Content-based Automatic Video Genre Identification

Faryal Shamsi[1], Sher Muhammad Daudpota[2], Sarang Shaikh[3]
Department of Computer Science
Sukkur IBA University
Sukkur, Pakistan

***Abstract*—Video content is evolving enormously with the heavy usage of internet and social media websites. Proper searching and indexing of such video content is a major challenge. The existing video search potentially relies on the information provided by the user, such as video caption, description and subsequent comments on the video. In such case, if users provide insufficient or incorrect information about the video genre, the video may not be indexed correctly and ignored during search and retrieval. This paper proposes a mechanism to understand the contents of video and categorize it as Music Video, Talk Show, Movie/Drama, Animation and Sports. For video classification, the proposed system uses audio and visual features like audio signal energy, zero crossing rate, spectral flux from audio and shot boundary, scene count and actor motion from video. The system is tested on popular Hollywood, Bollywood and YouTube videos to give an accuracy of 96%.**

***Keywords*—Motion detection; scene detection; shot boundary detection; video genre identification**

I. Introduction

The word genre is defined as socially agreed category of content. So, the term content-based automatic video genre identification means, to recognize the category of a video on basis of its contents. The heterogeneous nature of video contents, makes the genre identification a challenging job. With the evolution of internet and social networking websites, content sharing is becoming a popular trend [1]. The level of facilitation provided to user by such websites leads to increase the information overload, while organization of the contents is becoming a challenging task [2]. The most popular form of content on social media is videos [3]. The nature of video contents is diverse as it combines all other types of media such as text, audio and image [4]. The top ranking social networking sites like Facebook, YouTube allow users to explore billions of videos per day. Proper organization of such videos is therefore a necessary operation to ensure efficient indexing and searching. In spite of all the progress in the field of multimedia mining and contents based filtering, still there is need of a system which can automatically understand the contents of a video. A reliable automatic video genre identification system which can categorize any type of video, is yet to be proposed.

The existing video indexing and searching mechanisms available are fully at mercy of the information provided by up-loader. On the other hand, an up-loader enjoys full autonomy while generating and sharing any type of contents. The up-loader is not bound to provide necessary information about the content so that it can be utilized for the purpose of indexing. Also, there is no check and balance to ensure the integrity of the information provided by the up-loader. For example, an up-loader has complete freedom to give any title to the video, no matter how irrespective it is, with the actual contents of video. An up-loader might give his or her own name to a movie or can caption a talk show as a movie, in such cases a user may not be able to view these videos if his/her search string does not match with the information available with the video. So, there must be an identification system for video genre which considers its contents rather than the textual information provided by the up-loader.

This paper proposes a frame work to analyze contents of an input video and classify it in one of the five genres including music video, talk-show, movie/drama, animation and sports. We use audio and visual features to achieve this classification task.

Rest of the paper is organized as follows: Section II presents the literature review, Section III introduces different building blocks of the system, Section IV explains working of the audio classifier along with features used to work with audio of the input video, followed by discussion on video classifier in Section V. Finally, Section VI and Section VII report the results and conclude paper with discussion about future directions, respectively.

II. Related Work

Several attempts have been made in the recent years regarding video classification. Some of these attempts used text or audio based approaches while most worked on visual based or hybrid approaches. The text based approaches use the view-able text within a video to understand its contents [5]. A movie generally has textual information such as cast and subtitles which can give some insights about the video contents. Techniques extracting the text and TF-IDF (Term Frequency-Inverse Document Frequency) is discussed in [6].

Other than the text, audio contents within a video can also be used for video classification and genre identification. Various techniques are proposed in literature which uses audio information, like dialogues music or silence [7], [8].This information can be captured through ZCR - Zero crossing rate as proposed in [9], [10], [11], [12] for video content analysis. Similar type of audio classification algorithms are proposed in [13], [14], using HZCR (high zero crossing rate), spectrum flux and time energy of an audio signal to classify an audio into four distinct classes - speech, music, environment sound or silence. Similar type of algorithm is proposed in [15] to extract musical sequences, in Bollywood movies, and uses ZCR and RMS (Root-Mean-Square) of an audio signal as classifying features.

Although a video has several textual and audio features that can be used for classification but among all these features,

visual features are most dominant within the video contents. The colors (i.e. RGB values) [16], shapes (i.e. Image histogram) [17], [18], luminescence [19] and motion [20] are some of the commonly used visual features used to analyze the contents of video, as proposed in literature. The color and shape information can be statistically analyzed by generating the color histogram and shape histogram of images extracted from the frames of video [21]. MPEG motion vectors [22], [23] are used as features to examine the motion information available in a video. HMM - Hidden Markov Model combines various features (e.g. color, shape and audio) and the Gaussian Mixture Model the probability distribution and many similar techniques use hybrid approaches for video classification [24], [25], [26], [27].

The features used for video classification by the above methods are low level. There are also further high level features for video classification e.g. Shot Boundary Detection, Scene Detection [10], Shot Duration [28], [29], [30], Motion Detection [10], [29], [31] and so on. Shot Boundary detection is a major building block for video processing activities. A shot boundary is defined as one or more frames in a video where one shot ends and other starts. For the purpose of shot boundary detection various audio and visual features can be used such as colors, edges, luminescence and motion information available within the frames [32]. One of the benchmark data-set for shot boundary detection is TRECVid, developed by NIST (National Institute of Standards and Technology in Gaithersburg, USA). TRECVid contains many subsets of video test data and ground truth information targeting the shot boundary detection which are widely used to evaluate the performance of different shot boundary detection algorithms [33]. Significant advancements are made by [34], [35], [36], [37], [38] [39], [40] to perform shot boundary detection to achieve optimal accuracy. Different accuracy ranges have been reported on TRECVid dataset from 94% to 96%.

Our proposed system use an algorithm inspired by [34] that attempts to identify the similarity between each consecutive frames. The algorithm uses the SAD function that calculates the sum of absolute differences between the RGB color values of the corresponding pixels of each two adjacent frames. To identify the shot boundary, the SAD value is compared to a predefined threshold. If the difference is greater than that threshold the later frame will be considered as shot boundary, otherwise as a continuous shot. The algorithm can be summarized as:

$$d_{sad}(f_{i-1}, f_i) = \frac{1}{|F|} \sum_{i=2}^{F} |f_{i-1}(r) - f_i| \quad (1)$$

Where F is the total number of pixel in a frame, d_{sad} is the function calculating the sum of absolute differences of each two consecutive frames and r represents the coordinates of pixels. The above algorithm is able to categorize the shots as cut, fades and dissolves by using the luminescence information.

For content based analysis, a video can also be broken down into scenes rather than shots. Although some authors refer the terms shot and scene interchangeably, most consider shots as a subset of scene. A scene is a temporal segment of a video with repetitive shots. A scene may also be defined as continuous action with a specific event of a video. Scene detection is the process to automatically detect this repetitive pattern within a video. Scene detection is generally performed by using visual features like luminescence, motion detection and average length of shots to track the changes in lighting, environment and pace of the video in a movie [4]. Scene detection techniques may be rule based, where some predefined rule are constructed to analyze the frames and shots and decision is made if they belong to the same scene or not [5]. Some scene detection techniques may be graph based [41], stochastic based or cluster based [42], [43], [44] in contrast to the rule based [5], [45]. The accuracy of such algorithms range from 80% to 86% as reported these different attempts.

III. Video Genre Classification

The proposed system is able to classify a video into five distinct classes, including Music Video, Talk Show, Sports, Animation and Movie/Drama. The abstract model of the system is given in Fig. 1.

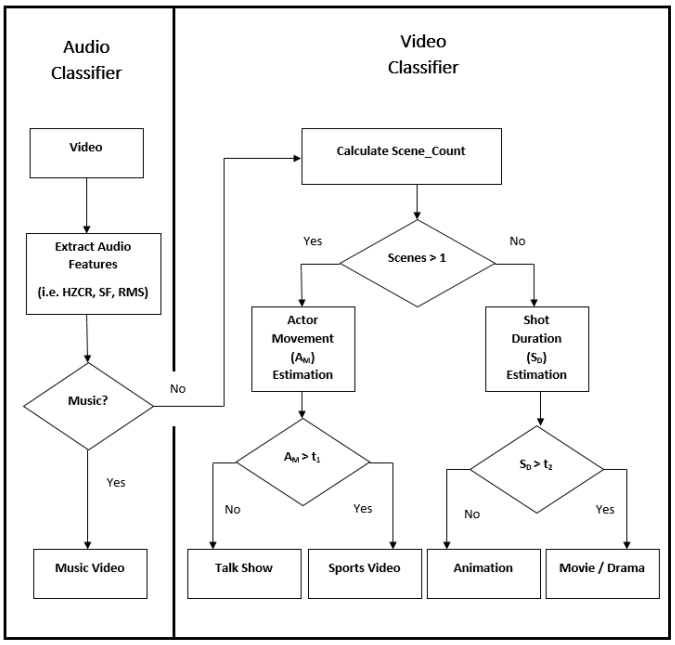

Fig. 1. Abstract Model of the Proposed System

The system is logically divided in two components, Audio Classifier and Video Classifier. The input video is fed to the Audio Classifier that extracts audio features to estimate music portion in the audio of input video. In case if majority of the frames are classified as having background music, the input video is classified as Music Video, else it is forwarded to the Video Classifier to check for the remaining genres of video.

The Video Classifier checks for the number of scenes in a video because it's a good indicator to separate Talk Show and Sports Video from Movies/Drama and Animations. As per video recording grammar, Talk Shows and Sports videos are comprising of only one scene i.e. for the whole duration of the video, all the shots belong to same scene, whereas for Movies/Drama and Animation there are multiple scenes

depending on the size of video file.

In case if there is only one scene in a video, we check for actor movement. If the actor movement is high with only one scene in the video, it is classified as a Talk Show, whereas, high actor movement results in a video being classified in Sport Video genre.

On the other hand, if there are multiple scenes in a video, we estimate shot duration in the input video. Usually, in order to increase excitement in Animation video, shot duration is kept very small, typically less than a second in many cases. Therefore, shot duration feature efficiently classifies input video in either Movie/Drama or Animation.

Rest of the paper describes working of Audio and Video Classifiers along with each components in both the classifiers.

IV. Audio Classifier

Classifying an audio input in music and non-music classes is an important task, many attempts have been made to use different audio features to accomplish this task [46], [47], [48], [49], [50], [51], [52], [53]. Lie Lu et al. [13] demonstrate that features like zero crossing rate, root mean square and spectrum flux are good at differentiating music signal from speech signal, however, the variations of these features in a defined window performs even better than the feature itself. They define high zero crossing rate ratio (HZCRR) as:

$$HZCRR = \frac{1}{2N}\sum_{n=0}^{N-1}[sgn(ZCR(n) - 1.5avZCR) + 1] \quad (2)$$

where $avZCR$ is defined as:

$$avZCR = \frac{1}{N}\sum_{n=0}^{N-1} ZCR(n) \quad (3)$$

Where N represents total number of frames in a one second window, n is frame index, $sgn[]$ is sign function and $ZCR(n)$ is zero crossing rate of n^{th} frame in one second window.

Zero crossings in music signal are usually very low with almost zero variations, whereas in speech signal, due to sudden drop in energy, zero crossing variations are significantly high.

Similarly, Low Short Time Energy Ratio (LSTER) is defined as:

$$LSTER = \frac{1}{N}\sum_{n=0}^{N-1}[0.5avSTE - sgn(STE(n)) + 1] \quad (4)$$

and,

$$avSTE = \frac{1}{N}\sum_{n=0}^{N-1} STE(n) \quad (5)$$

Experiments suggest LSTER value is significantly lower in music signals compared to speech signal, therefore it performs well in differentiating between music and speech signal.

Finally, Spectrum Flux (SF), which defines average variation value of spectrum between two adjacent frames in one second window is defined as:

$$SF = \frac{1}{(N-1)(K-1)}\sum_{n=1}^{N-1}\sum_{k=1}^{K-1} [log(A(n,k)+\delta) - log(A(n-k,k)+\delta]^2 \quad (6)$$

Where $A(n,k)$ represents Discrete Fourier Transform of n^{th} frame in a one-second window and defined as:

$$A(n,k) = \sum_{m=-\infty}^{\infty} x(m)W(nL-m)e^{j\frac{2\pi}{L}km} \quad (7)$$

Spectrum Flux values for speech are much higher than Music, therefore when combined with HZCRR and LSTER, it performs well for differentiating music from speech using SVM classifier.

In our task of genre identification, as we observe more seconds of music than speech, we classify input video in Music class. On the other hand, if there is low percentage of music seconds in the input video, the video is forwarded to Video Classifier for further processing.

V. Video Classifier

This section of our genre identifier receives only those videos in which music portion is low. As per grammar of video recording, which although is not documented but if violated would result in an ambiguous video, all the video genres, except music video contains lower portion of music than pure speech or environment sound. For example, imagine a talk show discussing on recent political affair with a loud background music, obviously, such a recording would not make sense for the viewers. A non-music video can belong to any of the later four genres including Talk-Show, Sports, Animation and Movie/Drama.

A Talk-Show video generally has only one scene where there are 5-7 distinct shots frequently repeating. There is a shot of anchor person. Then, there are separate shots of each of the invited guests. There may be some shots showing more than one people at a time. There is also a shot of showing the environment from a distance. Another common observation in Talk-Show is that the actor movement is very low, as people are generally sitting and rarely moving during the discussion.

Similarly, a Sports video also usually has only one scene but the factor which can easily distinguishes between a Talk-Show and a Sports video is the actor movement. Unlike a Talk-Show, in a sports video actor movement is significantly high. Fig. 2 and Fig. 3 illustrates the patterns observed in Talk-Shows and Sports videos, as described above.

In order to differentiate Movie/Drama and Animation from Sports and Talk-Show videos, we use number of scenes feature. As discussed earlier, in Sports and Talk-Show videos, generally, there is only one scene, whereas in Movie/Drama and Animation, the number of scenes are many, depending on the video length.

Fig. 2. Consecutive frames in a Talk-Show, demonstrating only one scene and low actor movement.

Fig. 3. Consecutive frames in a sports video demonstrating only one scene and high actor movement.

Fig. 4. Frames in a Movie, demonstrating more than 1 scene and long shot duration.

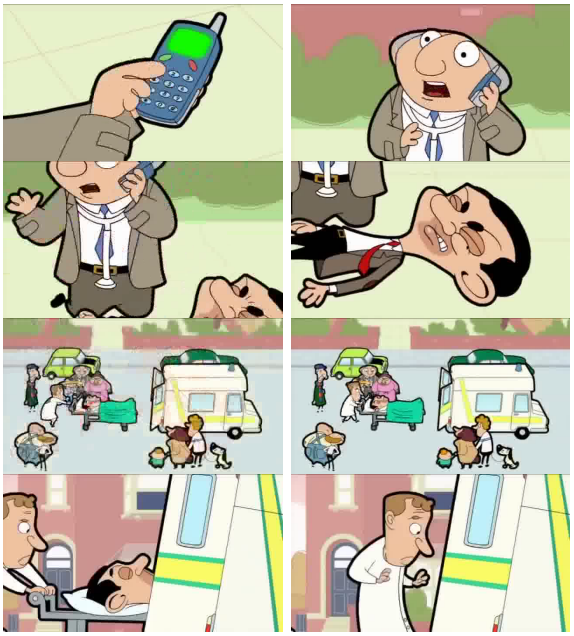

Fig. 5. Frames in an Animation, demonstrating more than 1 scene and small shot duration.

Finally, if input video comprises of many scenes, there is a possibility that the input video is either a Movie/Drama or an Animation. In order to differentiate between these two video genres, we use shot duration. It has been observed that shot duration is high when input video is a Movie/Drama, whereas the value of shot duration is low, typically less than a second for many shots, when input video is an Animation. Fig. 4 and Fig. 5 show consecutive frames from two popular Bollywood movies and two animations. It can be observed that all the frames in movies are belonging to same shot, whereas in animation, all the frames are formulating different shots.

Table I summarizes differences in different video features across all four genres.

In the next section, we describe different components of video classifier including shot detection, scene detection and actor movement.

A. Shot Boundary Detection

There are many attempts in the literature to detect shot boundary in video [33], [35]–[40]. Anil k. Jain et al. [54]

TABLE I. VIDEO GENRE IDENTIFICATION FEATURE VALUES

Video Genre	Scene Count	Actor Movement	Shot Duration
Talk Show	Low	Low	High
Sports	Low	High	Low
Drama/Movie	High	Low	High
Animation	High	High	Low

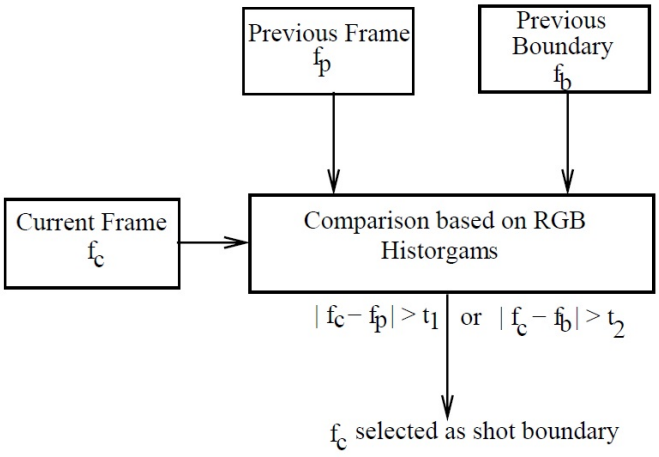

Fig. 6. Shot Boundary Detection [54]

proposed a method of shot boundary detection that is presented in Fig. 6.

A frame is defined as shot boundary if it is significantly different from either the previous frame or previous shot boundary. The different here refers to RGB histogram difference. One of the benchmark data-set for shot boundary detection is TRECVid, developed by NIST (National Institute of Standards and Technology in Gaithersburg, USA). TRECVid contains many subsets of video test data and ground truth information targeting the shot boundary detection which are widely used to evaluate the performance of different shot boundary detection algorithms [33].

One of the major issues in most of the shot detection approaches found in literature is requirement of learning threshold values for detecting shot boundary. Our experiments suggest that threshold values learned on a data-set do not perform well on other data-sets. For example, the threshold value for sum of absolute difference based on RGB color at pixel level learned on TRECVid 2007, as proposed by [34], does not perform well when used to detect shot boundaries from Bollywood movies. This might be because of different intensities in color values in two completely different data sets.

Inability of threshold values to make them independent of data-sets results in need of an algorithm that learns threshold values dynamically from the data set. Our approach of shot detection is based on clustering that eliminates the need of learning and applying static threshold values.

Suppose, we have a video V defined as set of n frames,

$$V = \{f_1, f_2, \cdots, f_n\} \tag{8}$$

Simple sum of absolute difference d_{sad} between f_{i-1} and f_i is defined as [34]:

$$d_{sad}(f_{i-1}, f_i) = \frac{1}{|F|} \sum_{r \in F} |f_{i-1}(r) - f_i(r)| \tag{9}$$

Were F is the total number of pixel in a frame, d_{sad} is the function calculating the sum of absolute differences of each two consecutive frames and r represents the coordinates of pixels. A shot boundary is detected if,

$$d_{sad} \geq T \tag{10}$$

Where T is a threshold value learned through experiments. This approach of threshold value was tested on TRECVid 2007 and the precision and recall was reported to be 94.4% and 90.5%, respectively. In our approach, we use a set D consisting of d_{sad} for all the frames in a video V,

$$D = \{d_{sad_1}, d_{sad_2} \cdots d_{sad_{n-1}}\} \tag{11}$$

Our approach is based on two simple assumptions:

1) If two consecutive frames f_i and f_{i-1} belong to same shot, their d_{sad} would be relatively low
2) Consequently, if both the frames, f_i and f_{i-1} belong to two different shots, that is $fi+1$ is actually representing a shot boundary, d_{sad} would be high

Based on above two assumptions, different approaches in the literature have learned threshold values. We believe that learning threshold values could be avoided if we employ a clustering approach. We use simple K-Means clustering algorithm which is defined by a centroid for each cluster. Any point is assigned to its nearest centroid based on Euclidean distance. In our case, the value of $K = 2$ which means two clusters. d_{sad} values in set D suggest that all the points are belonging to one of the two obvious clusters, suppose C is a set of centroids defined as:

$$C = \{c_i, c_j\} \tag{12}$$

Then each data point d_i in the set D is assigned to a cluster based on its squared distance from center c_i. Data suggests, the difference values fall in two distinct clustering categories:

1) The difference values are too low if consecutive frames are belonging to same shot.
2) The difference values are reasonably high if consecutive frames are belonging to two different shots.

Interestingly, in a video, non-representative frames to shot boundary are lower than the number of frames representing shot boundaries, thus, the cluster which holds lower number of frames is actually holding the shot boundaries. The advantage of our shot boundary detection over existing approaches is to avoid the need of learning threshold values. In a way, clustering mechanism is actually a dynamic threshold learning mechanism.

1) Shot Detection: Post-Processing Step for Increasing System Precision: Clustering based approach discussed in previous section, produces high recall values on TRECVID-2007, however, precision value is only 38%. It has been observed that simple clustering based algorithm is unable to distinguish well between two consecutive frames which are part of same shots but having high movement of high level objects like person, car, animal, etc. The absolute sum of difference as defined in previous section is high when there is significant movement between two consecutive frames belonging to same shot. In order to improve the system precision, we use RGB histogram values. No matter how much is the movement from one frame to next, RGB histogram remains almost same given both frames are belonging to same shot. We use this simple technique to eliminate false positives that are incorrectly clustered in the cluster which should hold shot boundary frames. With simple clustering algorithms for shot boundary detection followed by post-processing step, we achieved recall and precision of 97% and 92%, respectively on TRECVID-2007.

B. Scene Detection

A scene is a temporal fragment of a video where shots are frequently repetitive. For example, in a talk-show. most of the shots are part of same scene. If observed closely, these shots are part of a frequently repetitive pattern. Same is the case with sports videos, where similar shots tend to repeat themselves and overall environment remains same. Contrasting, in case of a movie, drama or animation the repetition of same shot is not demonstrated throughout the video. Rather, this repetition is merely confined within a specific fragment of that video. A scene can be detected just by capturing this repetition of shots by using the Algorithm 1.

C. Actor Movement Estimation

For video genre identification, actor movement is a strong predictor to classify between a talk-show and a sports video. In a talk-show the people and objects are generally stationary, while in a sports video, people and objects are continuously in moving state. The procedure followed to capture the actor movement is illustrated in Fig. 7. Here, two consecutive frames of video are broken down into 8x8 grid, resulting in 64 sub-images within each frame. The corresponding sub-images of both frames are compared. If there is low actor movement, many regions of the frame will remain same in both frames. On the other hand, if there is high actor movement, the sub-images will demonstrate significant level of dissimilarity.

Mathematically, actor movement A_m is defined as:

$$A_M = \frac{1}{n}\sum_{i=1}^{n} M_i \tag{13}$$

Where M represent average sum of difference between all the regions of a given $Frame_i$ to its corresponding regions in $Frame_j$ and defined as:

$$M = \sum_{i=1}^{64} |Frame_i - Frame_{i-1}| \tag{14}$$

Algorithm 1 Scene Boundaries Detection Algorithm

INPUT : Set of Shot Boundaries S_h
OUTPUT : Set of Scene Boundaries S_c

```
current ← 1, i ← 0, match ← 0
while current >= count(Sh) do
  current ← current + 2
  i ← current
  while current >= i do
    if Sh[current] == Sh[i − 2] then
      match ← 1
      break
    else
      match ← 0
      i − −
    end if
  end while
  if match ← 0 then
    while (current − 1) >= (i − 3) do
      if Sh[current − 1] == Sh[i − 3] then
        match ← 1
        Sc ← current
        break
      else
        match ← 0
        i − −
      end if
    end while
    if match ← 0 then
      Sc ← current − 1
    end if
  end if
end while
```

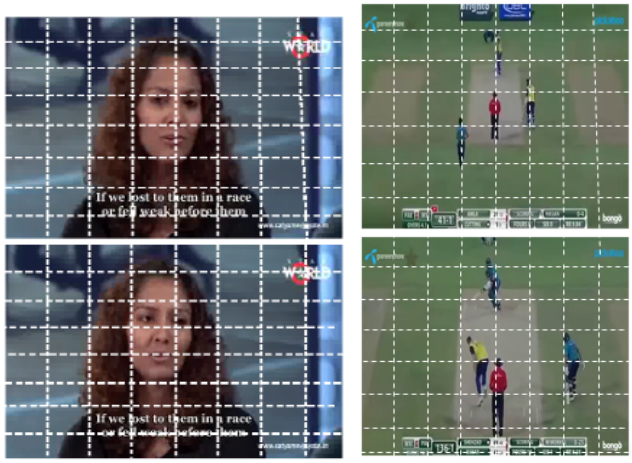

Fig. 7. Consecutive frames of 2 videos divided into 8x8 grid - on left are frames of a Talk Show (demonstrating low Actor Movement) and on right from a Sports Video (demonstrating high Actor Movement)

D. Shot Duration Estimation

To differentiate between a movie or drama and an animation, shot duration is used as a classifying feature. In a movie or drama, the shot duration is generally longer as compared to an animation. The proposed system performs shot boundary detection by using the cluster based approach as discussed earlier. The shot duration of i^{th} shot is defined as:

$$T_i = F_n - F_1 \tag{15}$$

Where F_n is last frame in a given shot and F_1 is the first frame. The average shot duration in a video, that we use as feature to differentiate between movie/drama and animation genres, is defined as:

$$S_D = \frac{1}{n}\sum_{i=1}^{n} T_i \tag{16}$$

VI. Experimental Results

The proposed system considerably depends on the intermediate procedures like shot detection, scene detection, motion detection and shot duration. Unfortunately, there is no benchmark data-set available with ground truth values recorded already. Therefore, for this study, a data-set comprising of 50 videos from the five desired genres (i.e. Music Video, Sports Video, Talk Show, Animation and Movie/Drama) was constructed. The number of scenes, and shot duration were observed manually to setup ground truth before applying the scene detection, motion detection and shot duration estimation. The manually calculated values were then compared to the values calculated by the system to evaluate its performance. The videos were selected carefully to avoid bias in the results. The data-set included 10 videos from each genre.

For music video, there is much signal variation even in case of pop, jazz and rock music and old classical songs or ballads. Thus the music videos were carefully selected to cover most of the possible classes. In this regard, the developed data-set included popular songs from Bollywood Movies and some international music channels from different categories.

The 10 sports videos in the data-set, all belonged to a distinct sports category: 1. Cricket, 2. Football, 3. Hockey, 4. Car Race, 5. Table Tennis, 6. Basketball, 7. Badminton, 8. Swimming, 9. Snooker and 10. Wrestling.

The 10 talk shows consisted 5 Pakistani shows, 3 talk shows from Indian TV channels and 2 American Talk shows. Similarly, the movies were selected both from Hollywood and Bollywood Industry, while the selected TV serials were taken from Pakistani, Indian as well as Turkish TV channels. The movies and dramas belonged to different genres like comedy, tragedy, romantic and action to cover the major variations of shot duration and camera movement. Also, the quality of graphics has drastically improved in past few decades, therefore to cover animations of all types, these were taken from old cartoons movies like pink panther to latest release of animated games like Assassin Creed and other animated movies in different foreign languages.

Table II shows the list of videos we have used in our data-set. The proposed system accurately predicts genre of 96% videos, the detailed results of genre identification are presented in confusion matrix of Table III. The results of scene detection are shown in Table IV.

TABLE II. Selected Videos in User defined Data-set for Genre Identification

Video	Genre
Capital Talk 29th May 2018	Talk Show
Capital Talk 31st May 2018	Talk Show
Live with Kelly 23rd April 2018	Talk Show
Live with Kelly 21st May 2018	Talk Show
Live with Kelly Priyanka	Talk Show
A+ Morning Show Hamid Mir	Talk Show
News Eye 29th May 2018	Talk Show
Coffee with Karan (Juhi & Madhuri)	Talk Show
Satyamev Jayete (Dangal)	Talk Show
Comedy Nights with Kapil	Talk Show
Cricket	Sports
Football	Sports
Tennis	Sports
Car Race	Sports
Badminton	Sports
Hockey	Sports
Snooker	Sports
Wrestling	Sports
Basketball	Sports
Swimming	Sports
Big Buck Bunny	Animation
Dreamworld - Ice Age	Animation
Memorable Moments - Rio	Animation
Tom and Jerry (Little Quacker)	Animation
Ferdinand - Escena Final Español	Animation0n
Mr. Bean -Nurse	Animation
JAN on See TV	Animation
Courage - The cowardly dog	Animation
Pink Panther	Animation
Assassin's Creed Origin	Animation
Secret Super Star Part 1/6	Movie/Drama
Andaz Apna Apna	Movie/Drama
Dhoop Kinare	Movie/Drama
Bulbulay	Movie/Drama
Diyar e Dil	Movie/Drama
Yaqeen Ka Safar	Movie/Drama
Dhamaal	Movie/Drama
Harry Potter	Movie/Drama
CID	Movie/Drama
Kosem Sultan	Movie/Drama
Everything I do, I do it for you	Music Video
I am Alive, Stuart Little 2	Music Video
It's my life, Bon Jovi	Music Video
Kung fu fighting, Kung fu Panda	Music Video

Continued on next page

Continued from previous page

Video	Genre
Allah hi dega, Ubaid Rana	Music Video
Sultan, Title Song	Music Video
Challa, Jab tak hai jan	Music Video
Hanikarak Bapu - Dangal	Music Video
Khamaj, Shafqat Amanat Ali	Music Video
Tere bina zindagi, Andhi	Music Video

VII. Conclusion

Different video genres have different recording rules, never documented but practiced so extensively that if violated, would result in a poor video. For example, imagine a movie action scene with high shot duration or low actor movement. Obviously, such videos won't inspire users. In this paper, we exploit these widely practiced recording rules for different genre videos and classify an input video in one of five genres including music, movie/drama, sports, talk-show and animation. Broadly, we use three visual features to achieve this classification task - 1) Shot Duration 2) Scene Count 3) Actor Movement along with three audio features including 1) High Zero Crossings Rate Ratio, 2) Spectrum Flux and 3) Low Short Time Energy Ratio. The audio features are used to separate music video from rest of the genres whereas scene count separates talk-show and sports video from movie/drama and animation. Finally, actor movement differentiates between talk-show and sports, and shot duration separates animation from movie/drama genre. We achieved classification accuracy of 96% on genre identification. Our shot boundary detection technique gives a precision and recall of 93.12% and 86.24%, respectively on TRECVID 2007 data-set. The advantage of our shot detection approach is elimination of need to learn threshold value for shot detection that varies in different data-sets.

VIII. Future Work

The current video genre identification system can further be extended to identify sub-genres and sub-categories. A music video can be classified as ballad, classic, rock, pop and so on. Literature suggest many approaches of music genre identification based on music audio, whereas there is a possibility to classify musical video in any particular category. Similarly, by applying different speech recognition techniques a Talk show can be further classified as political, business, or entertainment categories. By applying shape identification techniques on objects within a video, the type of sports can also be identified in future. The shape and size of ball differs from one sports category to another. Presence of bat, racket, boundary rope or hockey sticks can also reveal more about the sports category. The current system classifies movie and drama as a same category, a future system can differentiate even between these two categories by assessing video size. Movie and dramas can also be sub divided into popular genres like comedy, action, tragic or romantic. An animation can be differentiated with a gaming video by using low level features like frame rate and motion rate, etc.

References

[1] S. Asur and B. A. Huberman, "Predicting the future with social media," in *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, pp. 492–499, IEEE Computer Society, 2010.

[2] J. Tang, Y. Chang, and H. Liu, "Mining social media with social theories: a survey," *ACM SIGKDD Explorations Newsletter*, vol. 15, no. 2, pp. 20–29, 2014.

[3] D. H. Park, I. Y. Choi, H. K. Kim, and J. K. Kim, "A review and classification of recommender systems research," *International Proceedings of Economics Development & Research*, vol. 5, no. 1, p. 290, 2011.

[4] D. Brezeale and D. J. Cook, "Automatic video classification: A survey of the literature," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 38, no. 3, pp. 416–430, 2008.

[5] S. Wang and Q. Ji, "Video affective content analysis: a survey of state of the art methods," *IEEE Transactions on Affective Computing*, no. 1, pp. 1–1, 2015.

[6] A. G. Hauptmann, R. Yan, Y. Qi, R. Jin, M. G. Christel, M. Derthick, M.-y. Chen, R. V. Baron, W.-H. Lin, and T. D. Ng, "Video classification and retrieval with the informedia digital video library system.," in *TREC*, 2002.

[7] U. Srinivasan, S. Pfeiffer, S. Nepal, M. Lee, L. Gu, and S. Barrass, "A survey of mpeg-1 audio, video and semantic analysis techniques," *Multimedia Tools and Applications*, vol. 27, no. 1, pp. 105–141, 2005.

[8] G. Lu, "Indexing and retrieval of audio: A survey," *Multimedia Tools and Applications*, vol. 15, no. 3, pp. 269–290, 2001.

[9] S. M. Doudpota and S. Guha, "Mining movies to extract song sequences," in *Proceedings of the Eleventh International Workshop on Multimedia Data Mining*, p. 2, ACM, 2011.

[10] L. Canini, S. Benini, and R. Leonardi, "Affective recommendation of movies based on selected connotative features," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 4, pp. 636–647, 2013.

[11] J. H. French, "Automatic affective video indexing: Sound energy and object motion correlation discovery," in *Southeastcon, 2012 Proceedings of IEEE*, pp. 1–6, IEEE, 2012.

[12] M. Xu, J. Wang, X. He, J. S. Jin, S. Luo, and H. Lu, "A three-level framework for affective content analysis and its case studies," *Multimedia tools and applications*, vol. 70, no. 2, pp. 757–779, 2014.

[13] L. Lu, H. Jiang, and H. Zhang, "A robust audio classification and segmentation method," in *Proceedings of the ninth ACM international conference on Multimedia*, pp. 203–211, ACM, 2001.

[14] X. Ding, B. Li, W. Hu, W. Xiong, and Z. Wang, "Horror video scene recognition based on multi-view multi-instance learning," in *Asian Conference on Computer Vision*, pp. 599–610, Springer, 2012.

[15] S. M. Doudpota, S. Guha, and J. Baber, "Shot-based genre identification in musicals," in *Wireless Networks and Computational Intelligence*, pp. 129–138, Springer, 2012.

[16] Z. Cernekova, C. Kotropoulos, and I. Pitas, "Video shot segmentation using singular value decomposition," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 3, pp. III–181, IEEE, 2003.

[17] R. W. Lienhart, "Reliable dissolve detection," in *Storage and Retrieval for Media Databases 2001*, vol. 4315, pp. 219–231, International Society for Optics and Photonics, 2001.

[18] D. Lelescu and D. Schonfeld, "Statistical sequential analysis for real-time video scene change detection on compressed multimedia bitstream," *IEEE Transactions on Multimedia*, vol. 5, no. 1, pp. 106–117, 2003.

[19] J. Yu and M. D. Srinath, "An efficient method for scene cut detection," *Pattern Recognition Letters*, vol. 22, no. 13, pp. 1379–1391, 2001.

[20] B. Lehane, N. E. O'Connor, and N. Murphy, "Action sequence detection in motion pictures.," in *EWIMT*, Citeseer, 2004.

[21] J. Mas and G. Fernandez, "Video shot boundary detection based on color histogram," *Notebook Papers TRECVID2003, Gaithersburg, Maryland, NIST*, vol. 15, 2003.

TABLE III. VIDEO GENRE IDENTIFICATION RESULTS

		Predicted				
		Music	Talk Show	Sport	Movie/Drama	Animation
Actual	Music	10	0	0	0	0
	Talk Show	0	9	0	1	0
	Sports	0	0	10	0	0
	Movie/Drama	0	0	0	10	0
	Animation	0	0	0	1	9

TABLE IV. RESULTS OF SCENE DETECTION ON 40 VIDEOS OF USER DEFINED DATA-SET, (AS. SCENE DETECTION NOT APPLIED ON 10 MUSIC VIDEOS)

Video Name	Actual Scene Count	Predicted Scene Count	Precision	Recall
Capital Talk (Video 1)	1	1	100	100
Capital Talk (Video 2)	1	1	100	100
Live with Kelly (Video 1)	4	9	45	100
Live with Kelly (Video 2)	12	18	50	82
Live with Kelly (Video 3)	1	1	100	100
A+ Morning Show	1	1	100	100
News Eye (Dawn News)	1	1	100	100
Coffee with Karan	1	1	100	100
Satyamev Jayete	1	1	100	100
Comedy Nights with Kapil	1	1	100	100
Big Buck Bunny	5	5	100	100
Dreamworld - Ice Age	19	20	83	74
Memorable Moments - Rio	22	30	60	77
Tom and Jerry	14	17	76	93
Ferdinand	8	6	100	75
Mr. Bean	12	18	60	100
JAN on See TV	10	8	84	72
Courage - The cowardly dog	20	24	83	98
Pink Panther	11	10	90	82
Assassin Creed Origin	6	6	100	100
Secret Super Star, Bollywood	120	108	75	60
Andaz Apna Apna, Bollywood	40	12	90	41
Bulbulay, ARY Digital	18	35	52	100
Diyar e Dil, Hum TV	31	26	73	55
Yaqeen Ka Safar, Hum TV	38	26	92	70
Dhoop Kinare, PTV Classic	39	32	50	81
Dhamaal, Bollywood	53	102	55	100
Harry Potter, Hollywood	57	113	73	100
CID, Sony Entertainment	27	36	60	78
Kosem Sultan, Turkish	20	26	70	900
Cricket Video	1	1	100	100
Football Video	1	1	100	100
Tennis Video	1	1	100	100
Car Race Video	1	1	100	100
Badminton Video	1	1	100	100
Hockey Video	1	1	100	100
Snooker Video	1	1	100	100
Wrestling Video	1	1	100	100
Basketball Video	1	1	100	100
Swimming Video	1	1	100	100
Overall	504	615	85.5	90.7

[22] V. Kobla, D. S. Doermann, K.-I. Lin, and C. Faloutsos, "Compressed-domain video indexing techniques using dct and motion vector information in mpeg video," in *Storage and Retrieval for Image and Video Databases V*, vol. 3022, pp. 200–212, International Society for Optics and Photonics, 1997.

[23] H. Wang, A. Divakaran, A. Vetro, S.-F. Chang, and H. Sun, "Survey of compressed-domain features used in audio-visual indexing and analysis," *Journal of Visual Communication and Image Representation*, vol. 14, no. 2, pp. 150–183, 2003.

[24] D. PEH, "Ro duda, pe hart, and dg stork, pattern classification, new york: John wiley & sons, 2001, pp. xx+ 654, isbn: 0-471-05669-3," *Journal of Classification*, vol. 24, no. 2, pp. 305–307, 2007.

[25] R. O. Duda, P. E. Hart, D. G. Stork, *et al.*, "Pattern classification. 2nd," *Edition. New York*, vol. 55, 2001.

[26] C. M. Bishop, "Pattern recognition and machine learning (information science and statistics) springer-verlag new york," *Inc. Secaucus, NJ, USA*, 2006.

[27] S. Fischer, R. Lienhart, and W. Effelsberg, "Automatic recognition of film genres," *Technical reports*, vol. 95, 1995.

[28] A. Yazdani, E. Skodras, N. Fakotakis, and T. Ebrahimi, "Multimedia content analysis for emotional characterization of music video clips," *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, p. 26, 2013.

[29] M. Xu, C. Xu, X. He, J. S. Jin, S. Luo, and Y. Rui, "Hierarchical affective content analysis in arousal and valence dimensions," *Signal Processing*, vol. 93, no. 8, pp. 2140–2150, 2013.

[30] R. M. A. Teixeira, T. Yamasaki, and K. Aizawa, "Determination of emotional content of video clips by low-level audiovisual features," *Multimedia Tools and Applications*, vol. 61, no. 1, pp. 21–49, 2012.

[31] Y. Cui, S. Luo, Q. Tian, S. Zhang, Y. Peng, L. Jiang, and J. S. Jin, "Mutual information-based emotion recognition," in *The Era of Interactive Media*, pp. 471–479, Springer, 2013.

[32] C. Cotsaces, N. Nikolaidis, and I. Pitas, "Video shot boundary detection and condensed representation: a review," *IEEE signal processing magazine*, vol. 23, no. 2, pp. 28–37, 2006.

[33] A. F. Smeaton, P. Over, and A. R. Doherty, "Video shot boundary

detection: Seven years of trecvid activity," *Computer Vision and Image Understanding*, vol. 114, no. 4, pp. 411–418, 2010.

[34] Y. Kawai, H. Sumiyoshi, and N. Yagi, "Shot boundary detection at trecvid 2007.," in *TRECVID*, 2007.

[35] Z. Li, X. Liu, and S. Zhang, "Shot boundary detection based on multilevel difference of colour histograms," in *2016 First International Conference on Multimedia and Image Processing (ICMIP)*, pp. 15–22, IEEE, 2016.

[36] T. Kar and P. Kanungo, "Video shot boundary detection based on hilbert and wavelet transform," in *Man and Machine Interfacing (MAMI), 2017 2nd International Conference on*, pp. 1–6, IEEE, 2017.

[37] C. Pingping, Y. Guan, X. Ding, and Z. Yu, "Shot boundary detection with sparse presentation," in *Signal Processing (ICSP), 2016 IEEE 13th International Conference on*, pp. 900–904, IEEE, 2016.

[38] S. Domnic, "Walsh–hadamard transform kernel-based feature vector for shot boundary detection," *IEEE Transactions on Image Processing*, vol. 23, no. 12, pp. 5187–5197, 2014.

[39] J. Mondal, M. K. Kundu, S. Das, and M. Chowdhury, "Video shot boundary detection using multiscale geometric analysis of nsct and least squares support vector machine," *Multimedia Tools and Applications*, vol. 77, no. 7, pp. 8139–8161, 2018.

[40] M. Yazdi and M. Fani, "Shot boundary detection with effective prediction of transitions' positions and spans by use of classifiers and adaptive thresholds," in *Electrical Engineering (ICEE), 2016 24th Iranian Conference on*, pp. 167–172, IEEE, 2016.

[41] S.-B. Park, H.-N. Kim, H. Kim, and G.-S. Jo, "Exploiting script-subtitles alignment to scene boundary dectection in movie," in *Multimedia (ISM), 2010 IEEE International Symposium on*, pp. 49–56, IEEE, 2010.

[42] M. Ellouze, N. Boujemaa, and A. M. Alimi, "Scene pathfinder: unsupervised clustering techniques for movie scenes extraction," *Multimedia Tools and Applications*, vol. 47, no. 2, pp. 325–346, 2010.

[43] V. T. Chasanis, A. C. Likas, and N. P. Galatsanos, "Scene detection in videos using shot clustering and sequence alignment," *IEEE transactions on multimedia*, vol. 11, no. 1, pp. 89–100, 2009.

[44] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of molecular biology*, vol. 48, no. 3, pp. 443–453, 1970.

[45] A. Hanjalic, "Shot-boundary detection: Unraveled and resolved?," *IEEE transactions on circuits and systems for video technology*, vol. 12, no. 2, pp. 90–105, 2002.

[46] L. Lu, H.-J. Zhang, and S. Z. Li, "Content-based audio classification and segmentation by using support vector machines," *Multimedia systems*, vol. 8, no. 6, pp. 482–492, 2003.

[47] C. Panagiotakis and G. Tziritas, "A speech/music discriminator based on rms and zero-crossings," *IEEE Transactions on multimedia*, vol. 7, no. 1, pp. 155–166, 2005.

[48] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, vol. 2, pp. 1331–1334, IEEE, 1997.

[49] D. Li, I. K. Sethi, N. Dimitrova, and T. McGee, "Classification of general audio data for content-based retrieval," *Pattern recognition letters*, vol. 22, no. 5, pp. 533–544, 2001.

[50] H. Harb, L. Chen, and J.-Y. Auloge, "Speech/music/silence and gender detection algorithm," in *In Proceedings of the 7th International conference on Distributed Multimedia Systems DMS01*, Citeseer, 2001.

[51] A. Pikrakis, T. Giannakopoulos, and S. Theodoridis, "A speech/music discriminator of radio recordings based on dynamic programming and bayesian networks," *IEEE Transactions on Multimedia*, vol. 10, no. 5, pp. 846–857, 2008.

[52] L. Lu, S. Z. Li, and H.-J. Zhang, "Content-based audio segmentation using support vector machines," in *Proc. ICME*, vol. 1, pp. 749–752, 2001.

[53] J. Saunders, "Real-time discrimination of broadcast speech/music," in *icassp*, pp. 993–996, IEEE, 1996.

[54] A. K. Jain, A. Vailaya, and X. Wei, "Query by video clip," *Multimedia systems*, vol. 7, no. 5, pp. 369–384, 1999.