AUTOMATIC VIDEO GENRE CATEGORIZATION USING HIERARCHICAL SVM

Xun Yuan †, *Wei Lai* ‡, *Tao Mei* †, *Xian-Sheng Hua* ‡, *Xiu-Qing Wu* †, *Shipeng Li* ‡

† Dept. of EEIS, University of Science and Technology of China, Hefei 230027, P. R. China
‡ Microsoft Research Asia, 5F Sigma Center, 49 Zhichun Road, Beijing 100080, P. R. China
{yuanxun; meit}@mail.ustc.edu.cn; {weilai; xshua; spli}@microsoft.com

ABSTRACT

This paper presents an automatic video genre categorization scheme based on the hierarchical ontology on video genres. Ten computable spatio-temporal features are extracted to distinguish the different genres using a hierarchical Support Vector Machines (SVM) classifier built by cross-validation, which consists of a series of SVM classifiers united in a binary-tree form. As the order and genre partition strategy of the SVM classifier series affect the over performance of the united classifier, two optimal SVM binary trees, local and global, are constructed aiming at finding the best categorization orders, i.e., the best tree structure, of the genre ontology. Experimental results show that the proposed scheme outperforms C4.5 Decision Tree, typical 1-vs-1 SVM scheme, as well as the hierarchical SVM built by K-means.

Index Terms— Pattern Classification, Video Signal Processing

1. INTRODUCTION

With the proliferation of digital camcorder, HDTV and digital television, a rapid increasing number of digital videos are available to average users. Automatic categorization of mass videos into various genres and sub-genres is helpful for video management, searching, and browsing.

Many works have already been performed on this topic. In [2], a C4.5 decision tree is applied to build the classifier for video genre labeling, using a 10-dimension feature vector consisting of editing rules, motion and color features. An analogous work in [3] utilizes different sets of features and categorizes videos into more genres, where C4.5 decision tree is applied as well. The method in [7] extracts a set of features from the basic statistics on color, motion and audio, and then develops different independent classification modules to categorize five video genres. In [5], the well-established rules of film are analyzed and film previews are discriminated into four categories using mean shift classification. Camera motion parameters are employed to identify six major types of sports in [6].

This work was performed when the first and third authors were research interns at Microsoft Research Asia.

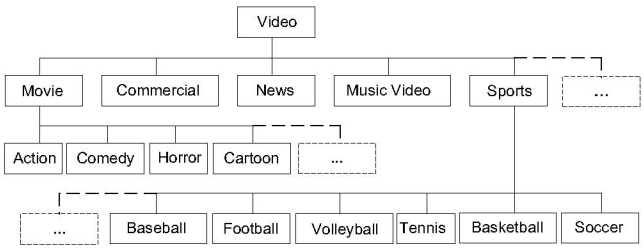

Fig. 1. Video genre hierarchical ontology.

Although several works have been performed, how to comprehensively categorize videos into different genres still remains a challenging issue. It is observed that most of the previous works only categorize videos based on a relatively incomplete ontology [2][3][7], or only focus on one kind of video genre such as sports [6] or film [5]. Based on previous literatures about video classification and the general video categories we typically applied in daily life, in this paper we attempt to propose a much more comprehensive hierarchical ontology for video genres and apply it to a much larger amount of video data. Moreover, in contrast to [2] and [3] which use C4.5 Decision Tree for video categorization, we will build a novel hierarchical classification scheme which can achieve better performance, regarding the hierarchical ontology on video genres proposed in this paper. As shown in Figure 1, videos are firstly divided into five basic genres: *news*, *movie*, *sports*, *commercials* and *music video*. *Movie* is then divided into *cartoon*, *action*, *comedy* and *horror*; and *sports* are then divided into *football*, *baseball*, *basketball*, *soccer*, *tennis* and *volleyball*. This hierarchical ontology is extensible and the majority of common videos can be loosely classified accordingly.

The rest of this paper is organized as follows. Section 2 analyzes the visual features for genre categorization. Section 3 introduces the hierarchical local and global SVM. Experimental results are shown in Section 4, followed by concluding remarks and future works in Section 5.

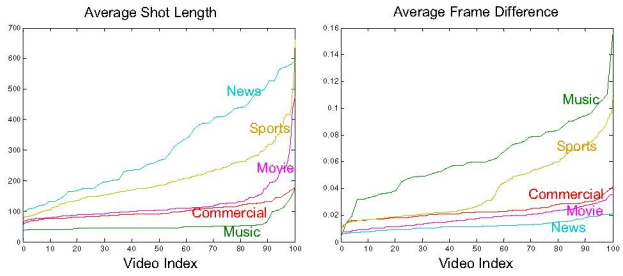

Fig. 2. Average shot length and average frame difference.

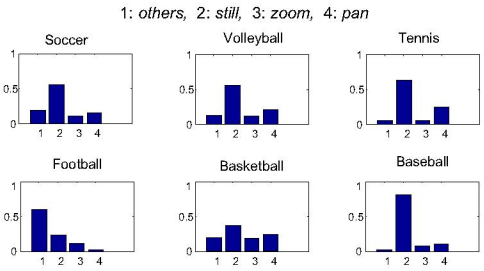

Fig. 3. Camera motion in the six sports genres.

2. SPATIO-TEMPORAL FEATURE ANALYSIS

Features from different modalities, such as surrounding text, transcript, audio etc., may help discriminate different video genres. In this paper, we focus on visual features only, while combining textual and aural features into this framework is our future work. Our feature selection and extraction are based on the following three considerations: firstly, the features should be as more comprehensive as possible to reflect high level semantics well; secondly, different features should be independent; lastly, the computation cost should be relatively low. To be exact, two types of features, including temporal features (average shot length, cut percentage, average color difference and camera motion) and spatial features (face frames ratio, average brightness and average color entropy) are applied in our work, which form a 10-dimensional feature vector.

2.1. Temporal Features

Temporal features include average shot length (1D), cut percentage (1D), average color difference (1D) and camera motion (4D). Here we apply the shot boundary detector in [9], which finds two types of shot boundaries, i.e. cut and gradual transition. The average shot length is computed by averaging all the shot lengths in a video. Additionally, we calculate the ratio of cut transition to the overall shot boundaries as a complementary temporal feature. Average color difference is computed by averaging all the color histogram differences over the whole video. The difference of two consecutive color histograms is defined as:

$$1 - \frac{\sum_{i=1}^{N} \min(H_j(i), H_{j-1}(i))}{\sum_{i=1}^{N} H_j(i)} \qquad (1)$$

where N is the number of color histogram bins, and H_j and H_{j-1} are the color histograms of frame j and $j-1$, respectively. Figure 2 shows the curves of the average shot length and average frame difference for the five different video genres. We extract these temporal features from 100 video clips for each genre and plot the curves by arranging their values in an ascending order. From the figure we can find several differences between genres in terms of such temporal features, such as generally *music video* has relatively larger frame difference and shorter shot length than the others because of its frequent change of content and flash light, and *sports* have the relatively long average shot length.

Camera motion is a distinctive feature for video genres, especially for *sports*. For example, both the "camera motion extraction ratio" and "camera motion transition" are taken into account for *sports* categorization in [6]. We classify the camera motion of each shot into four types: *still*, *pan*, *zoom*, and *others* [10]. We adopt the average extraction ratios of these four camera motion types to compose a 4D vector. It is observed that these 4D motion features are discriminative for the six sports sub-genres (refer to Figure 1). Figure 3 shows an example of camera motion vector in six sports sub-genres. It can be seen that the *still* motion in *soccer* and *baseball* is the predominant component, and *pan* is mostly frequent in *volleyball* and *basketball*, while others occupy the most in *football*.

2.2. Spatial Features

Spatial features consist of face frames ratio (1D), average brightness (1D) and average color entropy (1D). Although the temporal features of *news* and *movie* exhibit similar, they can be easily distinguished by face-related features as *news* usually contain many anchor person shots. We define the face frame ratio to be the ratio between the number of frames containing faces to the overall number of frames in a video. In addition, there are two other spatial features, i.e. average brightness and average color entropy. We calculate the mean μ and variance σ^2 of pixel values above a pre-defined threshold in HSV space. As a result, the average brightness for i-th frame is defined as [6] $B = \mu_i \times \sigma_i^2$. Color entropy is a reflection of the color uniformity, which is computed by multiplying the hue entropy and value entropy in HSV space. All the spatial features are detected in a fixed interval of 15 frames and averaged over the whole video.

Figure 4 shows the curves of average face percentage and brightness for the five video genres. From the figure, we can see that *news* has the highest face percentage, brightness is higher in *sports* and *news* while it has the lowest value in *horror*, and *basketball* has much higher entropy than *soccer*.

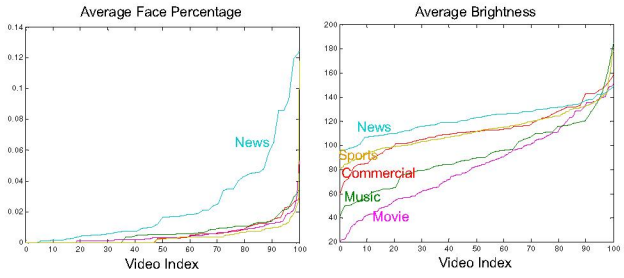

Fig. 4. Average face percentage and average brightness.

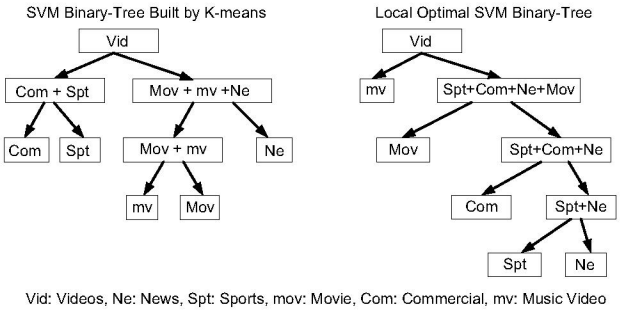

Fig. 5. SVM binary tree for our video categorization task.

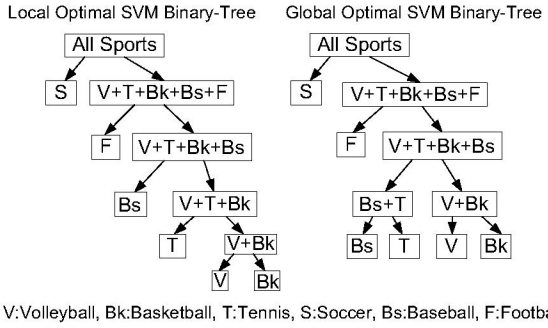

Fig. 6. Local and global optimal SVM binary-tree for sports categorization.

3. HIERARCHICAL SVM BINARY-TREE

SVM has been widely acknowledged for its strong theoretical foundations and good generalization capability [1]. Since it was originally designed for binary classification, there are various strategies currently to apply it in multi-class problem. These strategies can be mainly categorized into two types: methods that consider all classes at once and methods based on binary classifications (e.g. 1-vs.-1, 1-vs.-all, and DDAG) [4]. However, how to efficiently and effectively extending SVM for multi-class classification still remains an open research issue. On the other hand, in our video genre classification task, some genres exhibit similar in terms of features, while others exhibit distinctive. So a straightforward approach is to create a binary tree according to the feature characteristics between different genres, where each node in this tree represents two set of distinguished classes.

Motivated by the above analysis, we propose a novel hierarchical SVM binary-tree in this paper to deal with the multi-class problem, which is automatically and dynamically built up from the training set. Unlike the typical hierarchical binary-trees that use K-means [8] or C4.5 Decision Tree [2][3] to partition the feature space into two individual classes independently at each node in original feature space, our hierarchical SVM binary-tree is able to locally find the best separation at each node, or globally find the best order and structure of the whole tree. Therefore, two types of SVM binary-tree forms, called local and global, are proposed for our video genre categorization task.

3.1. Local Optimal SVM Binary-Tree

Liu *et al.* [8] proposed a method to build a hierarchical multi-class SVM binary tree, where K-means is employed to separate classes into two clusters at each node, and a SVM classifier is trained on these two clusters. This procedure is repeated until a binary tree is finally created. However, similar to C4.5 Decision Tree, K-means algorithm only clusters the features in the original feature space instead of separating the two clusters using the best separation boundary. Therefore, there is usually a classification accuracy decrease. Our local optimal SVM binary-tree attempts to find the best separation at each node using 2-fold cross-validation. The cross-validation process divides the training data into two parts: one is used to train a SVM classifier and the other is used to test the obtained classifier; then the two parts exchange. All possible separations at each node are tested using this cross-validation method, and the one with the highest accuracy is chosen as the separation at this node. This process is repeated to dynamically create a hierarchical SVM structure. The hierarchical SVM built by K-means and cross-validation for the first level video genres is shown in Figure 5. The different structures of these two schemes indicate that the separation obtained by K-means in the original feature space does not correspond to the optimal separation. The disadvantage of proposed method is that the training time increases. At each node there are N classes, so there are $(2^{N-1}-1)$ separations to be tested. However, when N is not very large, the training time is acceptable.

3.2. Global Optimal SVM Binary-Tree

The local optimal SVM binary-tree aims at finding the best separation at each node locally, instead of the optimal "order" or "structure" of the separation at all the nodes in the whole tree globally. Thus the structure of local optimal binary-tree may not be global optimal. To tackle this issue, we conduct a global optimization of the SVM binary-tree as well. Unlike the aforementioned local optimal scheme that performs cross-validation at each single node, the global scheme per-

Table 1. The comparison of precisions of the five classification schemes in the three classification tasks

Classifier	Video Categorization			Sports Categorization			Movie Categorization		
	Best%	Average%	Worst%	Best%	Average%	Worst%	Best%	Average%	Worst%
(1) C4.5 Decision Tree [2][3]	85.09	83.72	81.90	86.59	82.03	78.62	71.72	65.93	62.07
(2) Typical 1-vs.-1 and voting SVM [4]	88.00	86.02	84.33	97.06	93.09	90.44	78.38	73.49	70.42
(3) Hierarchical SVM built by K-means [8]	85.09	84.52	83.25	93.84	89.86	85.95	78.62	75.79	71.72
(4a) Local Optimal SVM Binary-Tree	**88.61**	**86.67**	**84.75**	**95.60**	**93.91**	**92.75**	**81.38**	**76.10**	**73.10**
(4b) Global Optimal SVM Binary-Tree	**87.77**	**86.97**	**85.59**	**96.74**	**94.71**	**94.57**	**80.00**	**76.21**	**71.03**

forms cross-validation for each possible binary-tree, and then chooses the tree with the highest prediction accuracy.

According to experimental results, for the first-layer genre classification (see the five genres in the top layer of Figure 1), and *movie* categorization tasks, the global optimal binary-tree is the same as the one built up by local optimal binary-tree. While for *sports* categorization, the binary-trees built up by local and global optimization are different (as shown in Figure 6).

4. EXPERIMENTAL RESULTS

To validate the efficiency of the proposed hierarchical SVM, we collect more than 60 hours' videos of various genres from TV recordings. The videos are cut into short clips with the duration from 3 to 10 minutes. There are totally 600 clips (200 for *sports* and 100 for the other four genres in the first ontology level, respectively). In each genre, 50% of the clips are randomly selected for training and the rest for testing.

We implement four classification schemes for comparison: (1) C4.5 Decision Tree [2][3]; (2) typical 1-vs.-1 and voting multi-class SVM [4]; (3) Hierarchical SVM built by K-means [8]; and (4) Hierarchical SVM Binary-Trees (Local and Global) proposed in this paper. We perform three categorization tasks according to the hierarchical video ontology in Figure 1, i.e. video categorization in the first level, as well as sports and movie categorization in the second level. The experimental results are listed in Table 1. It can be seen that the three SVM classifiers (2, 3 and 4) outperform C4.5 Decision Tree (1) in all the three classification tasks. Specifically in sports and movie categorization, typical 1-vs.-1 scheme (2) achieves around 11% and 8% improvement in terms of average precision compared with C4.5 Decision Tree (1), respectively. In video categorization, the improvement is only 2.5%, the reason lies in that the feature differences between different video genres are more distinctive in the first level than those in the second level. Therefore, Decision Tree is able to separate the features in the original space well. While for the second level, mapping the features into a higher dimension space by SVM help better separate the different classes. The performances of the proposed hierarchical SVM approaches are higher than the other two SVM classifiers (2 and 3) in terms of average precision, while the global optimal one (4b) is slightly better than the local optimal one (4a), in all the three classification tasks.

5. CONCLUSION AND FUTURE WORK

In this paper, we have presented a novel method for automatic video genre categorization using spatial-temporal low-level features. We first define a hierarchical and relatively comprehensive ontology for video genres, and then propose a novel hierarchical SVM scheme for genre categorization, in which a series of SVM classifiers are dynamically built up in a binary-tree form and optimized locally or globally. The extensive experiments have indicated that it remarkably outperforms traditional C4.5 Decision Tree and two other SVM schemes. In future work, we will incorporate multimodal information into our framework, such as audio, transcripts and surround text.

6. REFERENCES

[1] V. Vapnik, "Statistical Learning Theory," *New York: Wiley*, 1998.

[2] B.-T. Truong, S. Venkatesh, and C. Dorai, "Automatic Genre Identification for Content-Based Video Categorization," in *Proceedings of ICPR*, pp. 1-10, 2000.

[3] Y. Yuan, Q.-B. Song, and J.-Y. Shen, "Automatic Video Classification Using Decision Tree Method," in *Proceedings of ICMLC*, pp. 1153-1157, 2002.

[4] C.-W. Hsu, C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. on Neural Networks*, vol. 13, pp. 415-425, 2002.

[5] Z. Rasheed, Y. Sheikh, and M. Shah, "On the use of computable features for film classification," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 15, pp. 52-64, 2005.

[6] S. Takagi, S. Hattori, K. Yokoyama, A. Kodate, and H. Tominaga, "Sports video categorizing method using camera motion parameters," in *Proceedings of ICME*, pp. 461-464, 2003.

[7] S. Fischer, R. Lienhart, and W. Effelsberg, "Automatic Recognition of Film Genres," in *Proceedings of ACM Multimedia*, pp. 295-304, 1995.

[8] S. Liu, H. Yi, L.-T. Chia, and D. Rajan, "Adaptive Hierarchical Multi-class SVM classifier for Texture-based Image Classification," in *Proceedings of ICME*, pp. 1190-1193, 2005.

[9] H.-J. Zhang, A. Kankanhalli, and S. W. Smoliar, "Automatic Partitioning of Full-Motion Video," *Multimedia System Journal*, vol. 1, pp. 10-28, 1993.

[10] D. Lan, Y.-F. Ma, and H.-J. Zhang, "A Novel Motion-Based Representation for Video Mining," in *Proceedings of ICME*, pp. 469-472, 2003.