VIDEO ANNOTATION BY ACTIVE LEARNING AND SEMI-SUPERVISED ENSEMBLING

Yan SONG [1], Guo-Jun QI [2], Xian-Sheng HUA [3], Li-Rong DAI [1], Ren-Hua WANG [1]

Department of EEIS, Univ. of Sci&Tech of China [1]
Department of Automation, Univ. of Sci&Tech of China[2]
Microsoft Research Asia[3]

ABSTRACT

Supervised and semi-supervised learning are frequently applied methods to annotate videos by mapping low-level features into semantic concepts. Due to the large semantic gap, the main constraint of these methods is that the information contained in a limited-size labeled dataset can hardly represent the distributions of the semantic concepts. In this paper, we propose a novel semi-automatic video annotation framework, active learning with semi-supervised ensembling, which tries to tackle the disadvantages of current video annotation solutions. Firstly the initial training set is constructed based on distribution analysis of the entire video dataset. And then an active learning scheme is combined into a semi-supervised ensembling framework, which selects the samples to maximize the margin of the ensemble classifier based on both labeled and unlabeled data. Experimental results show that the proposed method performs superior to general semi-supervised learning algorithms and typical active learning algorithms in terms of annotation accuracy and stability.

1. INTRODUCTION

Video annotation aims at extracting metadata for describing the video content at both syntactic and semantic levels. It is an important step for fast and robust search of video or video clips in large video collections. Generally, to achieve good generalization ability, typical learning-based methods need a large labeled corpus to build statistic models for the semantic concepts, in which intensive manual labeling is required. On the contrary, unlabeled samples are plentiful and easy to be obtained via diverse ways. Thus semi-supervised learning techniques, which exploit unlabeled data as well, are proposed to improve the performance of annotation especially in case that the labeled dataset is limited. Existing approaches include semi-supervised ensemble methods [2][3], co-training[4], etc.

Ensemble methods, as the frequently-applied adaboost [5], use a base learning mechanism to iteratively construct a weak classifier and add it to the current ensemble classifier with an appropriate scalar multiplier (step-size). It is known that such algorithms are actually performing gradient descent of an error function in a function space [6]. D´Alch´e Buc et al [2] showed that the error function can be extended and applied in the case of semi-supervised learning. K.P.Bennett [3] further proposed the ASSEMBLE algorithm, which works with any cost-sensitive base learner and at the same time has a simple, adaptive step-size rule. Although semi-supervised ensemble methods perform well in the empirical tests on both two-class and multi-class problems, they fail on complex dataset [2], where the labeled data will really bring crucial information that can't be obtained in unlabeled data such as the case in video dataset.

To tackle this issue, in the novel framework presented in this paper, we argue that firstly the training set should be carefully constructed to ensure the performance of the initial classifier. That is, according to the distribution analysis of the entire video dataset, a "skeleton" of the semantic prototypes can be obtained in an initial training set represented by a limited number of samples. Then the active learning scheme is further introduced, which improves the performance by adding the informational details with least manual labeling.

The initial training set can be obtained by exploiting the clustering information, which is extracted according to the following observation on a large video collections (i.e., typically a semantic concept and its corresponding feature variation within a same video are relative smaller than those among different videos, as well as the concept drifting is gradual in most cases [1]). Thus, firstly the video shots are pre-clustered in an over-segmentation manner based on visual similarity and temporal order [9]. And then the initial training set is constructed by selecting samples according to the clustering information.

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 sample 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. In practice, most of active learning methods empirically apply "closest-to-boundary" criterion to choose the most uncertain samples [7][8]. The major limitation of existing active learning algorithms on video annotation is that the "closest-to-boundary" sample selection criterion

may not be able to tackle the large variations and complexity of typical semantic concepts in videos.

In this paper, a novel active learning scheme, ALBoostU, is proposed, which combines the two primary components of active learning into a semi-supervised ensembling framework. ALBoostU inherits good properties of ASSEMBLE (i.e. the unlabeled data can be assimilated into margin cost sensitive ensemble algorithms etc.[3]). Moreover, it takes advantages of active learning to accelerating the converging speed of the learning process. Based on the relationship between the margin of the ensemble classifier and the upper bound of the generalization error, the efficiency of the selected samples and learning process can be evaluated theoretically.

The rest of this paper is organized as follows. Section 2 introduces the proposed framework. In Section 3, the criterion of dynamic construction of the initial training is presented. In section 4, the proposed active learning algorithm with ASSEMBLE is detailed. Experiment results are presented in Section 5, followed by concluding remarks and future work in Section 6.

2. FRAMEWORK

Figure.1 illustrates the flow of video dataset pre-processing. Firstly, each video is segmented into shots according to timestamp (for DVs) or visual similarity (for analog videos). Each shot is represented by a certain number of frames uniformly excerpted from the shot. And then all the shots are time-constrained clustered, which is same as [9]. In the following process, each cluster is represented by the shot closest to its center in feature space and one cluster is regarded as one sample of dataset.

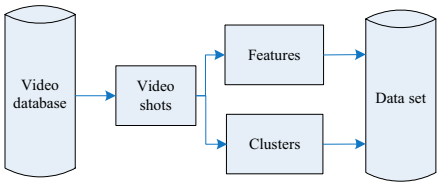


Figure.1. Pre-processing of the video dataset

As shown in Figure.2, the semi-automatic annotation process consists of two primary steps including *Construction of Initial Training Set* and *Active Learning with ASSEMBLE*.

The initial training set is constructed by selecting the "representative" samples (clusters) from dataset. And the initial prediction of unlabeled part of dataset is obtained by the Gaussian-RBF kernel SVM trained on the constructed dataset for multiple-class problems and is used in ASSEMBLE. Then, in active learning with ASSEMBLE, the base learner f_t is firstly trained on current labeled and unlabeled data (with initial prediction). Then the sample selection and learning process is embedded to further maximize the margin of the ensemble classifier. After that, the initial prediction of the unlabeled part will be updated according to current ensemble F. This learning process will iterate for several rounds.

As the dataset is consisted of the cluster centers, it is necessary to extend the annotation result of each sample to each shot in video dataset. Here, we simply make the shots in a cluster to take the same label as the label of the cluster center.

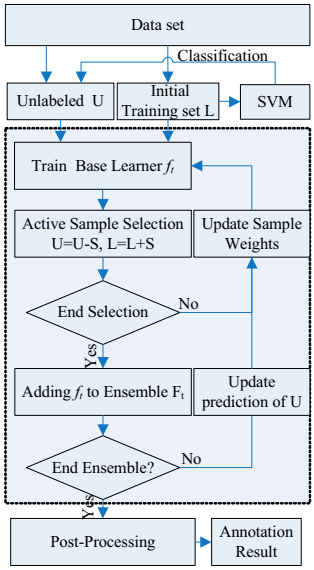


Figure.2. Framework of semi-automatic annotation process

3. CONSTRUCTION OF INITIAL TRAINING SET

As aforementioned, the initial training set is constructed to roughly represent the prototypes of the semantic concepts to be modeled from the video collections. As the distribution of these prototypes is locally consistent and globally dispersive in the low-level feature space, we propose to use the following properties, *Salience* and *Dispersiveness* to measure the representativeness of the training set.

The *Salience* property measures the local consistency of clusters. That is, the clusters with relatively large number of video shots should be selected firstly. The *Dispersiveness* property is defined from the intuition that as the clusters are obtained in an over-segmentation manner [9], the two salient clusters close to each other in temporal order may belong to the same concept with high probability. So generally selecting one of them is enough. That is to say, the sample to be selected should distribute dispersively through the whole video dataset thus more prototypes of the semantic concept can be included in training set.

The final criterion for constructing the initial training set is a linear combination of the *Salience* and *Dispersiveness* properties under the constraint of the fixed size of initial training set. In implementation, a heuristic searching scheme is used. Firstly, we select the samples according to their cluster size. If there are samples lying close to each other in temporal order, only one of these

samples is selected. This selecting and removing iteration will continue until the fixed size of training set is obtained.

4. ACTIVE LEARNING IN SEMI-SUPERVISED ENSEMBLE FRAMEWORK

In this section, firstly we will briefly introduce the criterion of training new weak learner f_{t+1} in ASSEBLE[3], and then detail the proposed active learning with ASSEMBLE, ALBoostU, which is based on the formal one.

4.1. Criterion of training new weak learner

In ASSEMBLE [3], let F_t represents the ensemble classifier after adding the t-th component classifier. The weak learner f_{t+1} is trained to minimize the margin cost function for labeled and unlabeled data as follows

$$C(F) = \sum_{i \in L} \alpha_i M(y_i F_t(x_i)) + \sum_{j \in U} \alpha_j M(|F_t(x_j)|) \tag{1}$$

where a_i,a_j are weights of labeled and unlabeled samples, L and U are labeled and unlabeled dataset and $M(z)=e^{-z}$ is the cost function. And the label for unlabeled data is $y_i=sign(F(x_i))$. Finding a possible new base classifier f_{t+1} to minimize the margin cost function is equivalent to maximize the inner product $J(F,f) = -\langle \nabla C(F), f \rangle$. Thus the criterion of training new weak learner is

$$\begin{aligned} f_{opt} &= \arg\max_f J(F,f) \\ &= \sum_{i \in L} a_i y_i f(x_i) M'(y_i F(x_i)) + \sum_{i \in U} a_i y_i f(x_i) M'(sign(F(x_i)) F(x_i)) \end{aligned} \tag{2}$$

where $M'(z)$ is the derivative of the margin cost function with respect to variable z. As $C(F)$ for unlabeled data is not differentiable, the sub-gradient of $C(F)$ is defined as

$$\nabla C(F_t(x_i)) = a_i y_i M'(F(x_i)) \quad \text{for } x_i \in U \tag{3}$$

4.2. Active learning with ASSEMBLE

Generally, the performance of ASSEMBLE is constrained by the labeled dataset and the complexity of the semantic concepts. To further improve the performance, active learning is introduced. For comparison, we firstly propose a straightforward active learning scheme, NALBoostU (Naïve Active Learning with AdaBoostU), which takes the closest-to-boundary criterion to select the most "informative" samples. However, this scheme does not take advantage of the margin cost function in ASSEMBLE.

To tackle this issue, we propose a novel active learning scheme, ALBoostU, which starts from the idea that the selected samples and the corresponding new weak learner should further minimize the margin cost function in equation(4). For simplicity, we take two-class problem as example to derive the sample selection and learning criterion. And this criterion can be generalized to the multiple-class problem by the schemes that use multiple two-class classifiers to solve multiple class issues.

In one round of active learning, a sample x_j, $j \in U$ is selected to be labeled. From equation(2), the unlabeled sample and corresponding base learner are selected to maximize the gradient of the cost function. That is

$$\begin{aligned} (f_{opt}, x_j, y_j) &= \arg\max_{f, x_j, y_j} (J(F, f_{opt}, x_j, y_j)) \\ &= \arg\max_{f, x_j, y_j} [\sum_{i \in L} a_i y_i f(x_i) M'(y_i F(x_i)) \\ &\quad + a_j y_j f(x_j) M'(sign(F(x_j)) F(x_j)) \\ &\quad + \sum_{i \in U - \{x_j\}} a_i y_i f(x_i) M'(sign(F(x_i)) F(x_i))] \end{aligned} \tag{4}$$

where a_i is the sample weights. The optimal solution may be obtained by searching through the base learner's function space and the unlabeled samples' feature space, which is a computation extensive process. Here, we take an iterative process similar to EM (Expectation Maximization) to find the suboptimal solution. As the samples in dataset are the clusters, the cluster size should be considered in the following EM-like process.

E-Step: In ASSEMBLE algorithm, the base learner f is trained according to current labeled and unlabeled dataset. We calculate the increment of expectation on the inner product by adding an unlabeled sample x_j with all possible label y_j^r, according to current ensemble classifier F and base learner f. The unlabeled sample is selected as follow,

$$\begin{aligned} (x_j, y_j) &= \arg\max_{x_j \in U, y_j} E[J(F, f, x_j, y^r{}_j) - J(F, f)] \\ &= \arg\max_{x_j \in U, y_j} n_j a_j E[(y^r{}_j f(x_j) M'(y^r{}_j F(x_j)) \\ &\quad - \mathrm{sgn}(F(x_j)) f(x_j) M'(\mathrm{sgn}(F(x_j)) F(x_j))] \end{aligned} \tag{5}$$

where n_j is the number of shots that the j-th cluster contains.

M-Step: After adding the newly labeled samples, the base learner f will be re-trained as follow

$$\begin{aligned} f_{opt} &= \arg\max_{f'} E[J(F, f', x_j, y_j) - J(F, f)] \\ &= \arg\max_{f'} n_j a_j E[(y_j f(x_j) M'(y_j F(x_j)) \\ &\quad - \mathrm{sgn}(F(x_j)) f(x_j) M'(\mathrm{sgn}(F(x_j)) F(x_j))] \end{aligned} \tag{6}$$

By this EM-like iterative process, the suboptimal solution of (f_{opt}, x_j, y_j) is obtained, which further maximizes the margin of ensemble classifier by sample selection and learning process of active learning. The remaining processes including prediction of unlabeled data, update of the sample weights and choice of step-size of newly obtained base classifier etc., are similar as ASSEMBLE [3].

According to the upper bound of generalization error function in terms of the margin of ensemble [10]

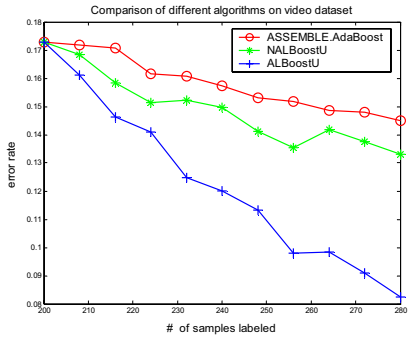
$$\hat{\Pr}(\mathrm{margin}_f(x, y) \le \theta) + \tilde{O}(\sqrt{\frac{d}{m\theta^2}}) \tag{7}$$

for any $\theta > 0$ with high probability, d is VC-dimension of base classifier space and m is the size of training set. The effectiveness ALBoostU may be that ALBoostU will reduce the first term in equation(7) maximally by adding new samples.

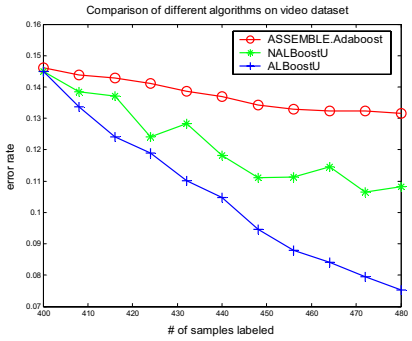
5. EXPERIMENT

To evaluate the performance of our proposed algorithms, we conduct several experiments on real video dataset, which contains about 50 videos with a wide variety

of contents. After pre-process of the video dataset, about 6400 shots are obtained. These shots are further pre-clustered into about 1400 clusters in an over-segmentation manner. Each shot is manually labeled as "indoor", "cityscape", "landscape" and "unknown" according to the definitions in TRECVid [12]. The base learner used is MLP (Multiple Layer Perceptron) as illustrated in TORCH [11], which consists of a linear layer, a tanh layer, a linear layer and a log-softmax layer. Low-level feature vector we used is 90-D, consisting of a 36-D HSV color histogram, a 9-D color moment and a 45-D block-wise edge distribution histogram [9]. Each shot is represented by a certain number (i.e. 10) of frames uniformly excerpted from the shot, and the shot closest to the cluster center is taken as the sample to form the dataset.



(a)



(b)

Figure.3. Comparison of different algorithms on video dataset. (a) Initial size of training set: 200. (b) Initial size of training set: 400

As aforementioned in Section 3, the initial training set with 200 and 400 samples is dynamically constructed according to representative criterion. In each round of ALBoostU, 8 samples are selected for user to label. In ASSEMBLE algorithm, the samples for labeling are randomly selected. In NALBoostU algorithms, the samples are selected according to the closest-to-boundary criterion. The experiment results are shown in Figure 3, where we can see that the proposed ALBoostU performs superior to both NALBoostU and ASSEMBLE. In Figure 2, we can also see that, when the size of initial training dataset is 200, the error rate is about 0.172; while when the size of initial training set is 400, the error rate is 0.146. That is to say, even using the representative criterion to constructing training set, adding 200 more training samples only reduce 0.026 in terms of error rate. While after adding 80 samples, the error rates of ASSEMBLE, NALBoostU and ALBoostU reduce 0.018, 0.03 and 0.08, respectively. One possible explanation is that by selecting 200 samples, the "skeleton" of the video dataset can be well drawn. So adding more such kind of labeled samples will be less helpful than those in active learning scheme

6. CONCLUSION AND FUTURE WORK

In this paper, we proposed a novel semi-automatic video semantic annotation framework, ALBoostU, which further maximizes the margin of ensemble classifier by labeling elaborately-selected new samples. The experiments on real video dataset have shown promising results. This framework can be further applied in building semantic indexing for large repository of video data on the Internet that enables real content-based video search. Future work will be to apply this scheme on multiple semantic concepts, more types of videos, and larger video database.

7. REFERENCES

[1] Jun Wu, Xian-Sheng Hua, Hong-Jiang Zhang, An Online-Optimized Incremental Learning Framework for Video Semantic Classification, ACM Multimedia, 2004.

[2] D'Alch´e Buc,F, Grandvalet,Y. and Ambroise,C., Semi-supervised MarginBoost, Advances in Neural Information Processing Systems MIT Press, 2002.

[3] K.P.Bennett, A.Demiriz and R.Maclin, Exploiting unlabeled Data in Ensemble Methods, SIGKDD'02,2002.

[4] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In COLT: Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers, 1998.

[5] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm, In International Conference on Machine Learning, pages 148–156, 1996.

[6] L. Mason, J.Baxter, P.Bartlett and M.Frean. Boosting algorithms as gradient descent, Advances in Neural Information Processing Systems 12, pages 512—518, MIT Press, 2000.

[7] Schohn, G. and D. Cohn, Less is more: Active learning with support vector machines, Proc. 17th International Conf. on Machine Learning, 2000: p. 839–846.

[8] Tong, S.K., D., Support vector machine active learning with applications to text classification, Journal of Machine Learning Research, 2001: p 45–66.

[9] Y.Song, X-S.Hua and L-R.Dai, etc. Semi-Automatic Video Annotation Based on Active Learning with Multiple Complementary Predictors, MIR'05, 2005.

[10] R. E. Schapire, The Boosting Approach to Machine Learning An Overview, MSRI Workshop on Nonlinear Estimation and Classification, 2002

[11] R.Collobert, S.Bengio, SVMTorch: Support vector machines for large-scale regression problems, Journal of Mach. Learning, 2001.

[12] Guidelines for the TRECVID 2003 Evaluation, http://wwwnlpir.nist.gov/projects/tv2003/tv2003.html