Academic Science

International Journal of Electronics, Electrical and Computational System
IJEECS
ISSN 2348-117X
Volume 6, Issue 6
June 2017

An Effective Way to Summarize Video through Musical Features

Ms.Shweta D. Kherde
Department of Electronics Engineering
MIT Academy of Engineering,
Alandi, Pune,India

Mrs. Dipti.Y.Sakhare
Department of Electronics Engineering
MIT Academy of Engineering, Alandi
Pune, India

Abstract— *Today, extensive measure of mixed media stockpiling area make the perusing, getting to and conveyance of video substance exceptionally steady and even extremely troublesome for the dealing with. To increment quicker perusing of extensive video accumulations and additional effective substance ordering and accessing,video rundown has been anticipated. This paper, gives a huge approach by utilizing melodic components for video rundown framework. Paper concentrates on the division of a specific gathering of sight and sound substance, varying media melodic streams, into little music pieces. Current methodologies comprise in direct grouping in a couple sound classes (music,speech,noise) and as indicated by study, on account of varying media melodic streams no reliable assessment has been performed yet.Detailed considers by experienced clients have been led to decide the nature of synopsis. Introduction of various strategies for highlight extraction and grouping, for example, MFCC and SVM has been completed. Extend point is to decrease the database required for the video by video summarization. The probes different types of melodic video and examinations with the rundowns just in view of music track and video track demonstrates that the consequences of synopsis utilizing proposed technique are extremely valuable and successful to help understand user's desire.*

Keywords—*MFCC; SVM; Segmentation; video summarization; features.*

I. INTRODUCTION

An immense measure of video information is caught or put away throughout the day for various reasons, because of the progression and accessibility of video technology, to get to and deal with these recorded recordings turns into a tremendous test day by day. Video rundown (VS) is great approach for separating valuable parts. In video summarization, a long video is exhibited in a short frame keeping critical substance and disregarding the undesirable occasions so that a watcher can comprehend about the entire video inside brief time. Video synopsis can be accomplished for the most part by an arrangement of key speaking to the whole substance of the first video utilizing a little arrangement of chose casings. Video rundown is likewise called a video skimming, which introduces a short video highlight of the first video.

One of surely understood sound component utilized for acknowledgment is Mel Frequency Cepstral Coefficients (MFCC)which is utilized as a part of my paper .MFCC had been utilized as one of the sound elements for highlight extraction.It likewise gives a decent order result. There are additionally other sound elements including pitch range, vitality and zero intersection rate. SVM is a two-class classifier in view of the standards of basic hazard minimization is utilized as a part of my proposed system.It has well speculation capacity when contrasted with neural system based classifier and concealed Markov show.

"Fig. 1" is the piece outline of the proposed approach. The music track and the video track is the dividation of music video. For the music track, by examining the music content utilizing music includes, a versatile bunching calculation, and music space data a music rundown is created.Shots are resolved and grouped utilizing visual substance examination, for the video track. In conclusion, the music video outline is delivered by exceptionally adjusting the music synopsis and bunched visual shots.VS gives preferred outcomes over audiocentric or imagecentric rundown techniques, in view of the criticism of clients in my review. My technique can boost the scope for both sound and visual substance without giving up them two.

In this paper extension of the rundown to music recordings by utilizing sound division in view of sound grouping and low-level sound element examination is done. firstly the video information is separated into sound and video stream. At that point, sound stream is ordered into two classes: music, non-music. During the time, visual investigation segments the video stream into shots. Finally, audio division and visual division are consolidated to recognize scene change. The rest of the paper is composed as takes after: Section II portrays the related research on video summarization. Section III shows the detail exchange of the proposed scheme. Section IV gives Experimental outcomes and dialog. Ultimately, Section V talk about a finish of this paper.

International Journal of Electronics, Electrical and Computational System
IJEECS
ISSN 2348-117X
Volume 6, Issue 6
June 2017

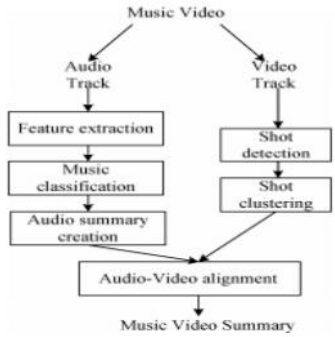

Fig 1. A block diagram of the proposed summarization system.

II. RELATED WORK

There are several researches has been carried out by multiple authors to address the video summarization technique:Tan Teng Teng , Lim Tien Sze , Ong Lee Yeng [1]. In this paper, to enhance recent CCTV system, they proposed microcontroller embedded system.To carry out the sound detection, audio processing and analysis abnormal sound embedded system is proposed. This study is using only single microphone for sound detection.Fast Fourier Transform (FFT) extracts audio amplitude as well as frequency range which are targeted feature. Yingying Zhu, Dongru Zhou[2] In this paper, they present a scene change detection method based on audio and visual features, which analyzes both auditory and visual sources and accounts for their inter-relation and benefits to semantically identify video scenes. Isao Otsuka, Hidetsugu Suginohara, Yoshiaki Kusunoki, Ajay Di-vakaran[3].This paper, describes a combination of three methods for detecting music/songs as a segment with high correctivity. First by using Gaussian Mixture Models audio classification was done. To establish the start/end of the music they detects the difference of audio energy between Right and Left channels,. Yuan-Shan Lee, Chia-Yung Hsu, Po-Chuan Lin, Chia-Yen Chen, and Jia-Ching Wang[4]. In this system first face recognition is performed. The Adaboost approach is adopted to find out the image areas which contain human faces.Performs the Non-negative Matrix Factorization (NMF) technique to decompose the face regions into basis and respective coefficients. Next, uses the coefficients as features to do classiffication by Support Vector Machine (SVM).During the same time, the voice part is used to do speaker verification via GMM-SVM approach. Md. Musfequs Salehin, Manoranjan Paul[5]. In this paper, a good method has been proposed combining area of foreground object obtained a Gaussian mixture-based dynamic background modelling and frame to-frame object motion. Gaussian mixture model (GMM) method has been used for this paper. James J. Deng, Clement H. C. Leung [6]. A several dynamic textures (MDT) model is proposed to model music as well as emotion dynamics over time,to estimate model parameters expectation maximization (EM) algorithm along with Kalman ltering and smoothing is used.

III. PROPOSED WORK

A. Feature selection

Mel-scale Frequency Cepstral Coefficients (MFCCs) is the benchmark highlight of my venture and the element logarithm of the vitality was recommended by creator A.Bazzica.These components are utilized widely in the discourse preparing area.In expansion with the pattern highlights, elective elements are proposed which are demonstrated as follows.

- Pitch and Energy highlights : principal recurrence of the speaker is spoken to by Pitch.It is every now and again utilized as a part of discourse based frameworks to separate discourse from non-discourse areas.The consonant item range (HPS) is utilized for pitch estimation and first request subordinate of the pitch shape.
- Spectral Flux : Spectral flux speaks to the adjustment in the power range and ghastly plentifulness of the flag between progressive casings. Tests which comprise of discourse signs are relied upon to have higher unearthly flux contrasted with quiet and group just examples.
- Spectral Centroid :Spectral centroid is the focal point of gravity of the extent spectrum.It is the normal for the perceptual brilliance of the sound signal.The centroid of the size range of the flag is relied upon to be situated distinctively for every sound occasion class,since the unearthly vitality is required to be amassed in various segments of recurrence.
- Low Short-time Energy Ratio of frames:LSTER can be effective element in separating the Speech only,Speech Over Crowd and Excited sound occasion classes.It speaks to the quantity of edges whose normal vitality is lesser than a large portion of the vitality in a more noteworthy window.The Excited class is relied upon to have the lower LSTER esteem as the majority of the edges are high vitality taken after by the Speech Over Crowd class.Speech Only class is required to have the higher LSTER esteem.
- Peaks in Magnitude Spectrum :It speak to the recurrence segments that add to the sufficiency of the flag. For analysis,the

International Journal of Electronics, Electrical and Computational System
IJEECS
ISSN 2348-117X
Volume 6, Issue 6
June 2017

highest 3 crests (recurrence esteems) in the greatness range are considered. They are relied upon to speak to the frequencies that command all through the entire clasp thus carry on diversely for speech,crowd and quiet.

B. Classification

Classifiers are intended to catch or to distinguish the examples accessible in information without being uncommonly programmed.Each classifier has its imbued approach to catch these patterns.Therefore an endeavor has been made to recognize the classiffier which can effectively characterize the sound portions accessible in videos.Following distinctive classifiers are utilized as a part of this work.But the benchmark classifier SVM(Support Vector Machine) is utilized as a part of my proposed framework :

- Support Vector Machines (SVMs)
- Self-Organizing Map (SOM)
- K-Mean

IV. METHODOLOGY

A. Feature Extraction

The point of Feature Extraction module is to separate the acoustic element vectors which are utilized to portray the unearthly properties of the time changing discourse flag .These component vectors are utilized for ID of speaker. There are a few methods existing for parametrically speaking to the discourse motion for the speaker recognition.Linear expectation coding (LPC), mel-recurrence cepstrum coefficients (MFCC) are the most frequently utilized technique.Known variety of the human ear's basic data transfer capacities with recurrence in view of the MFCCs.The MFCC system utilizes two sorts of channel, i.e. straightly dispersed channels and logarithmically separated channels. Mel recurrence scale is utilized as a part of MFCC.MFCCs are less inclined to the varieties in discourse waveform because of physical condition of speakers vocal rope.

1) MFCC Processor :

Mel frequency cepstral coefficients (MFCC) is no doubt the best known and most comprehensively utilized for both discourse and speaker recognition[10] .A unit of measure in light of human ear's apparent recurrence is a Mel.The Mel scale has around straight recurrence dispersing underneath 1000Hz and a logarithmic dividing over 1000Hz. The estimation of mel from recurrence can be communicated as

$$Mel(f)=2595*\log(1+f/700) \quad (1)$$

where f is the real frequency and mel(f) is the perceived frequency.The block diagram which shows the computation of MFCC is shown in below Figure.

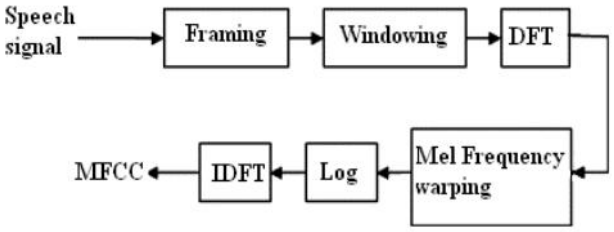

Fig.2 MFCC Extraction

In the principal arrange speech signal is partitioned into number of frames with the length of 20 to 40 ms and a cover of half to 75%. In the second stage windowing of each casing with some window capacity is done to reduce the discontinuities of the frames by narrowing the start and end of each edge to zero.Window is point insightful duplication of the confined frame and the window work in time domain.A great window work has a decreased primary projection and low side flap levels in their exchange function.To perform windowing capacity hamming window is utilized as a part of work. In third stage transformation of each casing from time space to frequency area is done in DFT piece. Hamming window is given as

$$W(n) = 0.54 - 0.46\cos\frac{2\pi n}{N-1} \quad (2)$$

where N denotes the width, in samples ,of a discrete-time, symmetrical window function w[n],0≤ n ≤N-1.

To transfer the real frequency scale to human perceived frequency scale called the mel-frequency scale,mel frequency warping is done in the next stage.The mel frequency warping is normally realized by triangular filter banks with the center frequency of the filter usually evenly spaced on the frequency axis which is shown in “fig 3”. In the fifth stage, log of the filter bank output is calculated and in final stage DCT (Discrete Cosine Transform) is calculated.

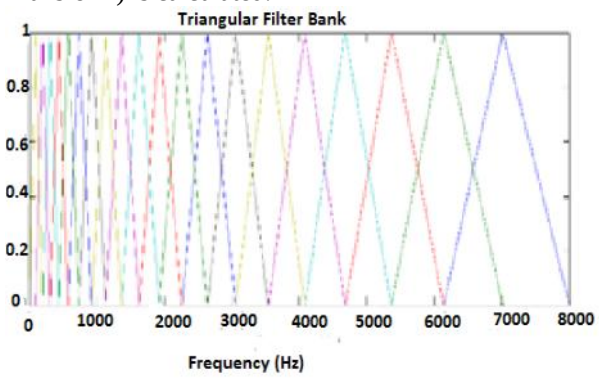

Fig 3. Triangular filter bank

International Journal of Electronics, Electrical and Computational System
IJEECS
ISSN 2348-117X
Volume 6, Issue 6
June 2017

B. Feature Matching :

The status of the art feature matching techniques used in speaker recognition include Hidden Markov Models (HMM), Dynamic time warping (DTW), Vector Quantization (VQ) and neural network techniques. The main idea of SVMs is to define a boundary between two classes by maximal separation of the nearest observations. In practice,SVMs are powerful algorithm on binary classification tasks. I have used Support Vector Machine because of its high accuracy.

2) Support Vector Machine :

SVM is a standout amongst the most essential advancements in example recognizable proof over the most recent 10 years. Different systems like Hidden Markov models (HMM) and Gaussian blend models (GMM) are bad over fitting and they don't specifically improve separation. SVM is a direct classifier[10]. For an arrangement of preparing illustrations, a SVM preparing calculation constructs a model that doles out new cases into one write or the other. It is vital to pick the correct separation line to evacuate the misclassification.It is required to broadens the edge between two classes. The most positive hyperplane is ascertained utilizing part works. The examples nearest to the isolating hyperplane are called bolster vectors.Support vectors are totally characterized by Optimal hyperplane.

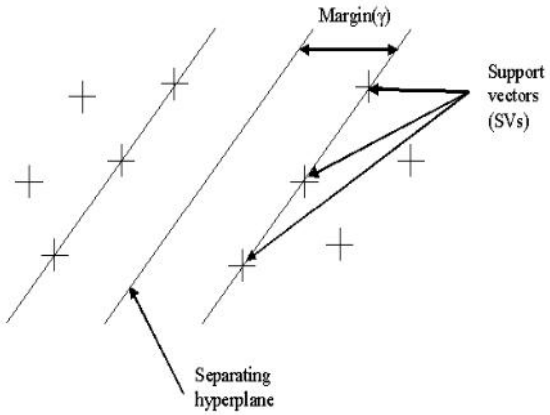

Fig 4. A linear support vector machine.

To find the optimal hyperplane, we have to solve the given optimization problem:

$$\text{Minimize} \quad \frac{1}{2}\|w\|^2 \qquad (3)$$

$$\text{Subject to } (x_i .w+b)\ y \geq 1$$

When the data are not linearly separable then no hyperplane exists for which inequality satisfied by all the points. Therefore slack variable ζi are included into the inequalities. This relaxes the inequality and some points are allowed to be misclassified. The objective function becomes:

$$\frac{1}{2}\|w\|^2 + C\sum_i(L)\,(\zeta i) \qquad (4)$$

$$\text{subject to } (x_i .w+ b)\ yi \geq i - \zeta i \text{ for all i.}$$

The second term of Equation 6 is the experiential risk associated with those points that are misclassified, C is a hyper parameter and L is the loss function (cost function).It trades off the effects of minimizing the experiential risk against maximizing the margin. To non-linearly map the input data to a high-dimensional space (feature space) Kernels are used. The new mapping is then linearly distinguishable.

V. RESULT AND PERFORMANCE ANALYSIS

Every one of the recordings utilized for preparing and assessment were gathered from You Tube® and from individual video recorder.While gathering, particular care was taken that the recordings are client generated,extensive altering and PC created impacts were not allowed.The video quality changes from low to medium. The run of the mill span of the recordings is few moments. An aggregate of 300 recordings were gathered –150 music and 150 non music. In the greater part of the recordings just a single substance is accessible however there are additionally some blended recordings in light of the way the recordings were gathered, a safe presumption can be made that the differences of the recording gadgets is great.The calculation was tried on every one of the recordings in the database where a forget one approach was utilized.

The component extraction was finished by utilizing MFCC (Mel Frequency Cepstral Coefficients).The recordings were displayed utilizing SVM. A video rundown framework utilizing melodic elements involves a preparation stage and a test stage. In the preparation stage the SVM models are made for every video. In testing stage the put away information are contrasted and the asserted SVM display and a choice is made. Before ordering a video from the database,frames from that video show in the preparation set are withdrawn, and SVM is performed on the rest of the element vectors as depicted in the previous section.This technique was rehashed for every last video in the database.

Testing Example of video from correctly classified videos is presented in the figure:

fig 5. Testing music video.

VI. CONCLUSION

A computationally productive music/non-music video arrangement for compelling video synopsis framework is exhibited. The calculation makes utilization of the MFCC elements of scenes in recordings SVM classifier for order of a picture. Since the choice is made for a total video grouping, the exactness of the single sub-square arrangement could be lower if the video changes definitely. This permits bringing down of the computational unpredictability of the grouping, while the aggregate execution of the framework is kept high. The framework was assessed on true recordings downloaded from You Tube and accomplished aggregate precision of 77.36% on a database of 300 recordings. Future work will address the issues all the more intently to expand the precision. Additionally, the calculation is tried on a database of video arrangements from a solitary source and in addition distinctive sources, for instance – groupings recorded with the camera of a different model of a cell phone .

ACKNOWLEDGMENT

The completion of this research was made possible by Technotraps Solutions, Dhayri. Thanks to this organization.I would like to express my gratitude to my guide, the head of the department and the principal of MITAOE, Alandi for the help and support given by them for completing this work.

REFERENCES

[1] T. T. Teng, L. T. Sze and O. L. Yeng, "Abnormal sound analytical surveillance system using microcontroller," *2016 IEEE 12th International Colloquium on Signal Processing & Its Applications (CSPA)*, Malacca City, 2016, pp. 162-166.

[2] Yingying Zhu and Dongru Zhou, "Scene change detection based on audio and video content analysis," *Proceedings Fifth International Conference on Computational Intelligence and Multimedia Applications. ICCIMA 2003*, 2003, pp. 229-234.

[3] I. Otsuka, R. Radhakrishnan, M. Siracusa, A. Divakaran and H. Mishima, "An enhanced video summarization system using audio features for a personal video recorder," in *IEEE Transactions on Consumer Electronics*, vol. 52, no. 1, pp. 168-172, Feb. 2006.

[4] Y. S. Lee, C. Y. Hsu, P. C. Lin, C. Y. Chen and J. C. Wang, "Video summarization based on face recognition and speaker verification," *2015 IEEE 10th Conference on Industrial Electronics and Applications (ICIEA)*, Auckland, 2015, pp. 1821-1824.

[5] M. M. Salehin and M. Paul, "An efficient method for video summarization using moving object information," *2015 18th International Conference on Computer and Information Technology (ICCIT)*, Dhaka, 2015, pp. 237-242.

[6] J. J. Deng and C. H. C. Leung, "Dynamic Time Warping for Music Retrieval Using Time Series Modeling of Musical Emotions," in *IEEE Transactions on Affective Computing*, vol. 6, no. 2, pp. 137-151, April-June 1 2015.

[7] Chen-Hsiu Huang, Chi-Hao Wu, Jin-Hau Kuo and Ja-Ling Wu, "A musical-driven video summarization system using content-aware mechanisms," *2005 IEEE International Symposium on Circuits and Systems*, 2005, pp. 2711-2714 Vol. 3

[8] G. Costantini, M. Todisco and R. Perfetti, "On the use of memory for detecting musical notes in polyphonic piano music," *2009 European Conference on Circuit Theory and Design*, Antalya, 2009, pp. 806-809.

[9] Y. Qian and M. Kyan, "Interactive user oriented visual attention based video summarization and exploration framework," *2014 IEEE 27th Canadian Conference on Electrical and Computer Engineering (CCECE)*, Toronto, ON, 2014, pp. 1-5.

[10] P. P. Dahake, K. Shaw and P. Malathi, "Speaker dependent speech emotion recognition using MFCC and Support Vector Machine," *2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*, Pune, 2016, pp. 1080-1084.

[11] U. Ben Simon, I. Lapidot and H. Guterman, "Comparison between normalizations for SVM — GMM supervectors speaker verification," *2010 IEEE 26-th Convention of Electrical and Electronics Engineers in Israel*, Eliat, 2010, pp. 000621-000625.

[12] U. Sharma, S. Maheshkar and A. N. Mishra, "Study of robust feature extraction techniques for speech recognition system," *2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE)*, Noida, 2015, pp. 654-658.