Mapping of Low Level to High Level Audio-Visual Features: A Survey of the Literature

Prof. Meenal N. Pande

***Abstract*-There is a great need to automatically segment,classify video data to develop efficient tool for searching and browsing. Researchers believe that categorization of video data can be achieved through the extracting the meaning of the videos.This task requires the bridging the gap between High and Low level features.**

***Keywords*- Visual features,audio features,computable features,high level semantic,film grammer.**

I. INTRODUCTION

A.OBJECTIVE OF THE MAPPING

Once a relationship between low level and high level is establish ,we can easily navigate through the video through the meaning(sematics).Example:-user extract the only those scene from action movie which contain only fights, rather than sequentially browsing the whole video.This relationship follow the norms of Human perception called Film Grammer.To understand the interpretation of grammer,first we need to understand the symbols and second is rules for combining the symbols to represent concept.It is necessary to relate these symbols of grammer to computable video features.

One method that viewers use to narrow their choices is to look for video within specific categories or gen re. Because of the huge amount of video to categorize, research has begun on automatically classifying video.

That automated methods of classifying video are an impor-tant and active area of research is demonstrated by the exis-tence of the TRECVid video retrieval benchmarking evaluation campaign [1]. TRECVid provides data sets and common tasks that allow researchers to compare their methodologies under similar conditions. While much of TRECVid is devoted to video information retrieval, video classification tasks ex ist as well such as identifying clips containing faces or on-screen text, distinguishing between clips representing outdoor or indoor scenes, or identifying clips with speech or instrumental sound Zeeshan[2] focus in this review on approaches to video classification,and distinguish this from video indexing. The choices of features and approaches taken for video classification ar e similar to those in the video indexing field. Much of the video indexing research is approached from the database perspective of being able to efficiently and accurately retrieve videos that match a user query . In contrast, video classification algorithms place all videos into categories, typically with a meaningful label associated with each (e.g., `sports video' or `comedy video').

COMPUTABLE FEATURES OF VIDEO

A computable features of Audio-visual data is defined as any statistics of data that can be easily extracted by any image signal processing and computer vision techniques.These features are global in nature and can be extracted using whole images,therefore they do not require any classification ,object detection,tracking etc.

These features contain video shots,shot length,shot motion content,color distribution,key Lightening and Audio energy.One can use these features to solve segmentation and classification of Talk and Game show,classification of movie type based on previews etc.

STRUCTURE OF FILM

High level symbol (emotions,irony,gesture etc)are very hard to represent in statistics as compare to Low level symbols(Lightining,shot length,background music etc).

Frames-are the smallest unit of the video.

A shot-A sequence of frames taken by a single camera with no major changes in the visual content.

Scenes-Similar shots make scenes.

Visual-Based Approaches

Most of the approaches for visual features are rely on visual elements either alone or combination with text and audio features.This corresponds that human can receive information through the sense of vision.

Most extract feature on a per frame or a per a shot basis.A video is collection of images known as Frames.

All of the frames within a single camera known as shots.A scene is a one more shots called a single unit.For Example conversion between host and guest may be shown such that only one person shown at a time.Each time camera appears to stop and move to the other person represents a shot change,but the collection of shots represents the entire talk show is a scene.Some researches use shot and scene term interchangeably.

A shot is natural way to segment a video and each of these segment represents a Higher level concept to humans.Also shot can be represents by a single frame called keyframe.Generally first frame is the keyframe of the shot. Shot are also associated with some cinematic principles. For example movie contain fights have shorter shots.

One of the major problem with shot based detection is that, analyzing the shot boundaries.Methods for automatically identifying shot boundaries are not always performed well .The use of cinematic features are more popular in visual based approach.These include color as a proxy for light levels,motion to measure actions and Average shot length to measure the rapidity of the video.One difficulty using visual based feature is large amount of potential data.This problem can be explore using keyframe to represent shots or with dimentionality reduction techniques.

Color based feature

A frame contain number of dots called pixel also.And the color of each pixel represents a specific values from a color space.There are two color space are more popular one is RGB(Red Green Blue) and HSV(Hue saturation value).In RGB color space represents each pixel is combination of red green blue individual color.HSV reprsents the color in terms of hue(wavelength of the color percept),saturation,value(brightness of the color).

The distribution of colors in video in color frame are also represented using color histogram that is,how many pixels exist in the frame for each possible color.color histogram also used to compare two frames for similarity,assume that both frame have similar count.

One drawback of color based approach is imeges represented in frames may have produced under different lightning conditions.so, comparision may not be correct.

One solution for this prosed byDrew and Au[3]normalize the color channel bans of each frame and then move into chromatography color space.

Motion Picture Expert Group

Consecutive frame within the same shot having similar and this temporal redundancy can be exploited using compression the video.This feature can be used to ,primarily extracting features directly from MPEG video are DCT coefficient and motion vector.This can improve the performance for video classification.No need to calculate the features and can be extracted without decoding the video.

Shot based Features

A.Shot Detection

To utilize the shot based features,shot must be detected first.Various ways of making transitions from one shot to the next Most of the shot transitions fall in to one of the ccategories:Hard cuts,fades,and dissolves.

Hard cut in which one can abruptly stop and another can start.fade-fades are further classify fade-in and fade-out,fade-in shots gradually fades into existence from a monochrome frame while fade-out shots gradually fades out of existence to a monochrome frame. Dissolves consist one shot fade-in while another fade-out.While it is important to identify the shot transition type in order to correctly identify the shot changes.

One of the simplest method to identify the shot is to take difference of color histogram of a consecutive frames.assume that the difference between frames are smaller within same shot[4].This method have potential problems,

1)Deciding what threshold defferences must exceed in order to declare a change in shot.

2)Shot contain lot of motion require high threshold value.The threshold value different for different videos.Within video no particular value correctly identify the shot changes.Too Low threshold value may identify shot changes but that doesn't exist while a threshold value is too high will miss some shot changes.

Iyengar and Lippman [5] detect shot changes using the Kullback-Leibler distance between histograms of consecutive frames that have been transformed to the rgb color space. The rgb values are calculated using

$r = R/R + G + B$
$g = G/R + G + B$
$b = B/ R + G + B$

where N is the number of bins in the histograms, $p(xi)$ is the probability of color xi for one frame and $q(xi)$ is the probability of color xi for the other frame.
Truong et al.[6]detect shot changes with shot transitions of the types hard cut,fade-in and fade-out,dissolves[7].

Rasheed and shah[2] detect shots using intersection of the histogram in HSV color space this method best for hard cuts[2]

Jadon et al.[8]detect shot changes as well as shot transition using fuzzy logic based approach.

Shot Length

Once a shot boundaries are identified, associate the shot length feature with shot.That will give number of frames within one shot.It also computable feature.Generally dialoguage shots are longer and span a large number of frames on the other hand shots for fight scene are change rapidly and last for fewer frames.

Shot Representation

A shot may span few or several frames,to compute frame feature like color distribution,key lightning effect required to process all frames within the shot boundaries.

Key Frame Selection

A method is to select multiple keys for each shot.Each shot is represented by a set of key frames such that all frames are distinct.Initially middle frame is selected and add to a empty set as a first key frame.The reason behind selecting middle frame is that frame must free shot transition effect.Next each frame within is shot is compared with each frame with in the previous frame.If frame differ from previously choosen key frames by more than a fixed threshold.It is added in the keyframe set otherwise ignored.

Shot Motion Content

This feature provide clues to the nature of the scene.Motion contains of shot also depend on the nature of shot.The dialogue shots are relatively calm on the other hand in fight scene motion are jerky and haphazard's with larger actor movements.

Color Variance

Zettle observation in[9],Variance in color can be represent a exploiting the film genre.Variance of color has a strong correlation structure.Comedy tend to have a large variety of bright colors,horror scene often adopt only darker hues,

Lightning Key

Lightning is an important dramatic agent. Generations of film makers have exploited luminance to evoke emotions.Using techniques[72],a deliberate relationship is exist between the lighting and film genre.Lighting can also be used to direct the attention of the viewer to a certain area of the important scene,It can also direct affect viewers emotions regardless of the scene

Audio based Low Level Features

Music and sound effect is often provide potential enegy to the scenes.Such as whether a situation is stable or unstable.For Example shot fighting and explosions are usually accompanied by the increase or sudden change in audio level.The enegy on the audio track can be used to identify such a event when the peak in the audio energy is relatively high

Audio-only approaches are found slightly more often in the video classification literature than text-only approaches. One advantage of audio approaches is that they typically require fewer computational resources than visual methods. Also, if the features need to be stored, audio features require less space. Another advantage of audio approaches is that the audio clips can be very short; many of the papers we reviewed used clips in the range of 1-2 seconds in length.

To produce features from an audio signal, the signal is sampled at a certain rate (e.g., 22050 Hz). These samples may then be grouped together into frames. Some authors choose to begin one frame where the last ended while others overlap the frames.

Features can be derived from either the time domain or the frequency domain. Fig. 1 is an example of the time domain, in which the amplitude of a signal is plotted with respect to time. Using the Fourier transform, a signal in the time domain can be transformed to the frequency domain, also known as the spectrum of the signal.

Mapping of Low level visual feature to High Level Semantic for a Movie Genre:

It is a great need to categorize movie genre in recent time.Zeeshan[2]computed a low level feature based on movie previews to classify movie genre.The commercial advertisement is created to attract people.A preview often provide the theme of a film and useful in film categorization.Zeeshan categories film in to four main genre:Commedies,action,horror, dramas .

Computable video feature are combined in a framework in a cinematic principle to provide mapping to these high level semantic classes.

In first step,Firstly classify film in to action and non-action categories based on the average shot-length feature and Motion content in the previews.Secondly classify Non-action film in to comedy,horror,drama by examining their Lightning-key feature.Lastly Action film ranked on the basis of number of explosions and gunfire events.

In second step,a mean shift classifier is used to discover the structure of the mapping between computed feature, Zeeshan[2] and high level movie genre.some researchers proposed space –based approach.For previews two features of the previews,shot length and shot activity is used.To categorize film they are used linear classifier in the two-dimentional feature space.By extending this feature Nam et al[10] able to detect violence in the previews. Zeeshan[2]uses Non-parametric approach,using Mean-shift clustering.Mean-Shift clustering have to be excellent properties and more powerful for real data.Further Zeeshan exploit the cinematic feature for classification.The performance of classification based on selection of exact feature.Below four computable[2] feature are best suitable for classifying movie genre.

Average Shot-Length,Average Motion content,color and Lighing key.

Applications of Mapping Low to high level features

This approach can also be broadens for many potential applications like scene understanding, building and updating of video databases with minimal human intervention, browsing and retrieval of video on the Internet.

CONCLUSION

Combination of audio-visual features and cinematic principles provides more powerful tools for film categorization.Classification based on four-dimentional feature space of Average shot length,average motion content,color and lighting key.This approach may support for full movie and explore the semantic from shot level to scene level.

REFERENCES

1] A. F. SMEATON, P. OVER, AND W. KRAAIJ, "EVALUATION CAMPAIGNS AND TRECVid," in *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*. New York, NY, USA:

[2] Z. Rasheed, Y. Sheikh, and M. Shah, "Semantic film preview classification using low-level computable features," in *3rd International Workshop on Multimedia Data and Document Engineering (MDDE-2003)*, 2003.

[3] M. S. Drew and J. Au, "Video keyframe production by efficien clustering of compressed chromaticity signatures," in *Poster session of the eighth ACM international conference on Multimedia (MULTIMEDIA '00)*, 2000, pp. 365–367B.

[4] H. Zhang, A. Kankanhalli, and S. W. Smoliar, "Automatic partitioning

of full-motion video," *Multimedia Systems*, vol. 1, pp. 10–28, 1993.

[5] G. Iyengar and A. Lippman, "Models for automatic classification of video sequences," in *Proceedings of SPIE Storage and Retrieval for Image and Video Databases VI*, I. K. Sethi and R. C. Jain, Eds., vol. 3312, 1997, pp. 216–227.

[6] B. T. Truong, C. Dorai, and S. Venkatesh, "Automatic genre identification for content-based video categorization," *Proc. 15th International Conference on Pattern Recognition*, vol. IV, pp. 230–233, 2000.

[7] "New enhancements to cut, fade, and dissolve detection processes in video segmentation," in *Proceedings of the eighth ACM international conference on Multimedia (MULTIMEDIA '00)*, 2000, pp. 219–227.

[8] R. Jadon, S. Chaudhury, and K. Biswas, "Generic video classification: An evolutionary learning based fuzzy theoretic approach," in *Indian Conference on Computer Vision, Graphics, and Image Processing (ICVGIP)*, 2002.

[9] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *AI*, vol. 17,no. 1-3, pp. 185–203, August 1981.

[10] J.Nam,M.Aghonymy,A.H.Tewfik Audi-visual content based violent scene characterization.In IEEE ,International conference on Image processing.

AUTHOR'S PROFILE

	Prof.Meenal Narayan Pande M.N.Pande has completed master's degree in COMPUTER APPLICATION(MCA) from PGDCST,HVPM,Amravat.Currently working as assistant professor in P.G.D.C.S.T,D.C.P.E,H.V.P.M college,Amravati