THE UNIVERSITY OF SHEFFIELD AND UNIVERSITY OF ENGINEERING & TECHNOLOGY, LAHORE AT TRECVID 2016: VIDEO TO TEXT DESCRIPTION TASK

Samyan Qayyum Wahla[†]*,Sahar Waqar*[†]*,Muhammad Usman Ghani Khan*[†] *and Yoshihiko Gotoh*[‡]

[†] AlKhawarizmi Institute of Computer Science,
UET Lahore Pakistan.
[‡] Department of Computer Science,
University of Sheffield, UK

ABSTRACT

In this paper, we aim at presenting our contribution related to video to text description task of TRECVID 2016. Our approach is based on the extraction of high level features using image processing and computer vision stage, scene recognition using HLFs with machine learning approach and generation of natural language is done using HLFs coming from different modules. Second approach is totally based on facial features and gestures. Third approach is similar to first approach but machine learning approach is replaced with deep learning.

Index Terms— Video description, Video to description task, Natural language generation , Emotion, Action , Age, Gender, Scene recognition

1. INTRODUCTION

Video to text description is a pilot task introduced by TRECVID this year[1]. Automatic annotation of videos using natural language text descriptions has been a long-standing goal of computer vision. The task involves understanding of many concepts such as objects, actions, scenes, person-object relations, temporal order of events and many others. In recent years there has been major advancement in computer vision techniques which enabled researchers to start working practically on solving such problem. A lot of use case application scenarios can greatly benefit from such technology such as video summarization in the form of natural language, facilitating the search and browsing of video archives using descriptions, can describe videos to the blind. In addition, learning video interpretation and temporal relations of events in the video will likely contribute to other computer vision tasks such as prediction of future events from the video [2, 3, 4]. A dataset of more than 30000 Twitter Vine videos have been collected. Each video has a total duration of about 6 sec. In this showcase/pilot task a subset of 2000 Vine videos was randomly selected and annotated accordingly. Each video is annotated twice by two different annotators. In total, 4 sets of non-overlapping 500 videos given to 8 annotators to generate a total of 4000 text descriptions. Those 4000 text descriptions split into 2 sets corresponding to the original 2000 videos.

2. VIDEO TO TEXT DESCRIPTION TASK

For Video to Text Description (VTD) task we submitted three runs. Following sections present detailed discussion of these runs.

2.1. Run 1: SAVOUR [2]

2.1.1. Framework Overview

The whole framework is shown in Figure 1. The first step is to convert a video into sequence of frames. Then these sequence of frames are passed through different modules. Modules are listed below:

- Face Detectio and Recognition (FDR)
- Age and Gender Detection (AGD)
- Emotion Detection and Recognition (EDR)
- Object Detection and Recognition (ODR)
- Action Detection and Recognition (ADR)
- Natural Language Generation (NLG)

2.1.2. Image Processing and Computer Vision Stage

Description is related to the humans, their actions, emotions and objects. This is generated by combining outputs of different modules into a sentence. This project involves several different modules and final output is totally dependent on output of all modules. So, system should wait until all modules finish. Synchronization is the major concern of this project as some modules work faster and finish before the others. For example, scene detection module assumes completion of

ODR module for its successful implementation. This shed lights on some of the restrictions that are placed on the time of output generation. A compromise has to be made between time and output generation.

In the start, frames are captured from the offline/online video. Roughly 24 frames are generated per second. It means 24 sentences should be generated per second. This would become too much redundant because no sudden change occurs in one second. Human actions usually take 0.5s to 3s to occur. An emotion would somewhat vary per 8 frames because of change in lip and eyes movement and object movement would also be not too sudden to be changed in 1s usually and scene would never change as camera is fixed. So, to reduce the redundancy, resources, time and complexity we generate sentence per specific number of frames which is discussed below.

2.1.3. Machine Learning Stage

This stage is given the input of high level features from the previous stage and system is trained using different machine learning algorithms for the scenario description. For example, if the extracted high level features are Emotion: Normal, Objects: chairs, table, projects, Age: 3 adults, 2 young , Action: sitting then the system can be trained to describe a meeting scenario [5].

2.1.4. Natural Language Generation Stage

Action is recognized per 8 frames which mean ADR gives no output per 7 frames and on 8th frame sentence contains a verb. This verb is maintained for the next 7 frames and next new verb is defined then, to be use in the sentence. This new verb can be same as previous verb as action of the person might not change for that interval. Emotion Recognizer can generate emotion string per frame if faces are detected but this would not be appropriate as passed minimum frames value when all outputs are available is 8. So, to overcome this issue, two sets of frames are created, one containing 1 to 4 frames and other set containing 5 to 8 frames. If face is detected among first or second set, EDR gives output. If no face is detected then empty string is generated as output. EDR is not supposed to detect emotion per frame because most of the frames are blurred or contains faces that are not recognizable. So, limited set is taken for the detection of emotion. So, if both sets give output, sentence generated after 8 frames gives both emotions as output. In the same fashion, Object Recognizer and Age & Gender module provide information per frame. This is also synchronized with the other modules EDR and ADR and in the end all information collected after 8 frames are expressed in the form of sentence using Natural Language Generation.

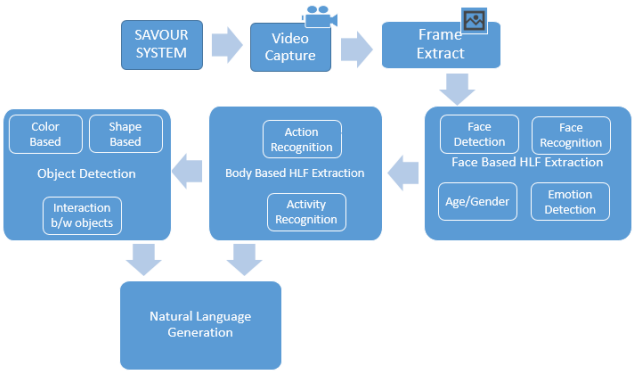

Fig. 1. Process of Video to Text Generation

2.1.5. Video Summarization Stage

Information generated is redundant if seen collectively after some time interval as most of the scenes dont change after 8 frames. So, to reduce this, information is passed to a summarizer which generates the summary of the video. This summary is stored in a text file with statistics about the number of times X action, Y emotion, Z objects came for a, b and c times etc. Final result is in the form of report.

2.2. Run 2: Textual Description from video using facial features and gesture [6]

In this approach, we consider the most important object in the video is human itself [7]. Considering human as most important object, we analyzed that most of the information of the human can be captured from human face and his hand and head based gestures. From face and gestured, we can detect the following high level features:

- Human face detection
- Age estimation
- Gender estimation
- Facial expressions
- Head Direction
- Position of hands

Detected list of high level features are enough to generate the textual description for a video sequence.

- Position of face not only implies the spatial position of human in the video frame but we can also estimate human posture whether human is standing or sitting.
- Facial expressions are helpful to evaluate the intention of a person.
- Position of hands describe the interaction with objects.

Similar to first run, this approach is also divided into multiple stages and modules. these stages are described below how these stages and modules interact with each other to generate video description in textual format.

2.2.1. image Acquisition

In this stage, video stream is captured from the live camera or from a recorded video. Video contains roughly 24 frames per second. The video is converted into sequence of frames. Then the individual frames are sent to individual frames.

2.2.2. Extraction of Facial Features

Facial features are extracted from a single frame, HLFs that are extracted in this stage are as follow:

- face in a frame
- Age of respective faces
- Gender of respective faces
- Facial expression and emotion

Face is extracted from a frame using haar cascade. The focal point premise for Haar classifier object distinguishable is the Haar-like features. These features, instead of using the power characteristics of a pixel, use the complexity changes between the pixel groups.the contrast changes between the pixel gatherings are used to center relative light and faint locales. A couple of neighboring gatherings with a relative intricacy distinction structure a Haar-like features. Haar-like qualities are used to perceive a picture. Haar aspects can without much of a stretch be scaled by growing or reducing the measure of the pixel gatherings being assessed. This grants attributes to be used to spot objects of diverse sizes. Then Geometric and Wrinkle features are extracted from the detected faces using sobel filters and aspect ratios. Emotions and facial expressions are calculated using bezier curve.

2.2.3. Gesture Recognition

In this module, SIFT features are combined with optical flow to estimate the motion of hands and heads. These motion patterns are used to train the system using KNN.

2.2.4. Object detection and Interaction of hands with objects

System was trained for different object and then the spatial and temporal information is used to understand the relationship between the human and different objects. This relationship is used to detect multiple activities.

2.2.5. Natural Language Generation

Similar to first run, extracted HLFs are passed to this stage to generate description.

2.3. Run 3: SAVOUR version with Neural Network [8]

Third run is similar to the first run except that we have replaced machine learning stage with neural network stage.

3. CONCLUSION

In this paper we presented our experiments performed in the TRECVID 2016 video to description tasks. This participation rewarded us an experience in our researches and in finding new ideas and directions in the domain of generation of textual description from videos.

4. ACKNOWLEDGMENTS

First three authors are working for the research grant "Automatic Surveillance System for Video Streams" provided by ICT RnD fund in Pakistan. We are thankful to ICT RnD for providing us research facilities and establishment of Computer Vision and Machine learning lab[1] at KICS.

5. REFERENCES

[1] George Awad, Jonathan Fiscus, Martial Michel, David Joy, Wessel Kraaij, Alan F. Smeaton, Georges Qunot, Maria Eskevich, Robin Aly, and Roeland Ordelman, "Trecvid 2016: Evaluating video search, video event detection, localization, and hyperlinking," in *Proceedings of TRECVID 2016*. NIST, USA, 2016.

[2] Muhammad Usman Ghani Khan, Nouf Al Harbi, and Yoshihiko Gotoh, "A framework for creating natural language descriptions of video streams," *Information Sciences*, vol. 303, pp. 61–82, 2015.

[3] Marcus Rohrbach, Wei Qiu, Ivan Titov, Stefan Thater, Manfred Pinkal, and Bernt Schiele, "Translating video content to natural language descriptions," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 433–440.

[4] Muhammad Usman Ghani Khan, "Natural language descriptions for video streams," 2012.

[5] Alexander Hauptmann, Rong Yan, and Wei-Hao Lin, "How many high-level concepts will fill the semantic gap in news video retrieval?," in *Proceedings of the 6th ACM international conference on Image and video retrieval*. ACM, 2007, pp. 627–634.

[6] Muhammad Usman Ghani Khan, Lei Zhang, and Yoshihiko Gotoh, "Human focused video description," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1480–1487.

[1] http://vision.fukatsoft.com/

[7] Muhammad Usman Ghani Khan Rao Muhammad and Adeel Nawab Yoshihiko Gotoh, "Natural language descriptions of visual scenes: Corpus generation and analysis," *EACL 2012*, p. 38.

[8] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko, "Translating videos to natural language using deep recurrent neural networks," *arXiv preprint arXiv:1412.4729*, 2014.