EAI Endorsed Transactions
on Creative Technologies

Research Article **EAI.EU**

Automatic Video Classification: A Review

Pooja Rani[1,*], Jaspreet Kaur[1] and Sahil Kaswan[1]

[1]JCDM College of Engineering, Sirsa Haryana 125055, India

Abstract

INTRODUCTION: In last few years number of internet users and available bandwidth has been increased exponentially. The availability of internet with such a low cost is making audiovisual content a more popular and easier form of information exchange. The internet is having a huge amount of this audiovisual content and to classify and choose a particular type of video is becoming a difficult task. A number of video classification methods (like text, audio and video feature extraction) have been proposed by researcher's community
OBJECTIVES: This work is carried out to give a review of different video classification techniques and give a comparative analysis of available video classification techniques and to suggest the most accurate and efficient method of video classification.
METHODS: Text, Audio and Visual video classification techniques.
RESULTS: It has been observed that a combination of audio and visual feature extraction can provide better results.
CONCLUSION: There are various methods of video classification either by using text, audio or video extraction. The text feature extraction is the least used method of video classification. The audio and visual feature extraction is being used in various applications but as we can understand that both the audio and visual feature extractions are having equal importance in video feature extraction but if we use combination of both these approaches, the results in form of accuracy of video classification can be further improved.

Keywords: Video Classification, Audiovisual, Feature Extraction, Bandwidth, Analysis, Review.

Received on 16 January 2020, accepted on 26 January 2020, published on 20 April 2020

doi: 10.4108/eai.13-7-2018.163996

1. Introduction

Now days, people have access to a huge amount of video content over internet and it is becoming a difficult task to find a video of interest from these videos. This increase in popularity of audiovisual content has created a need of such a system which can classify the desired video class from those tremendous videos. There are different methods to classify these videos classification are Text based approach, audio based approach and video based approach and these days combination of audio-visual feature extraction is being used. The first one make use text information of given input video file, the second type uses audio part of the input video file and classifies the system based on its audio information while on the other hand third method is to extract visual features of the input video file and categories that it belongs to which class. The combinational approach which is the recently developed type is based on the fact that both audio as well as visual part of any video contains almost equal information and both of them play an important role for video classification.

1.1. Text based Approach

This is the body text with no indent. The text based approach is further divided into two parts the first one is closed caption and the other is optical character recognizer (OCR) based type. The closed caption technique makes use of speaker recognition method and extracts the text from the speech. By Closed captioning

[*] Corresponding author. Email: poojamehta0193@gmail.com

technique, people with hearing disabilities can understand the video by displaying text of the speech screen. In second type the classifier make use of text written on various objects in the video for example if it is a sports video and players are wearing T-shirts of different numberings then it will extract that text information for classification purpose. The optical character recognizer id used to extract this text part on the object which is filmed. Optical character recognizer uses the text written on the screen like score board, the wordings like Out, Not Out, Over etc.

1.2. Audio Based Approach

In video classification techniques audio based approaches are more popular than text- based approaches. The main advantage audio based approaches is its complexity because it requires fewer computational resources than video based approach. Secondly, if it is required to store the features, audio features are small in size and due to this need less space [1].

Audio features can be derived by the time domain or the frequency domain. The features of time domain are further divided into two parts: I) Root Mean Square (RMS) and II) Zero Crossing Rate (ZCR).The volume standard deviation and dynamic range of volume can be used to detect the volume of any sound signal which is the root mean square of sound signal. Different categories of sounds have many sub bands and energy of each sub band can be measured separately. Zero crossing rates (ZCR) is a rate at which signal changes from minimum to maximum and maximum to minimum level in a current frame. ZCRs are higher in higher frequency signals. Variability of ZCR is higher in speech as compared to music. Any frame can be defined as a silence frame when its loudness as well as ZCR are below threshold level. Frequency Domain based approach can be further divided into 3 different types: I) Frequency Centroide, II) Bandwidth and III) MFCC.

1.3. Video Based Approach

Most of the approaches of video classification which we have studied in the literature work on visual elements, either alone or in combination with other approaches, text or audio features. Because humans receives most of their information in world is by their sense of vision. A video is a collection of images also known as frames. Video based Approach is further divided into 3 parts: I) Color Based II) Shot Based III) Object Based.

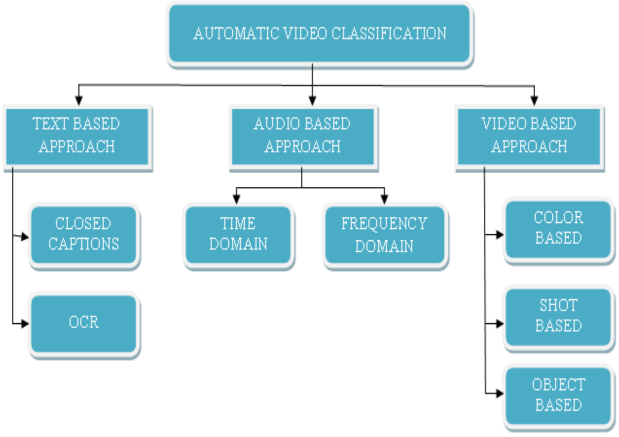

Figure 1. Automatic video classification techniques

2. Related Work

In everyday life videos appear in video games, televisions, CD players, radios and cellular telephones. Despite their vast applications, some problems linger, which should be addressed. Anjum & Cavallaro (2010) [5] note that recently there is a decline in the costs encountered on video equipment while at the same time there is a dramatic increase based on storage capabilities, leading to widespread use of video recording as well as video analytics. Additionally, these video recordings and video analytics are used in different fields such as visual surveillance, sports event coverage, remote sensing as well as home videos.

A study carried out by Ho (2005) [6] on the advances in multimedia information processing mentions that retrieving audio sounds can also be problematic especially when considering the recognition pattern. Ho classifies such problems into two fundamental issues. The first issue is the feature of selection and the second is the feature of classification using selected elements. Aizawa, Nakamura, & Satoh (2004) [7] propose a pioneer method where acoustic features – pitch, bandwidth, harmony, brightness as well as loudness – can be extracted from the known audio signals. According to Anjum & Cavallaro (2010) [8], feature spaces are generated from computing both the mean and covariance from a training set of the relevant multi-dimensional features. Camastra and his colleagues propose a more straightforward method by which pure thresholds can be employed from an average zero-crossing rate as well as energy features. The reason for implementing this technique is to discriminate speech as well as music signals. Bahatti et al. (2016) [9] carry out a study to increase the performances of audio classifications by focusing on the choice of efficient designations and relevant features.

A study conducted by Casey (2013) [10], classifies sounds as a low-level description. Audio spectrum basis as well as audio spectrum projection (based on the Continuous Hidden Markov Model and Probability Classification Model DSs) is defined in the Multimedia Description Schemes (MDS) document. Casey's study later describes automatic classification as the set of categories and relationships. The primary purpose of classification is, however, to provide a semantic relationship between different sets of groups. It is notable that as the classification scheme gets larger and fully connected the utility of the category relationship increases. The study also mentions that more massive classification schemes include sounds such as musical instruments, people, sounds made by animals as well as sounds from film and television.

Another research conducted by Camastra, Vinciarelli, & Yu (2009) [11] on machine learning and video analysis notes that due to increased technology videos produced need to be digitized and quantized to make appropriate classifications. Camastra and his colleagues' primary goal was to come up with techniques for transforming the central input waveform to specific and unique acoustic features of vectors where in real life situation are used as a representation of the information into a small-time window of the same signal. Regardless of high demand to transform the video forms from analog to digital conversations, the processes ought to be more simplified through sampling and quantization.

Yu-Fei Ma and Hong-Jiang Zhang (2003) [12] in their work used motion features which is one of the most important features of any video file. They used motion pattern descriptor and text features to characterize the motion features of a video. To map motion texture of the video kernel support vector machines (SVMs) method is used. A new audio classification method has been introduced by Yu Song et al. (2009) [13]. In this work, the first function was to divide the useful features into frequency and time domain and then the audio features were defined in different categories namely music, speech (pure and non pure), silence. It has been observed from results that the proposed technique provides reasonable accuracy for audio classification in news video.

A method to classify sports type videos by using multiple videos and multiple styles was introduced by Francesco Cricri et. al. (2014)[14].A comparative analysis was done using fuses of modalities separately for each class between multi-class SVM and a modified version of multi-class SVM. In several cases another 2D majority voting technique is introduced to achieve the highest event-level accuracy. The contribution of sensor, audio, spatial visual and spatiotemporal visual modalities in the fusion are dynamically adapted by considering their quality estimates given the input data.

Simon Jones, Ling Shao (2013) [15] have presented a variety of different approaches to content-based human action retrieval, including relevance feedback techniques such as SVMs, ABRS-SVMs and the simple maximum of similarities technique. The experiment showed that simpler methods tend to perform better. The RF algorithm – the maximum of similarities – proved surprisingly effective. Another method for classifying consumer video clips based on their soundtracks was presented by Keansub Lee, Daniel P. W. Ellis (2010) [16]. A set of 25 overlapping semantic classes, is adopted for their usefulness to users, the annotator-enabled labeling, viability of automatic detection and a sufficient amount of representation in available video data collections. By using these concepts groupings of 1873 videos from real users has been annotated. Each video clip has a sequence of Mel- frequency cepstral coefficient (MFCC) frames; the experiment was taken out with three, clip-level renderings.

Matthew Roach (2003) [17] used video types like news, sports, cartoon, music etc. in their work for video classification. The achieved results provided an average error rate (ERR) was 16%, 15%, and 10% respectively for audio and visual, and for combined modes. The best result performance for the visual mode has an EER of 4% which was for cartoons, and the audio mode has shown best result for news with an EER of 0.6%. While on the other hand, the combinational approach provided an EER of 10%and also shown more consistent accuracy across the five genres.

Li-Qun Xu and Yongmin Li (2003) [18] have made three key contributions. 1) For describing contents of a video in more efficient manner they integrated acoustic-visual features; 2) To insert appropriate temporal dynamics, a segment level "concatenated" feature vector within a transitional window; 3) To remove the spatial-temporal redundancy PCA has been applied in the low-level audio-visual descriptors. The video data base was having five categories sports, commercial, news, cartoon and music which were tested. An average classification rate of 86.5% has been achieved, given a 40 seconds decision window.

A new technique with SVMs for hyper spectral data has been introduced by Jin Chen and Cheng Wang (2009) [19]. In this method, magnitude and shape feature spaces of the complementary discrimination information and the ability of generalization of the stacking method were used to enhance the accuracy of the system. There were two levels of SVM classifiers and decision values which were outputs of level-0 SVMs were used as the input of level-1 classifiers. The reason behind using SVMs is their good generalization as compare to other classifiers.

Vakkalanka Suresh et. al.(2005) [20] used HMM and SVM in there video classification model. The classifier extracts spatial and temporal features for video classification and the basic idea was to use multiple classifiers for combining the evidences. For training and testing the model, a video data set with six program categories (cricket, football, tennis, cartoon, commercial and news) was used. Obtained results have shown that the combined approach provides better performance as compared to individual classifiers, and an accuracy of 93.12 was obtained. Further, to improve the performance of classifier other features like audio and text need to be combined with visual features because only visual feature of video are not sufficient for classification purpose. On the other hand, researchers like Aggelos Katsaggelos et. al. (2015) [21] summarized their views by analyzing the work of different researchers and concluded that in recent past years there was not any big contribution in this research area.

The fusion of audio and visual features extraction for video classification were used by researchers like Keiko Sato and Yasue Mitsukura (2011) [22]. Images and piano music was used for visual and audio features respectively. The correlation of impression value and magnitude of physical feature have been calculated to find the relationship in impression of an image and music. color information was considered as physical feature of image and frequency characteristics of music which were obtained by FFT were used in this work. Obtained results have shown that there is a strong correlation between colour information of the image and power of the music with adjective expressing "activity". Amal Dandashi et. al. (2017) [23] explained the requirement of automatic video classification. Their proposed technique is based on a multimodal approach in which they have used Arabic NER to extract text feature from speech which can help to find basic information related to events, location and persons involved. Along with these NER results they have used a visual based and audio processing component for key frame based event detection and to extract noise pattern among events respectively.

Xiantong Zhen et. al.(2014) [24] spatiotemporal steerable pyramid used for human action recognition. To evaluate the performance of STSP many experiments have been done. The max pooling operation, the difference of frames (DoF), and dimensionality reduction techniques were also investigated and their results were validated. The existing representation methods depends upon accurate and tuned tracking algorithms but the proposed method have no bounding boxes. The spatiotemporal filtering was efficiently performed with the use of the three-dimensional separable steerable filters. Obtained results proved that the proposed STSP is an efficient method for human action recognition.

The literature is having different methods of video classification based on text, audio and video feature extraction. Different algorithms HMM, ANN, SVM and RNN, all have their own advantages and disadvantages. If it is possible to combine any of these two or more approaches, then there are advantages of both the methods in one scheme.

3. Conclusion

There are various methods of video classification either by using text, audio or video extraction. The text feature extraction is the least used method of video classification. The audio and video feature extraction is being used in various applications but as we can understand that both the audio and visual feature extractions are having equal importance in video feature extraction but if we use combination of both these approaches, the results in form of accuracy of video classification can be further improved.

Acknowledgements.

This present paper has not been completed without the help of many generous and inspiring people. At this point I want to thank and recognize for their valuable contributions to the following individuals.

References

[1] Brezeale D, Cook J. D. Automatic Video Classification: A Survey of the Literature. IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews. 2008; 38 (3).

[2] Srinivasan. U, Pfeiffer. S, Nepal. S, M. Lee, Gu. L, Barrass. S. A survey of MPEG-1 audio, video and semantic analysis techniques. Multimedia Tools Appl... 2005; 27(1) 105–141.

[3] Gu. L: Indexing and retrieval of audio: A survey. Multimedia Tools Appl. 2001; 15(3), 269– 290.

[4] Davy. M and Godsill. J. S. Audio information retrieval: A bibliographical study. Cambridge Univ. Eng. Dept., Tech. Rep. CUED/FINFENG/ TR.429. 2002.

[5] N. Anjum, Cavallaro, A. Trajectory clustering for scene context learning and outlier detection. In Video search and mining, Springer, Berlin, Heidelberg. 2010; 33-51.

[6] Ho, S. Y. Advances in Multimedia Information Processing-PCM .2005; 6th Pacific Rim Conference on Multimedia. Proceedings. Springer Science & Business Media.

[7] Aizawa, K., Nakamura, Y, Satoh, S. I. Advances in Multimedia Information. Processing-PCM 2004; 5th Pacific Rim Conference on Multimedia, Tokyo, Japan, (2004) Proceedings (3332). Springer (2004).

[8] N. Anjum, & Cavallaro, A. Trajectory clustering for scene context learning and outlier detection. 2010; In Video search and mining, Springer, Berlin, Heidelberg. 33-51

[9] Bahatti, L., Bouattane, O., Echhibat, M. E., Zaggaf, M. H. An Efficient Audio Classification Approach Based on Support Vector Machines. 2016; International journal of advanced computer science and applications, 7(5), 205-211.

[10] Casey. M. A. Sound Classification and Similarity. 2002; Introduction to MPEG-7: Multimedia Content Description Interface 309-317.

[11] Camastra, F., Vinciarelli, A., Yu, J. Machine learning for audio, image and video analysis. 2009; Journal of Electronic Imaging, 18(2).

[12] Ma. Yu-Fei and J. Hong. Zhang Motion Pattern- Based Video Classification And Retrieval. 2003; EURASIP Journal On Applied Signal Processing 199-208.

[13] Song. Yu, Wang.Wen-Hong, Feng-Juan Guo. Feature Extraction and Classification for Audio Information In News Video. 2009; Proceedings Of The 2009 International Conference On Wavelet Analysis And Pattern Recognition, Baoding, 43-46.

[14] Cricri, F., Roininen, M. J., Leppanen, J., Mate, S., Curcio, I. D., Uhlmann, S., Gabbouj, M. Sport type classification of mobile videos. 2014; IEEE Transactions on Multimedia, 16(4) 917-932.

[15] Jones. Simon and Shao. Ling. Content-based retrieval of human actions from realistic video databases. 2013; Information Sciences 236, 56-65.

[16] Keansub L. and Daniel P. W. Ellis. Audio-Based Semantic Concept Classification for Consumer Video. 2010; IEEE Transactions on Audio, Speech, and Language Processing, 18(6) 1406-1416.

[17] Roach. Matthew, Mason. John. and Xu. Li-Qun. Video Genre Verification using both Acoustic and Visual Modes. 2003; IEEE Workshop on Multimedia Signal Processing.

[18] Xu. Li-Qun and Li. Yongmin. Video Classification Using Spatial-Temporal Features And PCA. 2003;International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698), 485-488.

[19] Chen. Jin, Cheng. Wang, Runsheng. Wang. Using Stacked Generalization to Combine SVMs in Magnitude and Shape Feature Spaces for Classification of Hyperspectral Data. 2009; IEEE Transactions On Geoscience And Remote Sensing, 47(7), 2193-2205.

[20] K. Aggelos, B. Sara, M. Rafael. Audiovisual Fusion: Challenges and New Approaches. 2015; In Proceedings of the IEEE, 103(9) 1635-1653.

[21] S. Keiko, M. Yasue. Effects of Music on Image Impression and Relationship between Impression and Physical Feature. 2011, IEEJ Transactions on Electronics, Information and Systems, 131(8) 1451-1458.

[22] D. Amali, A. Jihad, F. Sebti. Audio-Visual Video Classification System Design: For Arabic News Domain. 2016; International Conference on Computational Science and Computational Intelligencen 745-751.

[23] Z. Ning, Ling Guan. An Efficient Framework on Large-scale Video Genre Classification. 2010; IEEE International Workshop on Multimedia Signal Processing, 481-486.

[24] X. Zhen, S. Ling, X. Li. Action recognition by spatio-temporal oriented energies. 2014; Information Sciences, (281) 295-309.