AN AUTOMATIC VIDEO SEMANTIC ANNOTATION SCHEME BASED ON COMBINATION OF COMPLEMENTARY PREDICTORS

Yan SONG [1], *Xian-Sheng HUA* [2], *Li-Rong DAI* [1], *Meng WANG* [1] , *Ren-Hua WANG* [1]

[1]Department of EEIS, University of Sci&Tech of China
[2]Microsft Research Asia, 5F Sigma Center, 49 Zhichun Road, Beijing, China

ABSTRACT

Given a large set of video database, how to connect video segments with a certain set of semantic concepts with least manual labors is an elementary step for video indexing and searching. Due to the large gap between high-level semantics and low-level features, automatic video annotation with high accuracy is a challenging task. In this paper, we propose a novel automatic video annotation framework, which improves the annotation performance by learning from unlabeled samples and exploring temporal relationship in video sequences. To effectively learn from unlabeled data, a sample selection scheme based on combining a set of complementary predictors is proposed, which iteratively refines the performance of the initial predictors. A filtering-based method is applied to further improve the annotation accuracy as well, in which video temporal relationship is sufficiently exploited. Experiment results show that the proposed automatic video annotation method performs superior to general supervised learning methods and co-training.

1. INTRODUCTION

Video annotation is to assign certain predefined semantic concepts to video segments, which is an important step for further content analysis and semantic-level process [1]. As there is a large gap between high-level semantics and low-level features, typically we use learning-based methods to learn statistic models of the semantic concepts using a large labeled corpus aiming at achieving good generalization ability. However, due to the large data size of videos, manual labeling of video data is much more labor-intensive and subject to human errors, as well as it is difficult to obtain large number of labeled samples. On the contrary, unlabeled samples are plentiful and easy to be obtained. Therefore, many semi-supervised learning techniques, which exploit the information hidden in the unlabeled data, are proposed in recent years [2][3][4].

Co-training is a frequently applied method among these works, which utilizes the complementarity of two predictors to explore the unlabeled samples either through different features extracted on same data [2] or through different supervised learning algorithm [3]. In a typical co-training process, each predictor iteratively selects a set of unlabeled samples, which are considered to be correctly predicted, to teach the other predictor. However, there is a trade-off between the number of selected samples and the possible performance improvement after introducing these samples. In a round of co-training, if the number of newly added samples is much less than that of the original training set, co-training can not improve the initial predictors obviously. On the contrary, the more samples are selected, the more misclassifications may be introduced in, which may lead to unstable or even degraded performance. To tackle this issue, a better approach is to fix the number or ratio of selected samples and then try to reduce the misclassifications in these samples. A possible method for reducing misclassification is to combine each individual prediction by comparing their confidence levels [4]. However, due to the different feature spaces or learning methods, it is not appropriate to compare the confidence levels of different predictors directly.

In this paper, we propose a novel automatic video annotation approach derived from the typical co-training algorithm, which exploits both the model and the feature complementarity to improve the annotation performance. In this scheme, a generative model based predictor and two discriminative model based predictors generated on complementary feature sets are selected as the base learners. All the predictors are refined progressively by learning from unlabeled samples, in which the labels given by different predictors are combined through a voting method, so that the problem of confidence level comparison is avoided.

On the other hand, it is observed that although the semantic concept has large variance, locally it is consistently distributed. That is, typically the variation of a concept and its corresponding features within the clip of video is relatively smaller than those among different videos. Moreover, the concept drifting is gradual in most cases [1]. Such characteristics are beneficial to further improve the annotation accuracy. In this paper, a filtering based method is applied to exploit the video temporal consistency.

The remainder of the paper is organized as follows. In Section 2, the overview of the automatic video annotation

1-4244-0469-X/06/$20.00 ©2006 IEEE

framework is described. Then the confidence level calculation in different learning algorithms (i.e. GMMs, SVMs) is presented in Section 3. Next, we detail the typical co-training scheme based on complementary SVMs and our proposed scheme in Section 4. In Section 5, the experiment results are shown to prove the efficiency of proposed scheme, followed by the concluding remarks and the future works in Section 6.

2. FRAMEWORK OVERVIEW

The proposed automatic annotation framework is illustrated in Figure 1. Firstly, the videos are segmented into shots according to timestamp (for DVs) or visual similarity (for analog videos) [4]. In the following process, each shot is represented by a certain number of frames uniformly excerpted from the shot.

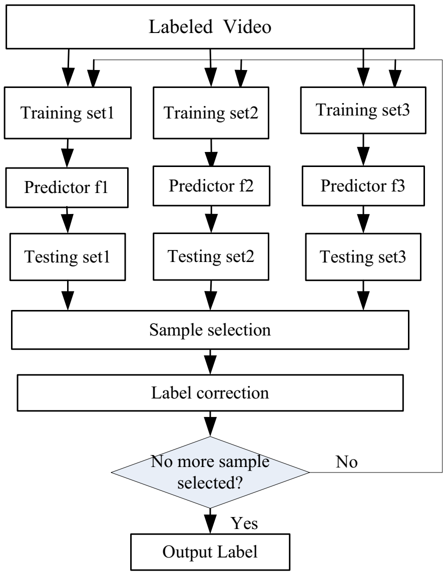

Figure.1 Flowchart of the automatic annotation

The complete low-level feature set we used is a 90-D feature (consisting of 45-D color histogram and 45-D layout edge histogram [4]). Predictor f_1, represented by a GMM (Gaussian Mixture Model), is trained on the 90-D complete feature set. Predictor f_2 and f_3, represented by Gaussian RBF kernel SVMs (Support Vector Machines), are trained on a natural split of the complete feature set used by predictor f_1. That is, predictor f_2 is trained on the color histogram features and predictor f_3 is trained on the layout edge histogram features. The complementarity of color features and edge features is shown in [4].

The annotation process consists of three primary steps, including *Model Training/Retraining*, *Sample Selection*, and *Label Correction*. Given the training set, each predictor is trained or retrained independently. Predictor f_1 is obtained by using a typical EM (Expectation Maximization) algorithm, while predictor f_2 and f_3 are trained by using the algorithms described in SVM-torch [9].

In one round of the learning process, each sample in the test set is assigned three labels (may be different) by the three different predictors, denoted by $f_i(\mathbf{x}_{i,k})$, i=1,2,3 ($\mathbf{x}_{i,k}$ is the k-th test sample in testing set i), with corresponding confidence levels of the predictions. Then, each predictor selects a certain number of samples, which are considered to be the most "confident" ones, as the candidates to be added to the training set for updating the initial predictors. Their labels are decided by the voting result of the current predictions, that is, $f_i(\mathbf{x}_{i,k})$ i=1,2,3.

In the label correction step, some isolated misclassified samples are further corrected by taking advantage of the temporal relationship among consecutive video shots. The learning process is terminated if there are no more samples selected for next iteration, and the final predictions of the testing set are taken as the annotation results.

To clarify the idea, we take two simple semantic concepts, indoor and outdoor, as an example to present our proposed automatic annotation scheme.

3. CONFIDENCE LEVEL CALCULATION FOR DIFFERENT LEARNING METHODS

In the sample selection step, each predictor needs to select samples from the testing set according to the prediction confidence levels. Typically the prediction of the test sample is decided based on MAP (Maximum a Posterior) criterion.

For a GMM model, the confidence level is defined by

$$C(\mathbf{x}) = P(f(\mathbf{x}) = L(\mathbf{x})) = \max_{l\in\{0,1\}} P(l(\mathbf{x})\,|\,\mathbf{x}) \qquad (1)$$

where $P(l(\mathbf{x})\,|\,\mathbf{x})$ is the a-posterior probability of a predictor, and $L(\mathbf{x})$ is the true label of the test sample $\mathbf{x}$.

Given a test sample $\mathbf{x}$, the output of SVM classifier f(x) is

$$f(\mathbf{x}) = h(\mathbf{x}) + b \qquad (2)$$

where $h(\mathbf{x}) = \sum_i y_i \alpha_i k(\mathbf{x}_i, \mathbf{x})$ lies in a Reproducing Kernel Hilbert Space (RKHS) induced by kernel k [7]. Although $f(\mathbf{x})$ provides the distance of $\mathbf{x}$ from the separating hyper-plane and its sign determines the class label, it can not be directly translated to a probability value that is useful for estimating the prediction confidence level.

In [7], $f(\mathbf{x})$ is mapped to the posterior probability $P(l(\mathbf{x}) = 1\,|\,f(\mathbf{x}))$ by a parametric form of sigmoid as follows

$$P(l(\mathbf{x}) = 1\,|\,f(\mathbf{x})) = \frac{1}{1 + \exp(A \times f(\mathbf{x}) + B)} \qquad (3)$$

where A, B are parameters of sigmoid function, which are fitted using maximum likelihood estimation from the training set. With the trained sigmoid function, the posterior probability and the confidence level of SVM-based predictors are calculated as those of GMM-based predictor.

4. SAMPLE SELECTION AND LABEL CORRECTION

In this section, we briefly introduce a typical co-training scheme firstly, and then detail the proposed sample selection and label correction scheme.

4.1. Co-training based on complementary SVMs

The typical co-training scheme is illustrated as follows:

```
Input:
    Two feature sets $V_1$ and $V_2$; a set of labeled samples $L$; and a set of unlabeled samples $U$.
While $U$ is not empty Do
    $C_1$ teaches $C_2$:
    (a) Train classifier $C_1$ based on feature sets $V_1$ on training data set $L$.
    (b) Classify all samples in $U$ using classifier $C_1$.
    (c) Move the top-$n$ samples from $U$, on which $C_1$ makes the most confident predictions, to $L$.
    $C_2$ teaches $C_1$:
    (a) Train classifier $C_2$ based on feature sets $V_2$ on training data set $L$.
    (b) Classify all samples in $U$ using classifier $C_2$
    (c) Move the top-$n$ samples from $U$, on which $C_2$ makes the most confident predictions, to $L$.
End While
Output: Classifiers $C_1$ and $C_2$
```

Co-training algorithm is effective to combine labeled and unlabeled data by taking advantage of different views of the same data. Here, we take the CO-SVM (CO-training with SVM) as a baseline learning algorithm.

4.2. Proposed sample selection scheme

As aforementioned, it is significant to select enough test samples with high classification accuracy, to improve the annotation performance of co-training algorithm.

In the proposed sample selection scheme, each predictor selects a certain number of test samples according to the rank list of the confidence levels, denoted by $K_i, i \in \{1,2,3\}$. For each $k \in K = K_1 \cup K_2 \cup K_3$, where K_i contains the indices of selected samples by predictor f_i, we redecide the label of the corresponding sample ($l(k)$) as

$$l(k) = \begin{cases} f_1(\mathbf{x}_{1,k}) & \mathit{if}\ f_1(\mathbf{x}_{1,k}) = f_2(\mathbf{x}_{2,k}) \\ & \mathit{or}\ f_1(\mathbf{x}_{1,k}) = f_3(\mathbf{x}_{3,k}) \\ f_2(\mathbf{x}_{2,k}) & \mathit{elseif}\ f_2(\mathbf{x}_{2,k}) = f_1(\mathbf{x}_{1,k}) \\ & \mathit{or}\ f_2(\mathbf{x}_{2,k}) = f_3(\mathbf{x}_{3,k}) \\ f_3(\mathbf{x}_{3,k}) & \mathit{otherwise} \end{cases} \tag{4}$$

Then, the newly selected samples, denoted by $S_i = \{(\mathbf{x}_{i,k}, l(k)) \mid k \in K, i = 1,2,3\}$ is added to the corresponding training set for next iteration of the learning process. Similar to co-training, this method also takes advantage of complementarity of selected samples, while the prediction accuracy of selected samples is further improved by voting the results of the three predictions.

4.3. Label correction based on filtering operation

To further improve the annotation accuracy, the temporal relationship between video shots is exploited. For example, it is less possible for such label sequence "…I, I, O, I, I… ", where symbol "I" means indoor, and symbol "O" means outdoor, to appear within a video clip, especially when these shots are similar in terms of colors or close in terms of timestamp. That is, the isolated "O" in the middle of a series of "I" may be misclassified with high probability in such case.

To correct this type of errors, a filtering operation on the label sequence is applied as follows: A window with fixed length (say, 3 or 5) is sliding along the label sequence. If the label at the middle of window is different with those at other positions, it is regarded as misclassified.

A more sophisticated method we may employ is based on pre-clustering results. The clusters of test video shots are obtained based on visual similarity and temporal order in an over-segmentation manner [4], in which the shots in a certain cluster mostly belong to a same semantic concept. Then, the sliding window is restricted in each cluster. Alternatively, we may use voting to redecide the labels of samples in each cluster. In experiments, we obtain very close results for all these methods.

5. EXPERIMENTS

In order to evaluate the performance of the proposed video annotation framework, several experiments are conducted on a home video database, which contains about 60 home videos. Each video is about one hour in duration and each shot is manually labeled as "indoor" or "outdoor" according to the definition in TRECVid [8]. The video database contains a wide variety of contents including wedding, vacation, meeting, party, and so on. Ten videos are randomly selected as the initial training set, and the others are taken as the testing set.

As mentioned in Section 2, predictor f_2 and f_3 are generated based on the complementary feature sets (45-D color features and 45-D layout edge features), respectively, while predictor f_1 is trained on the 90-D complete feature set.

Experiment 1: Classification results of supervised learning on different learning algorithms and feature sets.

Firstly, we take the supervised learning algorithm on different training sets as the baseline results (i.e., predictors are trained on labeled training set and then are used to classify each test sample). Also, the low-bound error rates are obtained by taking the whole dataset as training set (i.e., training error on the whole dataset). The results are shown in Table 1.

Table 1: *Classification results of supervised learning scheme based on different feature sets and learning algorithms.*

Models	*Feature set*	*Error rate*	*Low-bound error rate*
SVM model	*Color*	*0.239*	*0.0873*
SVM model	*Edge*	*0.298*	*0.0625*
SVM model	*Color+Edge*	*0.208*	*0.005*
GMM model	*Color+Edge*	*0.214*	*0.1236*

Experiment 2: Classification results of CO-SVM.
In this experiment, we apply CO-SVM scheme (mentioned in Section 4.1) for video classification/annotation. The SVMs trained on color and edge feature sets are selected as the initial predictors. The classification results are shown in Figure 2. Obviously the CO-SVM is more effective than the general supervised learning method.

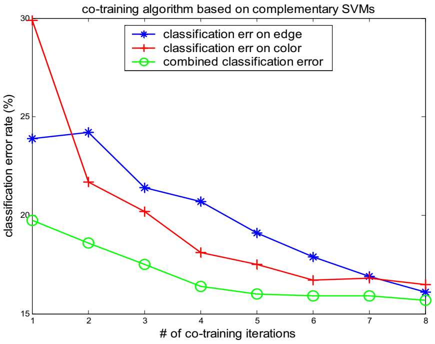

Figure.2. Classification error rate of CO-SVM

Experiment 3: Classification results of proposed scheme based on combining complementary predictors and label correction.
In this experiment, we apply the proposed sample selection and label correction scheme on video classification. The classification results of each learning round are shown in Figure 3. It can be seen that the performance of the proposed scheme outperforms CO-SVM scheme with the same training set in terms of classification error rate, which is close to the low-bound error rate of GMM model.

6. CONCLUSION AND FUTURE WORK

This paper has proposed a novel automatic video annotation framework based on learning from unlabeled video data and exploring temporal relationship of video shots. Experiment results have shown that the annotation accuracy is remarkably improved by the proposed scheme.

Future work will be to apply the scheme on multiple semantic concepts and a larger video database. Also, in current scheme, the annotation accuracy is mainly constrained by the initial training set. The methods for dynamically constructing more efficient training set will be studied in next step.

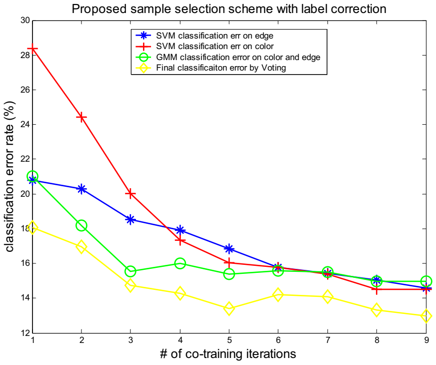

Figure.3. Classification error rate of proposed sample selection scheme with label correction.

7. REFERENCES

[1] J. Wu, X.-S. Hua, B. Zhang, H.-J. Zhang, "An Online-Optimized Incremental Learning Framework for Video Semantic Classification," ACM Multimedia pp 320-323, October 10-16, 2004.

[2] A. Blum, T.M., Combining labeled and unlabeled data with co-training. Proceedings of the Workshop on Computational Learning Theory, pp 92-100, 1998.

[3] S. Goldman, Y.Z., Enhancing supervised learning with unlabeled data. In Proceedings of the Seventeenth International Conf. on Machine Learning, 2000: p. 327-334.

[4] Y. SONG, X.-S. HUA, L.-R. DAI, M. WANG. "Semi-Automatic Video Annotation Based on Active Learning with Multiple Complementary Predictors," 7th International Workshop on Multimedia Information Retrieval, Singapore. Nov 10-11, 2005.

[5] Y. Wang and H.J Zhang, "Content-Based Image Orientation Detection with Support Vector Machines", IEEE Workshop on Content-based Access of Image and Video Libraries, pp.17-23, 2001.

[6] John C. Platt, "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods" in Advances in Large Margin Classifier, MIT Press, March 26, pp 61-74, 1999.

[7] Guidelines for the TRECVID 2003 Evaluation http://www-nlpir.nist.gov/projects/tv2003/tv2003.htm

[8] R.Collobert, S.Bengio, "SVMTorch: Support vector machines for large-scale regression problems". Journal of Mach. Learning, pp 143–160. 2001.