Int J Multimed Info Retr (2013) 2:289–297
DOI 10.1007/s13735-013-0034-8

SHORT PAPER

Genre-specific modeling of visual features for efficient content based video shot classification and retrieval

Chiranjoy Chattopadhyay · Amit Kumar Maurya

Received: 25 December 2012 / Accepted: 1 February 2013 / Published online: 12 March 2013

Abstract This paper presents a genre-specific modeling strategy capable of improving the task of content based video classification and the speed of data retrieval operations. With the ever increasing growth of video data it is important to classify video shots into groups based on its content. For that reason, it is of primary concern to design systems that could automatically classify videos into different genres based on its content. We consider the genre recognition task as a classification problem. We use support vector machines to perform the classification task and propose an improved video classification method. The experimental results show that genre-specific modeling of features can significantly improve the performance. Results have been compared with two contemporary works on video classification, to demonstrate the superiority of our proposed framework.

Keywords Content-based video classification · Genre-specific modeling · Visual feature · Shape · Texture · Kinematics

1 Introduction

There is a rapid growth of the amount of multimedia data that is obtained from real-world multimedia sharing websites like Google video, Yahoo videos, Youtube, etc. Moreover, easy availability of video capturing devices (camcorders, smart phones) has also increased the production of video data by a significant amount. To search videos from a database, a straight forward approach is to perform a linear search, which takes a lot of time. It is important to categorize these huge amount of videos into different genres so that end users can search, choose or verify a desired video based on its content. The work presented in this paper aims at automating the task of content based classification of pre-segmented video shots into various genres, to bring down the retrieval time.

We have developed a genre-specific feature modeling strategy to address this automatic classification problem. Specifically, our system was designed to categorize videos into different genres [24], and facilitates fast retrieval of video shots. Although experiments reported in this paper only covers a few video genres, the system can be scaled up to handle other categories also. It has been found that features specific to a particular genre are sometimes more discriminative than other features, and if used judiciously, may lead to a robust classification framework. Using different combination of intrinsic low-level features can boost the performance of the classification task and thereby retrieval of video shots. Hence the appropriate representation of the potential information in the related features is crucial for video-content understanding. Though in some cases the audio or metadata can provide additional distinguishing information, they are either not readily available or can be confusing at times. Hence their utility is still limited. Therefore, in this paper, we only consider the visual information for the classification of various genres of video shots.

Recently, researchers have proposed techniques [11,30,31] for automatic video genre classifications. However, all these required a sufficient amount of metadata for satisfactory performance. In case of a content-based approach one do not have to worry about the manual tagging of video shots. Automatic content extraction will help in identification of genre-specific characteristics of video shots for proper categorization. Researchers have used low-level and

C. Chattopadhyay (✉) · A. K. Maurya
Indian Institute of Technology Madras, Chennai600036, India
e-mail: cchatto@cse.iitm.ac.in

A.K. Maurya
e-mail: peaceamit@gmail.com

high-level, task-specific [13,15] features, as well as a combination of them for content-based classification of video shots. In another approach [27] semantic aspects of a video genre, such as editing, motion and color distribution has been used as features and the decision tree algorithm was used to build the classifier. In [19] motion pattern (block motion estimation algorithm) from the compressed domain features has been used for video classification and retrieval. Support vector machines (SVM) have been used for sports video classification in [25]. Techniques for extraction of cuts, fades, motion, etc, lighting conditions of videos have been used in [22] for film classification. In [18] an automatic technique has been reported for sports video classification using shots length, facial close up shots, texture of human face as features. Very recently, in [29] genre-specific concept models were used for semantic video indexing. In [14] techniques for domain specific features for effective shot classification techniques are discussed. Researchers have also looked into the possibilities of exploiting features from multiple modality, viz. visual, audio, texts present in a video shot for video genre classification. In [3] both audio and text-based features are used for tagging and retrieving video shots. Video genres are identified using only audio information from TV shows in [26]. A novel method to identify the violent videos only with audio features is introduced in [17]. Very recently, in [16] SIFT features are extracted from the video and a BOVW (Bag of Visual words) approach has been used with SVM for video concept detection.

In our approach, we have adapted a genre-specific modeling of visual feature for video shot classification. Detection of shot boundary from a video stream is still an active area of research. Nevertheless, performance of video shot segmentation is not always satisfactory. This acts as a bottleneck for the performance of effective video classification task. Since the primary focus of this work is to find proper visual features for content based video classification, we have adapted a hierarchical approach to classify videos into different genres. To support our work we have organized our database into a dendogram (see Fig. 1), where each parent node represents a generalized class of its children. This type of database organization is important for a successful classification task. We have exploited the fact that video shots belonging to different genres manifest different discriminatory characteristics when compared to other genres and also different categories within that particular genre. For example, human beings and vehicles can be distinguished based on the shape characteristics. However, while trying to differentiate between human motion activities like running and walking, kinematic features are more relevant. Again, cartoon shots contain visual areas with high quality stock and seamless texture, and dominance of a particular set of colors. On the other hand, as compared to this a natural video contains varying, non-uniform textures and a relatively uniform distribution of colors in general. This motivates us to perform a genre-specific modeling of features, in which different models are trained with different features. Rational behind this approach is to capture the genre-specific semantics of different video shots. As compared to two recent works [14] and [31], our framework shows superior classification accuracy.

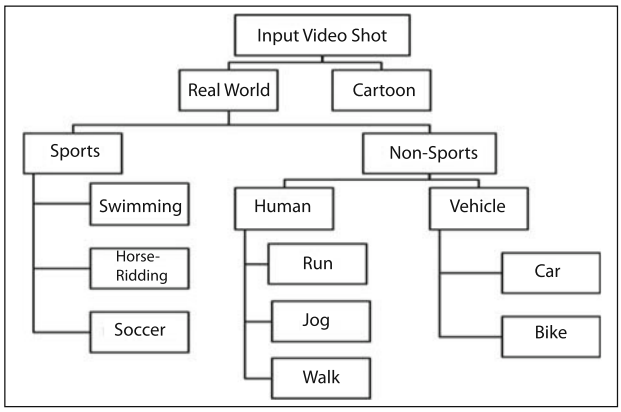

Fig. 1 Different categories of video shots

The rest of this paper is organized as follows. In Sect. 2, we give a brief account of the proposed methodology for content-based video categorization. In Sect. 3, we describe different categories of features used for the classification task. Video content modeling and classification strategies are discussed in Sect. 4. In Sect. 5 we describe the experimental results and provide comparative study with two existing works in literature. In Sect. 6, we provide a detailed explanation as to how the retrieval time of our system is significantly less as compared to a linear search system or systems which perform redundant feature computation. Finally, we provide conclusions in Sect. 7 and also suggest some ideas for future research in this area.

2 Brief description of the proposed method

Only limited groups of heterogeneous features distinguish certain semantics from others. Visual features constitute important cues to the human perception system so as to extract salient information from a video shot. The main focus of our system is to exploit the visual (both spatial and temporal) features present in a video shot and use it to categorize them into different genres based on their content. This categorization helps us in efficient retrieval of video shots from the database (gallery of video shots). This has been explained in Sect. 6. Figure 2 depicts the overall framework of our proposed method.

Usually video streams contains multiple events within it. All the frames within a single camera action are called a shot. Researchers have devoted considerable amount of effort to

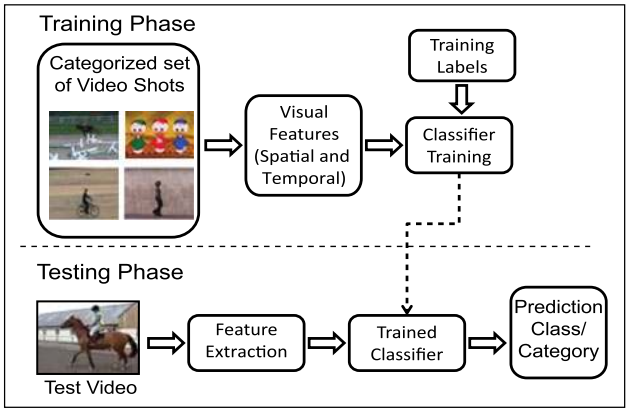

Fig. 2 Overall framework for proposed method

segment videos into shots. For example, suppose a person is driving a car and this situation has been filmed in such a way that the camera always follows the car. After some time the car stops and the man opens the door and comes out of the car and goes away. The camera stops as the car stops and then follows the person. The collection of frames that contains only the car constitutes a video shot where the object of interest is the car. Afterwards, the attention shifts to the person and that becomes another shot. In literature the term shot and scene has been used interchangeably. Since our work is concentrated on classification of video shots, we assume that videos of longer duration are already segmented into shots of relatively small duration (approx. 5–10 s). This assumption is very much pertinent because of the fact that an user essentially searches for a particular event from a gallery of video shots. Therefore, grouping videos at the shot level will give improved performance at the time of search, as compared to grouping the actual video stream.

In our proposed framework, we have selected features which are best suited for classifying between two given genres of videos and trained SVM [7] for classification purpose using those features. We have adopted a hierarchical approach to classify the video shots into different genres based on their content. At first we categorize the video shots into coarser groups (e.g., real world vs. cartoon). Later, at a lower level of hierarchy we classify them into finer categories (e.g., videos of vehicle category are further classified into videos containing car or bikes, etc). To perform this task we have arranged our database into a hierarchical structure (dendogram). A parent node in this tree denotes a super-category and child nodes depicts the sub-categories. In this work we have used SVM as classifiers due to its strong theoretical basis and generalization properties. Section 4 discusses more on the classifier organization. As discussed earlier, we have focused only on the visual features (both spatial and temporal) for the classification task. We have empirically determined the feature(s), which have shown enough discriminatory properties between two classes and used them to train our classifiers. We have compared our result with a very recent work on video categorization [14] and got an improved result in terms of classification accuracy. Next section describes the features used for classification in details.

3 Feature extraction

Features used for describing the content play a pivotal role in the overall success of any classification task. In this paper we have focused only on the visual features, both spatial (color, texture, shape) and temporal (motion kinematics) for the classification task. Following subsections describe the features used in our framework and their significance.

3.1 Spatial feature descriptors

We used three different low-level spatial features, which represent color, shape of the segmented foreground object and texture information in the video. Following subsections give details of the feature computation process.

3.1.1 Color descriptor

Usually the color images are converted to gray scale for computational reasons and also interest in the intensity values of the pixels in the given image. In case of recognition based on other contexts such as shape or texture, color information is not needed. We have used color video frames for processing and computed Color Layout Descriptor (CLD), which is standardized as a color descriptor in MPEG-7 [20]. Since color is not uniform over the images, we have transformed the images to other color spaces. The images are converted from RGB color space to YCbCr color spaces, so that variance in color becomes observable. We have also converted the video key frame into HSV color space and computed the average hue (H_{avg}) and the maximum saturation (S_{max}) level as a feature. The value of S_{max} is used as one of the feature to distinguish natural scenes from cartoons, which has more saturation. Moreover, cartoons usually have more pixels belonging to a particular intensity. To capture that we compute the percentage of pixels (I_P) above a particular predefined intensity threshold (I_{th}). For our experiments we have empirically determined the value of I_{th} to be 0.45.

3.1.2 Shape descriptor

Studies [5] have shown that shape is an important cue to the human perception system for object recognition. The perceptual recognition of objects in a video shot is conceptualized to be a process in which the input frame is segmented into regions (foreground blobs) and then the shape, motion

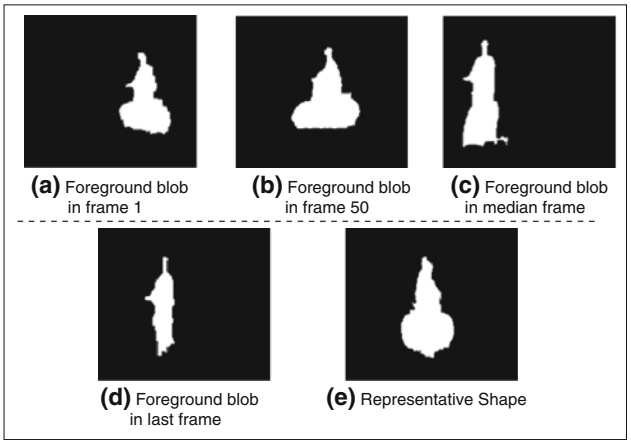

Fig. 3 Representative shape for a cycle video

characteristics are extracted by tracking the foreground blob for content analysis. In [10] only the foreground blob of the median frame has been taken as the representative shape for the entire video. But, this technique falls short in cases where extracted foreground blobs appear similar for two entirely different video shots (mostly due to pose change). To compensate for this drawback we compute a representative shape from all the frames instead of only the median frame. At first foreground blobs are extracted from the videos using the technique reported in [2]. These foreground blobs are overlaid upon each other, by aligning them with respect to the centroid. The resulting image captures the overall shape of the object and is extremely robust to orientation changes. The representative shape is smoothed by a Gaussian filter to give a better overall representation. Figure 3 shows the effectiveness of this technique in capturing the overall shape of the object. Segmented foreground object of a cycle video at different time instances are shown in Fig. 3a–d. Figure 3e depicts the final representative shape after overlaying the foreground blobs. Once we get the representative shape for a particular video shot, we calculate the HOG [8] feature from it. The rational behind the selection of HOG is that features like in [4,21], which works well under different challenging scenarios, are invariant to rotation. We purposefully wanted our system to be sensitive to rotation because two different objects at a specific orientation may appear similar.

3.1.3 Texture descriptor

Texture features are also an important group of image descriptors. We have computed Edge Histogram Descriptor (EHD) and Edge Intensity Histogram (EIH) as textural descriptors from a key frame. Since, some genres (e.g., cartoons) will have homogeneous textured areas, this is a very good discriminatory feature. We also detect the presence of prominent straight lines using the Hough Transform (HT) [9] and use it to classify between sports videos where the playing field (football, swimming) shows distinct characteristics due to the presence of lines on it. Moreover, natural scenes exhibit heterogeneous texture features as compared to the cartoons, which has a relatively homogeneous distribution of textures. EHD, which is an 80-dimensional feature vector, has been standardized as a texture descriptor in MPEG-7 standard [20] and or similarity search and retrieval.

To generate EIH, at first we gradient intensities in the vertical (G_V) and horizontal (G_H) directions. Then the intensity (A) of the gradient at each points in the image was calculated using $A = \sqrt{G_V^2 + G_H^2}$. After that an eight element histogram (EIH) was calculated for the values in this edge intensity image. The values were also normalized with respect to the image size to make them invariant to the image size.

3.2 Motion feature descriptor

There are mainly two sources of motion or dynamics in a video shot: foreground object motion and camera motion. In this work we have considered video shots having very little or no camera motion. There may be another source of dynamics as the rate of scene change, which occurs mainly due to video editing. Since we are working with pre-segmented video shots, this category is not applicable to our case. To capture the motion of the moving foreground object we segment the foreground object using the technique reported in [2]. Then we track the centroid of the moving object to extract the trajectory of the moving object. From the extracted trajectory we compute the direction and rate of change of motion of the moving foreground object.

But this information alone is not sufficient for classifying the motion of the objects. There are instances where the foreground object moves diagonally across the video frame. For example, there is a possibility of a diagonal jog having the same slope as that of a horizontal walk. The reason being the fact that distance traveled in case of diagonal jogging will be more, so the velocity $\left(\text{i.e.}\frac{\text{distance}}{\text{time}}\right)$ will be similar to that of horizontal walk. Therefore, the trajectory of the object also has to be considered for classification. The displacement of the centroid of the object in the vertical direction helps us in distinguishing these two scenarios. A similar problem also occurs between a diagonal run and a horizontal jog which can also be solved by the same technique. Therefore, we have determined a set of thresholds, one for the slope of the distance versus time plot and another threshold that distinguishes diagonal motion from horizontal motion. Figure 4 depicts the overall classification process based on displacement and velocity. Since the centroid tracking approach gave the best results, it was chosen to classify human motion.

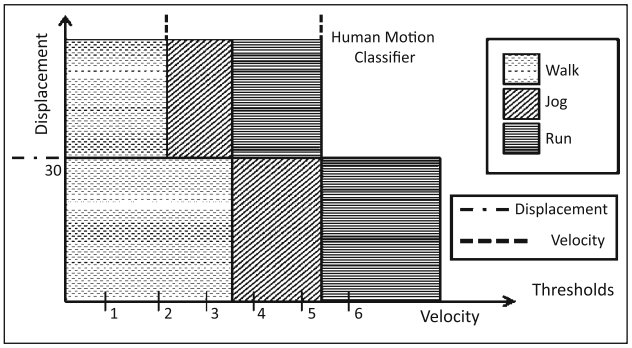

Fig. 4 Heuristic classifier for categorizing motion of videos with human objects

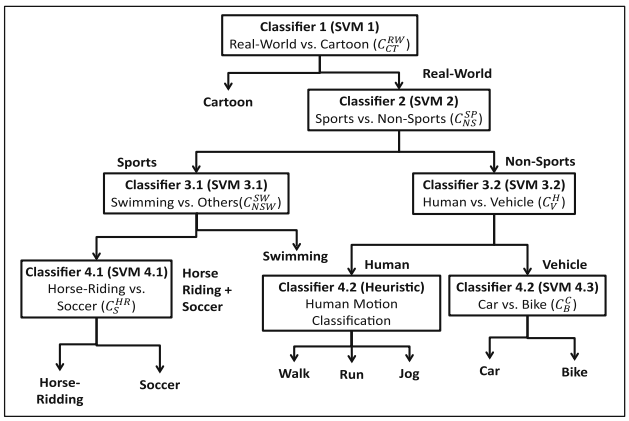

Fig. 5 Organization of classifiers at different levels of video categories.

4 Classification methodology

After feature extraction, the next step in video classification task is the video content modeling. Many effective modeling techniques have been proposed in the literature. The effectiveness of the classification task depends on the classifier chosen. In literature there are various classification algorithms. In this work, we have chosen SVM to model the video content since, it has been well known for better generalization capabilities. The learning of model involves discrimination of each class against all other classes. It has been found that SVM performs well for binary classification. There exists strategies to make it work for multi-class classification task as well. In our video genre classification task, there can be a set of features which helps us to distinguish between different genres. 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 sets of distinguished classes.

We first determine a particular super-category of a video shot and then use features specific to that particular genre to further classify into sub-genres. Figure 5 depicts the hierarchical organization of classifiers used for our experiments. We train all SVM [6] based classifiers using features specific to that particular genre. Details of the features used for a particular class is discussed in Sect. 5. We have adopted a twofold cross-validation method. 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.

5 Experimental evaluation

In Sec. 5.1, we have discussed about the dataset used for our experimental purposes. From Sects. 5.2 to 5.8, we present the classification accuracy at each step of the classification task, as shown in Fig. 5. In Sect. 5.9, we present a comparison of our approach with two existing techniques, for each step of the classification task.

5.1 Dataset creation

Our video dataset is diverse, both in terms of source as well as content. We have created a collection of videos from publicly available datasets [1,12,23] for different genres of videos given in Fig. 1 to evaluate the proposed video genre classification system. We have also recorded real-world video shots consisting of different outdoor locations, using a still handheld Sony camcorder and downloaded videos from internet. The collection of dataset is available in [28]. This emphasizes the diversity in terms of source. The ground truth for the class of each video was hand labeled by the authors. As previously discussed, the assignment was done based on the dominant content present in the video. For simplicity, in all the videos used for experimental purposes, there is only one content. Thus, each video will have genre label(s) depending upon its position in the hierarchy, e.g., a car video will be labeled with real-world, vehicle and car. In our video database we have scenes from the campus, moving car, different human actions, sport actions, cartoons, etc. This provides content diversity to our database. This work is motivated by the way human being perceives the content in a video shot. We first identify the genre of the video and then using our previous experience on that particular category, we extract further information to detect the sub-genre, e.g., at first, we detect whether the video shot is a real-world video or cartoon. If it is a real-world video then only we process it further and detect whether it belongs to sports genre or not, and so on.

5.2 Real world versus cartoon

At first, we classify the video shots into two broad categories, namely real-world or cartoon. For classification purpose, we have used both color and texture features which are $S_{\max}$,

Table 1 Classification Accuracy at different levels of hierarchy

(a) Classification accuracy for C_{CT}^{RW}

Samples/class	Class	Real-world	Cartoons	Accuracy (%)
360	Real-world	327	33	91
280	Cartoon	42	238	85

(b) Classification accuracy for C_{NS}^{SP}

Samples/class	Class	Sports	Non-sports	Accuracy (%)
99	Sports	92	7	95
144	Non-sports	4	140	97.2

(c) Classification accuracy for C_V^H

Samples/class	Class	Human	Vehicle	Accuracy (%)
66	Human	60	6	91
36	Vehicle	3	33	92

(d) Classification accuracy for C_B^C

Samples/class	Class	Car	Bike	Accuracy (%)
21	Car	17	4	81
21	Bike	1	20	95

(e) Classification accuracy for C_{NSW}^{SW}

Samples/class	Class	Swimming	Non-swimming	Accuracy (%)
45	Swimming	42	3	95
47	Non-swimming	6	41	88

(f) Classification accuracy for C_S^{HR}

Samples/class	Class	Horse ridding	Soccer	Accuracy (%)
25	Horse ridding	22	3	96
22	Soccer	2	20	91

I_P, CLD, EIH, and EHD. We compute these features from the training samples and create a single feature vector of 98 dimensions (1-$S_{\max}$, 2-I_P, $3-10$ CLD, $11-18$ EIH and $19-98$ EHD). An SVM with quadratic kernel has been trained with these features. We have used 150 and 110 training samples for real-world and cartoon videos, respectively. Table 1a shows the accuracy of the real-world versus cartoon classifier $\left(C_{\mathrm{CT}}^{\mathrm{RW}}\right)$. It can be observed that, since the real-world videos contain non-homogeneous texture patterns across the frames as compared to the homogeneous patterns present in the cartoons, use of the above mentioned features gives a reasonable performance. Cartoon videos having non-uniform texture pattern similar to natural scenes are wrongly classified as real-world videos.

5.3 Sports versus non-sports

Once video shots are identified as real-world video, in the next level they are classified as sports or non-sports videos. A SVM was trained based on the CLD, extracted from the video. All the 150 real-world videos used for training in the previous level were subdivided into two parts, consisting of 60 sports videos and 90 non-sports videos selected randomly. Table 1b shows the performance accuracy for the sports versus non-sports classifier $\left(C_{NS}^{SP}\right)$. It can be observed that the color features, which are already computed are sufficient to distinguish between these two genres of video. At the time of testing there is no need to recompute the features at this level, which results in a faster classification process.

Table 2 Performance of motion classification

Samples/class	Class	Walk	Run	Jog	Accuracy (%)
100	Walk	100	0	0	100
100	Run	8	90	2	90
100	Jog	0	5	95	95

5.4 Human versus vehicle

Shape is an important discriminatory feature to classify between humans and vehicles. We have used HOG feature for classification. As discussed in Sect. 3.1.2, we compute the HOG feature from the representative shape. We have trained the level 3 SVM classifier with this feature using the 90 non-sports video shots using a quadratic kernel. The framework was tested using a total of 102 videos. Table 1c shows the performance accuracy for the human versus vehicle classifier (C_V^H). At the time of testing, representative shape and HOG features are computed from the real-world video shots only if it belongs to the non-sports category. It can be observed from the result that only the shape feature is sufficient to distinguish between these two categories. Moreover, our proposed representative shape is also able to distinguish between two video categories with high accuracy.

5.5 Run versus Jog versus Walk

The centroid tracking method was used to classify the kinematics of the human object as it exhibited the best performance. We have computed the thresholds as discussed in Sect. 3.2. It can be observed from Table 2 that our heuristics

Table 3 Comparison of classification accuracy of the proposed technique with two other contemporary techniques of video classification

Class	Classification accuracy		
	Proposed (%)	Yuan et al. [31] (%)	Hasan et al. [14] (%)
Real-world versus cartoon	86	82	80
Sports versus non-sports	96	91	89
Human versus vehicle	91	85	76
Human motion classification (run, walk, jog)	95	91	78
Vehicle sub-category(car vs. bike)	88	80	78
Swimming versus non-swiming	95	89	90
Horse-ridding versus soccer	91	85	86

based classifier is able to distinguish between these three classes of actions with a very high accuracy. The classifier gets confused between the two classes Run and Jog, which is quite natural even from the human point of view, but was able to distinguish them from the more obvious category of walking. Moreover, at this level, we just need to compute the distance traveled by the person and the average velocity from the trajectory, which has already been extracted at the time of foreground blob extraction, which saves the time for feature recomputation.

5.6 Car versus bike

We have experimentally determined that shape features work best for the classification of these two categories of video shots. As discussed earlier HOG feature was extracted from the representative shape and used in the classification process. 50 videos were used for training the level 4 SVM using quadratic kernel. 19 of these were car videos and the remaining 31 were bike videos. Table 1d shows the performance accuracy of the car versus bike classifier (C_B^C).

5.7 Swimming versus non-swimming

Classification between swimming and other sports categories (horse-riding and soccer) has been done by training an SVM using quadratic kernel. In sports video classification, video frames contain the playing field where most of the action is happening. This gives significant discriminating cue among the two classes of sports categories. Swimming video shots contain distinct appearance with a dominant color and presence of non-homogeneous texture due to the presence of water ripples. Hue value (H_{avg}) from bottom half of the key frame and the number of prominent straight lines using HT [9] is obtained. These two features are used to train the swimming versus non-swimming classifier ($C_{\text{NSW}}^{\text{SW}}$) SVM. Table 1e shows the details of performance accuracy of $C_{\text{NSW}}^{\text{SW}}$.

5.8 Horse-riding versus soccer

To classify between horse-riding and soccer video shots, we have used the same set of features used for swimming and non-swimming video shot classification. Football ground contains more homogeneous pattern as compared to horse-riding where a number of prominent edges are more due to the presence of fences. All the remaining training video shots were divided into two classes and the horse-riding versus soccer classifier $\left(C_S^{\text{HR}}\right)$ has been trained. Classification accuracy of C_S^{HR}, as given in Table 1f, shows high accuracy.

5.9 System performance

To experimentally verify the performance of our proposed framework, we have compared our results (classification accuracy) with a very recent work [14] on video shot classification for movie management and another work [31] on automatic genre classification using hierarchical SVM. The proposed technique in [14] uses a spatial (key frame based approach) feature and computes a 48-dimensional feature vector to classify different video shots. On the other hand [31] uses both spatial and temporal features and uses hierarchical SVM binary-tree approach for video genre classification. Table 3 shows the comparison of classification accuracy of our proposed framework, for each step of the classification task, with [14] and [31]. It can be observed that our proposed method outperforms both the techniques proposed in [14] and [31] in almost all the cases. It shows the importance and effectiveness of judiciously selecting features at every level of the hierarchy and the hierarchical organization of video shots to incorporate the semantic information for improved performance. The improvement in classification emphasizes the superiority of genre-specific feature modeling as compared to using a single feature vector for all genres.

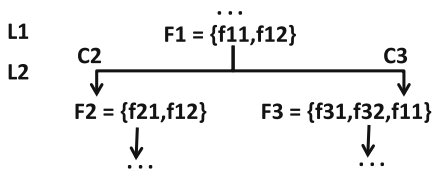

Fig. 6 Hierarchical ordering of features

6 Retrieval efficiency

Our proposed categorization framework also facilitates efficient content-based video retrieval (CBVR). There are two reasons behind this, (i) linear versus hierarchical search and (ii) conditional feature computation. To create a rank-ordered list of videos from a linearly organized database requires comparison with all the videos in the database and is time consuming. But our framework first determines the genre of the query video shot and then compares with videos with that particular genre. This reduces the the search space and thereby the search time. Furthermore, we achieve better retrieval efficiency by conditionally computing the feature from a video shot. We compute feature specific to a genre if at all that video has been found out to be so. Moreover, we do not recompute the features already computed at a higher level of the hierarchy and needs to be reused (see Fig. 6). Suppose at a hierarchy level L1, we have computed feature set $F1 = \{f11,\ f12\}$ and the classifier determines a class level G2. Again the classifier for genre G2 was trained using the feature set $F2 = \{f21,\ f12\}$. Since, $f12$ has already been computed, in the next iteration, we will only compute the feature $f21$ and determine the genre level for the video shot. Moreover, we need not compute other features ($f31,\ f32$, etc.) since the query video shot does not belong to that genre. This helps us in achieving significant speedup in our retrieval process.

7 Conclusion and future work

This paper has presented a framework for video classification based on genre-specific modeling of visual features using SVM models. Experimental results have shown that the genre-specific modeling of spatial and temporal features can provide useful information for video content understanding and can be used as discriminatory criteria to achieve an improved classification performance on a video database of diverse categories. However, it is also to be stated that use of visual features alone may not be sufficient for better classification accuracy. Studying the feasibility of genre-specific modeling of multimodal features like audio, text along with the visual features for content-based video genre classification provides a good scope for future research. As a part of our ongoing work, we are planning to work on videos with camera movement so as to incorporate more video genres.

Acknowledgments We are thankful to the members of Visualization and Perception Lab for their support in data acquisition and comments. Additional thanks to Prof. Sukhendu Das and Debarun Kar for their insightful suggestions. We would like to express our special appreciation to Prashant B for his implementation of some experimenting modules.

References

1. UCF50 http://www.cs.ucf.edu/kreddy/Datasets.html
2. Barnich O, Van Droogenbroeck M (2011) Vibe: a universal background subtraction algorithm for video sequences. IEEE TIP 20(6):1709–1724
3. Bartolini I, Patella M, Romani C (2011) Shiatsu: tagging and retrieving videos without worries. MTA 15(1):52–64
4. Belongie S, Malik J, Puzicha J (2002) Shape matching and object recognition using shape contexts. IEEE TPAMI 24(4):509–522
5. Biederman I (1987) Recognition-by-components: a theory of human image understanding. Psychol Rev 94:115–147
6. Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM TIST 2:27:1–27:27
7. Cortes C, Vapnik V (1995) Support-vector networks. ML 20(3):273–297
8. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: CVPR, vol 1, pp 886–893
9. Duda RO, Hart PE (1972) Use of the hough transformation to detect lines and curves in pictures. Commun ACM 15(1):11–15
10. Dyana A, Das S (2010) MST-CSS (multi-spectro-temporal curvature scale space), a novel spatio-temporal representation for content-based video retrieval. In: IEEE TCSVT, pp 1080–1094
11. Fischer S, Lienhart R, Effelsberg W (1995) Automatic recognition of film genres. In: ICM, pp 295–304
12. Gorelick L, Blank M, Shechtman E, Irani M, Basri R (2007) Actions as space-time shapes. IEEE TPAMI 9(12):2247–2253
13. Haering N, Qian R, Sezan M (2000) A semantic event-detection approach and its application to detecting hunts in wildlife video. IEEE TCSVT 10(6):857–868
14. Hasan MA, Xu M, He X, Chen L (2012) Shot classification using domain specific features for movie management. In: DASFAA, pp 314–318
15. Ide I, Hamada R, Tanaka H, Sakai S (1998) News video classification based on semantic attributes of captions. In: Proceedings of the 6th ACM international multimedia conference, pp 60–61
16. Jianxun Z, Bo W (2012) Video semantic concept detection based on multi-modality fusion. In: ICCSEE
17. Li L (2012) A novel violent videos classification scheme based on the bag of audio words features. IJCST 2(1):4–15
18. Li L, Zhang N, Duan LY, Huang Q, Du J, Guan L (2009) Automatic sports genre categorization and view-type classification over large-scale dataset. In: ICM, pp 653–656
19. Ma YF, Zhang HJ (2003) Motion pattern-based video classification and retrieval. EURASIP JASP 2003:199–208
20. Manjunath B, Salembier P, Sikora T (2003) Introduction to MPEG-7-multimedia content description interface. Wiley, New York
21. Mokhtarian F, Mackworth A (1986) Scale-based description and recognition of planar curves and two dimensional shapes. IEEE TPAMI 8(1):34–43
22. Rasheed Z, Sheikh Y, Shah M (2005) On the use of computable features for film classification. IEEE TCSVT 15(1):52–64
23. Schuldt C, Laptev I, Caputo B (2004) Recognizing human actions: a Local SVM Approach. In: ICPR, pp 32–36
24. Snoek CGM, Worring M (2005) Multimodal video indexing: a review of the state-of-the-art. MTA 25(1):5–35
25. Suresh V, Mohan CK, Swamy RK, Yegnanarayana B (2004) Content-based video classification using support vector machines. In: ICONIP, pp 726–731

26. Tiwari R, Zhang C (2011) Video genre detection using a multi-modality approach. In: ICM, pp 879–880
27. Truong BT, Dorai C (2000) Automatic genre identification for content-based video categorization.In: ICPR, pp 230–233
28. VP lab video collection. http://www.cse.iitm.ac.in/~vplab/videoc.html
29. Wu J, Worring M (2012) Efficient genre-specific semantic video indexing. IEEE TM 14(2):291–302
30. Yang L, Liu J, Yang X, Hua XS (2007) Multi-modality web video categorization. In: Workshop onMIR, pp 265–274
31. Yuan X, Lai W, Mei T, Hua XS, Wu XQ, Li S (2006) Automatic video genre categorization using hierarchical SVM. In: ICIP, pp 2905–2908