Automatic Video Genre Classification

Kandasamy.K[1], Subash.C[2],
Department of ECE[1,2],
University College of Engineering Villupuram[1, 2],
kkandasamy.03@rediffmail.com[1],
mssubash92@gmail.com.

Dr.Rajaram.M[3],
Vice-Chancellor[3],
Anna University[3],
Chennai - 25[3].
vc@annauniv.edu[3].

***Abstract*—** Video Classification has been an active research area for many years. In this work a novel method has been proposed for the classification of videos into different categories. Video Classification algorithms can be broadly classified into two types. The first type of classifier is a category specific video classifier, which classifies videos of various sports categories into classes such as tennis, baseball, etc. The second type of classifier is a generic video classifier, which classifies the videos into generic categories such as sports, commercials, news, animation etc. This work aims at generic video classification and exploits motion information and cross correlation measure for classification.

Keywords-Video classification; Edge detection; Color Histogram;

1. INTRODUCTION

The explosive growth of the video usage in the past two years was a direct result of the huge increase in the internet bandwidth speeds. According to the Com Sore, a leading statistics reporting company for digital media, 150 million U.S. internet users have watched 96 videos in average per viewer in December 2008. [1] This number is a resultant of the 13 percent increase of US online audience in the month of December 2008 compared to the previous month. The underline factor beneath this huge increase in online video usage is the surge in online traffic due to the increasing bandwidth speeds all across the globe.

Cisco Visual Networking Index projects global IP traffic will increase at a rate of 46 percent from year 2007 to 2012, which is essentially doubling the traffic every two years [2]. To put this in perspective, in year 2012 internet video traffic alone will be equal to the 400 times the total traffic carried by the U.S. internet backbone in year 2000 [2]. Furthermore the number of online video portals has drastically increased in the last two years. Currently there is a vast amount of web sites available on the Internet which includes some kind of streaming video contents.

The combination of Google owned sites which include youtube.com, have streamed over ten billion videos in the month of August 2009. The YouTube web portal alone accounts for more than 99 percent of these streaming hits. Altogether a total of twenty five billion online videos have been viewed by the Internet users in the U.S. in August 2009 [3]. The following table from Com Score, Summarizes the top U.S. online video portals with regard to the number of videos viewed by the internet users in August 2009.

2. ONLINE DOCUMENTS TO ONLINE VIDEO

All these statistics pinpoint one absolute truth the usage of online videos is growing in exponential numbers. We can compare this exact phenomenon to the text based documents when the Internet was at its infant age in the late 90's and early 2000 time period. The amount of online text based documents were increasing in as to nothing rates during the early days of the Internet resulting in a vast amount of cluttered data, scattered around all across the net. If it's not for the web crawlers and the search engines, these data would have been of little or no importance, simply because no one can find it and use it when there is a need. The same argument can be applied to the growing amount of online videos. If these online videos weren't properly tagged and categorized the benefit of having these videos online would be greatly reduced.

However there is a fundamental difference when it comes to automatic object classification of these two worlds, the text documents world and the video world. With the text based documents, any simple web crawler carries the competency to go through those documents and extract the words and basic meanings. In essence, given a search query, a search engine would be able to utilize the extracted words to output reasonably accurate results. Unfortunately, the case is not so simple and straight forward with online videos. It is impossible to go through a huge amount of videos, such as the videos available on the internet and extract meanings out of them, so that a search engine would use those results to respond to a query. For example, a practical use case scenario would be an internet user uploading a sports video clip to the YouTube. After uploading the video to the YouTube website that user has the option to categorize the video to appropriate sections, in this case the "Sports" category. Also, that user can define the appropriate keywords and tags such as: "Baseball", "Red sox", "Word Series". If the user decided not to manually categorize and define the tags and the keywords, any other users would not have been able to retrieve this video.

Hence, the automatic video classification problem is fundamentally different from the classical document classification problem. As mentioned earlier this is mainly due to the semantic differences between a text file and a video file. In easiest terms we can define a text file as a one dimensional file which contains only the text dimension. On the other hand, a video file can be defined as a three dimensional file which contains, all three of the dimensions: audio, visual and text. There has been a significant progress in document classification research work using various different methods.

According to Allamraju & Chun [4] related work section, different authors have proposed various different systems for document classification and clustering problem. Budzik et.al

[5] have proposed a keyword extracting system which extracts keywords form a document that are representative of the document's content.

These keywords then get fed in to a web search engine and the results are generated according to this extracted keyword. In another type of research that utilizes documents, an automatic summarizer has been built by the authors based on the frequency of the words, cue phrase, location and title query method. In the word frequency method, each sentence is assigned a score based on the relevant words in that sentence. In the cue phrase method, each sentence was assigned a cue score based on the presence of relevant and important phrases. In the location method, a score is assigned to the sentence based on its location in a paragraph or proximity to headings.

In this method, sentences containing words present in the document's title are given a higher score. In the query method, sentences matching the query words are given more importance. The final decision is based on the weighted sum of the frequency, cue phrase, location, and title and query method [4]. By varying and multiplexing the decision among five categories these authors were able to gain a high granularity to the expected results because the weighted final score is a representation of the sentences that are most important and most representative to the content of the original document.

However it is not impossible to generate a video classification solution that will classify videos to different categories based on the content of that particular video. There has been a considerable amount of research and ground works done by different people and organizations regarding the video classification problem.

3. RELATED WORKS

Automatic classification of digital video into various genres or categories such as sports, news, commercials and cartoons is an important task and enables efficient cataloging and retrieval with large video collections. The method described in [1] uses different types of spatial and temporal features. The features are modeled using two different classifier methodologies are used. In [2] the authors address the problem of video genre classification for four classes using a set of visual features is used for classification. In [3] the Authors proposed a novel hierarchical SVM scheme for genre categorization, in which a series of SVM classifier are dynamically built up in a binary tree form and optimized. High-level MPEG-7 features were extracted and applied in multi-modality classifiers in [4]. The best classification result at the moment is with an accuracy of 95% using a dataset of 8 different genres [5].In [6] a video genre can be discriminated from the other based on the analogous features and attributes that is disparate from other genres. The technique described in Suresh, Krishna Mohan, Kumaraswamy and Yegnanarayana [2004] uses different types of spatial and temporal features. In [7] a new feature called block intensity comparison code (BICC) for video classification is proposed. The feature is tested on four different video genres viz., cartoon, sports, commercial, and news. The TREC (Text Retrieval Conference) conference, sponsored by National Institute of Standards and Technology (NIST) is an important part of the video classification field. In fact TRECVid (TREC Video Retrieval Evaluation) was branched off from the original TREC as a result of this growth in the field starting from year 2003. Today TRECVid has become a bench marking and evaluation cam pain for the automatic video classification field. However, most effort from TRECVid is focused on retrieving video information and using that information in a search query so that a user would be able to search through a video for a specific content.

The automatic video classification problem is a slightly different problem when compared to the video retrieval problem. For instance, an automatic video classifier will define a particular video as a sports video or a news video, where as a video retrieval machine will focus on indexing each and every portion of the video for future retrieval. One very good example of a video retrieval system is "Gaudi" system (http://labs.google.com/gaudi) that has been developed by Google.

As mentioned earlier, a video file contains three dimensions of data portions: visual, and text. According to Breezeless and Cook [13] previous video classification efforts can be classified in to three approaches that go in par with those three dimensions. Namely, the four categories are text-based approaches; visual based approaches and mixed approaches. Most authors have used a combination of these approaches because semantics of the video has audio, visual and text components in it. Furthermore, video classification can be applied in different manners. Some authors may choose to classify the video as a whole while others may choose to classify specific feature or a component of a video. For example, while one author tries to classify a whole video segment as a news video, another might focus on identifying and classifying the business news section of that particular video. Continuing on this trend of classification, while most authors may focus on classifying videos in to rather broad categories such as action movies, comedy movies, romantic movies, some authors have attempted a narrow categorization methods. For example instead of classifying videos as sports videos, they have attempted to classify videos as basketball, baseball videos. Many of these efforts have incorporated cinematic principles or concepts from the film theory.

For example, when compared with comedy movies, horror movies have low lighting levels in the scenes. If you get an aggregated number of well-lighted scenes vs. dark scenes there is a higher chance that a horror movie containing a larger number of dark scenes. Also when comparing action movies vs. romantic movies it is apparent that action movies contain higher ratio of fast moving sceneries. Utilizing these kind of cinematic principles tend to yield very accurate results in classifying videos. However, focusing only on cinematic principles in visuals may not be sufficient to get all rounded results.

4. AUDIO BASED VIDEO CLASSIFICATION

Audio based video classification is another form of categorization method that has gained a significant popularity when it comes to video classification technologies. Compared with pure text based classification, audio based classification methods tend to yield more accurate results. Because of this reason, many video classification methods are based on audio based approaches rather than text based approaches. Furthermore, compared to solely visual based approaches, audio based approaches seem to in considerably lower

processing cost. Based on the differences between audio and video files, audio based approaches require less computational power than the visual based approaches.

If a particular scene separated into an audio based scene and a visual based scene, the audio based scene tends to carry more information than its video counterpart. This is also quite apparent when the file sizes are examined. For video based analysis, at least 10 – 15 seconds worth of visuals are needed to identify some key characteristics. However, in an audio clip, it may be sufficient to use 1 – 2 second clips for characterization. In fact, many researchers have used audio clips ranging around these lengths for their work. In order to examine the necessary components in an audio file, it is necessary to process the signal using a steady sample rate. Some of the commonly used sampling rates are 44.1 kHz and 22050 Hz. After deriving the samples using the sampling rate, those individual samples can be gathered to form frames.

This frame-forming process is highly analogous to how visual scenes are processed. Additionally, to further enhance the process it is possible to collect these frames and define frame boundaries so that those frames can be identified using a key frame. After collecting these frames audio files are processed in two different domains: frequency domain and the time domain [13]. In the time domain the amplitude of a signal with regards to the time is analyzed, where as in the frequency domain, the amplitude with regards to the frequency is considered.

This time domain to frequency domain conversion can be done with the Fourier transformation. When comprehending audio-based features, it is much more beneficial to take the human interpretation about sound into account. For example by using the audio information we can derive three different layers of information: low-level acoustics, midlevel sounds, and high level sounds. [15] Lui & Wang et al. have worked on a video classification scheme that is solely based on audio feature extracting, with a sampling rate of 22050 Hz for the audio sampling [15]. They have focused on audio attributes such as non-silence ratio, standard deviation and dynamic rate of the volumes, frequency, noise to voice ratio, standard deviation of the pitch and energy levels, etc. To detect the frames that are silent they have compared the volume and zero crossing rate (ZCR – the times that an audio waves crosses the zero axis) of each frame to preset thresholds. In this work authors have figured out that, if both volume and ZCR are less than the threshold values the frame can be declared as silent. Throughout their research the importance of using ZCR measurement is highlighted. Furthermore ZCR values help to avoid the low energy speech frames from being classified as 'silent'.

5. VISUAL BASED VIDEO CLASSIFICATION

Many video classification efforts that have been done so far was based on some kind of a visual based approach. This is very intuitive given that anyone would agree the visual element is the most important dimension out of the three dimensions of a video. Therefore in order to gain from these visual clues most researchers have incorporated visual based approaches to their work. Most of these research that use visual features tend to extract features on a per frame or per shot basis.

Basically, a video is a collection of images commonly referred to as frames. All of the frames within a single camera action comprise a shot. A scene is one or more shots that form a semantic unit [13]. To clarify this point let's consider an example: a scene in a baseball game. The moment when the pitcher starts to pitch the ball to the moment when someone catches the ball can be considered as one semantic unit, in other words one whole scene. Even though the camera moves from one angle (from pitcher) to another angle where it focuses on the batsman the whole unit can be considered as a scene. Sometimes authors refer to a scene as a shot**.** Instead of analyzing a whole video one frame at a time, it makes logical Sense to analyze it scene by scene. A whole scene can be represented as one logical unit of the whole story. For example in an action movie there can be a scene where two people are fighting, or a car chasing scene. We can represent these scenes using one key frame.

As mentioned earlier even though a particular scene has multiple frames, in order to get a representative picture only one key frame is being used. When it comes to these scenes, the above-mentioned cinematic principles can be applied as well. The scenes which are extracted from a horror film may contain much more low-lighted scenes compared to scenes from a comedy movie. Also scenes from an action movie may contain lots of fast moving frames compared to scenes from a romantic movie.

Nevertheless, these types of scene-based approaches have their own disadvantages. First of all direct identification of scenes in a given video clip can be a tedious task. Unless it's done manually, it may not be intuitive to develop an automatic scene selection algorithm for a video clip. The major reason for this is that scene boundaries can be very hard to identify for some video clips. Sometimes a whole video clip may only contain one logical shot, and at other times it can be multiple shots.

The definition of a shot boundary can be different from one person to another based on multiple factors such as their taste, the way they analyze a clip, etc. Anyhow, we cannot undermine the importance of these shot based approaches. After all it is one of the major factors that can contribute to a very good classification scheme. Girgensohetalin their paper describes a video classification method based on the visual features only. They have chosen to classify video frames into six different classes: presentation graphics, long shots of the projection screen lit, long shots of the presentation screen unlit, long shots of the audience, medium close-ups of human figures on light backgrounds, and medium close-ups of human figures on dark backgrounds. Frames have then been extracted from MPEX videos every 0.5s. Each frame is converted to a 64 x 64 gray scale intensity image. After extracting the frames they have used a various different transform algorithms to transform the vectors and perform the classification.

6.HYBRID APPROACHES FOR THE VIDEO CLASSIFICATION

As emphasized above, a video has a three dimensional nature to it and most of the work that has been done so far incorporates hybrid classifying methods. Utilizing

characteristics of video, audio, and text attributes tend to produce much more efficient results than incorporating only one such feature. This is because most of the time in order to get a coherent meaning out of a video, the audio, visual and textual features of a particular video file must be taken into account. Most researchers have utilized at least one more combination of video, text and audio along with their primary choice to overcome unnecessary fluctuation of final results.

Wei et al. [14] have developed a classification method based on face and text processing for different types of TV programs. The features used for this distinct classification were obtained from the tracking of faces and of super-imposed text in the video stream. This face and text tracking considered the basis of their classification method and they have put a great deal of dependability to this tracking system.

The entire classification method is based on how well this tracking system behaves. They have identified two issues involved in such object tracking methods: the detection of the targets in each frame and the extraction of object trajectories over frame sequences to capture their movements. For the face tracking purpose authors have used a tracking scheme that utilized YUV color coordinates (We used YUV as one of the metrics. It will be described later in this report) for skin-tone region segmentation to adapt to the MPEG-1 and MPEG-2 framework. Authors claim that their system was accurate owing to the utilization of different features and the novel iterative partition process. Interestingly, they haven't applied these face detection techniques to every single frame.

It would have been somewhat in efficient to apply high intensity face detection processing to every single frame of every single video in both training and testing sets. To improve the speed of the face tracking, they have taken advantage of the content continuity between consecutive frames by considering the joint detection of faces and trajectory extraction. More importantly, they have utilized one very significant cinematic principle: the variation of faces within a continuous shot is usually small. Hence they have only applied the face detection to the first few frames of each shot. For each detected face, the mean and the standard deviation in color, the height, the width and the center position have been computed. Authors have picked four types of TV programs to do their classification: news, commercials, sitcoms and soap operas. To classify a given video segment into one of these four categories, they have mapped it into the same feature space and evaluated its probability of being each category by the weighted distances to the centers of the news, commercial, sitcom and soap clusters.

7. BASICS OF VIDEO CLASSIFICATION

Video emphasizes the visual rather than the audio aspects of the television medium. Video is defined as a series of framed images put together, one after another, to simulate edge interactivity. (Normally 25 to 30 frames per second). A shot is defined as unbroken sequence of frames taken from one camera. It is the building block and a physical entity of a video and is delimited by shot boundaries.

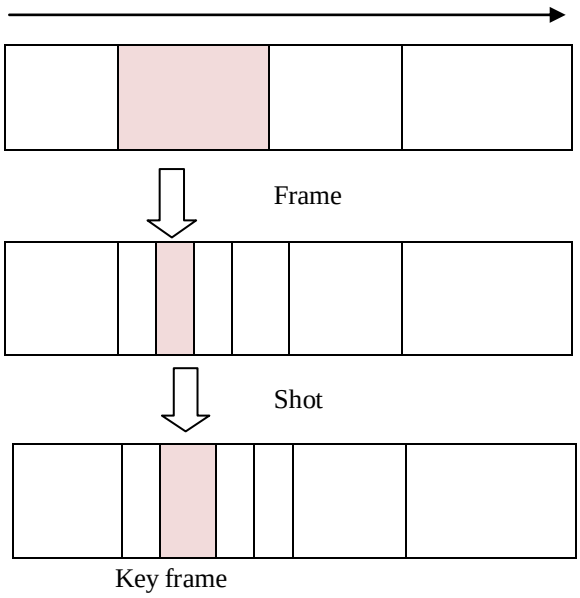

Fig 1.1 Hierarchical Structure of Video

8. VIDEO GENRE

Broadcast video can be regarded as being made up of genre. The genre of a video is the broad class to which it may belong e.g., sports, cartoon, commercials etc., Genre can themselves in turn made up of genre. Genre classifications at the same levels are manually exclusive. So the video genres can be regarded as a four structure. The following figure 1.2 shows the various levels of video genre.

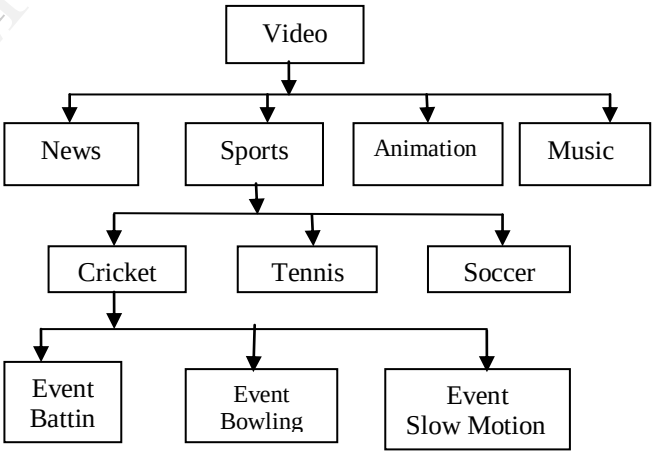

Fig 1.2 Levels of Video Genre

9. OVERVIEW OF COLOR HISTOGRAM

Color histograms are flexible constructs that can be built from images in various color spaces, where RGB, rg chromaticity or any other color space of any dimension. A histogram of an image is produced from by discretization of the colors in the image into a number of bins, and counting the number of image pixels in each bin. For example, a Red- Blue chromaticity histogram can be formed by first normalizing color pixel values by dividing RGB values by R+G+G, then

quantizing the normalized R and B coordinates into N bins each; say N=4, which might yield a 2D histogram.
A histogram can also be made three dimensional, it is harder to display.

The histogram provides a compact summarization of the distribution of data in an image. The color histogram of an image is relatively invariant with translation and rotation about the viewing axis, and varies only slowly with the angle of view. By comparing histograms signature of two images and matching the color content of one image with the other, the color histogram is particularly well suited for the problem of recognizing an object of unknown of an RGB image into the illumination invariant rg chromaticity space allows the histogram to operative well in varying light levels.

The main drawback of histogram for classification is that the representation is dependent of the color of the object being studied, ignoring its shape and texture. Color histograms can potentially be identical for two images with different object content which happens to share color information. Conversely, without spatial or shape information, similar objects of different color may be indistinguishable based solely on color histogram comparisons. There is no way to distinguish a red and white cup from a red and white plate. Put another way, histogram based algorithms have no concept of a generic cup and a model of a red and white cup is no use when an otherwise identical blue and white cup.

A. *Emotion Recognition Methods*

Motion based recognition techniques try to recognize the human activity by analyzing directly the motion itself, without referring it to any static model of the body.

1. Shape based (e. g ,background subtracted human silhouettes)
2. Flow based (e. g, motion field's generated using optical flow).

Based on the metrics used to detect the difference between successive frames, the algorithms can be divided broadly into two categories:

i. Pixel
ii. Block based

Block based approaches use local characteristic to improve the robustness to camera and object movement. Comparing the histograms of successive frames can do a step further towards reducing sensitivity to camera and object movements. As a disadvantage one can note that two images with similar histograms may have completely different content.

B. *Pixel-wise Differencing*

Many types of features have been proposed to characterize motions of various objects and activities with pixel level precision. In the project, pixel wise differencing method is used to compute motion.

Motion is important dynamic attributes of video. Different video genres present diverse motion patterns. In this project, motion is extracted by pixel wise differencing of consecutive frames. Difference image is first obtained from two consecutive frames. A 25 dimensional features vector is extracted by dividing each difference into 5x5 blocks.

The detection of motion is the first stage in many automated visual surveillance applications. It is always desirable to achieve very high sensitivity in the detection of moving objects with the lowest possible false alarm rates.

Motion information T_k or difference image is calculated using

$$T_{k(i,j)}=\begin{cases}1 \; if \; D_{kT(i,j)} > t \\ 0 \quad otherwise\end{cases} \tag{1}$$

Where t is the threshold, t=30 has been used in the experiment

$$D_{k(i,j)}= | I_{k(i,j)}-I_{k+1\,(i,j)} | \; ; \; 1 \leqslant i \leqslant w, \; 1 \leqslant j \leqslant h \tag{2}$$

$I_{k\,(i,j)}$ is the intensity value of the pixel(i,j) in the k^{th} fram
W & h are the width and height of the image

$$d^2_{f,t}(u,v)=\sum_{x,y}\left[f(x,y)-t(x-u,y-v)\right]^2 \tag{3}$$

$$d^2_{f,t}(u,v)=\sum_{x,y}\left[f^2(x,y)-2f(x,y)t(x-u,y-v)+t^2(x-u,y-v)\right] \tag{4}$$

$$c(u,v)=\sum_{x,y} f(x,y)t(x-u,y-v) \tag{5}$$

$$\gamma(u,v)=\frac{\sum_{x,y}\left[f(x,y)-\bar{f}_{u,v}\right]\left[t(x-u,y-v)-t\right]}{\left\{\sum_{x,y}\left[f(x,y)-\bar{f}_{u,v}\right]^2\sum_{x,y}\left[t(x-u,y-v)-t\right]^2\right\}^{0.5}} \tag{6}$$

C. *Cartoon Genre*

Fig1.4Two consecutive frames and their difference image

D. *Edge Feature*

The Canny Edge Detector is one of the most commonly used image processing tools, detecting edges in a very robust manner. It is a multi-step process, which can be implemented on the GPU as a sequence of filters. The approach proposed in this phase, uses an image size of 320x320 to extract the edges of different genres. Canny edge detection algorithm is used to obtain the edge information. To extract the edge information, each edge image is divided into 5x5 blocks, each of size 64x48, to extract a 25 dimension feature vector. It is important that edges occurring in images should not be missed and that there be no responses to non-edges Figure4.4 shows an illustration of edge information extracted for edge feature.

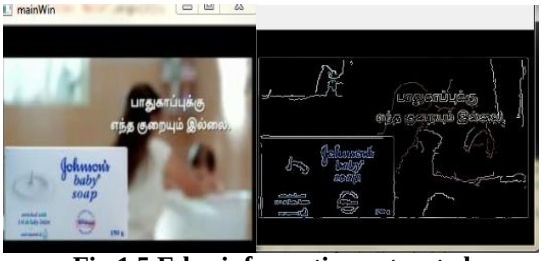

Fig 1.5 Edge information extracted

E. 64 BIN Color Histogram

Color histogram is used to compute images in many applications. This work uses color histogram as another static feature. The RGB (888) color space is quantized into 64 colors by considering only the two most significant bits from each plane. For each frame of the video sequence 64 bin color histogram is obtained.

Further research into the relationship between color histograms data to the physical properties of the objects in an image has shown they can represent not only object color and illumination but relate to surface roughness and image geometry and provide improved estimate of illumination and object color.

For each frame, a 64 dimensional feature vector is extracted. In order to reduce the dimension of the feature vector, only the dominant top 16 values are extracted using PCA (Principal Component Analysis) and are used as features in experiments and is shown in Figure 1.6

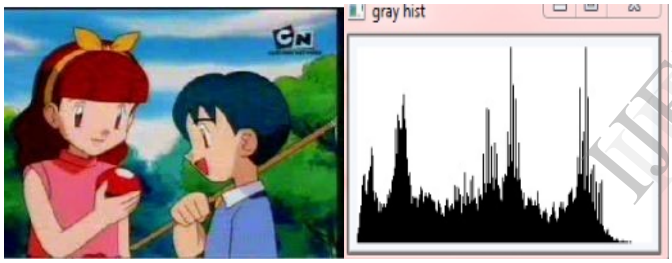

Fig1.6 The dominant 16 features extracted from Color Histogram

10. EXPERIMENTAL RESULTS

Experiments are performed to evaluate the accuracy and efficiency of the method. Motion features are extracted and Edge Feature is used to calculate the similarities between the frames in video sequence, which helps for categorizing the different videos. The result plot shows the accuracy for different video genres such as news, cartoon, sports and commercials. The following table shows the result obtained the distance measure:

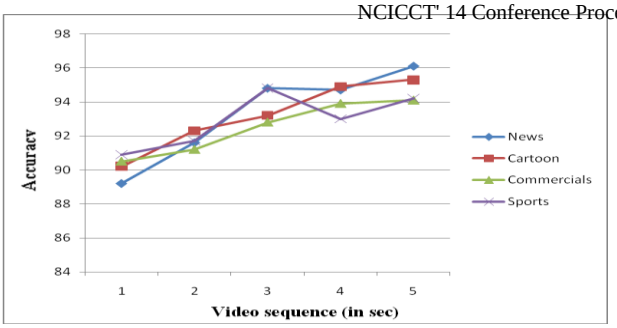

Fig 1.7 Video sequence Vs Accuracy

11. RESULTS

In the first phase, individual frames are considered, edge features are extracted from the frames, and Cross Correlation method is used to compare the results. In the second phase, color histogram features are considered and Cross Correlation method is used to compare the difference between the features. Since there is no specific algorithm/ method to fix up the number of status to be used in Cross Correlation. Initially separate distance is used for each one of edge and histogram features.

Finally, they are combined using 'sum rule' to generate the final output. Cross Correlation is found to give better results and for edge and color histogram features. The experimental results are tabulated Table 1, Table 2 and Table 3. Edge features are extracted using canny edge detector. It is noted from Table 1, the overall classification accuracy using edge information alone is 93.91%. It is seen that cartoon and news genre can be reliably identified using edge features.

The static feature performs well for classifying sports and commercials genre. Table 2 illustrates that color histogram feature has an accuracy of 86.65% It classifies cartoon, commercials and news with reliability, since they exhibit diversified color features. A small percentage of cartoons are misclassified in commercials since some commercials are animated cartoons. From table 3, it's seen that standard deviation method shows accuracy of 96.14% the overall accuracy improves to 94.5% further, at low dimension, the overall performance of Cross Correlation is also found to be exceptional.

Fig1.7 OUTPUT DIAGRAM

Table 1 Edge Feature (93.91%)

Events	News	Cartoon	Commercial	Sports
News	93.8	0	4.2	1.4
Cartoons	0	93.1	5.1	0
Commercial	0	4.9	93	0
Sports	3.4	1	0	93.3

Table 2 Color Histogram Features (86.65%)

Events	**News**	**Cartoon**	**Commercial**	**Sports**
News	96.2	0	0	3.8
Cartoon	0	98.03	1.97	0
Commercial	0	0	100	0
Sports	33.48	14.15	0	52.37

Table 3 Color Histogram Features using Standard deviation Method (96.14%)

Events	**News**	**Cartoon**	**Commercial**	**Sports**
News	97.55	0	0	2.44
Cartoon	0	98.03	1.97	0
Commercial	0	0	100	0
Sports	4.4	2.6	0	89

12. CONCLUSION

Automatic classification using edge and color histogram feature to identify the genre of video was achieved. Experiments are conducted on popular four video genres namely cartoon, sports, commercials and news. The approach initially evaluates the performance of two features individually by using cross correlation as the classifier method. Experiment shows that the approach is promising to classify the video genres. Finally by combining the individual results the system gives a good classification accuracy of 94.5%.

It is observed that even after combining the individual results, the system could not distinguish news and sports with high accuracy, which in turn leads to further retrieval process.

REFERENCES

[1] akkalanka Suresh, C.Krishna Mohan, R. Kumaraswamy and B.Yegnanarayana, Combining multiple evidence for video classification, IEEE Int. Conf. Intelligent Sensing and Information Processing (ICISIP-05),Chennai, India, pp. 187-192, Jan. 4-7, 2005.

[2] Vakkalanka Suresh, C.Krishna Mohan, R. Kumaraswamy, and B.Yegnanarayana, Content-Based Video Classification using SVMs , in Int. Conf. on Neural Information Processing , Kolkata, pp.726-731, Nov. 2004.

[3] Automatic Video Genre Categorization Using Hierarchical SVM. 1-4244-0481-9/06/$20.00 @2006 IEEE.

[4] R. Glasber, S. Schmiedeke, M. Mocigemba, and T. Sikora, "New Real-Time Approaches for Video-Genre-Classification Using High- Level Descriptors and a Set of Classifiers," in 2008 IEEE International Conference on Semantic Computing, 2008, pp. 120–127.

[5] ComScore Press Release. U.S. Online Video Viewing Surges 13 Percent in Record-Setting December.[Online] 2009.

[6] News Cisco. Cisco Visual Networking Index Projects Global IP Traffic to Reach Over Half Zettabyte (1) in Next Four Years. newsroom.cisco.com. [Online] June 16, 2008.

[7] ComScore Press Release. Google Sites Surpasses 10 Billion Video Views in August.

[8] Sri Harsha Allamraju, Robert Chun. Enhancing Document Clustering through Heuristics and Summary based Pre-processing. San Jose State University, Department of Computer Science, San Jose, CA: s.n.,2009.

[9] Information access in context, Knowledge Based Systems. J.Budzik, K.J. Hammond, L.Birnbaum. 2001,Vol. 14, pp. 37-53.

[10] Automatic Video Classification: A Survey of the Literature. Darin Brezeale, Diane J. Cook. s.l. : IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews, May 2008, Vol. 38.

[11] Automatic News Video Segmentation and Categorization Based on Closed-Captioned Text. Weiyu Zhu,CandemirToklu, Shih-Ping Liou.s.l.: IEEE International Conference on Multimedia and Expo, 2001.

[12]Audio feature extraction and analysis for scene segmentation.