A REVIEW OF MACHINE LEARNING TECHNIQUES USED FOR VIDEO CLASSIFICATION

Seetha Parameswaran[1], Dr. Shelbi Joseph[2]
[1]Research Scholar, [2]Assistant Professor
School of Engineering, Cochin University of Science and Technology

Abstract
The last decade has witnessed a boom in the number of internet users. This also has led to the proliferation of content circulating in the internet including video, audio, text and images. With the ever growing amount of data in the internet demands the need of a system that could classify the different types of data. Among this video classification is a very significant work which will be able to automatically classify videos based on its content. This paper has reviewed many popular video classification algorithms and also has proposed a system which is expected to improve the process of video classification with improved accuracy. The proposed system makes use of the strength of supervised models and fuzzy property to improve the classification accuracy. And also there is very little work which uses supervised algorithms for video classification.

Keywords: Video, Video classification techniques, Visual based approach, Deep learning models

I. INTRODUCTION

With the widespread penetration of Internet worldwide and with the improved availability of cheap connectivity and storage, users have shifted towards using more and more video in their daily use. The large penetration of Social media like YouTube, Facebook, Instagram, etc and decentralized communication paradigms like Whatsapp has significantly altered user behavior, with multimedia content being generated and shared with greater ease than in the past. Video provides a much richer experience to the user and enables users to visualize progression of events over time. Thus, it is easier to show using video how to prepare a chocolate cake instead of giving a written text of its recipe. Or, it is easier to understand a lecture from video than it is to read from text. Video contains both spatial and temporal information. Objects in a video carry meaning by the way they are ordered in a frame. They also carry additional meaning when they move or don't move from one frame to another. While extracting information from a video, it is imperative to identify and classify objects and their transformation.

Availability of large video databases offers another major understanding. Now, we can identify and group similar videos and name or classify them according to the action taking place within a video. For example, there are many videos of birthday parties shared by family and friends online. By identifying the action of blowing birthday candles, all such videos can be grouped together and named as birthday videos.

Video classification is the first step toward multimedia content understanding [6]. The classification of video will enable users to find the information they are looking for much quickly. It will be a great challenge to identify

the type of a video and classify it solely by evaluating the content of the video. A lot of works have been going on this area and many methods are automatically able to classify videos based on the content of the video.

In general, two types of features are used for video classification. [2] The first type of feature is audio feature computed from low-level acoustic properties. The second type of feature is visual feature including color and motion.

Machine learning involves observing a set of tasks and learning from them so that there is a tangible improvement in user goals.

The effectiveness of a video classification system is determined by the feature vector and the classifier. This paper attempts to review the various feature extraction methods used for video classification and the classifiers used. Many papers in the area of video classification published in the past years were reviewed. This work will be helpful to those who intend to start a research in video classification to have an insight into the popular existing works on the field.

The contribution of this paper is organized in the following sections as follows; the section II of this paper discusses the various feature extractors and the classifiers used. The next section lists the taxonomy. And finally in section IV the conclusion and future scope is presented.

II. BACKGROUND STUDY

Early works on video classification was based on Hidden Markov Models (HMM) [2]. Audio and visual features were integrated and fed into two-stage HMM. A new HMM was built for each new feature discovered. The concatenation approach gives better performance only if the features are highly correlated. Face and text trajectories [3] were extracted from video clips and HMM were used to classify a given video clip into predefined categories like, commercial, news, sitcom and soap. Zhou et.al use clustering techniques and a supervised rule based system for video classification [4].

A Gaussian Mixture Model (GMM) based classifier can be used on motion information, extracted from video clips using foreground object motion and background camera motion [5]. Lin et.al [6] combine text feature from captions and visual features from the video clips and use Support Vector Machine (SVM) classifier. Another work by Xu and Li captures a Spatio-temporal audio-visual feature vector and processes it using Principal Component Analysis (PCA) which is then classified using GMM classifier [7]. Using Expectation-Maximisation (EM) algorithm, the parameters of GMM are estimated. In [9] sparse reconstruction of low-dimensional features is used with SVM classifier.

A Recurrent Neural Network (RNN) containing Long Short-Term Memory (LSTM) neurons is trained using SIFT features to categorize actions in sports video sequences [8]. This work by Baccouche et. al. was well appreciated for the accuracy it delivered.

With the advancement in deep learning models and architecture, the feature extraction gets automated. 3D Convolutional Neural Networks were used action recognition [9]. 3D convolutions effectively incorporate spatial and motion information. Long-term RNN models directly map variable-length video frames to variable length natural language text and model complex temporal dynamics [10]. RNN can be optimized with back propagation.

In paper [11] convolutional neural networks (CNN) is used for large scale video classification and it is found that slow fusion model performs consistently better than early or late fusion models. Ng et al[12] evaluate CNN with feature pooling and LSTM RNN and found that Recurrent Convolutional Neural Network may be able to generate better features.

Two-stream CNN architecture, one for spatial features and other for temporal features, is used in [13]. The work in [14] uses rank pooling coupled with CNN on activity and action recognition tasks. The rank pooling encodes temporal information by ordering frames of videos in a chronological order of their appearance. Yoo J. H proposed that the learning algorithm uses a bilevel optimization method using convolutional neural networks. LSTM feature extractor on CNN and batch normalization to improve performance can also be used [15]. Non-linear context gating was introduced in [16] to model interdependencies between features and it was the used to classify videos.

A. ***Taxonomy***

Taxonomy of Video classification algorithms is shown in figure 1.

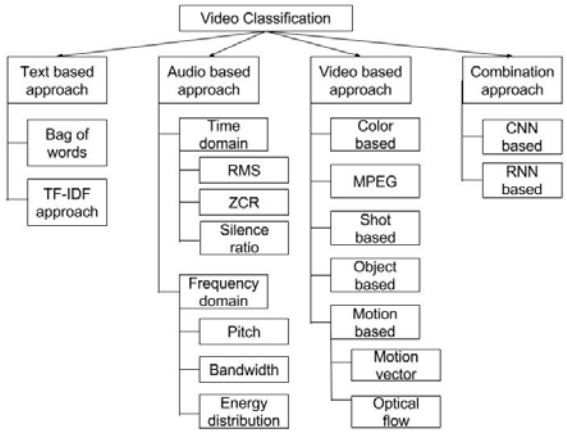

Figure 1: Taxonomy of Video Classification

a) Text based approach

Text-only approaches are the least common in the video classification literature. [1] Text in video can be either viewable text or transcripts. Optical character recognition (OCR) is used convert viewable text into usable text. In case of transcripts in the form of close captioning can be extracted using speech recognition techniques. open captioning requires text detection methods and OCR. Advantages of text based approaches include the easy understanding of relationship between features and genre and application of text mining techniques. The disadvantages outweigh the advantages. Transcript text need not capture the video correctly. Closed captions are not available for all videos. Generation of transcript is difficult in dialog less videos. Generating feature vectors from closed captions is computationally expensive and are error-prone.

b) Audio based approach

Audio features are obtained by sampling [17], [18] and grouping. Audio features can be either time domain or frequency domain. Time domain low-level features are root mean square (RMS) [21], zero crossing rate (ZCR) [20] and silence ratio. Frequency domain low-level features include energy distribution, bandwidth, pitch, mel-frequency cepstral coefficients (MFCC) [19]. Audio features are computationally less expensive and require less storage space when compared to visual features.

c) Visual based approach

Visual based approach relies on visual elements for video classification, either alone or along with text and audio features [1]. Visual features used are color based [25] [26], motion vectors [27], shot based [22] [23] [24], object based [24] [26] and motion based [22] [26]. Figure 2 shows various supervised learning classifiers.

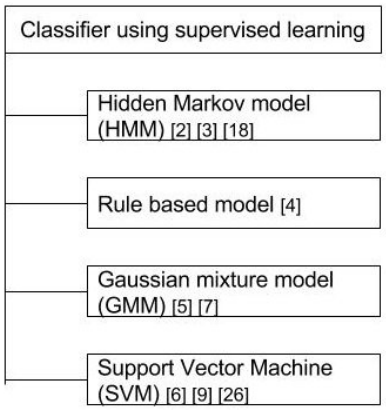

Figure 2: Supervised Learning Classifiers

d) Combination approach

Some combinations of text, audio and visual features are used. A feature super vector is build using convolution or product methods. Most commonly used architecture for a combination approach is by using deep learning method. The commonly used techniques by deep learning methods are shown in figure 3.

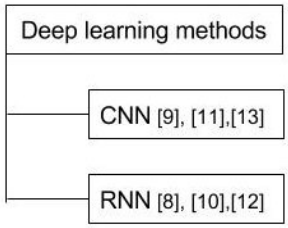

Figure 3: Deep Learning Methods

III. PROPOSED WORK

The methods reviewed so far in this paper uses unsupervised learning to develop video classification methods. This work uses the strength of unsupervised learning method and fuzziness to attain better classifications. In this proposed work the video data is first divided into key clips and then we extract the visual features like color, texture and motion vectors. These features are then subjected to supervised deep learning models and the then results then undergo Fuzzy C means clustering. With the objective of further improving the results we apply an optimization algorithm which then categorizes the video under a label. The proposed system is shown in figure 4.

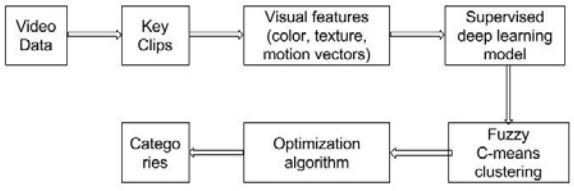

Figure 4: Proposed Method

IV. CONCLUSION AND FUTURE WORK

The paper has discussed various video classifications methods that were proposed in the last few years. The methods were closely reviewed for their strengths and weaknesses. The paper was also form a taxonomy of the techniques available for video classification. The contribution of the paper is in proposing a new method for video classification which uses deep learning models that uses supervised learning methods. The fuzzy property is also used in clustering process using fuzzy C means clustering. It is expected that the proposed system will give more accurate classification of videos over the existing methods. A possible future extension of the work is to use unsupervised training of CNN or RNN and evaluate its behaviour for its accuracy.

V. REFERENCES

[1] Brezeale, D., & Cook, D. J. (2008). Automatic video classification: A survey of the literature. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 38(3), 416-430.

[2] Huang, J., Liu, Z., Wang, Y., Chen, Y., & Wong, E. K. (1999). Integration of multimodal features for video scene classification based on HMM. Multimedia Signal Processing, 1999 IEEE 3rd Workshop on (pp. 53-58). IEEE.

[3] Dimitrova, N., Agnihotri, L., & Wei, G. (2000, September). Video classification based on HMM using text and faces. Signal Processing Conference, 2000 10th European (pp. 1-4). IEEE.

[4] Zhou, W., Vellaikal, A., & Kuo, C. C. (2000, November). Rule-based video classification system for basketball video indexing. Proceedings of the 2000 ACM workshops on Multimedia (pp. 213-216). ACM.

[5] Roach, M. J., Mason, J. D., & Pawlewski, M. (2001). Video genre classification

using dynamics. In Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on (Vol. 3, pp. 1557-1560). IEEE.

[6] Lin, W. H., & Hauptmann, A. (2002, December). News video classification using SVM-based multimodal classifiers and combination strategies. In Proceedings of the tenth ACM international conference on Multimedia (pp. 323-326). ACM.

[7] Xu, L. Q., & Li, Y. (2003, July). Video classification using spatial-temporal features and PCA. In Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on (Vol. 3, pp. III-485). IEEE.

[8] Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., & Baskurt, A. (2010). Action classification in soccer videos with long short-term memory recurrent neural networks. Artificial Neural Networks–ICANN 2010, 154-159.

[9] Tzagkarakis, G., Charalampidis, P., Tsagkatakis, G., Starck, J. L., & Tsakalides, P. (2012, May). Compressive video classification for decision systems with limited resources. In Picture Coding Symposium (PCS), 2012 (pp. 353-356). IEEE.

[10] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2625-2634).

[11] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 1725-1732).

[12] Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., & Toderici, G. (2015). Beyond short snippets: Deep networks for video classification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4694-4702).

[13] Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In Advances in neural information processing systems (pp. 568-576).

[14] Fernando, B., & Gould, S. (2016, June). Learning end-to-end video classification with rank-pooling. In International Conference on Machine Learning (pp. 1187-1196).

[15] Yoo, J. H. (2017). Large-scale Video Classification guided by Batch Normalized LSTM Translator. arXiv preprint arXiv:1707.04045.

[16] Miech, A., Laptev, I., & Sivic, J. (2017). Learnable pooling with Context Gating for video classification. arXiv preprint arXiv:1706.06905.

[17] Liu, Z., Wang, Y., & Chen, T. (1998). Audio feature extraction and analysis for scene segmentation and classification. The Journal of VLSI Signal Processing, 20(1), 61-79.

[18] Liu, Z., Huang, J., & Wang, Y. (1998, December). Classification TV programs based on audio information using hidden Markov model. In Multimedia Signal Processing, 1998 IEEE Second Workshop on (pp. 27-32). IEEE.

[19] Roach, M., & Mason, J. S. (2001). Classification of video genre using audio. In Seventh European Conference on Speech Communication and Technology.

[20] Dinh, P. Q., Dorai, C., & Venkatesh, S. (2002). Video genre categorization using audio wavelet coefficients. ACCV 2002.

[21] Moncrieff, S., Venkatesh, S., & Dorai, C. (2003, July). Horror film genre typing and scene labeling via audio analysis. In Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on (Vol. 2, pp. II-193). IEEE.

[22] Iyengar, G., & Lippman, A. (1998, January). Models for Automatic Classification of Video Sequences. In Storage and Retrieval for Image and Video Databases (SPIE) (pp. 216-227).

[23] Jadon, R. S., Chaudhury, S., & Biswas, K. K. (2010). Generic video classification: An evolutionary learning based fuzzy theoretic approach. small, 7, 0-13.

[24] Wei, G., Agnihotri, L., & Dimitrova, N. (2000). TV program classification based on face and text processing. In Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on (Vol. 3, pp. 1345-1348). IEEE.

[25] Truong, B. T., & Dorai, C. (2000). Automatic genre identification for content-based video categorization. In Pattern Recognition, 2000. Proceedings. 15th International Conference on (Vol. 4, pp. 230-233). IEEE.

[26] Yuan, X., Lai, W., Mei, T., Hua, X. S., Wu, X. Q., & Li, S. (2006, October). Automatic video genre categorization using hierarchical SVM. In Image Processing, 2006 IEEE International Conference on (pp. 2905-2908). IEEE.

[27] Brezeale, D., & Cook, D. J. (2006, August). Using closed captions and visual features to classify movies by genre. In Poster session of the Seventh International Workshop on Multimedia Data Mining (MDM/KDD2006).