

Automatic Genre Classification from Videos

S. Karthick, S. Abirami, S. Murugappan, M. Sivarathinabala
and R. Baskaran

Abstract In recent decades, there has been a huge growth in the amount of multimedia content stored in networked repositories. Many video hosting websites exist in today's scenario such as youtube, metacafe and google video etc., where people are uploading and downloading their videos. At present, indexing and categorization of these videos is a tiresome job. Either the system asks the user to suggest tags for the videos which they upload or people are employed to tag the video manually. Manual tagging has been done based on the views of users, search terms, etc. In order to eliminate this problem, this paper proposes a model that automatically categorizes the videos based on their genres. The main aim of this work was to categorize the videos broadly on major domains such as sports, music and news using temporal, textural, motion, and color features. In sports, the videos have been classified further into cricket and football. The hierarchical SVM has been used for automatic training and selection of the genre of the video. A total of 350 videos from various Web sites have been used for training the classification system. This system achieves an overall average detection ratio up to 98 % while maintaining very low false detection rate of 2 %.

Keywords Video genre classification • Video categorization • Image processing

S. Abirami (✉) · M. Sivarathinabala
Department of Information Science and Technology, College of Engineering,
Anna University, Chennai, India
e-mail: abirami_mr@yahoo.com

R. Baskaran
Department of Computer Science and Engineering, College of Engineering,
Anna University, Chennai, India

S. Murugappan
Department of Computer Science and Engineering, Tamilnadu open University, Chennai,
India

S. Karthick
Department of Information Science and Technology, Anna University,
Chennai, India

1 Introduction

Rapid increase in the amount of multimedia content stored in Internet in today's world has led to many researches such as storage optimization, faster video content transmission, and multimedia content retrieval especially videos. Among these popular researches, video content retrieval requires many tagging for categorization of videos. Because of the massive amount of video contents, automatic classification is an essential area which needs to be actively researched. In literatures, three automated methods have been demonstrated by the existence of the TRECVID video retrieval benchmarking evaluation campaign. TRECVID provides data sets and common tasks that allow researchers to compare their methodologies under similar conditions. While much of the TRECVID has been devoted to video information retrieval, video classification tasks exist by limiting it by identifying clips containing faces or on-screen text, distinguishing between clips representing outdoor or indoor scenes, or identifying clips with speech or instrumental sound. Therefore, existence of fully automated video classifier on this genre becomes mandatory. In order to address these shortcomings, this research concentrates on the development of an automatic genre classification system.

2 Previous Works

Large number of approaches has been attempted in the literature for performing automatic classification of videos. These approaches could be broadly divided into four groups: Text-based approaches, audio-based approaches, visual-based approaches, and hybrid features approaches. These approaches draw features from three different modalities such as texts, audio, and visual. Text-only approaches are uncommon method. Here, viewable text could be text on objects that are filmed (scene text), such as an athlete's name on a jersey or the address on a building, or it could be text placed on-screen (graphic text), such as the score for a sports event or subtitles [1]. Text features are produced from this viewable text by identifying text objects followed by the use of optical character recognition (OCR) [2] to convert these objects to usable text. Zhu et al. [3] classified news stories using features obtained from closed captions. Audio-only approaches are found slightly more often than text-only approaches. Roach and Mason [4] used audio, in particular MFCC, from videos for genre classification. Moncrieff et al. [5] classified the videos into commercial, basketball game, football game, news report, and weather forecast based on their audio features. Roach et al. [6] used audio-based cinematic principles to distinguish between horror and non-horror movies. In visual-based approaches, frame differencing and optical flow techniques have been used in the classification of videos. Fakheri et al. [7] investigated found that the frame differencing [1, 8] method has lower computational requirements than optical flow method in classification of video. Wei et al. [9] classify four types of TV programs:

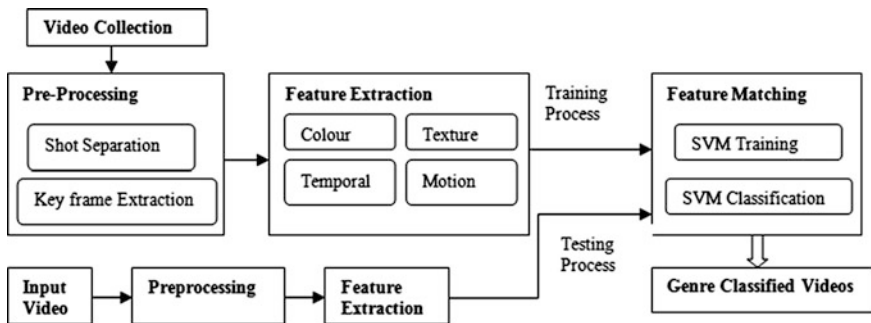


Fig. 1 Automatic video genre classification system

news, commercials, sitcoms, and soap operas. Roach et al. classified the video as either cartoon or non-cartoon. Yuan et al. [10] proposed two new classification methodologies using a series of support vector machine classifiers arranged in a hierarchy. This paper attempts to classify the different genres of videos using a hybrid approach of visual, temporal, and motion features.

3 Video Genre Classification System

In this paper, a novel method has been attempted to automatically classify the videos. News, cartoon, music, cricket, and football are the major genres taken and classified using this system. The overall genre classification system is shown in Fig. 1. The input video has been preprocessed. Next to it, the features are extracted from the key frames extracted. Color, texture, temporal, and motion features are extracted. Once the features are extracted, SVM classifier is used to classify the video.

3.1 Preprocessing

Initial step in video analysis is to preprocess the input video. In preprocessing, the video has been split into several shots/frames for further processing. Shot separation needs to be done after preprocessing.

3.1.1 Shot Separation

A video shot is a small image sequence captured continuously by a single camera. There are two major approaches available for shot separation such as pixel-difference method and histogram-based method. Both the methods involve a threshold to

be fixed in advance. Fixing a single threshold for all the videos are a poor technique and leads to erroneous results. The different threshold values are required for different categories of video. Hence, we attempted a saturation-based approach in this paper to separate shots.

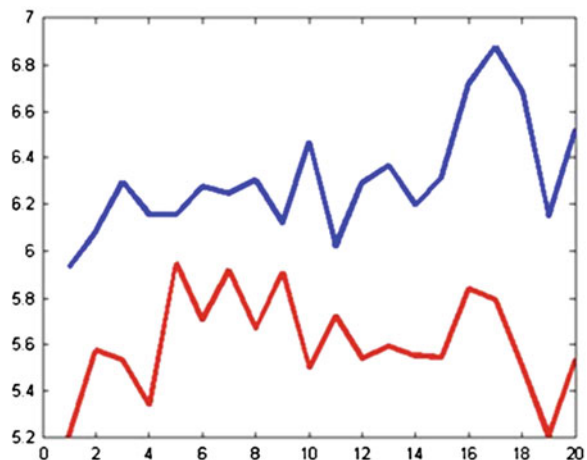
Saturation-based Approach

Saturation-based approach finds the frames when there is an abrupt change in the frame, by calculating the saturation value for each frame. The saturation values change gradually either in an increasing or decreasing nature, when the scenes did not have much of changes. The saturation value changes abruptly when there is a huge change in the scene. This is used to obtain the shots from the video. In this system, we make use of the saturation-based shot separation approach to determine the shots.

3.1.2 Key Frame Extraction

Processing all the frames resulting shot separation is very expensive and time consuming made use in further processing. We make use of the entropy difference method which is used in [11] to identify the key frames from each shot. More similar frames and duplicate key frames have been removed by similarity match. Sample key frames extracted for a shot length of 180 frames have been shown in Fig. 2.

Fig. 2 Entropy comparisons for football and cricket



3.2 Feature Extraction

There are several features [12], which can be extracted from the key frames. In this system, four major features are concentrated as they help in providing improved results and easier categorization of genres in videos. The four major features extracted are: Temporal features, motion features, texture features, and color features.

3.2.1 Temporal Features

Temporal features of class populations define some time-dependent constraints that the population must satisfy. Hence, temporal feature can be considered as an important descriptor in genre classification [+1]. It is clearly seen that there will be larger number of shots when there is a frequent scene change. News genres do not have many changes as the news reader is stationary and there is not much of motion. It has observed that the number of shots obtained for news videos will be comparably lesser than the other genres. Comparably, music videos have frequent shot changes. As a result, it has been proposed here to calculate the shot rate, key frame rate, and average length from the input videos as the features for texture measurement. The number of shots per 1,000 frames is calculated and has been given in Eqs. (1) and (2):

$$\text{Shot rate} = (\text{No. of Shots found} / \text{Total no. of frames}) \times 100 \quad (1)$$

The average number of frames in a shot is defined as the shot size. Number of key frames per 100 shots gives the key-frame rate and it is calculated as follows:

$$\text{Key-frame rate} = \text{Average key-frames per shot} \times 100 \quad (2)$$

3.2.2 Motion Features

The effective use of motion information is essential for content-based video retrieval. In this system, foreground motion activity is extracted using the pixel-difference method [13]. The rate of change in the successive key frames presents the foreground motion activity. The percentage of the pixels whose values have been changed will be high in case of frames involving motion activity than the frames which have low movements. This factor has been used to estimate the amount of motion activity in videos.

3.2.3 Texture Features

Texture features give information about the spatial arrangement of color or intensities in an image or selected region of an image. A statistical method of examining texture that considers the spatial relationship of pixels is the gray-level co-occurrence matrix (GLCM), also known as the gray-level spatial dependence matrix. The texture of an image has been characterized by the GLCM functions. The GLCM functions have been calculated how often pairs of pixel with specific values and in a specified spatial relationship occur in an image. From the GLCM function, statistical measures have been extracted from this matrix. The GLCM is created by calculating how often a pixel with the intensity (gray-level) value 'i' occurs in a specific spatial relationship to a pixel with the value 'j.' By default, the spatial relationship is defined as the pixel of interest and the pixel to its horizontally adjacent. Other spatial relationships can also be represented between two pixels. Each element $p(i, j)$ in the resultant GLCM is simply the sum of the number of times that the pixel with value i occurred in the specified spatial relationship to a pixel with value j in the input image. Energy, entropy, and homogeneity features are extracted from GLCM which could better represent the textural values. The energy also known as the angular second moment is the sum of the square values of all the elements in the calculated gray-level co-occurrence matrix. The nature of the gray-level co-occurrence matrix shows that the energy for the uniform image would be maximum and would reduce as the uniformity in the image decreases. It is calculated using the formula given in Eq. (3) as:

$$\text{Energy} = \sum_{i,j} p(i,j)^2 \quad (3)$$

Entropy is a statistical measure of randomness that can be used to characterize the texture of the input image. Entropy has been calculated as in Eq. (4) follows:

$$\text{Entropy} = -\sum (p \cdot \log_2(p)) \quad (4)$$

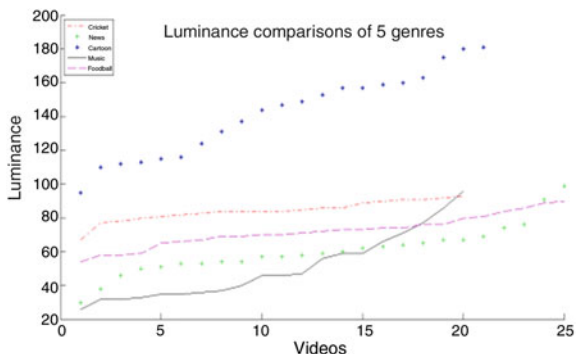
Homogeneity measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal. An image which is uniform has its values concentrated over the diagonal of the gray-level co-occurrence matrix. The homogeneity is calculated using the formula given in Eq. (5) as:

$$\text{Homogeneity} = \sum_{i,j} p(i,j)/(1 + |i - j|) \quad (5)$$

Brightness of the each key frame has been calculated and averaged over the entire video. The average brightness for a key frame is calculated using Eq. (6) as follows:

$$B = \mu_i \times \sigma_i^2 \quad (6)$$

Fig. 3 Luminance comparisons for various genres



where B is average brightness of the frame, μ_i is the mean of the pixels above a threshold, and σ_i^2 is variance of the pixels above a threshold. Here, energy, entropy, brightness, and homogeneity have been calculated for all videos. It has been observed that the brightness is comparatively low in cricket and football. Energy is greater in football than in cricket because football videos possess greater uniformity in their frames resulting in concentration of values in GLCM. Similarly, entropy is lesser in football than in cricket because cricket videos have high camera motion and less uniformity. The randomness of the pixel values is greater in cricket than football videos. Entropy comparison of cricket and football genres has been shown in Fig. 2. Here, cricket has been represented in blue and football in red color. The comparison of luminance values for videos of different genres is shown in Fig. 3.

3.2.4 Color Features

Color feature is another important feature used in video categorization. Most of the early systems [14, 15] have color as its main or at least a component of their feature set. This shows the importance of color features. In this system, three major color features have been extracted, namely luminance, pitch factor, and grass factor. Among the color descriptors, we retain color histogram which captures global color distribution in an image. Here, the color space selected is RGB space, since it is the most commonly used color space. Luminance of a pixel is calculated and represented as in the following Eq. (7):

$$Y = 0.3R + 0.59G + 0.11B \quad (7)$$

The video frames of different genre have been processed and the maximum and the minimum bin values of the frames are obtained. It has been found that the cartoon videos have a higher luminance when compared to the videos from other genres. The reason behind this is that the cartoon videos are luminant. The major steps involved in cartooning an image are luminance quantization and edge blurring [16]. Thus, this effect is taken into account in this system. The grass factor

calculates the amount of predominant green pixels contained in the image. Foot ball and cricket are kind of sports that are played in lush green grounds. So, the key frames extracted would contain a huge amount of green pixels. This can be extracted to separate these two genres from the rest of the genres. The green pixel can be identified in a RGB model by the dominance carried by the ‘g’-value of the pixel which is more than both ‘r’ and ‘b’ pixels. The pitch factor finds the presence of the cricket pitch in the key-frame. The pitch is characterized by the dominant ‘r’-value in the RGB representation of the image surrounded by the dominant g-values of the grass. This identification of the pitch helps in separating the football videos from the cricket videos, since no pitches exist in a foot ball field. The algorithm for detecting the pitch is as follows:

```

Input : Videos
Output : Keyframes to detect pitch
Algorithm:
function detect_pitch(frame)
Step 1: Identify the red dominant pixels in the frame.
Step 2 : Compute the centroid of the regions found.
Step 3 : if( any region's centroid is in the center of the ground)
        check if the area of the region is greater than threshold
            May be a pitch frame.
        else
            Not a pitch frame.
Step 4 : End
Step 5 : Identify green dominant pixels in the frame using detect_grass()
        if there is a layer of grass below the pitch
            Not a pitch frame
        else
Step 6 : if(green_count>45 % && pitch_count/green_count>20%)
        Pitch detected.
        else
            Grass value too low to distinguish Pitch.
Step 7 : End

```

A sample set of key frames with high grass content from the entire key frames is shown in Fig. 4 and the sample set of key frames with pitch from the entire key frames is shown in Fig. 5.

3.3 SVM Training and Classification

Once the features have been extracted, they are fed as input to the SVM classifier for genre classification. The model chosen here for training and classification of videos is the support vector machine (SVM). The local optimal SVM tree approach is used for training and classification of the videos. The news seems to be the most easily separable genre from the rest. Initially, we check for news video versus the rest. The feature set used for training SVM in step 1 are foreground motion, shot rate, shot size,

Fig. 4 Key frames to detect grass

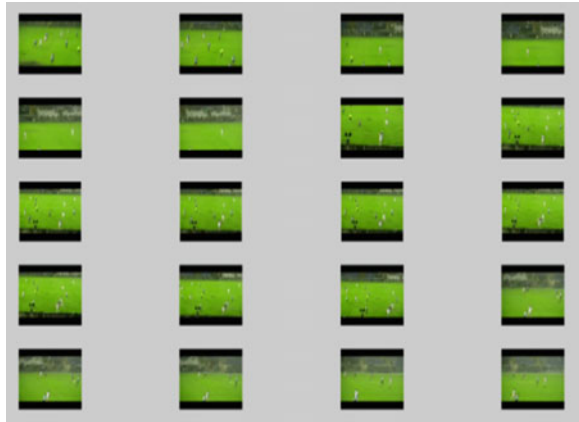
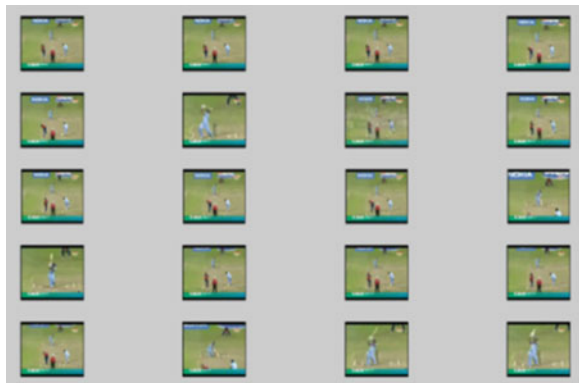


Fig. 5 Key frames to detect pitch



key frame rate, and homogeneity. The cartoon is the next genre to be classified from the rest of genres. This is made by making use of the luminance factor. In the step 3, we aim to separate music genre from the rest. Shot rate, shot size, key frame rate, brightness, and grass factor are used for training. Let the cricket genre is separated from others using the features pitch factor, energy, and entropy. Next to this, the football is compared with the others to separate football genre from the remaining genres and out of boundary conditions. The energy and entropy features are made use of in this step. The SVM-trained objects are stored and used for classification. During the testing phase, the preprocessing is done on the input video and all the eleven features are extracted from the key frames. The feature vector is given as input to the trained SVM in a five step processes are as follows:

1. Foreground motion, shot rate, shot size, key frame rate, and homogeneity features are given as input to the SVM. The SVM feature is lying to the training object of layer 1 for classification. If the classification output is positive, then news genre is confirmed, else step 2 is proceeded.

2. Luminance feature of input video is given as input to the SVM and classified against the SVM training object in layer 2. If positive classification is achieved, then cartoon genre is separated, else proceed to step 3.
3. Input video's shot rate, shot size, key frame rate, brightness, and grass factor features have been given as input to the SVM against SVM training object of layer 3. In case of positive classification, music genre is confirmed, else go to step 4.
4. Pitch factor, energy, and entropy features of input video are considered as input to the SVM against the SVM training object in layer 4. Cricket genre is confirmed if positive, else proceed to step 5.
5. The input videos energy and entropy features have been given as input to the SVM to check against the SVM training object in level 5. If positive, then football genre is confirmed else the input video may be probably out of boundary.

4 Results and Performance Analysis

The videos have been collected from various sources such as YouTube and Video. fy. A total of 350 videos have been processed. Among them 65 videos considered from news, 65 videos from cartoon, 70 videos from music, 85 videos from cricket, 65 videos from football, and 150 videos have been chosen randomly from all genres and used for training SVM. The Table 1 below shows the basis of confusion matrix. The basic metrics used for analyzing the performance are listed in Table 1: Apart from that, fall metrics have also used.

$$\text{Specificity} = \frac{\text{TN}}{(\text{TN} + \text{FP})} \quad (8)$$

$$\text{Sensitivity} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \quad (9)$$

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FP} + \text{TN} + \text{FN})} \quad (10)$$

$$\text{Hit Rate} = \frac{\text{No. of videos classified}}{\text{Total no. of videos to be classified}} \quad (11)$$

Table 1 Confusion matrix

		Predicted class	
		Yes	No
Actual prediction	Yes	True positive (TP)	False negative (FN)
	No	False positive (FP)	True negative (TN)

Table 2 Hit rate and miss rate

Genres	Hit rate (%)	Miss rate (%)
News	100	0
Cartoon	97	3
Music	97.5	2.5
Cricket	98	2
Football	99	1

Table 3 Specificity, sensitivity, and accuracy

Times	P	N	TP	FP	TN	FN	Specificity (%)	Sensitivity (%)	Accuracy (%)
News	30	50	30	0	50	0	100	100	100
Cartoon	30	60	29	3	57	1	95	96.67	95.56
Music	100	75	92	5	70	8	93.33	92	92.57
Cricket	65	50	63	3	47	2	94	96.92	95.65
Football	30	50	29	4	46	1	92	96.67	94.35

$$\text{Miss Rate} = \frac{\text{No. of videos not classified correctly}}{\text{Total no. of videos to be classified}} \tag{12}$$

The hit rate defines the number of correct predictions and miss rate defines the number of wrong predictions. The hit rate and miss rate for the five genres have been shown in Table 2.

The specificity, sensitivity, and accuracy for the five genres tested have been given in Table 3. The news videos are most efficiently classified, while football suffered significantly in classification.

Shots comparison for various methods such as the saturation-based method, the histogram-based, and pixel difference-based methods with respect to the hit rate/miss rate of shot detection are shown in the Table 4.

Table 4 Miss rate and hit rate of the video shots

Genre	Hit rate of video shots			Miss rate of video shots		
	Saturation based (%)	Histogram based (%)	Pixel difference based (%)	Saturation based (%)	Histogram based (%)	Pixel difference based (%)
News	99.79	98.07	98.46	0.03	0.05	0.10
Cartoon	99.74	98.14	97.55	0.3	0.5	0.54
Music	98.69	96.30	94.06	0.5	1.78	2.8
Cricket	99.32	97.04	98.43	0.47	0.68	0.56
Football	98.40	98.37	97.72	0.29	0.34	0.31

5 Conclusion and Future Work

In this work, the videos have been categorized into five genres namely, news, cartoon, music, cricket, and football. A total of 350 videos have been trained and 150 videos are tested properly. The videos have been collected from various sources such as YouTube. The videos have been first preprocessed and shots have been separated. From the key frames extracted, four major categories of features namely, temporal, motion, texture, and color features have been extracted. The eleven dimensional feature vectors have been given as input to the SVM with linear kernel for training and classification of genres from videos. The combination of the chosen feature vector improves the specificity, accuracy, and sensitivity of this system compared to the existing system [14]. This system achieved overall average correct detection ratio up to 98 % while maintaining very low false detection rate, namely 2 %. In future, this system could be extended to more number of genres and spatial-related values.

References

1. D. Gowsikhaa, A.S. Manjunath, Suspicious human activity detection from surveillance videos, *Int. J. Internet Distrib. Comput. Syst.* **2**(2), 141–149 (2012)
2. A. Hauptmann, R. Yan, Y. Qi, R. Jin, M. Christel, M. Derthick, M.-Y. Chen, R. Baron, W.-H. Lin, T.D. Ng, Video classification and retrieval with the informedia digital video library system, in *Text Retrieval Conference (TREC02)* (2002)
3. W. Zhu, C. Toklu, S.-P. Liou, Automatic news video segmentation and categorization based on closed-captioned text, in *IEEE International Conference on Multimedia and Expo (ICME 2001)* (IEEE, 2001), pp. 829–832
4. M. Roach, J. Mason, Classification of video genre using audio. *Eurospeech* **4**, 2693–2696 (2001)
5. S. Moncrieff, S. Venkatesh, C. Dorai, Horror film genre typing and scene labeling via audio analysis (2003)
6. M.J. Roach, J.S. Mason, M. Pawlewski, Motion-based classification of cartoons, in *Proceedings of the International Symposium on Intelligent Multimedia* (2001), pp. 146–149
7. M. Fakheri, M.C. Amirani, T. Sedghi, Gabor wavelets and GVF functions for feature extraction in efficient content based colour and texture images retrieval. (IEEE, 2011), pp. 1–5
8. D. Gowsikhaa, S. Abirami, R. Baskaran, Automated human behavior analysis from surveillance videos, in *Artificial Intelligence Review* (2012), pp. 1–19
9. G. Wei, L. Agnihotri, N. Dimitrova, TV program classification based on face and text processing, in *IEEE International Conference on Multimedia and Expo*, vol. 3 (IEEE, 2000), pp. 1345–1348
10. X. Yuan, W. Lai, T. Mei, X.-S. Hua, X.-Q. Wu, S. Li, Automatic video genre categorization using hierarchical SVM, in *IEEE International Conference on Image Processing (ICIP)* (IEEE, 2006), pp. 2905–2908
11. M. Mentzelopoulos, A. Psarrou, Key frame extraction algorithm using entropy difference, in *MIR'04* (2004)
12. D. Gowsikhaa, S. Abirami, R. Baskaran, Construction of image ontology using low level features for image retrieval, in *Proceedings of the International Conference on Computer Communication and Informatics (ICCCI 2012)* (2012), pp. 1–7

13. M.J. Roach, J.S. Mason, M. Pawlewski, Video genre classification using dynamics, in *the proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP '01)* (2001)
14. B. Ionescu, C. Rasche, L. Florea, C. Vertan, P. Lambert, Classifying documentary, music, news and animated genres with temporal, color and contour information (IEEE, 2011), pp. 1–4
15. M. Guironnet, et al., Video classification based on low-level feature fusion model. Author manuscript, published in *European Signal Processing Conference (EUSIPCO'2005)* (2005)
16. M.-H. Han, H.N. Song, K.-H. Yoon, Cartooning using reference images, in *Proceedings of 2010 IEEE 3rd International Conference on Human-Centric Computing* (2010), pp. 1–5