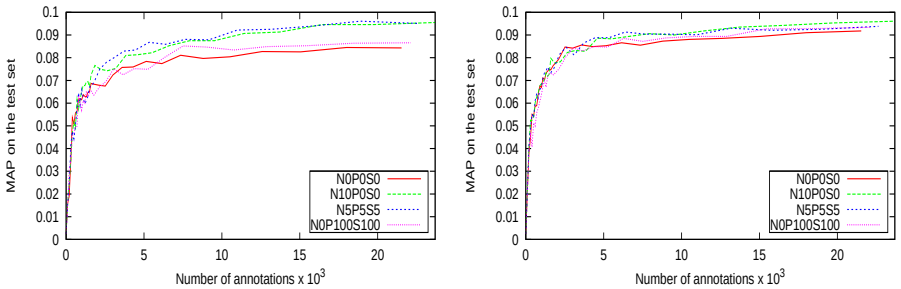


Fig. 2. Active cleaning strategies: Cleaning *MCG* annotations by *CA* in left, and in right Cleaning *CA* by *MCG* annotations.

Figure 2 shows the effectiveness of the active cleaning strategies E4 and E5 compared to the baseline (E1) and the *Skip-Pos* (E8) strategy, with the two considered experiments, the MCG-CA (left) and CA-MCG (right). As we can see in this figure, in both experiments, the system performance (using the MAP) was increased when the cleaning system considered the re-annotations of negative samples, as in E4 and E5. Hence, the Cross-Val strategy E4 works in re-annotating only 10% of the negative samples, and E6 re-annotating 5% of each type of the annotations (positive, negative, skipped). Moreover, the active cleaning maintains the purpose of using the active learning approaches to annotate large scale image/video databases. Thus, the best performance could be obtained when annotating only 15-30% of the development set. The enhancement in the performance is more important when cleaning the lower-quality annotations

than better-quality annotations. Furthermore the active cleaning can better enhance the performance under the condition that the number of annotations is the same.

4.4 A posteriori cleaning effectiveness

Table 3 shows the same results in the case of a posteriori cleaning. The results are similar to the results obtained by active cleaning, as shown in the previous section, but Active cleaning is is more effective and efficient. In this table, as we can see, using the full three annotations (N100P100S100) leads to a better performance than using different fractions as in table 1. Even though, it requires three times as many annotations as the baseline, while each of the other combinations requires only few more annotations than the baseline. This is due to either the fraction is small (e.g. N5* or N10*) or because the target concepts are sparse.

	MCG-CA	CA-MCG
E1=N0P0S0	0.0840	0.0910
E2=N0P10S0	0.0833	0.0917
E3=N0P0S10	0.0847	0.0927
E4=N10P0S0	0.0858	0.0917
E5=N5P5S5	0.0841	0.0921
E6=N10P10S10	0.0852	0.0910
E7=N20P20S20	0.0877	0.0921
E8=N0P100S100	0.0866	0.0931
Full3=N100P100S100	0.0962	0.0962

Table 3. The result of the posteriori cleaning with the eight experiments described in table 1.

5 Conclusions

We have described the active cleaning approach that was used to complement the active learning approach in the TRECVID collaborative annotation. The actual impact of the active cleaning approach was evaluated on TRECVID 2007 collection. The evaluations were conducted using complete annotations that were collected from different resources, including the TRECVID collaborative annotations and the MCG-ICT-CAS annotations.

From our experiments, a significant improvement of the annotation quality was observed when applying the cleaning by cross-validation strategy, which selects the samples to be re-annotated. Experiments show that higher performance can be reached with minimum double annotations of 10% of negative samples or 5% of all the annotated samples selecteded by the proposed cleaning strategy using cross-validation. It has been shown that, with an appropriate strategy, using a small fraction of the annotations for cleaning improves much more the system's performance than using the same fraction for adding more annotations.

6 Acknowledgments

This work was partly realized as part of the Quaero Program funded by OSEO, French State agency for innovation. Experiments presented in this paper were carried out using

the Grid'5000 experimental testbed, being developed under the INRIA ALADDIN development action with support from CNRS, RENATER and several Universities as well as other funding bodies (see https://www.grid5000.fr).

References

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. J. Howe. The rise of crowdsourcing. *Wired Magazine*, 14(6), June 2006.
4. A. J. Joshi, F. Porikli, and N. Papanikolopoulos. Multi-class active learning for image classification. In *CVPR*, pages 2372–2379, 2009.
5. A. Kumar and M. Lease. Modeling annotator accuracies for supervised learning. In *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM) at the Fourth ACM International Conference on Web Search and Data Mining (WSDM)*, pages 19–22, Hong Kong, China, February 2011.
6. 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*, 13:86–91, July 2006.
7. 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.
8. G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, and H.-J. Zhang. Two-dimensional active learning for image classification. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:1–8, 2008.
9. G. Quénot, B. Delezoide, H. le Borgne, P.-A. Moëllic, D. Gorisse, F. Precioso, F. Wang, B. Merialdo, P. Gosselin, L. Granjon, D. Pellerin, M. Rombaut, H. Bredin, L. Koenig, H. Lachambre, E. E. Khoury, B. Mansencal, J. Benois-Pineau, H. Jégou, S. Ayache, B. Safadi, J. Fabrizio, M. Cord, H. Glotin, Z. Zhao, E. Dumont, and B. Augereau. Irim at trecvid 2009: High level feature extraction. In *TREC2009 notebook*, 16-17 Nov 2009.
10. B. Safadi and G. Quénot. Active learning with multiple classifiers for multimedia indexing. In *CBMI*, Grenoble, France, June 2010.
11. V. S. Sheng, F. Provost, and P. G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 614–622, New York, NY, USA, 2008. ACM.
12. 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.
13. C. G. M. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 25(1):5–35, January 2005.
14. R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 254–263, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.
15. S. Vijayanarasimhan and K. Grauman. Multi-level active prediction of useful image annotations for recognition. In *NIPS*, pages 1705–1712, 2008.