Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2019)
IEEE Xplore Part Number: CFP19K34-ART; ISBN: 978-1-5386-8113-8

# An overview of machine learning techniques applicable for summarisation of characters in videos

Gayatri Nair, Karishma Elsa Johns
*TKM College Of Engineering*
Kollam, Kerala, India
{gayatrivikraman & karishmaelsa23march}@gmail.com

Shyna A, Ansamma John
*Department of Computer Science and Engineering*
*TKM College Of Engineering* Kollam, India
{s4shyna & ansamma.john}@gmail.com

***Abstract*—Machine Learning and its applications have been developed and utilised in implementing a wide array of functionalities. Its usage in the area of image and video processing have resulted in an improved mechanism for video analysis. It can further be used in the development of a system that helps summarise the role of characters in a video. In this paper we analyse the existing techniques that can help formulate the required model.**

## I. INTRODUCTION

Machine Learning and its applications are being excessively utilised in different fields as the auto learn technology helps in reducing the manual effort required to perform complex tasks. It now has applications in a variety of fields ranging from medical diagnosis, speech recognition, image analysis. The image and video analysis functionality proves to be very useful in the entertainment industry which consists of movie production and web content streaming platforms. The machine learning applications that are currently in use in these industries include predicting the movie box-office collection, auto recommendation systems, movie content categorisation etc. The video and image processing capability can be improved to implement the functionality of character summarisation. This is useful for the web streaming platforms, as it contributes in enhancing the overall viewing experience for the user.

Character summarisation is the process of identifying details about the role played by a character in the video. These details include description of the emotions portrayed by the character, whether the character is part of the main lead, the screen time of the character, a description of the scenes which has the character's presence. Each of these functionalities can be implemented by modifying and integrating some of the existing methods and applications of machine learning. All the functional modules begin by identifying the character's presence on screen. This is done using the machine learning technique of face recognition. Character detection techniques make use of various face tracking algorithms to obtain the frames consisting of a given character. Once the character is detected the other steps, like identifying whether the character is part of the lead or not, can be followed. The next step of character summarisation is summarising the frames, this can help in a better analysis of the video, as the processing time for summarised frame is less. The various summarisation techniques presently in use make use of both audio and images to examine the important frames and reject the rest. Next step is the process of emotion recognition. It is an important part of the summarisation technique. It helps detect the dominant emotion of a character based on the frequency of all the emotions depicted in the video. The last functionality of getting a brief description of the scenes which a character is a part of, can be administered by performing image captioning on the needed frames. Thus the individual systems on integration can model a character summary module which will make the manual reviewing task automated.

In this paper we present a review of the techniques that are currently used to implement the individual models. We analyse the different methods of Character tagging, Video Summarisation, Emotion Detection and Image Tagging in section 2. We also analyse the current methods that are being used by the various web streaming platforms to improve the user viewing experience using machine learning. Section 3 describes the conclusion of the paper.

## II. TECHNIQUES EMPLOYED

### A. Character Tagging

In order to summarise a character's role we first need to identify the frames in which the character appears. This can be implemented using face detection and character tagging. Character Tagging is the process of detecting the actor and identifying the person in the video based on facial characteristics. Then the frames containing the required character can be used as required. Adaptive appearance model tracking for still-to-video face recognition [1] uses segmentation to obtain the region of interest. A face tracking module then uses a set of pre-loaded gallery face models to identify the character. It makes use of the Sequential Karhunen-Loeve technique which minimizes the total mean squared error. Adaptive skew sensitive ensemble for face recognition on video surveillance [2] also makes use of a set of modules to divide the work, including the tracker, a classification system and a fusion module that improves the performance by utilising spatio-temporal characteristics. One of the biggest problem witnessed during face recognition is Single Sample Face Recognition, where the recogniser is trained using a single image of the target and it needs to identify the target in different poses and lighting condition later. An Approach to Improving Single Sample Face Recognition Using High

978-1-5386-8113-8/19/$31.00 ©2019 IEEE

Confident Tracking Trajectories [3] addresses the problem by using one detector per target individual where each detector has a target individual's face model which is updated with time. Even with enough training data set, variability of appearance poses a problem. The test individual may wear spectacles, hat, beard, moustache, a different hairdo etc. The camera may also have a different angle, it may be moving with different illuminations. Incremental Learning for Robust Visual Tracking [4] makes use of likelihood estimates instead of using gradient descent for optimistically tracking the location, hence providing a possible solution to appearance variability of objects or characters. Support Vector Machines have also been gaining popularity in face detection modules. Online Multi-Face Detection and Tracking using Detector Confidence and Structured SVMs [5] implements Face Structured Detection and Tracking(FAST-DT) method by using a support vector machine based tracker, but it is better implemented on short videos when compared to long ones. The data thus obtained may still have non-uniformity due to noise. Face recognition for video surveillance with aligned facial landmarks learning [6] uses active shape model to align and convert the data into a form that removes the irregularities present.

Some methods uses a large amount of data to improve accuracy. In Offline automatic actor tracking in a movie [7] the authors created a large volume of dataset where they could extract and annotate 2002 different face tracks and it overcomes the demerits of other recognition methods wherein the temporal density of data is less. The models may also use features other than just facial characteristics to track a new individual. Mean shift face tracking with dynamic target model update using Bayesian skin classifier [8] uses the classifier which extracts new skin colours in the frames and help with identification. Similarly End-To-End Face Detection and Recognition [9] uses R CNN which makes the system robust and implementable for any feature. Automatic Actor Recognition for Video Services on Mobile Devices [10] follows a different method and it develops a system which once trained produces a list of frames in which a particular actor is present. A more interesting approach to summarising the movie is recognition of lead actor. In Main Character recognition in the task of movie annotation [11] the authors uses the method of frequency and close up analysis to determine the number of times a character appears on the screen. A threshold based on previous movies can then be used to determine whether the character appears in lead role or not. Most of the methods have shortcomings while dealing with problems related to quality of video, occlusion etc.

### B. Video summarisation

After the corresponding frames are obtained it can be summarised as that helps in efficient storage, quick browsing, and retrieval of large collection of video data without losing important aspects. There are two methods associated with summarisation. In the first method key frames which best represents the video can be selected using either a static or a dynamic approach. In the second method video skimming can be done to summarise. An Improved Algorithm for Video Summarization- A Rank Based Approach [12] divides the process into 3 stages, it calculates the score for frames based on a set of features and keeps the frames with higher scores after eliminating the duplicate ones in the process. The issue here is that the frames thus selected for summary may sometimes have dialogue beginning in the middle of a sentence, to avoid such problems Automated Video Summarization Using Speech Transcripts [13] groups the frames together based on the audio transcript, it uses the inter word pause to understand the difference in sentences and summarises the videos accordingly to avoid long shots. One of the main factors that influence a good summarisation technique is efficient utilisation of storage. Near-Lossless semantic video summarisation [14] uses subshots to extract the representative information. They determine the subshots based on camera motions. It improves upon the storage consumption as compared to others by using compression techniques on the non silence parts of the video. Frame Clustering Technique Towards Single Video Summarisation [15] uses the similarity between a set of frames to group them together as one cluster in an agglomerative manner. Accurate representation of the summarised frames is necessary. Video Summarization from Spatio-Temporal Features [16] works in a similar manner to object identification using spatio temporal features. It uses the Hessian matrix which provides a good accuracy in object recognition. Video Summarization by Learning Deep Side Semantic Embedding [17] uses autoencoders to encode the frame inputs and then build upon it. A summarised video provides a better representation of the data by focusing on the important frames and saving resources in the form of time and storage.

### C. Emotion Detection

Emotion detection is the process of recognising emotions from the facial expressions, body language, as well as verbal language. The different approaches to identify emotions in a video include facial expression analysis, text analysis and sound analysis.

Facial expression analysis methods are popular because the expressions are consistent irrespective of language barriers. Automatic emotion recognition in video [18] uses a tracking model based approach. The existing weaknesses of emotion recognition is overcome using a model based methods by making use of the relationship among the different action units and then recognising the action units simultaneously. An Automatic Emotion Detection System from Face Image of a Video Using Bacterial Foraging Optimization [19] uses the BFO algorithm which helps minimize the mean square error between the noisy and clean image obtained after filtering of frames and hence improves the frame quality. The ability of LSTM to remember while processing sequences increases its usefulness in identifying emotions. Video-based emotion recognition using CNN-RNN and C3D hybrid networks [20] utilises the property of LSTM but it chooses a window of few successive faces and the model is implemented on the middle frame of the window which might not have the required emotion during training. Temporal Multimodal Fusion for Video Emotion Classification in the Wild [21] overcomes the limitation of 3D convolution neural network by analysing every window and then weighing it based on the scores given. A Sentiment-and-Semantics-Based Approach for Emotion Detection in Textual Conversations [22] develops an LSTM based model, it handles textual conversation which may consists of emojis along with the usual text. It implements normalisation techniques to classify the emoticons as out of vocabulary.

Another method of emotion recognition is through text analysis. Textual documents consist of words that may directly convey an emotion. Scene Emotion Detection using Closed Caption based on hierarchical attention [23], it applies a Hierarchical Attention Network model which perform document classification when applying document hierarchical structure. Detecting Implicit Expressions of Sentiment in Text Based on Commonsense Knowledge [24], they created a model by developing knowledge base based on the reaction an actor has to a series of events.

It is possible to extract emotion from audio input also. Real-Time Speech Emotion and Sentiment Recognition for Interactive Dialogue Systems [25] aims at improving the user experience while conversing with a chatbot by infusing emotion identification skills. Their results showed that CNN outperformed the SVM model for emotion recognition from speech.

### D. Image Captioning

It is the process of identifying what the frame represents and producing a text that describes the frame, i.e. it maps the image with its natural language description. It is useful for analysing the content and can be utilised to generate a general description of all the frames in which the character appears. This can be then summarised to obtain character description.

Any deep learning framework will require a large amount of paired dataset. In Unpaired Image Captioning by Language Pivoting [26], the characteristics in source language is used and the required alignment is produced in machine language which works well with the validation test. Most of the RNN models suffer form short term memory issue. Minds Eye: A Recurrent Visual Representation for Image Caption Generation [27] uses bidirectional method that can produce description from image and also a visual image given the description. Their dynamic representation model provides a long term storage solution. In Recurrent Fusion Network for Image Captioning [28] an encoder and decoder module is used. The encoder is implemented using a CNN trained for image classification and the decoder is implemented by LSTM. The captioning power is increased by the usage of Discriminative Supervision. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering [29] implemented both bottom up and top down approaches on the image and concluded that bottom up approach improves the accuracy on the image. Another approach to the captioning of images is usage of text along with visuals. In Text-Guided Attention Model for Image Captioning [30], guidance caption extraction is used along with an image and caption encoder. It may suffer from over-fitting if selection of caption is done on the basis of consensus score.

### E. Current Trends

Amazon Prime Video, an online web streaming platform provides a variety of services to its users, a lot of which is based on machine learning. It recommends new shows to the user based on their previous viewing pattern. A genre of show similar to the one most watched is suggested by the system. The platform also provides an X ray feature which detects the actor present on the screen using Celebrity Recognition and then gets the past working data of the actor from IMDB. Netflix is also a popular web streaming platform that makes use of machine learning techniques to provide the users with a personalised experience which in turn increases the retention rate of customers. Netflix provides personalised poster artwork for the user, so that the probability of the user watching the show increases. They also provide personalised home pages and they aim at using data analysis to improve the quality of videos streamed on their platform. However, these platforms do not provide an auto generated description of the characters in the video.

## Conclusion

In the paper we have reviewed some existing techniques and applications of machine learning that can be improved and integrated to develop a character summarisation module. This is useful for the streaming platforms that are currently limiting their machine learning services to on screen character detection, series suggestion and personalisation based on previous usage pattern. These character summary functionality can be further developed to provide a wholesome viewing experience to users.

## References

[1] M. A. A. Dewan, E. Granger, G.-L. Marcialis, R. Sabourin, and F. Roli, “Adaptive appearance model tracking for still-to-video face recognition,” *Pattern Recognition*, vol. 49, pp. 129–151, 2016.

[2] M. De la Torre, E. Granger, R. Sabourin, and D. Gorodnichy, “Adaptive skew-sensitive ensembles for face recognition in video surveillance,” *Pattern Recognition*, vol. 48, pp. 3385–3406, 11 2015.

[3] M. A. A. Dewan, D. Qiao, F. Lin, D. Wen, *et al.*, “An approach to improving single sample face recognition using high confident tracking trajectories,” in *Canadian Conference on Artificial Intelligence*, pp. 115–121, Springer, 2016.

[4] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, “Incremental learning for robust visual tracking,” *International journal of computer vision*, vol. 77, no. 1-3, pp. 125–141, 2008.

[5] F. Comaschi, S. Stuijk, T. Basten, and H. Corporaal, “Online multi-face detection and tracking using detector confidence and structured svms,” in *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6, IEEE, 2015.

[6] W. T. Lin J, Xiao L, “Face recognition for video surveillance with aligned facial landmarks learning,” *Technology and health care : official journal of the European Society for Engineering and Medicine*, vol. 26, no. S1, pp. 169–178, 2018.

[7] Q. Galvane, J. Fleureau, F.-L. Tariolle, and P. Guillotel, “Automated cinematography with unmanned aerial vehicles,” *arXiv preprint arXiv:1712.04353*, 2017.

[8] M. Pawar, "Mean shift face tracking with dynamic target model update using bayesian skin classifier," in *2012 IEEE International Conference on Computational Intelligence and Computing Research*, pp. 1–5, IEEE, 2012.

[9] L. Chi, H. Zhang, and M. Chen, "End-to-end face detection and recognition," *CoRR*, vol. abs/1703.10818, 2017.

[10] L.-T. Cheok, S. Y. Heo, D. Mitrani, and A. Tewari, "Automatic actor recognition for video services on mobile devices," in *2012 IEEE International Symposium on Multimedia*, pp. 384–385, IEEE, 2012.

[11] G. Dubrovskiy, "The main characters recognition in the task of movie annotation," in *Proceedings of the 2014 IEEE NW Russia Young Researchers in Electrical and Electronic Engineering Conference*, pp. 109–111, Feb 2014.

[12] M. Srinivas, M. M. Pai, and R. M. Pai, "An improved algorithm for video summarization  a rank based approach," *Procedia Computer Science*, vol. 89, pp. 812 – 819, 2016.

[13] C. M. Taskiran, A. Amir, D. B. Ponceleon, and E. J. Delp, "Automated video summarization using speech transcripts," in *Storage and Retrieval for Media Databases 2002*, vol. 4676, pp. 371–383, International Society for Optics and Photonics, 2001.

[14] L.-X. Tang, T. Mei, and X.-S. Hua, "Near-lossless video summarization," in *Proceedings of the 17th ACM international conference on Multimedia*, pp. 351–360, ACM, 2009.

[15] P. R. Sachan *et al.*, "Frame clustering technique towards single video summarization," in *2016 Second International Conference on Cognitive Computing and Information Processing (CCIP)*, pp. 1–5, IEEE, 2016.

[16] R. Laganière, R. Bacco, A. Hocevar, P. Lambert, G. Païs, and B. E. Ionescu, "Video summarization from spatio-temporal features," in *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop*, pp. 144–148, ACM, 2008.

[17] Y. Yuan, T. Mei, P. Cui, and W. Zhu, "Video summarization by learning deep side semantic embedding," *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.

[18] R. KalaiSelvi, P. Kavitha, and K. L. Shunmuganathan, "Automatic emotion recognition in video," in *2014 International Conference on Green Computing Communication and Electrical Engineering (ICGCCEE)*, pp. 1–5, March 2014.

[19] A. r. Xavier, K. Mehata, and M. Ponnavaikko, "An automatic emotion detection system from face image of a video using bacterial foraging optimization," vol. 11, pp. 4288–4295, 04 2016.

[20] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using cnn-rnn and c3d hybrid networks," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pp. 445–450, ACM, 2016.

[21] V. Vielzeuf, S. Pateux, and F. Jurie, "Temporal multimodal fusion for video emotion classification in the wild," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pp. 569–576, ACM, 2017.

[22] U. Gupta, A. Chatterjee, R. Srikanth, and P. Agrawal, "A sentiment-and-semantics-based approach for emotion detection in textual conversations," *CoRR*, vol. abs/1707.06996, 2017.

[23] C.-U. Kwak, J.-W. Son, A. Lee, and S.-J. Kim, "Scene emotion detection using closed caption based on hierarchical attention network," in *2017 International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 1206–1208, IEEE, 2017.

[24] A. Balahur, J. M. Hermida, and A. Montoyo, "Detecting implicit expressions of sentiment in text based on commonsense knowledge," in *Proceedings of the 2Nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '11, (Stroudsburg, PA, USA), pp. 53–60, Association for Computational Linguistics, 2011.

[25] D. Bertero, F. B. Siddique, C.-S. Wu, Y. Wan, R. H. Y. Chan, and P. Fung, "Real-time speech emotion and sentiment recognition for interactive dialogue systems," in *EMNLP*, 2016.

[26] J. Gu, S. Joty, J. Cai, and G. Wang, "Unpaired image captioning by language pivoting," *Lecture Notes in Computer Science*, p. 519535, 2018.

[27] X. Chen and C. L. Zitnick, "Mind's eye: A recurrent visual representation for image caption generation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2422–2431, June 2015.

[28] W. Jiang, L. Ma, Y. Jiang, W. Liu, and T. Zhang, "Recurrent fusion network for image captioning," *CoRR*, vol. abs/1807.09986, 2018.

[29] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and VQA," *CoRR*, vol. abs/1707.07998, 2017.

[30] J. Mun, M. Cho, and B. Han, "Text-guided attention model for image captioning," *CoRR*, vol. abs/1612.03557, 2016.