Semi-Automatic Video Annotation Based on Active Learning with Multiple Complementary Predictors

Yan SONG¹, Xian-Sheng HUA², Li-Rong DAI¹, Meng WANG¹

¹ Department of EEIS, University of Sci&Tech of China, No 4 Huangshan Road, Hefei 230027, China
² Microsoft Research Asia, 49 Zhichun Road, Beijing 100080, China
¹ {songy,lrdai,wangmeng}@ustc.edu.cn, ² xshua@microsoft.com

ABSTRACT
In this paper, we will propose a novel semi-automatic annotation scheme for video semantic classification. It is well known that the large gap between high-level semantics and low-level features is difficult to be bridged by full-automatic content analysis mechanisms. To narrow down this gap, relevance feedback has been introduced in a number of literatures, especially in those works addressing the problem of image retrieval. And at the same time, active learning is also suggested to accelerate the converging speed of the learning process by labeling the most informative samples. Generally an active learning scheme includes a sample selection engine and a learning engine. In this paper, we will discuss the limitations of existing active learning algorithms and propose a novel active learning scheme based on multiple complementary predictors and incremental model adaptation, which improves the efficiencies of both of the primary components of active learning. Firstly, an efficient sample selection scheme is proposed, in which multiple predictors are applied to find most informative samples. Then an incremental model adaptation technique, maximum likelihood linear regression (MLLR), is used to update the classifiers which tackle the issue of unbalance between the original training set and the newly labeled data. It is proved that the samples selected by the proposed scheme are more representative than general active learning scheme, as well as the incremental model adaptation scheme is effective especially when the newly added data size is small.

Categories and Subject Descriptors
H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing-indexing methods; I.2.10 [Artificial Intelligence]: Vision and Scene Understanding-video analysis.

General Terms
Algorithms, Experimentation.

Keywords
Active Learning, Video Annotation, Semantic Classification.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MIR '05, November 10–11, 2005, Singapore.
Copyright 2005 ACM 1-59593-244-5/05/0011…$5.00.

1. INTRODUCTION
Automatic content analysis techniques are widely applied to extract metadata from videos that aims at describing the content of videos at both syntactic and semantic levels. With the help of these metadata, tools and systems for video retrieval, summarization, delivery and manipulation can be created effectively. Video annotation, which can also be formulated as a video semantic classification problem, is an elementary step for obtaining these metadata.

However, great difficulties are encountered in automatically bridging the large gap between high-level semantics (what we expect) and low-level features (what we can obtain). The results of automatic video annotation techniques are far from satisfactory due to the lack of training data compared with the large variations of the video semantic concepts that are desired to be modeled, as well as the large gap between low-level features and high-level semantics. While at the same time, manual annotation is not only labor-intensive and time-consuming, but also subjects to human errors. Recently, relevance feedback is considered to be an effective method to narrow down this gap. And, to accelerate the convergence speed of learning process, several active learning schemes, in which the most informative samples are chosen to be labeled, have been proposed.

Generally, active learning is a repetitive process comprising two primary components: a sample selection engine and a learning engine. In one round of an active learning process, the selection engine selects samples from unlabeled sample pool and requests user to label them before passing to the learning engine. The learning engine then uses a supervised learning algorithm to train or update the classifier with these newly labeled samples.

For classical active learning approaches [1], Cohn et al. suggest to select samples that minimize the expected classification error defined by

\[
\int_{x} E[(\hat{y}(x) - y(x))^2 \mid x] p(x) dx
\]

where \(y(x)\) is the true label of \(x\), \(\hat{y}(x)\) is the classifier output and \(p(x)\) is the prior density of all samples. Due to the computation complexity of integral operation, direct implementation of Equation (1) is usually difficult. In practice, most of active learning methods empirically apply closest-to-boundary criterion to choose the most uncertain samples [2][3][4].

The major limitation of existing active learning algorithms is the efficiency of the sample selection criterion, which may not be able to tackle the large variations and complexity of typical semantic concepts in videos. In[3], Goh et al considered the complexity of to-be-modeled concepts and proposed a multimodal learning
approach using images’ semantic labels to guide a concept-dependent active-learning process. In our previous work [12], a representative active learning scheme was proposed, in which local consistency of video content is effectively taken into consideration. In [6], Fang Qian et al. proposed a relevance feedback approach, which is capable of collecting representative samples by exploiting complementary views of the same data. Typically, the complementarity is achieved through either redundant views of the same data [9] or different supervised learning algorithms[10].

Generally, for the learning engine, the newly labeled data will be added to original training set and then the classifier will be retrained on the updated training set. Compared with the original training set, the number of newly labeled samples is much smaller, thus the performance improvement of the retrained classifier is generally not obvious.

In this paper, we propose a novel semi-automatic semantic annotation framework based on active learning with multiple complementary predictors, termed ALCP (Active Learning with Complementary Predictors). Firstly, an efficient sample selection scheme is proposed, in which multiple predictors are applied to find the most informative samples. Then an incremental model adaptation technique based on Maximum Likelihood Linear Regression (MLLR) is used to update the classifiers which tackle the issue of unbalance between the original training set and the newly labeled data.

The rest of this paper is organized as follows. Section 2 introduces the overview of the proposed framework of active learning with complementary predictors. In section 3, the model adaptation using MLLR technique is briefly presented. In section 4, the proposed active learning scheme based on multiple predictors and MLLR is detailed. Experiment results are presented in Section 5, followed by concluding remarks and future work in Section 6.

2. FRAMEWORK OVERVIEW

Figure 1 illustrates the flowchart of the proposed active learning scheme. Firstly, the videos are segmented into shots according to timestamp (for DV) or visual similarity (for analog videos). And multiple key-frames are extracted (based on the method presented in [11]) from each shot and regarded as a compact representation of the corresponding shot.

A relatively small number of training video shots are then labeled manually. In the learning step of the active learning process, two initial predictors, \( f_1 \) and \( f_2 \), represented by GMMs (Gaussian Mixture Model), are trained on two complementary feature sets, such as color feature set and edge feature set, which are extracted from the same video data (i.e., the manually labeled video shots). Given the training set, GMMs are obtained by typical EM (Expectation Maximization) algorithm and considered as rough distribution representations of the certain semantic concepts in low level feature spaces.

Then the testing video shots are pre-clustered based on visual similarity and temporal order in an over-segmentation manner [12], in which the shots in a certain cluster mostly belong to the same semantic concept. In the process of classification and active learning, one cluster is regarded as one sample, instead of using one shot as an individual sample. And each cluster is represented by the shot closest to its center in feature space. This scheme significantly reduces the number of shots that need to be corrected or labeled by users. For the indoor and outdoor concepts, clustering based on color similarity and timestamp (which means the shots will be more likely clustered together if they are taken in the same time period) is an effective method, as illustrated in Figure 2. For other concepts, different features, clustering methods, or similarity measures may be required.

After clustering, the testing samples (i.e., clusters that represented by the “central” shots) are assigned semantic labels by the two initial predictors. In the sample selection step of the active learning process, an effectiveness measure is defined, which takes into consideration both the closest-to-boundary criterion (typically applied in general active learning schemes) and the misclassification probabilities of the multiple complementary predictors. Higher effectiveness measure means that the corresponding sample will provide larger amount of performance enhancement after being manually labeled. So the samples with high effectiveness measures (topmost \( n \)) are selected for user to do manual labeling. The newly labeled samples and those with the misclassification probabilities lower than a predefined threshold are both taken as the adaptation data for online model adaptation using MLLR technique. The updated predictors are regarded as initial predictors for next round of active learning. This iterative process will stop after reaching a predefined number of rounds, and thereafter outputs the final prediction (classification) results.

To clearly present the proposed idea, as well as to prove the effectiveness of this scheme, we take concept indoor and outdoor annotation for home videos as an example. And two predictors...
are trained on two complementary feature sets, including color features and layout edge features (according to [8]). It is easy to apply this method on other video concept annotation, such as cityscape and landscape. And multiple predictors/classifiers may also be applied using similar methods (refer to Section 4).

3. ONLINE MODEL ADAPTATION

This section briefly introduces the basic idea of online model adaptation, that is, the approach to updating the classifiers (models to be extracted) after sample selection and labeling. As aforementioned, a method based on MLLR is applied in our proposed framework.

Suppose a training data set with semantic label \( c \) is denoted by \( \mathbf{x}_c = [\mathbf{x}_{c,i}, \mathbf{x}_{c,2}, ..., \mathbf{x}_{c,n(c)}]^T \) (where \( n(c) \) is the number of samples in the training set). GMM model for this concept is trained (using EM algorithm) from this data set and represented by

\[
f_{\mathbf{x}}(\mathbf{x}_c | \Theta_{\mathbf{x},c}) = \sum_{m=1}^{M} \alpha_{m,c} N(\mu_{m,c}, \Sigma_{m,c})
\]

where \( \Theta_{\mathbf{x},c} = [\mu_{1,c}, \Sigma_{1,c}, ..., \mu_{n(c),c}, \Sigma_{n(c),c}] \) is the parameter set defining a given mixture, \( N(\mu_{m,c}, \Sigma_{m,c}) \) is a Gaussian component, \( \alpha_{m,c} \) are the mixture weights that satisfy \( \sum_{m=1}^{M} \alpha_{m,c} = 1 \).

With the newly labeled adaptation data (obtained from sample selection process), the means and variances are adapted in two separated stages. The adaptation of the mean vector is achieved by applying a transformation matrix \( W \) to the mean vector,

\[
\mathbf{\mu}' = W\mathbf{\mu} + \mathbf{B},
\]

where \( \mathbf{B} \) is the bias vector and \( W \) is the transformation matrix. Both \( W \) and \( \mathbf{B} \) are calculated based on the adaptation data according to the MLE (Maximum Likelihood Estimation) criterion. With the newly obtained means, the variances are also updated[16]. The MLLR method is able to calculate the transformation matrix efficiently when the number of adaptation data is small [2].

4. ACTIVELearning WITH MULTIPLE COMPLEMENTARY PREDICTORS

In this section, firstly we will briefly introduce a classification scheme based on complementary predictors without active learning (i.e., without manual labeling during the learning process except a set of training samples are pre-labeled from the beginning), and then detail the proposed ALCP scheme, which is based on the formal one.

As aforementioned, each sample in the test set is assigned two labels (may be different), denoted by \( f_i(x), i \in \{1, 2\} \), by the two initial predictors (classifiers) based on MAP (Maximum A Posterior). The corresponding misclassification probabilities, \( P_{1, err} \) and \( P_{2, err} \), can be represented by

\[
P_{i, err} = P(f_i(x) \neq L(x)) = 1 - \max_{c \in \{0, 1\}} P(c | x), \quad i \in \{1, 2\}
\]

where \( P(c | x) \) is the a-posterior probability of predictor \( i \) (i.e. 1 or 2) and \( L(x) \) is the true label of sample \( x \).

4.1 Classification Based on Complementary Predictors

The complementarity of different predictors is obtained through different feature sets, which can be considered as different “viewpoints” to the same data set. Here, predictor \( f_1 \) is based on a set of color features, while predictor \( f_2 \) is based on edge features. The effectiveness of this scheme is based on the assumption that the selected feature sets are complementary.

Two 45-dimensional features are applied as the complementary feature sets in our indoor/outdoor classification problem. The first feature set is a 45-dimensional color feature consisting of a 36-dimensional color histogram in HSV space and a 9-dimensional color moment features. The other 45-dimensional feature set is a block-wise edge distribution histogram (EDH). A video frame is represented in terms of \( 2 \times 2 + 1 \) blocks as shown in Figure 3. In each block, 9 dimensional EDH features (8
directions for edge pixels and 1 total number of non-edge pixels) are extracted (according to [14]).

Figure 3. The five blocks in a video frame (The frame is firstly divided into 2×2 non-overlapping blocks. And fifth block is the central block overlapping with those four blocks. All blocks are in same size).

The correlation map of these two feature sets extracted from a training video (see Section 5) is shown in Figure 4. It can be seen from this figure, the correlation between different feature sets is low while the correlation within same feature set is much higher. Experiment results also support the assumption that these two predictors are complementary.

Figure 4. Correlation matrix between color feature set and edge feature set (features are extracted from a training video applied in Section 5).

Based on these complementary predictors, we are able to improve the classification performance even without active learning, as the method detailed below.

1. The two complementary predictors, \( f_1 \) and \( f_2 \), are trained on the corresponding feature sets extracted from the same training data set.

2. Use predictor \( f_1 \) and \( f_2 \) to predict the labels of each sample in the test set. The label of sample \( L(x) \) is determined as follows:
   
   (2.1) If \( f_1(x) = f_2(x) \), then \( L(x) = f_1(x) = f_2(x) \)
   
   (2.2) If \( f_1(x) \neq f_2(x) \), then

   \[
   L(x) = \begin{cases} 
   f_1(x) & \text{if } P_{1,err}(x) \leq P_{2,err}(x) \\
   f_2(x) & \text{if } P_{2,err}(x) < P_{1,err}(x) 
   \end{cases}
   \]

   (The misclassification probability of \( L(x) \) takes the minimum of \( P_{1,err} \) and \( P_{2,err} \)).

3. The samples whose misclassification probabilities are less than a predefined threshold are selected as adaptation data to update the classifiers.

4. Apply the adaptation data to do online model adaptation using the method mentioned in Section 2. Thus the new predictors are modified to fit the newly discovered distribution obtained from these adaptation data.

5. Repeat (2)-(4) for several times and then output the final results.

4.2 Active Learning Process

A straightforward active learning scheme with complementary predictors can be described as follows (similar as [15]).

1. Train two initial complementary predictors based on the complementary feature sets extracted from the same training data set.

2. Predict the labels of testing sample using the two predictors. The label of each testing sample is determined using the method mentioned in section 4.1.

3. Select the most “uncertain” samples for each predictor, and ask user to confirm their labels.

4. Take the manually labeled samples and those predicted with low misclassification probabilities as the adaptation data set to do model adaptation. (The updated predictors are taken as the initial predictors for next round of active learning).

5. Repeat (2)-(4) for predefined number of rounds and then output the final results.

Although the above scheme may be better than the active learning scheme using only one predictor, there still has space to improve the performance by combining the prediction results during the sample selection process. It comes from the intuition that the most uncertain samples for one predictor may be “confident” ones for the complementary predictor as it is from another viewpoint to “understand” the same data. Based on this observation, we propose the ALCP scheme, which is similar to the method we have presented in Section 4.1, except a more sophisticated sample selection approach is applied.

To select more informative samples for labeling, an effectiveness measure \( Es(x) \) for each testing sample is defined, where \( x \) is a
sample in the test set. This measure reflects the estimated contribution of a certain sample in the subsequent supervised learning process when it is labeled manually. Recall that the testing samples are the centers of shot clusters, so the cluster size is another important factor to be considered.

General closest-to-boundary criterion shows that the samples closest to current classification boundary should be selected firstly, while several different cases are further considered in our proposed framework, detailed as follows.

(1) When two predictors have same prediction on a certain test sample, i.e., \( f_1(x) = f_2(x) \): If either of the predictors’ misclassification probability is low, the prediction is considered as correct; otherwise this sample is considered as a possible informative one and may be selected to be labeled manually. That is, the \( Es(x) \) is high for the later case.

(2) When two predictors have different predictions on a under-test sample, i.e., \( f_1(x) \neq f_2(x) \): If the misclassification probability of one of the two predictors is lower than a predefined threshold, then its prediction result on this sample is considered to be correct; otherwise \( x \) is a possible informative sample and need to be manually labeled.

(3) As aforementioned, one test sample actually represents one cluster of samples (shots). When a cluster represented by sample \( x \) has large number of shots, it is better to be manually labeled unless the predictor has very high prediction confidence.

According to these considerations, we quantitatively define an effectiveness measurement for each test sample \( x \) as a conditional probability, which indicates the “degree” that the predicted label conflicts with the true label:

\[
Es(x) = nP(\hat{L}(x) \neq L(x) \mid f_1(x), f_2(x)) \tag{5}
\]

where \( n \) is the number of shots in the cluster represented by the test sample \( x \). \( \hat{L}(x) \) is the predicted label of sample \( x \), (refer to Section 4.1), \( L(x) \) is the true label. \( f_1(x), f_2(x) \) are the predicted labels with misclassification probabilities \( P_{1,\text{err}} \) and \( P_{2,\text{err}} \), respectively.

The following derivations of \( Es(x) \) are based on the assumption that the prediction results of these two predictors and the corresponding misclassification probabilities are independent.

(1) If \( f_1(x) = f_2(x) \), then \( \hat{L}(x) = f_1(x) = f_2(x) \), and we have

\[
Es(x) = nP(\hat{L}(x) \neq L(x) \mid f_1(x) = f_2(x))
= \frac{nP(f_1(x) \neq L(x), f_2(x) \neq L(x))}{P(f_1(x) = f_2(x))}
= \frac{nP_{1,\text{err}} P_{2,\text{err}}}{P_{1,\text{err}} P_{2,\text{err}} + (1-P_{1,\text{err}})(1-P_{2,\text{err}})} \tag{6}
\]

(2) If \( f_1(x) \neq f_2(x) \), then

\[
\hat{L}(x) = \begin{cases} f_1(x) & \text{if } P_{1,\text{err}}(x) \leq P_{2,\text{err}}(x) \\ f_2(x) & \text{if } P_{2,\text{err}}(x) < P_{1,\text{err}}(x) \end{cases}
\]

and we have

\[
Es(x) = nP(\hat{L}(x) \neq L(x) \mid f_1(x) \neq f_2(x))
= \frac{nP_1(f_1(x) \neq L(x), f_2(x) = L(x)))}{P(f_1(x) \neq f_2(x))}
= \frac{nP_2(f_1(x) \neq L(x), f_2(x) = L(x)))}{P(f_1(x) \neq f_2(x))}
= \frac{n(P_{1,\text{err}} P_{2,\text{err}} - P_{1,\text{err}} P_{2,\text{err}})}{(1-P_{1,\text{err}}) P_{2,\text{err}} + (1-P_{2,\text{err}}) P_{1,\text{err}}}
\]

Consequently, we obtain the following result by combining (6) and (7):

\[
Es(x) = \begin{cases} \frac{nP_{1,\text{err}} P_{2,\text{err}}}{P_{1,\text{err}} P_{2,\text{err}} + (1-P_{1,\text{err}}) P_{2,\text{err}}} & \text{if } f_1(x) = f_2(x) \\frac{n(P_{1,\text{err}} P_{2,\text{err}} - P_{1,\text{err}} P_{2,\text{err}})}{(1-P_{1,\text{err}}) P_{2,\text{err}} + (1-P_{2,\text{err}}) P_{1,\text{err}}} & \text{if } f_1(x) \neq f_2(x) \end{cases} \tag{8}
\]

Accordingly the proposed ALCP scheme is described as follows:

(1) From a training set, two predictors \( f_1 \) and \( f_2 \) are trained based on the complementary feature sets.

(2) Use \( f_1 \) and \( f_2 \) to predict the labels for all samples in the test set, and two misclassification probabilities are also obtained for each sample at the same time. For each testing sample, the effectiveness measure score \( Es(x) \) is calculated.

(2.1) The testing samples will be predicted using the scheme mentioned in section 4.1, and the corresponding effectiveness measures are calculated according to Equation(8).

(2.2) The top \( m \) samples (\( m \) is predefined) with highest effectiveness measures are selected for manual labeling.

(2.3) The manually labeled samples and those predicted by complementary predictors with misclassification probabilities less than predefined threshold are selected as the adaptation data.

(3) Apply the adaptation data to do online model adaptation using the method mentioned in Section 2.

(4) Repeat (2)-(3) for a predefined number of rounds and then output the final results.

5. EXPERIMENTS

In order to evaluate the efficiency of the proposed semi-automatic annotation framework, we test it on nearly 10 home videos, which contain about 3000 shots. Each shot is manually labeled as
indoor or outdoor according to the definition in TrecVid [13]. One home video is randomly selected as the training set for training the initial predictors (based on complementary feature sets) and the remaining 9 videos are used for testing.

Low-level features used in the experiment include 36-D HSV color histogram, 9-D color moment and 45-D block-wise edge distribution histogram, as described in Section 4.1.

The test videos are segmented into shots, and then clustered into groups based on visual similarities. Each cluster is represented by the shot nearest to the cluster center. All the representatives of these clusters are taken as the actual testing set.

Two sets of experiments are implemented to evaluate different characteristics of the proposed ALCP scheme as follows.

A.1. Comparison of different model adaptation methods: In this experiment, we compare the MLLR and EM algorithm for model adaptation after obtaining the newly labeled samples. In the comparison, we iteratively select one video as training data and the other 9 videos as test videos (so totally there are $90 = 10 \times 9$ sets of results). The average classification accuracy is shown in Figure 5. From this figure, it can be seen that when the adaptation data set size is small (which is the typical case in our framework), the performance of MLLR is better than EM method. With the increasing of the adaptation data size, the performance of EM algorithm becomes better.

![Figure 5. Average classification accuracy comparison on model adaptation method: MLLR and EM.](image)

A.2. Comparison of different sample selection criterions: Three sample selection schemes are compared, including random-selecting scheme, general active learning scheme and proposed active learning scheme (ALCP). The general active learning scheme and the proposed active learning scheme have been described in Section 4.2. Random selecting scheme means we ask user to label randomly selected samples. Figure 6 shows the average classification results (average on 90 sets of results, similar as A.1) of three sample selection schemes. From Figure 6, we can see that the proposed method performs superior to the other two methods.

![Figure 6. Comparison of average classification accuracy for random sample selection, general active learning, and ALCP.](image)

6. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a novel active learning scheme for semi-automatic video semantic annotation based on multiple complementary predictors and MLLR. Multiple predictors improve the efficiency of sample selection in the active learning process, and MLLR tackled the issue of data unbalance in online model adaptation.

Future work would be to apply the scheme on multiple semantic concepts, more types of videos, and larger video database. For different semantic concepts, corresponding features should be applied. So it is important to do feature extraction and selection based on concepts. Theoretical analysis on the strategy of predictor combination, as well as the effectiveness of selected samples, is also necessary.

7. ACKNOWLEDGMENTS

We specially thank Xiao-Bing LI for providing the idea of MLLR technique applied in online model adaptation and some of experiment results.

8. REFERENCES


dependent active learning for image retrieval. ACM MM,
2004.

in Relevance Feedback. 1998.

regression for speaker adaptation of continuous density
Hidden Markov Model. Computer Speech and Language.

[8] Jun Wu, Xian-Sheng Hua, Hong-Jiang Zhang, An Online-
Optimized Incremental Learning Framework for Video

[9] A. Blum, T.M., Combining labeled and unlabeled data with
co-training. Proceedings of the Workshop on Computational

[10] S. Goldman, Y.Z., Enhancing supervised learning with
unlabeled data. In Proceedings of the Seventeenth

Graph Analysis. Computer Vision and Image understanding,
1998.

[12] Yan Song, Xian-Sheng Hua, Li-Rong Dai, Ren-Hua Wang,
Semi-Automatic Video Semantic Annotation Based on
Active Learning. VCIP, 2005.


Detection with Support Vector Machines”, in Proc. IEEE
Workshop on Content-based Access of Image and Video

Sampling with Co-Testing: Preliminary Results”,

[16] M.J.F.Gales, P.C.Woodland, “Mean and Variance
Adaptation within the MLLR Framework”, Computer Speech
and Language Volume 10. 1996