AN INTERACTIVE VIDEO ANNOTATION FRAMEOWRK WITH MULTIPLE MODALITIES*

Meng Wang[†], *Xian-Sheng Hua*[‡], *Yan Song*[†], *Li-Rong Dai*[†], *Ren-Hua Wang*[†]

[†] University of Science and Technology of China, Hefei 230027, P.R. China
[‡] Microsoft Research Asia, Beijing 100080, P.R. China
wangmeng@mail.ustc.edu.cn, xshua@microsoft.com, {songy, lrdai, rhwang}@ustc.edu.cn

ABSTRACT

Active learning and semi-supervised learning methods are frequently applied in multimedia annotation tasks in order to reduce human labeling effort. However, in most of these methods only single modality is applied. This paper presents an interactive video annotation framework, which is based on semi-supervised learning and active learning with multiple multimodalities. In the proposed framework, unlabeled samples are iteratively selected to be annotated manually according to certain strategy which has taken the potentials of different modalities into account, and then a graph-based semi-supervised learning algorithm is conducted on each modality. This process repeats for several rounds, and the results obtained from multiple modalities are then fused to generate final output. The proposed framework is computationally efficient, and the experimental results on TRECVID 2005 benchmark show that the proposed framework considerably outperforms previous approaches.

Index Terms— *Video annotation, active learning, multimodality*

1. INTRODUCTION

Semantic annotation of video sequences is an elementary step to obtain video metadata, which facilitates content-based video retrieval, summarization, and other manipulations. Ideally, video annotation task is formulated as a classification problem and it is accomplished by learning based methods. However, due to the large gap between low-level features and high-level semantic concepts, typically learning based methods require a large labeled training set to achieve a satisfactory performance. As manual annotation is labor-intensive and time-consuming (e.g., experiment in [4] proves that typically annotating 1 hour of video takes 2 to 18 hours), many methods to help reduce human effort have been proposed [5， 9].

One way to deal with the problem is to apply Semi-Supervised Learning (SSL) methods, which can tackle the training data insufficiency problem by leveraging large amount of unlabeled data. Another approach is to utilize active learning, which selects the most informative samples for further labeling so that the training set is more effective.

Although many different SSL and active learning methods have been successfully applied in multimedia classification, most of the existing works neglect context of "multimodality", i.e., these methods are only applied with single modality. Here multimodality is defined as multiple features of multimedia content (such as color, edge, texture, audio, and text), and multimodality fusion is defined as combining classification results obtained with these features. As demonstrated by a great deal of experiments, multimodality fusion can improve multimedia classification performance compared with simply concatenating different features into a large vector, since the latter method usually encounters *dimensionality curse* problem, which may introduce performance degradation [14].

In [5, 6], Chen et al. proposed a simple multimodality active learning method with SVM, which selects a certain number of samples closest to the hyperplane of each sub-model (i.e., model trained based on each individual modality). Experiments have demonstrated its superiority over single-modality based active learning. However, this method neglects the "discriminative ability" of different modalities. For example, some features may not be discriminative enough for a giving concept, and consequently the corresponding sub-models can hardly be improved by active learning process, thus selecting more samples for other modalities may be more promising.

In this paper, we propose an efficient interactive video annotation framework based on SSL and active learning with multiple modalities. In this framework, we present a novel strategy to select samples for manual labeling in the active learning process, where the numbers of selected samples for different modalities are adapted to the performance variations of their corresponding sub-models. For each sub-model, we choose samples according to the following three criteria: *informativeness*, *density*, and *diversity*. With the labeled samples, manifold-ranking (a graph-based SSL method) is applied with each modality feature.

2. THE PROPOSED FRAMEWORK

In this section, we introduce the proposed interactive video annotation framework illustrated in Figure 1. Firstly, videos are segmented into small temporal units (shots or sub-shot). Then several feature sets are extracted from these units. Typically each feature set is regarded as one modality, but alternatively we can also use other modality generation method, such as the method proposed in [14].

In the interactive annotation process, we iteratively employ manifold-ranking (a graph-based SSL method) [9, 17] to learn new model for each modality, and in each iteration a number of unlabeled samples are selected for manually annotation according to certain strategy. The manifold-ranking process and our sample selection strategy are introduced in the following sub-sections.

After several iterations, the results obtained from multiple sub-models are fused by linear combination with the weights decided by cross-validation on labeled set, since existing works prove this method is both effective and efficient.

* This work was performed when the first author was visiting Microsoft Research Asia as a research intern.

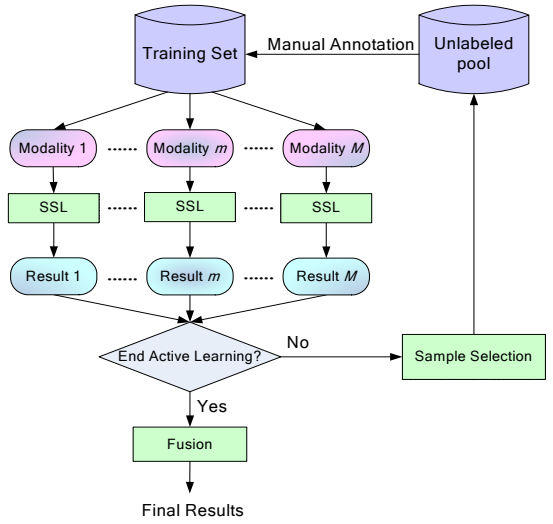

Figure 1. Proposed interactive video annotation framework

2.1. Manifold-Ranking

Manifold-ranking is a graph-based SSL method, which is based on an assumption that the labels of nearby samples are close [17]. Let $L = \{x_1, x_2, \ldots, x_l\}$ be labeled set and $U = \{x_{l+1}, x_{l+2}, \ldots, x_n\}$ be the unlabeled set. Considering there are M modalities, each sample x_i is represented by $\{x_i^1, x_i^2, \ldots, x_i^M\}$, where x_i^m is the feature representation for x_i in the m-th modality (similarly, for other notations we also use superscript to denote modality index). We define a vector $\boldsymbol{y}^+ = \{y_1^+, y_2^+, \ldots, y_n^+\}$, where $y_i^+ = 1$ if x_i is a labeled positive sample, and $y_i^+ = 0$ otherwise. Conversely, we define $\boldsymbol{y}^- = \{y_1^-, y_2^-, \ldots, y_n^-\}$, where $y_i^- = -1$ if x_i is a labeled negative sample, and $y_i^- = 0$ otherwise. Then we implement the manifold-ranking process for the m-th modality as follows:

1. Define a sparse graph above all samples in m-th modality: x_i^m and x_j^m are connected if x_i^m belongs to the K-nearest neighborhood of x_j^m, and vice versa.
2. Define affine matrix $\mathbf{W}^m$ by letting $W_{ij}^m = \exp(-|\mathbf{x}_i^m - \mathbf{x}_j^m|/2\sigma)$ if x_i and x_j are connected and $i \neq j$, and otherwise $W_{ij}^m = 0$.
3. Construct a matrix $\mathbf{S}^m = \mathbf{D}^{-1/2}\mathbf{W}^m\mathbf{D}^{-1/2}$ in which $\mathbf{D}$ is a diagonal matrix with its (i, i)-element equals to the sum of the i-th row of $\mathbf{W}^m$.
4. Initialize $[\boldsymbol{f}^{m+}, \boldsymbol{f}^{m-}]$. Then iterate $[\boldsymbol{f}^{m+}, \boldsymbol{f}^{m-}] = \alpha S \times [\boldsymbol{f}^{m+}, \boldsymbol{f}^{m-}] + (1-\alpha)[\boldsymbol{y}^+, \boldsymbol{y}^-]$ for T times, where α is a parameter in (0, 1).
5. Combine $\boldsymbol{f}^{m+}$ and $\boldsymbol{f}^{m-}$ as $\boldsymbol{f}^m = \beta\boldsymbol{f}^{m+} + \boldsymbol{f}^{m-}$ for output, where β is a positive weight.

The above manifold-ranking process is similar to the one adopted in [9]. There are several points that need to be addressed in the above process. The first one is that we adopt a sparse graph. This implementation significantly reduces computational and storage cost while retaining close performance [9]. Another issue is that we have chosen L_1 distance metric in the affine matrix $\mathbf{W}$. This is because experiments demonstrate that L_1 distance better approximates the perceptual difference for many visual features [9]. In step 5 we output results as $\boldsymbol{f}^m = \beta\boldsymbol{f}^{m+} + \boldsymbol{f}^{m-}$. This is because positive samples are always more scarce and compact than negative samples, and typically they contribute more in concept learning [9]. Thus we set a weight β to modulate the effect of positive samples and typically we set $\beta > 1$.

2.2. Sample Selection Strategy

Firstly we consider the sample selection problem for individual modality. We propose an approach based on three criteria: *informativeness*, *density*, and *diversity*. Then we present our multimodality sample selection strategy, where the numbers of selected samples for different modalities are adapted according to their performance variations, so that different potentials of multiple modalities are taken into account.

2.1.1. Informativeness

This sampling criterion aims at selecting the unlabeled samples that can add most information to the current model. Generally, the most "uncertain" samples in the classification process are selected, such as the samples near the hyperplane in SVM [13]. However, in terms of video annotation, as for most concepts the positive samples are much less than negative ones, then labeling a positive sample has much larger effect than a negative one. In this case we should select the samples that are more likely to be positive, similar to the relevance feedback process in CBIR [12]. To make a trade-off between these two criteria, we build a linear combination of them, where the weights are decided by the *frequency* measure of the concept. We define the *frequency* measure of a concept as the percentage of positive samples in labeled training set, i.e.,

$$frequency = \frac{\sum y_i^+}{l} \tag{2}$$

We limit the *frequency* measure to [0, 0.5], and then define the *informativeness* measure of unlabeled sample x_i as follows

$$\begin{aligned} &informativeness(x_i^m) = \\ &2 \times frequency \times (1 - |f^m(x_i)|) + (1 - 2 \times frequency) \times f^{m+}(x_i) \end{aligned} \tag{3}$$

If *frequency* is near 0, i.e., the positive samples are very scarce, so that they contribute much more than negative samples in concept learning. Then $f^{m+}(\mathbf{x}_i)$ has a weight near 1, and the samples more likely to be positive are selected for manually annotation. If *frequency* is large, i.e., the positive and negative samples are balanced (such as several frequent concepts: *indoor*, *outdoor*, *people*, *face*, etc), then $1-|f^m(\mathbf{x}_i)|$ has larger weight and the samples closer to classification boundary are selected.

2.1.2. Density

Prior works indicate that prior density distribution $p(x)$ can be utilized in active learning. Cohn et al. have demonstrated its usefulness in theory [7]. Wu et al. define a *representativeness* measure for each sample according to its distance to nearby samples, and take it as a criterion of sample selection [15]. Zhang et al. estimate data distribution $p(x)$ by Kernel Density Estimation (KDE) [11], and then take it into account in sample selection [16].

Here we define *density* measure based on KDE, by which $p(\mathbf{x})$ can be estimated as follows

$$p(x^m) = \frac{1}{n} \sum_{x_i \in L \cup U} K(x^m - x_i^m) \tag{4}$$

where $K(x)$ is a kernel function, which satisfies $K(x) > 0$ and $\int K(x)dx = 1$. We use exponential kernel (i.e., $K(x)=\exp(-|x|/2\sigma)$,

and for each estimated point we only consider the nearby samples. Then according to the definitions in Section 3.1 we have

$$p(x_i^m) = \frac{1}{n}\sum_{j=1}^{n} W_{ij}^m \tag{5}$$

Consequently we define *density* measure by normalizing $p(x)$ to [0, 1] as follows

$$Density(x_i^m) = \frac{\sum_{j=1}^{n} W_{ij}^m}{\max_i \sum_{j=1}^{n} W_{ij}^m} \tag{6}$$

2.1.3. Diversity

Previous studies demonstrate that the selected samples should be diversified [3, 14]. Thus we define a *diversity* measure for unlabeled data similar to these existing works and incorporate it into our sample selection strategy. Given kernel K, the angle between two samples x_i^m and x_i^m is defined as

$$\cos(< x_i^m, x_j^m >) = \frac{\left|K(x_i^m, x_j^m)\right|}{\sqrt{K(x_i^m, x_i^m)K(x_j^m, x_j^m)}} \tag{7}$$

We adopt exponential kernel again and ignore the faraway sample pairs, thus it is easy to derive that the *diversity* measure for sample x_i can be defined as

$$Diversity(x_i^m) = 1 - \max_{x_j \in L} W_{ij}^m \tag{8}$$

2.1.4. Multimodality Sample Selection

In this sub-section we discuss our multimodality sample selection strategy based on the above three criteria. According to theoretical analysis [7], it is more rational to use *density* measure as a weight of *informativeness* measure than linearly combing them. So here we weight *informativeness* measure by *density*, and then linearly combine them with *diversity* measure, i.e.,

$$\begin{aligned} &effectiveness(x_i^m) = \\ &\gamma \times density(x_i^m) \times informativeness(x_i^m) + (1-\gamma) \times diversity(x_i^m) \end{aligned} \tag{9}$$

Experiments will demonstrate that this strategy is more effective than building a linear combination of all three criteria.

Up to now we have addressed sample selection criteria for individual modality. A remained problem is how to select samples for multiple modalities.

In [5], Chen et al. analyze that samples should be selected for each individual modality to keep the specificity of multiple feature sets, so that the selected samples will not be constrained in a more and more restricted area. Thus they select equal number of samples for each modality. Although this method has shown appealing performance, it ignores the different potentials of multiple modalities. For several modalities that are not discriminative enough for the giving concept, it will achieve a "saturation" state after several active learning iterations, i.e., selecting more samples for this modality can hardly improve its corresponding sub-model. In this case, we should select more samples for other modalities. Thus we construct our strategy based on a notion of *performance gain*. For each modality, we define its performance gain as its performance variation between the latest two learning iterations, which can be estimated from the latest two selected sample batches as follows

$$\Delta perf^m = \begin{cases} 0, & if\ \ performance^m(t) < performance^m(t-1) \\ performance^m(t) - performance^m(t-1), & else \end{cases} \tag{10}$$

where t is active learning iteration index. Then we let the numbers of selected samples be proportional to the performance gains of multiple modalities, i.e.,

$$h^m = \frac{\Delta perf^m}{\sum_{m=1}^{M} \Delta perf^m} \times h \tag{11}$$

The above strategy is based on an assumption that the performance gain of a modality is larger if the modality is further from saturation. If a modality has a large performance gain, then more samples are selected for this modality in the next iteration; otherwise, if a modality achieves saturation state, then few samples are selected for it. Experiments demonstrate that this adaptive approach outperforms selecting fixed number of samples for each modality.

3. EXPERIMENTS

To evaluate the performance of the proposed framework, we conduct experiments which follow the guideline of TRECVID 2005 high-level feature extraction task. TRECVID 2005 dataset consists of 273 news videos and is about 160 hours in duration [1]. The dataset is split into a development set and a test set. The development videos are segmented into 49532 shots and 61901 sub-shots, and the test videos are segmented into 45766 shots and 64256 sub-shots. A key-frame is selected for each sub-shot, and from the key-frame we extract the following six feature sets: (1) block-wise color moment based on 5 by 5 division of the image (225D); (2) HSV correlogram (144D); (3) HSV histogram (64D); (4) wavelet texture (128D); (5) co-occurrence texture (16D); and (6) lay-out edge distribution histogram (75D). Each feature set is regarded as a modality, and thus we obtain six modalities.

We take sub-shot as the unit for interactive annotation, and limit the selected samples in the development set, so that the test set is only used for performance evaluation. After several active learning iterations, we fuse the results on the test set from sub-shots to shots by *max* aggregation, and then evaluate *average precision* of the first 2000 shots, which follows the guideline of TRECVID benchmark [2]. In this way, we make our results comparable with reported results for the TRECVID task [10]. In experiments, we set parameters α, β, K, and T in manifold-ranking process to 0.9, 10, 25, and 20, respectively (see Section 2.1). We set parameter γ in Eq. (9) to 0.7.

Firstly, we conduct several experiments with regarding full development set as training data. We compare the following three methods: (1) SVM + single modality (SVM+SM), i.e., SVM with all features concatenating into a large vector; (2) manifold-ranking + single modality (MR+SM); and (3) manifold-ranking + multimodality (MR+MM), i.e., conduct manifold-ranking for each modality and then fuse the results. We illustrate the experimental results in Table 1. From the results we can see that MR+MM remarkably outperforms SVM+SM and MR+SM, which indicates that multimodality fusion is critical to annotation performance. The performance of MR+MM is comparable to the reported leading results in the TRECVID task [10].

Table 1. Experimental results with full development set

Concept	SVM+SM	MR+SM	MR+MM
Walking_Running	0.206	0.204	0.221
Explosion_Fire	0.087	0.069	0.070
Maps	0.455	0.472	0.473
Flag-US	0.097	0.092	0.114
Building	0.483	0.449	0.461
Waterscape_Waterfront	0.382	0.367	0.396
Mountain	0.319	0.354	0.381
Prisoner	0.0003	0.002	0.0041
Sports	0.402	0.383	0.405
Car	0.314	0.265	0.286
MAP	0.274	0.266	**0.281**

Then we conduct experiments to demonstrate the effectiveness of our active learning strategy. We compare our approach with the following four different schemes:

Scheme 1:

Integrate a global *effectiveness* measure as $effectiveness(x_i) = \sum\{\Delta perf^m \times effectiveness(x_i^m)\}$, and then select h samples according to this measure.

Scheme 2:

Select h/M samples for each modality, i.e., the method proposed in [5].

Scheme 3:

Replace Eq. (9) by defining effectiveness measure as a linear combination of *informativeness*, *density*, and *diversity* measures, where the weights are set to 0.4, 0.3, and 0.3 respectively.

Scheme 4:

Randomly select samples.

We set h = 500 (i.e., select 500 samples in each iteration). The results are illustrated in Fig. 2. From the figure we can see that the proposed approach remarkably outperforms the other four schemes. It can achieve about twice MAP compared with random sample selection, and obtains a comparable performance with MR+SM on full development set when 15% of the original development set are labeled (about 10000 samples).

According to the description in Section 2, it is easy to derive that the computational cost of each active learning iteration scales as $O(M \times T \times K \times n)$, where M is the number of modalities, T is the propagation time in manifold-ranking process, K is the neighborhood size, and M is the number of modality. In practical experiments the response time for interactive annotation is about 15 seconds (Pentium 1.8G Hz, 512M RAM). Thus the proposed framework is more efficient and practical compared with previously proposed methods (such as SVM + active learning).

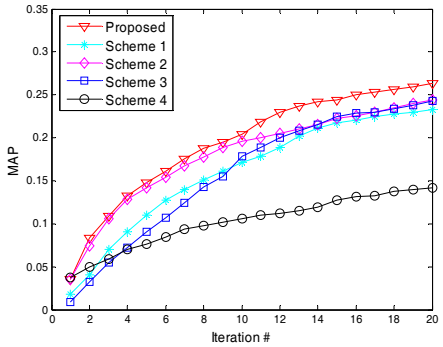

Figure 2. Active learning performances with different strategies

4. CONCLUSION AND FUTURE WORK

In this paper we proposed an efficient interactive video annotation framework based on semi-supervised learning and active learning. In the proposed framework, semi-supervised learning is conducted for each modality, and a novel sample selection strategy is adopted with multiple modalities. The experiments on TRECVID dataset have shown promising results.

In this study we only consider annotating different concepts independently. However, in practice it may be more efficient to annotate multiple concepts simultaneously [8]. In the future we will take simultaneous annotation of multiple concepts into account in our framework.

5. ACKNOWLEDGMENTS

This work is partly supported by NSFC under contract No. 60632040.

6. REFERENCES

[1] TRECVID: TREC Video Retrieval Evaluation, http://www-nlpir.nist.gov/projects/trecvid

[2] TREC-10 Proceedings appendix on common evaluation measures, http://trec.nist.gov/pubs/trec10/appendices/measures.pdf

[3] Klaus Brinker. Incorporating diversity in active learning with support vector machines. *ICML*, 2003

[4] C.-Y. Chen, B. L. Tseng and J. R. Smith. Video collaborative Annotation forum: establishing groundtruth-labels on large multimedia datasets. *TRECVID 2003 Workshop Notebook Papers*, 2005.

[5] M. Chen, M. Christel, A. Hauptmann and H. Wactlar, Putting active learning into multimedia applications: dynamic definition and refinement of concept classifiers. *ACM Multimedia*, 2005.

[6] M. Chen and A. Hauptmann. Active learning in multiple modalities for semantic feature extraction from video. *AAAI workshop on learning in computer vision*, 2005.

[7] D. A. Cohen, Z. Ghahramani and M. I. Jordan. Active learning with statistical models. *JAIR*, 1996.

[8] M. Naphade and J. R. Smith. Active learning for simultaneous annotation of multiple binary semantic concepts. *ICIP*, 2004.

[9] J. He, M. Li, H.-J Zhang, H. Tong and C. Zhang. Manifold-Ranking based image retrieval. *ACM Multimedia*, 2004.

[10] P. Over, T. Ianeva, W. Kraaij and A. F. Smeaton. TRECVID 2005: An Overview. *TRECVID Workshop*, 2005.

[11] E. Parzen. On the estimation of a probability density function and the mode, *Annals of Mathematical Statistics*, 1962

[12] Y. Rui, T. S. Huang, M. Ortega and S. Mehrotra. Relevance feedback: a power tool for interactive content-based image retrieval. *IEEE trans. CSVT*, 1998.

[13] S. Tong and E. Y. Chang. Support vector machine active learning for image retrieval. *ACM Multimedia*, 2001.

[14] Y. Wu, E. Y. Chang, K. C.-C. Chang and J. R. Smith. Optimal multimedia fusion for multimedia data analysis. *ACM Multimedia*, 2004.

[15] Y. Wu, I. Kozintsev, J.-Y Bouguet and C. Dulong. Sampling strategies for active learning in personal photo retrieval. *ICME*, 2006.

[16] C. Zhang and T. Chen. Annotating retrieval database with active learning. *ICIP*, 2003.

[17] D. Zhou, O. Bousquet, T. N. Lal, J. Weston and B. Scholkopf. Learning with local and global consistency. *NIPS*, 2004.