Open access • Proceedings Article • DOI:10.1109/ICPR.2000.902901

Automatic genre identification for content-based video categorization — Source link

Ba Tu Truong, Chitra Dorai

Institutions: Curtin University, IBM

Published on: 01 Jan 2000 - International Conference on Pattern Recognition

Topics: Categorization and Music information retrieval

Automatic Genre Identification for Content-Based Video Categorization

Ba Tu Truong
Svetha Venkatesh
Department of Computer Science
Curtin University of Technology
GPO Box U1987, Perth, 6845, W. Australia
{truongbt, svetha}@cs.curtin.edu.au

Chitra Dorai
IBM T.J. Watson Research Center
P.O. Box 704, Yorktown Heights
New York 10598, USA
dorai@watson.ibm.com

Abstract

This paper presents a set of computational features originating from our study of editing effects, motion, and color used in videos, for the task of automatic video categorization. These features besides representing human understanding of typical attributes of different video genres, are also inspired by the techniques and rules used by many directors to endow specific characteristics to a genre-program which lead to certain emotional impact on viewers. We propose new features whilst also employing traditionally used ones for classification. This research, goes beyond the existing work with a systematic analysis of trends exhibited by each of our features in genres such as cartoons, commercials, music, news, and sports, and it enables an understanding of the similarities, dissimilarities, and also likely confusion between genres. Classification results from our experiments on several hours of video establish the usefulness of this feature set. We also explore the issue of video clip duration required to achieve reliable genre identification and demonstrate its impact on classification accuracy.

1. Introduction

Automatic classification of digital video into various genres, or categories such as sports, news, commercials, music, cartoons, documentaries, and movies is an important task, and enables efficient cataloging and retrieval with large video collections. At the highest hierarchy level, film and video collections can be categorized into different program genres. Video classification into TV genres is discussed in [2, 6, 5]. Approaches such as [3, 11] classify movie trailers using film genre labels. At the next level of the hierarchy, domain videos such as sports can be classified into different sub-categories [1, 8]. At a much finer level of resolution, a video sequence itself can be segmented and each segment can then be classified according to its semantic content. Events in a baseball telecast [4] or newscasts [12] can be indexed in this manner.

Our work addresses the problem of video classification at the highest level of abstraction: Genres. In particular, we examine a set of features that would be useful in distinguishing between sports videos, music, news, cartoons, and commercials. In contrast to [2, 6] we concentrate on features that can be extracted only from the visual content of a video. Rather than learning features from video data sets, we use human perception and discernment of video genre characteristics as a starting point, and extract computational features that would reflect those visual characteristics such as editing, motion, and color. Some of our features are similar to those proposed in [2] and some are *new*, but we also go beyond [2] to show classification results on several hours of video. In addition, we address the important related issue of the length of a clip required to be processed for reliable genre identification and its impact on the classification performance using proposed features, since this issue has not been studied elsewhere.

2. Proposed Feature Set and Trends

Consider a video clip, V, a contiguous sequence of $n+1$ frames, $V = \{f_1, f_2, ..f_{n+1}\}$. A frame transition vector, $T = \{t_1, t_2, \dots, t_n\}$ is first computed from V, where each t_i is a feature set, $\{t_i^{\mu}, t_i^{v}\}$ computed jointly from frames f_i and f_{i+1}. Specifically, $t_i^{\mu} = |f_{i+1}^{\mu} - f_i^{\mu}|$ and $t_i^{v} = |f_{i+1}^{v} - f_i^{v}|$, where f_i^{μ} and f_i^{v} denote the mean and variance of luminance values of pixels in frame f_i, respectively.

2.1. Shot Processing

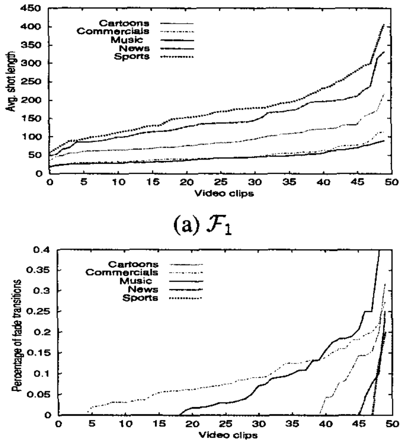
The video sequence is automatically segmented into shots using the method detailed in [10] which performs segmentation by detecting effects such as cuts, fades, and dissolves in the video. After this step, each member t_i of T receives a label, x from the set, $\mathcal{L} = \{shot, cut, fade, dissolve\}$ depending on whether f_i and f_{i+1} are part of a pure shot, cut, fade or dissolve transition respectively. The transition vector T is then grouped into k segments, $\{S_1, S_2, \dots, S_k\}$, where S_i is a set of consecutive elements, $\{t_j, t_{j+1}, \cdots\}$ that have the same label. Each S_i is also assigned a label from $\mathcal{L}$ according to the label type of frames it contains. Let Γ^x and Ω^x denote sets containing segments S_i and t_i of label type x, respectively. Let Δ^{shot} denote the set of pure shot frames in the video sequence.

2.2. Feature Extraction

We now extract a set of features that capture distinctive cinematic aspects of a video genre such as editing (features

0-7695-0750-6/00 $10.00 © 2000 IEEE

1 and 2), motion (features, 3 to 6), and color (features, 7 through 10). The precise definitions of these features together with their intuitive meanings follow. Accompanying each feature is a plot of the values of the feature computed for 50 video samples (60sec each) randomly selected from each genre (c.f. Section 3), after sorting the values in ascending order.



(a) $\mathcal{F}_1$

(b) $\mathcal{F}_2^{fade}$

Figure 1. Editing feature values in ascending order for 50 video samples from each genre.

Average shot length $\mathcal{F}_1$ is a useful feature in video characterization, since it is fundamental to our perception of scene pace and content. Therefore, short-duration shots are often used in commercials and music videos with fast music. In contrast, longer shots are used in sports to maintain the continuity of actions (see Figure 1a). Shot length is measured as the number of frames between the last frame of the preceding transition and the first frame of the succeeding transition. So if a shot contributes to the existence of a segment S_i then its length would be $|S_i| + 1$. The average shot length computed from the whole clip, V, is used as a classification feature:

$$\mathcal{F}_1 = \frac{\sum_{i=1}^{k} \theta_i}{|\Gamma^{shot}|}, \text{ where } \theta_i = \begin{cases} |S_i| + 1 & \text{if} S_i \in \Gamma^{shot} \\ 0 & \text{otherwise.} \end{cases}$$

The percentage of each type of transition used for editing can also identify a video genre. For example, while *fade* transitions are common in commercials and sometimes in music, they are rarely used in sports and news (see Figure 1b). We compute the percentage of each type of transitions x, $x \in \{cut, fade, dissolve\}$ as:

$$\mathcal{F}_2^x = \frac{|\Gamma^x|}{|\Gamma^{cut}| + |\Gamma^{fade}| + |\Gamma^{dissolve}|}.$$

Camera movement influences the narration of scene content. In sports such as soccer and rugby fixed cameras are positioned around the field, and since the ball changes its position continuously, a lot of camera movement is needed to track the ball continuously. In contrast, in newscasts, the object of interest such as an anchor person or a reporter remains relatively static (see Figure 2a). Camera motion magnitude of a frame, f_i is computed using two consecutive frames, f_i and f_{i+1} with the method proposed in [9], and the overall amount of camera movement of a video segment is computed using frame tilt and pan:

$$\mathcal{F}_3 = \frac{\sum_{i=1}^{n} \theta_i}{|\Delta^{shot}|}, \text{ where } \theta_i = \begin{cases} |f_i^{\text{tilt}}| + |f_i^{\text{pan}}| & \text{if } f_i \in \Delta^{shot} \\ 0 & \text{otherwise.} \end{cases}$$

In music videos, there are often special effects such as quick changes of lighting and flash lights causing a large change in the variance of pixel luminance between two consecutive frames. We measure the prevalence of these effects as the number of shot features, t_i whose pixel luminance variance is above a certain threshold:

$$\mathcal{F}_4 = \frac{|\Omega_1|}{|\Omega^{shot}|}, \text{ where } \quad \Omega_1 = \{t_i \in \Omega^{shot} \mid t_i^v > T_1\}.$$

Figure 2b shows $\mathcal{F}_4$ as being distinctly higher for music videos than news.

The rate of "quiet" visual scenes, where both camera and object motion are very little, varies between different video categories. We expect music videos to be rather dynamic, while anchor shots in newscasts are rather static (see Figure 2c). The prevalence of static scenes in videos is measured using the number of frame transitions t_i whose mean and variance in pixel luminance are both less than certain thresholds.

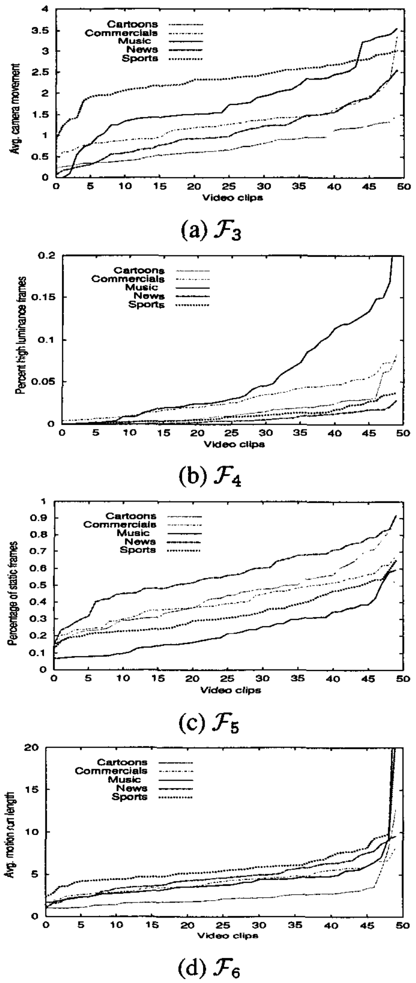
$$\mathcal{F}_5 = \frac{|\Omega_2|}{|\Omega^{shot}|}, \; \Omega_2 = \{t_i \in \Omega^{shot} \mid t_i^{\mu} < T_2 \text{ and } t_i^{v} < T_3\}.$$

A new feature proposed based on motion is the average length of *motion runs*. A motion run R_i is defined an unbroken sequence of those frames, f_i whose sum of absolute pixel-wise luminance differences between f_i and f_{i+1} exceeds a certain threshold, T_4. Let $|R_i|$ be the length of this run. Let R denote the set of all motion runs in the video clip. Then

$$\mathcal{F}_6 = \frac{\sum_i |R_i|}{|R|}.$$

Figure 2d shows $\mathcal{F}_6$ for the five genres. $\mathcal{F}_6$ is consistently high for sports when compared against cartoons. The main reason for this is that motion in sports tends to occur continuously in time, while in the normal production process of a cartoon a single drawing may be exposed a number of times resulting in a lower pixel-wise difference between consecutive frames.

There are also the distinctions in the distribution of color histograms between different video genres. Let $f_i^{\mathcal{H}}$ denote the luminance histogram of frame f_i and $f_i^{\mathcal{H}_k}$ denote the histogram of k largest bins in the color histogram, i.e., the



(a) $\mathcal{F}_3$

(b) $\mathcal{F}_4$

(c) $\mathcal{F}_5$

(d) $\mathcal{F}_6$

Figure 2. Motion features for 50 video samples from each genre.

set of k most prevalent luminance levels. We measure the coherence of these k bins based on δ, the standard deviation of indices of $f_i^{\mathcal{H}_k}$ as a new feature. Thus:

$$\mathcal{F}_7 = \frac{\sum_{i=1}^{n+1}\theta_i}{|\Delta^{shot}|}, \text{ where } \theta_i = \begin{cases} \delta(f_i^{\mathcal{H}_k}) & \text{if } f_i \in \Delta^{shot} \\ 0 & \text{otherwise.} \end{cases}$$

Figure 3a shows that sports have the lowest value of $\mathcal{F}_7$, since the color of the playing field tends to be highly homogeneous, while music videos tend to have high values of $\mathcal{F}_7$ indicating high color variability.

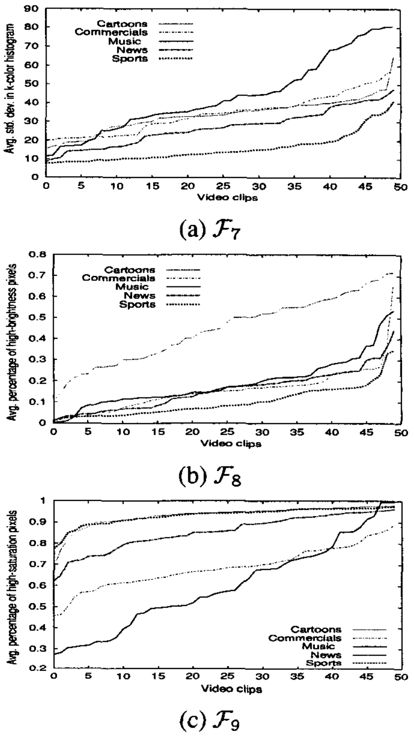
The HSV color space provides two other interesting features. For example, the average brightness for cartoons is much higher than other video genres (see Figure 3b). We compute $\mathcal{F}_8$ based on $f_i^{\mathcal{V}}$, the percentage of pixels having brightness above a certain threshold T_5.

$$\mathcal{F}_8 = \frac{\sum_{i=1}^{n+1}\theta_i}{|\Delta^{shot}|}, \text{ where } \theta_i = \begin{cases} f_i^{\mathcal{V}} & \text{if } f_i \in \Delta^{shot} \\ 0 & \text{otherwise.} \end{cases}$$

We compute $\mathcal{F}_9$ based on $f_i^{\mathcal{S}}$, the percentage of pixels having saturation above a certain threshold T_6.

$$\mathcal{F}_9 = \frac{\sum_{i=1}^{n+1}\theta_i}{|\Delta^{shot}|}, \text{ where } \theta_i = \begin{cases} f_i^{\mathcal{S}} & \text{if } f_i \in \Delta^{shot} \\ 0 & \text{otherwise.} \end{cases}$$

Figure 3c shows that the average saturation for cartoons and sports is much higher when compared against commercials and music videos.



(a) $\mathcal{F}_7$

(b) $\mathcal{F}_8$

(c) $\mathcal{F}_9$

Figure 3. Color statistics for 50 samples from each genre.

3. Experimental Results

We collected about 8 hours of TV material and digitized and encoded it in the MPEG-1 format. To ensure the variety of data, news and commercials from different channels on different days and at different times of the day were used. Some clips were in fact recorded more than 5 years ago. Sports clips came from different sub categories such as soccer, Australian football, rugby, and motor racing. Music clips were extracted from different dance music videos.

The C4.5 decision tree [7] is used to build the classifier for genre labeling. All the video material is first divided into units of approximately equal duration. The system was tested with features computed for different basic clip durations of 40sec, 60sec, and 80sec. During feature extraction, all the six thresholds were determined empirically and used stably and consistently across all clips for all durations. During each classification experiment for a clip duration d, 60% of all the clips obtained by segmenting the eight-hour material into d-long units were randomly selected for training while the remaining 40% of the clips were used for testing. For each duration, 100 such sets were randomly derived and used for classification, and the overall classification results are presented in Table 1. We measure in percentage the best, worst, average classification across 100 runs and the standard deviation for each duration, as we expect that slightly different decision trees would be built with different data combinations.

Stats.	Dur. *d*	*All*	-Ca.	-Co.	-Mu.	-Ne.	-Sp.
Best %	40'	84.6	88.4	87.2	85.4	89.2	85.3
	60'	86.2	88.3	**92.3**	90.3	**91.5**	**89.2**
	80'	**89.7**	**91.4**	90.0	**90.4**	91.2	89.0
Worst %	40'	78.4	83.4	83.1	81.1	83.5	80.3
	60'	**81.0**	83.1	**85.3**	**85.5**	**85.2**	**82.7**
	80'	79.5	**83.6**	83.1	83.7	82.3	80.5
Avg. %	40'	80.0	84.8	84.5	82.2	85.2	82.0
	60'	**83.1**	85.3	**86.8**	**87.2**	**87.4**	84.8
	80'	81.7	**85.7**	85.0	86.1	85.2	**87.4**
Stdv.	40'	**1.41**	**1.21**	**1.05**	**1.07**	**1.40**	**1.45**
	60'	1.66	1.54	1.46	1.28	1.90	1.81
	80'	1.91	1.56	1.73	1.80	2.17	2.09

Table 1. Genre classification results.

The *All* category in the table represents the classification results when samples from all genres were used in training and testing, while others such as -{Ca} represent classification results obtained omitting samples from one given genre, say Ca, cartoons during training (-{Co} is for omitting commercials, -{Mu}, music, -{Ne}, news, and -{Sp} for sports). The best result in each group are typeset in bold. In the best case for *All*, the classification rates are 86% (60sec) and 90% (80sec). The average classification for *All* is between 80 % and 83%. Examination of the standard deviation of the results implies that using video clips of 60 sec duration is the most appropriate strategy, as it offers the best trade off in terms of high classification and low standard deviation.

The best classification rate rises when one genre is omitted to around 92% due to patterns that exist in the genre confusion matrix. It is useful to analyze Figs. 1, 2, and 3. The average shot length, ($\mathcal{F}_1$) and its trends across samples are similar for ⟨commercials & music⟩, and also for ⟨sports & news⟩. Cartoons fall somewhere in between, but can be confused with either of the four genres. The motion feature, ($\mathcal{F}_3$) is similar for ⟨news & commercials⟩ and is close to but lower than music. However, all three categories are close. Further, cartoon features are close to those of news. Feature, $\mathcal{F}_4$ is high for music, but is still close to commercials. However, $\mathcal{F}_4$ well separates out news from ⟨commercials & music⟩. $\mathcal{F}_5$ clearly separates out ⟨news, commercials, & music⟩ and thus complements motion features $\mathcal{F}_3$ and $\mathcal{F}_4$. Features, $\mathcal{F}_6$, $\mathcal{F}_8$, and $\mathcal{F}_9$ separate out cartoons from all other categories. $\mathcal{F}_7$ separates out sports from music. A high degree of confusion can exist for news and sports since they are close in all features other than motion. Similarly, music and commercials have almost identical shot length and similar motion, and can lead to a mix-up.

4. Conclusion

We have presented a set of features that embody the visual characteristics of a video sequence for video genre identification. The experimental results on several hours of videos indicate that these features perform well in classifying videos into sports, news, commercials, cartoons, and music, thus enabling automatic genre-based filtering during categorization and search. Our study on the length of a clip needed to recognize its genre indicates that 60sec can serve as the most appropriate video duration to achieve reliable classification accuracy. Future work will investigate temporal sequencing of shots and their semantics to further improve the performance of our system.

References

[1] Y. Ariki, A. Shibutani, and Y. Sugiyama. Classification and retrieval of TV Sports News by DCT features. In *IPSJ International Symposium on Information System and Technologies for Network Society*, pages 269–272, Sept. 1997.

[2] W. Effelsberg, S. Fischer, and R. Lienhart. Automatic recognition of film genres. In *The Third ACM International Multimedia Conference and Exhibition (MULTIMEDIA '95)*, pages 367–368, New York, Nov. 1995. ACM Press.

[3] G. Iyengar and A. B. Lippman. Models for automatic classification of video sequences. In *Storage and Retrieval VI*, San Jose, Jan. 1998.

[4] T. Kawashima, K. Tateyama, T. Iijima, and Y. Aoki. Indexing of baseball telecast for content-based video retrieval. In *IEEE 1998 International Conference on Image Processing ICIP'98*, pages 871–874, Chicago, Oct. 1998.

[5] V. Kobla, D. DeMenthon, and D. Doermann. Detection of slow-motion replay sequences for identifying sports videos. In *IEEE 1999 International Workshop on Multimedia Signal Processing*, Copenhagen, Denmark, Sept. 1999.

[6] Z. Liu, J. Huang, and Y. Wang. Classification of TV programs based on audio information using hidden Markov model. In *IEEE Signal Processing Society 1998 Workshop on Multimedia Signal Processing*, pages 27–32, Dec. 1998.

[7] J. R. Quinlan. *C4.5: Programs for machine learning*. Morgan Kaufmann Publishers, San Mateo, California, 1993.

[8] E. Sahouria and A. Zakhor. Content analysis of video using principal components. In *IEEE 1998 International Conference on Image Processing ICIP'98*, volume 3, pages 541–545, Chicago, Oct. 1998.

[9] M. Srinivasan, S. Venkatesh, and R. Hosie. Qualitative extraction of camera parameters. *Pattern Recognition*, 30(4):593–606, 1997.

[10] B. T. Truong, C. Dorai, and S. Venkatesh. New enhancements to cut, fade, and dissolve detection processes in video segmentation. In *ACM Multimedia (Submitted)*, 2000.

[11] N. Vasconcelos and A. Lippman. Towards semantically meaningful feature space for the characterization of video content. In *International Conference on Image Processing ICPR'97*, volume 1, pages 25–28, Santa Barbara, California, June 1997.

[12] H. Zhang, S. Y. Tan, S. W. Smoliar, and G. Yihong. Automatic parsing and indexing of news video. *Multimedia Systems*, 2(6):256–266, 1995.