Semantic Video Classification by Integrating Unlabeled Samples for Classifier Training

Jianping Fan
Dept of Computer Science
UNC-Charlotte
Charlotte, NC 28223, USA
jfan@uncc.edu

Hangzai Luo
Dept of Computer Science
UNC-Charlotte
Charlotte, NC 28223, USA
hluo@uncc.edu

ABSTRACT

Semantic video classification has become an active research topic to enable more effective video retrieval and knowledge discovery from large-scale video databases. However, most existing techniques for classifier training require a large number of hand-labeled samples to learn correctly. To address this problem, we have proposed a semi-supervised framework to achieve incremental classifier training by integrating a limited number of labeled samples with a large number of unlabeled samples. Specifically, this semi-supervised framework includes: (a) Modeling the semantic video concepts by using the finite mixture models to approximate the class distributions of the relevant salient objects; (b) Developing an ***adaptive EM algorithm*** to integrate the unlabeled samples to achieve parameter estimation and model selection simultaneously; The experimental results in a certain domain of medical videos are also provided.

Categories and Subject Descriptors
I.5.1 [**Pattern Recognition**]: Models - *statistical*; I.2.6 [**Artificial Intelligence**]: Learning - *concept learning*

General Terms
Algorithm, Experimentation

Keywords: Semantic video classification, finite mixture models, unlabeled samples, adaptive EM algorithm.

1. INTRODUCTION

The explosive growth of video sources has created an urgent need of video indexing and retrieval at the semantic level, semantic video classification is a promising solution to this challenging problem [1]. The statistical approaches can support more effective semantic video classification by discovering both the obvious video making rules and the hidden correlations among different video patterns [3]. The major difficulty with the statistical approaches is that a large number of hand-labeled training samples are required to learn correctly [2]. This problem is becoming critical for some video domains such as medical videos because of our limitation of domain knowledge and lack of available labeled video resources, and thus expensive medical expertise is required to derive the semantic labels for a large number of training samples. Given this costly labeling problem, it is very attractive to design the classifier training techniques that can take advantage of unlabeled samples.

2. SEMANTIC CONCEPT MODELING AND VIDEO CLASSIFICATION

The performance of the semantic video classifiers largely depends on two key issues: (1) The effectiveness of video patterns for video content representation and feature extraction; (2) The significance of the algorithms for semantic video concept modeling and classifier training. To enhance the quality of features on discriminating among different semantic video concepts, we used the salient objects for video content representation and feature extraction [4].

To interpret the contextual relationship between a certain semantic video concept C_j and its relevant salient objects, the class distribution of the relevant salient objects is approximated by using a finite mixture model with κ_j mixture components:

$$P(X, C_j, \Theta_{c_j}) = \sum_{i=1}^{\kappa_j} \omega_{s_i} P(X, C_j | S_i, \theta_{s_i}) \qquad (1)$$

where κ_j indicates the optimal number of mixture components, $\Theta_{c_j} = \{\kappa_j, \omega_{c_j}, \theta_{s_i}, i = 1, \cdots, \kappa_j\}$ is the set of the parameters for these mixture components. To achieve more effective classifier training, the central objective of this paper is to automatically acquire the optimal model structure κ and model parameters ω, θ simultaneously by integrating a limited number of labeled samples with a large number of unlabeled samples. We have proposed an ***adaptive EM algorithm*** to achieve model selection and parameter estimation simultaneously [3].

After the weak classifiers are learned by using a limited number of labeled samples, we use the Bayesian framework to achieve a "soft" classification of the unlabeled video clips, thus the confidence score for an unlabeled video clip (i.e., unlabeled sample) $\{X_l, S_l\}$ to be classified into one certain semantic video concept can be defined as:

$$\psi(X_l, S_l) = \sqrt{\psi_\alpha(X_l, S_l)\psi_\beta(X_l, S_l)} \qquad (2)$$

where $\psi_\alpha(X_l, S_l) = max\{P(C_j|X_l, \Theta_{c_j})\}$ is the maximum

SIGIR'04 July 25-29, 2004, Sheffield, South Yorkshire, UK

posterior probability for the unlabeled sample $\{X_l, S_l\}$, $\psi_\beta(X_l, S_l) = \psi_\alpha(X_l, S_l) - max\{P(C_j|X_l, \Theta_{c_j})|P(C_j|X_l, \Theta_{c_j}) \neq \psi_\alpha(X_l, S_l)\}$ is the multi-concept margin for the unlabeled sample $\{X_l, S_l\}$.

Based on their confidence scores, the unlabeled samples can be categorized into two groups: (a) *known classes of existing semantic video concepts*; (b) *uncertain samples*. By adding the unlabeled samples with high confidence scores for incremental classifier training, the confidence scores for the uncertain samples can be updated over time. Thus the uncertain samples can further be categorized into two groups according to their updated confidence scores: (1) The uncertain samples that come from the *unknown classes of existing semantic video concepts*; (2) The uncertain samples that come from the *uncertain concepts*.

The uncertain samples with a significant change of confidence scores come from the unknown classes of existing semantic video concepts, and thus they should be included for incremental classifier training because they can provide some additional video contexts to achieve more accurate modeling of the semantic video concepts. On the other hand, the uncertain samples without a significant change of confidence scores may come from the uncertain concepts, and thus they should be eliminated from the training set. By performing the merging, splitting and death operations automatically, our adaptive EM algorithm has the following advantages [4]: (a) It does not require a careful initialization of the model structure by starting with a reasonably large number of the mixture components, and the model parameters are initialized directly by using the labeled samples; (b) It is able to take advantage of the negative samples to achieve discriminative classifier training; (c) It is able to escape the local extrema and enable a global solution by re-organizing the distributions of the mixture components.

Given a test video clip V_i, the salient objects are first detected automatically by using our detection functions. It is important to note that one certain test video clip may consist of multiple salient objects. Thus the test video clip V_i can be described by using the relevant salient objects, $V_i = \{S_1, \cdots, S_l, \cdots, S_n\}$. The class distribution of the relevant salient objects $V_i = \{S_1, \cdots, S_l, \cdots, S_n\}$ is then modeled as a finite mixture model $P(X, C_j, \Theta_{c_j})$ [3]. Finally, the test video clip V_i is assigned to the best matching semantic video concept C_j that corresponds to the maximum posterior probability. Some semantic video classification results are given in Fig. 1.

Given the limited sizes of the labeled training samples, we have tested the performance differences of our classifier training algorithm by using different sizes of unlabeled samples. The average performance differences are given in Fig. 2 for some semantic video concepts. One can find that the unlabeled samples can improve the classifier's performance significantly when only a limited number of labeled samples are available for classifier training. The reasons are that a limited number of labeled samples cannot interpret the necessary video contexts for semantic video concept interpretation and the unlabeled samples have the capability to provide additional video contexts to learn the finite mixture models more correctly.

3. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a novel framework to achieve more effective classifier training by integrating a large number of unlabeled samples with a limited number of labeled samples. Integrating the unlabeled samples for classifier training not only dramatically reduces the cost for labeling sufficient samples required for accurate classifier training but also increases the classifier accuracy significantly. It is worth noting that the proposed semantic video classification techniques can also be used for other video domains when the labeled training samples are available.

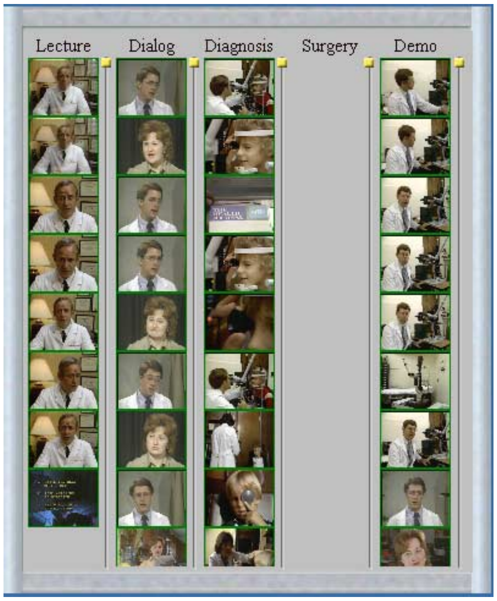

Figure 1: The semantic video classification results for a 1.5 hours medical video clip.

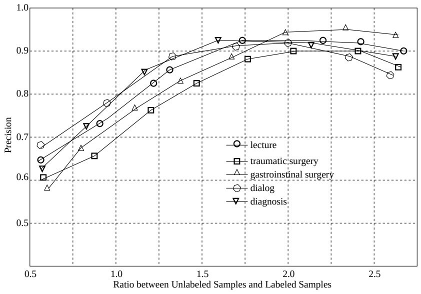

Figure 2: The relationship between the classifier performance (i.e., precision) and the percentage of the unlabeled samples for classifier training.

This work is supported by National Science Foundation under IIS-0208539. The online demo for this work can be found at: **http://www.cs.uncc.edu/~jfan/project4.html**.

4. REFERENCES

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

[2] K. Nigam, A. McCallum, S. Thrun, T. Mitchell, "Text classification from labeled and unlabeled documents using EM", *Machine Learning*, vol.39, no.2, 2000.

[3] J. Fan, H. Luo, A.K. Elmagarmid, "Concept-Oriented indexing of video databases towards more efficient retrieval and browsing", *IEEE Trans. on Image Processing*, vol.13, no.6, 2004.

[4] J. Fan, Y. Gao, H. Luo, G. Xu, "Automatic image annotation by using concept-sensitive salient objects for image content representation", SIGIR, 2004.