ENHANCED SEMI-SUPERVISED LEARNING FOR AUTOMATIC VIDEO ANNOTATION

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

[1] Department of EEIS, University of Sci&Tech of China
Huang Shan Road No.4 Hefei Anhui 230027, China
wangmeng@mail.ustc.edu.cn

[2] Microsoft Research Asia, 5F Sigma Center
49 Zhichun Road, Beijing 100080, China
xshua@microsoft.com

ABSTRACT

For automatic semantic annotation of large-scale video database, the insufficiency of labeled training samples is a major obstacle. General semi-supervised learning algorithms can help solve the problem but the improvement is limited. In this paper, two semi-supervised learning algorithms, self-training and co-training, are enhanced by exploring the temporal consistency of semantic concepts in video sequences. In the enhanced algorithms, instead of individual shots, time-constraint shot clusters are taken as the basic sample units, in which most mis-classifications can be corrected before they are applied for re-training, thus more accurate statistical models can be obtained. Experiments show that enhanced self-training/co-training significantly improves the performance of video annotation.

1. INTRODUCTION

With advances in storage devices, networks and compression techniques, large-scale video data is available to average users. How to browse and search these data has become a challenging task. To deal with this issue, it has been a common theme to develop automatic analysis techniques for deriving metadata from videos, which describes the video content at both syntactic and semantic levels. With the help of these metadata, the tools and systems for video retrieval, summarization, delivery and manipulation can be created effectively.

Semantic concept annotation is an elementary step for obtaining these metadata. As manual annotation for large video archive is labor-intensive and time-consuming, efficient automatic annotation methods are desired. For general automatic video annotation methods, statistical models are built from manually pre-labeled samples, and then labels can be assigned to unlabeled samples by these models. In this process, the lack of labeled samples is a major obstacle, which usually leads to inaccurate annotation results.

Semi-supervised learning algorithms, which attempt to learn from both labeled and unlabeled data, is one approach to deal with the lack of labeled samples. However, the improvements are limited as they are only based on certain assumptions of data set structure, such as decision boundary should avoid high density region and similar data samples mostly have a same label [2, 3].

Meanwhile, video sequence has a property named temporal consistency, which has already been commonly utilized in shot grouping and scene detection [7]. That is, the variation of semantic concept within one continuous video segment is much smaller compared to that in different video segments. In this paper, we will focus on how to enhance two general semi-supervised learning algorithms, self-training and co-training, by exploring temporal consistency in video sequences.

The remainder of this paper is organized as follows. In Section 2, general self-training and co-training are briefly introduced, and how to enhance them for video annotation is then discussed in Section 3. A framework of automatic video annotation based on enhanced self-training/co-training is presented in Section 4. Experiments are introduced in Section 5, followed by concluding remarks in Section 6.

2. SELF-TRAINING AND CO-TRAINING

Semi-supervised learning, a family of algorithms that take advantage of both labeled and unlabeled data, has been studied for a couple of years [2-6]. Among them, self-training, co-training, transductive SVM, and graph-based methods are frequently applied ones.

For self-training, firstly a classifier is trained from a small amount of labeled samples, which is then used to classify unlabeled samples. Typically the classified samples with high confidence levels are added to the training set. For co-training, it is assumed that the features can be split into two sets that are conditionally independent given the class, and each feature set is sufficient for training a “good” classifier. Initially two separate classifiers are trained based on these two feature sets with a set of labeled samples respectively. Each classifier then classifies unlabeled samples, and adds those with high confidence levels to the training set, which is applied to “teach” the other classifier. Afterwards two classifiers are re-trained from the new training set based on the corresponding feature sets, and the process repeats. Details about self-training and co-training can be found in [4] and [6].

[*] This work is performed when the first author was a visiting student in Internet Media Group, Microsoft Research Asia.

1-4244-0367-7/06/$20.00 ©2006 IEEE

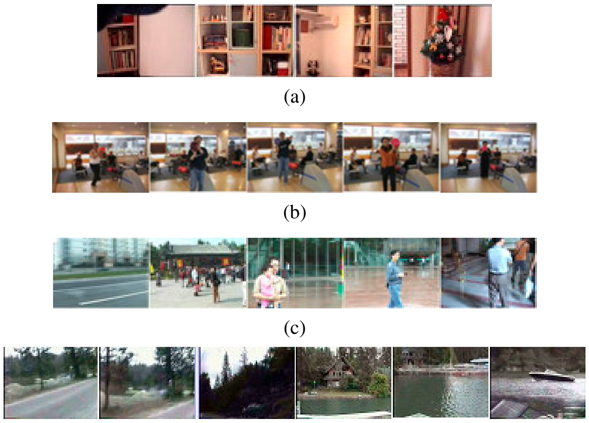
3. ENHANCED SELF-TRAINING AND CO-TRAINING

The basic ideas of both self-training and co-training are to iteratively expand the training set from classified samples with high confidence levels. Therefore, in these two algorithms, inaccurate prediction of the newly added training samples forms a bottleneck of performance improvement. However, as to be detailed in this section, some mis-classifications can be corrected by taking into account temporal consistency of semantic concept in video sequences, thus the performances of self-training and co-training can be improved.

To achieve the target, time-constraint clustering [7, 8] is applied. Considering the shots in each cluster typically have a same label (i.e., temporal consistency of semantic concept), the isolated mis-classifications can be corrected by an appropriate cluster unification process (e.g., label voting) after classification. After that, the shot clusters are taken as basic sample units and added into training set. This is the main idea of the enhanced self-training and co-training proposed in this paper.

3.1. Cluster label unification

As aforementioned, time-constraint clustering is applied to explore temporal consistency within video sequence. By selecting an appropriate window parameter that measures the time-constraint degree [7], the shots in one cluster typically have a same label, as shown by the examples in Fig. 1. Therefore, isolated mis-classified shots in a cluster can be corrected by a label unification process, only based on an assumption that the original classification accuracy is not too low (say, above 60%, which is easy to achieve).



(a)

(b)

(c)

(d)

Figure 1: Shots in a certain cluster can mainly be considered to have same label: (a) Room (b) Hall (c) Cityscape (d) Landscape

Considering a cluster C that contains N shots $\{x_1, x_2, \ldots, x_N\}$, the task of cluster label unification is to assign a label to the cluster from label set $\{l_1, l_2, \ldots, l_M\}$ that corresponds to a set of different semantic concept classes. Here we make a simple assumption that feature vectors of shots in the cluster are conditionally independent, which is expressed as

$$p(x_i, x_j \mid l_r) = p(x_i \mid l_r)\, p(x_j \mid l_r) \qquad (1)$$
$$where\ \ 1 \le r \le M;\ 1 \le i, j \le N$$

Consequently, posterior class probabilities of the cluster can be estimated as follows

$$\begin{aligned} P(l_i \mid C) &= P(l_i \mid x_1, x_2, \ldots, x_N) = \frac{P(l_i)\, p(x_1, x_2, \ldots, x_N \mid l_i)}{\sum_{j=1}^{M} P(l_j)\, p(x_1, x_2, \ldots, x_N \mid l_j)} \\ &= \frac{P(l_i)\prod_{k=1}^{N} p(x_k \mid l_i)}{\sum_{j=1}^{M} P(l_j)\prod_{k=1}^{N} p(x_k \mid l_j)} = \frac{P(l_i)}{\sum_{j=1}^{M} P(l_j)\prod_{k=1}^{N} \frac{p(x_k \mid l_j)}{p(x_k \mid l_i)}} \\ &= \frac{P(l_i)}{\sum_{j=1}^{M} P(l_j)\prod_{k=1}^{N}\left(\frac{P(l_i)}{P(l_j)}\frac{P(l_j \mid x_k)}{P(l_i \mid x_k)})\right)} = \frac{\dfrac{\prod_{k=1}^{N} P(l_i \mid x_k)}{P(l_i)^{N-1}}}{\sum_{j=1}^{M}\left(\dfrac{\prod_{k=1}^{N} P(l_j \mid x_k)}{P(l_j)^{N-1}}\right)} \end{aligned} \qquad (2)$$

where $P(l_i)$ is the prior probability of label l_i and $P(l_i|x_k)$ is the posterior probability of label l_i for shot x_k. Based on the estimated posterior probabilities $P(l_i|C)$, the unified cluster label is decided according to MAP criterion, and its confidence score is estimated according to [11] as follows

$$\psi(C) = \sqrt{P_{\max} P_{margin}} \qquad (3)$$

where $P_{max} = \max\{P(l_i|C),\ i=1,\ldots,N\}$ is the maximum posterior probability of the cluster label, $P_{margin} = P_{max} - \max\{P(l_i|C) | P(l_i|C) \neq P_{max},\ i=1,\ldots,N\}$ is the multi-class margin of the unlabeled sample.

Generally the assumption indicated by equation (1) will not be strictly satisfied. However, approximate posterior class probabilities based on equation (1) are generally sufficient for deciding the unified cluster label, which is similar as independence assumption in naive Bayesian Classifier [12]. As to be detailed in Section 5, the significantly improved performances of the enhanced self-training and co-training also prove the rationality of the assumption.

3.2. Enhanced Self-training and Co-training

As aforementioned, general self-training and co-training can be enhanced based on time-constraint clustering and cluster label unification, The process of enhanced self-training and co-training are shown in Fig. 2 and Fig. 3 respectively.

```
Input:
  A feature set $V$; a set of labeled samples $L$; a set of unlabeled samples $U$; the number of iterations $T$; and a set of clusters $UC$ generated from time-constraint clustering on unlabeled samples $U$.
For $t$ = 1, 2, …$T$
  (a) Train classifier $C$ based on feature set $V$ on training set $L$.
  (b) Classify all samples in $U$ using classifier $C$.
  (c) Unify labels for each cluster and calculate their confidence scores.
  (d) Move the samples in the top-$n$ clusters that have highest confidence scores in $UC$ to $L$ with the corresponding predicted labels.
Output:
    Classifier $C$.
```

Figure 2: Process of enhanced self-training

```
Input:
  Two complementary feature sets $V_1$ and $V_2$; a set of labeled samples $L$; a set of unlabeled samples $U$; the number of iterations $T$; and a set of clusters $UC$ generated from time-constraint clustering on unlabeled samples $U$.
For $t$ = 1, 2, …$T$
  $C_1$ teaches $C_2$:
  (a) Train classifier $C_1$ based on feature set $V_1$ using training set $L$.
  (b) Classify all samples in $U$ using classifier $C_1$.
  (c) Unify labels for each cluster in $UC$ and calculate their confidence scores.
  (d) Move the samples in the top-$n$ clusters that have highest confidence scores in $UC$ to $L$ with the corresponding predicted labels.
  $C_2$ teaches $C_1$:
  (a) Train classifier $C_2$ based on feature set $V_2$ on training set $L$.
  (b) Classify all samples in $U$ using classifier $C_2$.
  (c) Unify labels for each cluster in $UC$ and calculate their confidence scores.
  (d) Move the samples in the top-$n$ clusters that have highest confidence scores in $UC$ to $L$ with the corresponding predicted labels.
Output:
    Classifiers $C_1$ and $C_2$
```

Figure 3: Process of enhanced co-training

4. PROPOSED AUTOMATIC VIDEO ANNOTATION FRAMEWORK

The automatic video annotation framework based on enhanced self-training/co-training is illustrated in Figure 4. Firstly, one or two sets of predictors, which correspond to self-training and co-training respectively, are trained on one or two feature sets, which are extracted from the pre-labeled video data set (i.e., the labeled video shots).

As shown in Fig. 4, the test process consists of five steps. Firstly, the unlabeled videos are segmented into shots. Then, all these shots are time-constraint clustered, which is as same as in [7]. In enhanced self-training/co-training, pre-trained predictors are refined by learning from unlabeled samples as presented in above sections. Then more accurate predictors are obtained and applied to classify unlabeled shots. Of course for co-training there are two sets of results, which are then combined according to their confidence levels, which is as same as in [1]. The final step is cluster label unification again, which has been proved helpful for generating more accurate labels as final output.

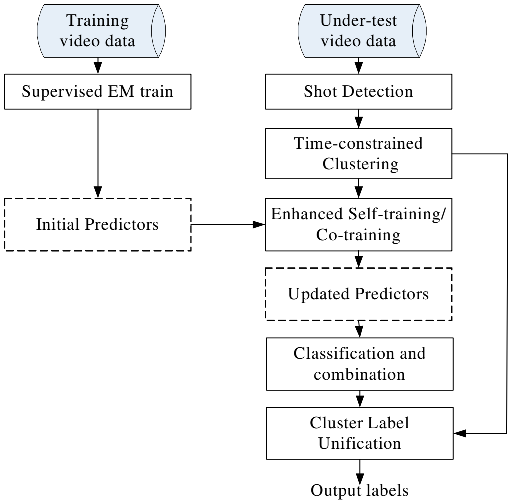


Figure 4: Proposed automatic video annotation framework based on self-training/co-training

5. EXPERIMENTS

To evaluate the performance of proposed video annotation framework, a couple of experiments are conducted on 60 home videos, which contains about 10000 shots. Each shot is manually labeled as *room*, *hall*, *cityscape* or *landscape*. "Cityscape" and "landscape" are already defined in TRECVID [9]. Here we further divide the concept of "indoor", which is also defined in TRECVID, into "room" and "hall". The label "room" mainly corresponds to smaller rooms, such as apartment, house and office, while the label "hall" mainly corresponds to public and large rooms, such as church, shop and restaurant.

Ten videos, which are about 15% of the whole data set, are chosen randomly to be the training set, and the others are test set. All of the results in this section are the average of 5 such runs.

The feature sets are formed by a color feature set, which includes 36-dimensional HSV histogram and 9-dimensional color moment features, and an edge feature set that includes 45-dimensional block-wise edge distribution histogram (EDH). GMMs (Gauss Mixture Model) are used to model different concepts. For self-training, GMMs are built on the whole 90-dimensional feature set. For co-training, two sets of GMMs are built on 45-dimensional color feature set and 45-dimensional edge feature set respectively, which is the same as our previous work [1].

The annotation results based on both general and enhanced self-training/co-training with different learning iterations (i.e., parameter *T* in Fig. 2 and 3) are presented in Fig. 5 and Fig. 6. We use an evaluation of average "***RP***" of four labels. Here ***RP*** is defined by *2rp/(r+p)* (as described in [10]), where *p* means precision and *r* means recall. More detailed results are illustrated in Table 1.

From Fig. 5 and Fig. 6 we can see that the increases of ***RP*** curves from enhanced self-training and co-training are mostly sharper, where the main reason is the reduction of inaccurate predicted labels in newly added training samples. Obviously the performances of annotation based on enhanced self-training and co-training significantly outperform those based on general self-training and co-training

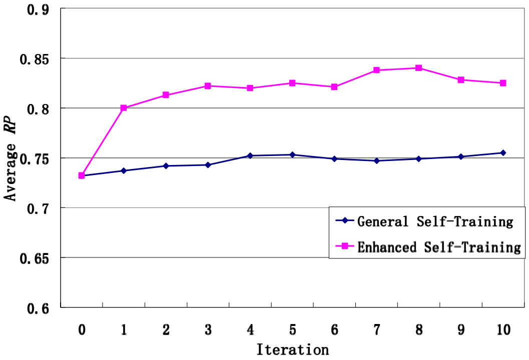


Figure 5: Average ***RP*** of annotation based on general and enhanced self-training

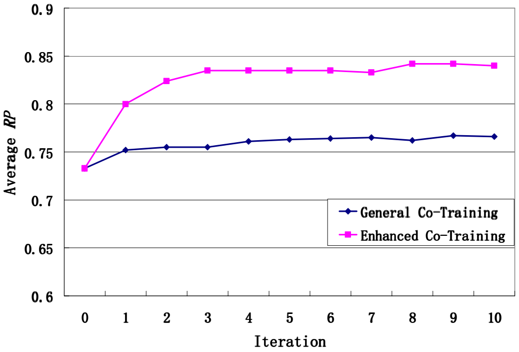


Figure 6: Average ***RP*** of annotation based on general and enhanced co-training

6. CONCLUSIONS

This paper has proposed a novel enhanced self-training/co-training algorithm for automatic video annotation. Compared with general ones, enhanced self-training and co-training explore temporal consistency of semantic concepts in video sequences based on time-constraint clustering and cluster label unification. Experiment results have shown that the enhanced self-training and co-training algorithms significantly improved the annotation performance.

7. REFERENCES

[1] Y. SONG, X.-S. HUA, L.-R. DAI and 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

[2] I. Cohen, F.G. Cozman, N. Sebe, M.C. Cirelo and T.S. Huang, "Semi-supervised Learning of Classifiers: Theory, Algorithms and Their Application to Human-Computer Interaction", *IEEE trans on PAMI*, vol. 26, no. 12, 2004

[3] X. Zhu "Semi-supervised Learning with Graphs", Doctoral dissertation, Carnegie Mellon University, CMU_LTI_05_192

[4] C. Rosenberg, M., Heberg, and H. Schneiderman "Semi-supervised self-training of object detetion models", *7th IEEE Workshop on Applications of Computer Vision*. Jan, 2005

[5] A. Blum and T. Mitchell, "Combining Labeled and Unlabeled Data with Co-Training," *Proc. COLT*, Madison, WI, 1998, pp. 92-100

[6] D. Zhang and W.S. Lee, "Validating Co-Training Models for Web Image Classification". *Proc. SMA Annual Symposium*, NUS, Jan 2005

[7] M. Yueng, B.L. Yeo and B. Liu, "Extracting story units from long programs for video browsing and nevigation", *Proc. IEEE International Conference on Multimedia Computing and Systems*, June, 1996

[8] D. Zhong, H.J. Zhang and S.F. Chang, "Clustering Methods for Video Browsing and Annotation", *Proc. SPIE Conf. on Storage and Retrieval for Image and Video Databases IV*, San Jose, February, 1996

[9] Guidelines for the TRECVID 2003 Evaluation http://www.itl.nist.gov/iaui/894.02/projects/t2002v/t2002v.html

[10] S. Raaijmakers, J.D. Hartog and J. Baan, "Multimodal topic segmentation and classification of news video", *Proc. ICME 2002*, vol.2, pp 33-36

[11] B. Li, K. Goh and E. Chang, "Confidence-based dynamic ensamble for image annotation and semantic discovery", *Proc. ACM MM*, 2003

[12] P. Domingos and M. Pazzani, "Beyond Independence: Conditions for the optimality of the simple Bayesian classifier", *Proc. International Conference on Machine Learning*, 1996

Table 1: Detailed results of annotation based on general and enhanced self-training/co-training

	Performance of general and enhanced self-training									Performance of general and enhanced co-training								
	Beginning			After General Self-training With 10 iterations			After Enhanced Self-training With 10 iterations			Beginning			After General Co-training With 10 iterations			After Enhanced Co-training With 10 iterations		
	p	*r*	***RP***	*p*	*r*	***RP***	*p*	*r*	***RP***	*p*	*r*	***RP***	*p*	*r*	***RP***	*p*	*r*	***RP***
Room	0.66	0.61	0.63	0.67	0.71	0.69	0.79	0.74	0.76	0.69	0.68	0.69	0.69	0.71	0.70	0.85	0.70	0.77
Hall	0.66	0.67	0.66	0.65	0.76	0.70	0.78	0.87	0.82	0.57	0.72	0.64	0.63	0.72	0.67	0.80	0.87	0.84
Cityscape	0.76	0.77	0.77	0.82	0.75	0.78	0.83	0.84	0.83	0.82	0.75	0.78	0.83	0.78	0.81	0.84	0.89	0.86
Landscape	0.89	0.83	0.86	0.90	0.83	0.87	0.94	0.84	0.89	0.9	0.77	0.83	0.92	0.86	0.89	0.94	0.87	0.90
Average	0.74	0.72	**0.73**	0.76	0.76	**0.76**	0.84	0.82	**0.83**	0.75	0.73	**0.74**	0.77	0.77	**0.77**	0.86	0.83	**0.84**