STUDY OF VIDEO CAPTIONING TECHNIQUES AND APPROACHES FOR GESTURE RECOGNITION

[1] Ms. Nisha R. Birajdar, [2] Mrs. Ranjeeta B. Pandhare

[1 2] Department of Computer Science and Engineering,
[1 2] Kolhapur Institute of Technology’s College of Engineering (Autonomous) Kolhapur, Maharashtra, India

Abstract -:
Captioning subtitles to video directly, it has improved its engagement on social media. Video Captioning stations meaning to action in video. Hand motion also termed as gesture in communication forms non-vocal communication. Video Captioning represents the non-vocal communication in words to sentence. Hand Gestures provide importance for communication in deaf mutes and is acquiring much importance in market of Entertainment, Education, Healthcare Industry. Artificial Intelligence and Robotics are wide areas used by Hand Gesture Recognition. This paper presents study of different video captioning techniques with approaches of gesture recognition.

Keywords: - Gesture Recognition, Video Captioning, Human Computer Interaction.

I. INTRODUCTION

Communication is increasing it is driving its importance in day to day life of computer generation, also any change in way with Human Computer Interaction (HCI) is facilitating. The main goal is to bring Human Computer interaction at its finest and acceptable level.

1.1 What is Video Captioning?

Instinctively generating the textual explanation from a non-natural system is the chore of image captioning, in identical mode a glance at order of images with respect to time to distinguish an action is termed as Video Captioning. A quick glance is sufficient for us to understand and describe what is happening in the picture or video.
Captioning is of two types, Open or Closed. Open Captions always can be seen and can’t be turned off, whereas Closed Captioning is displayed on television and can be turned on and off by the viewer.

Captioning is classified into two types Image Captioning and Video Captioning, Image captioning states what is depicted in the picture which is in static format and it does not include any action or motion. Video Captioning states motion or ongoing action in video, video is again a set of images which run hand in hand in order to generate particular action.

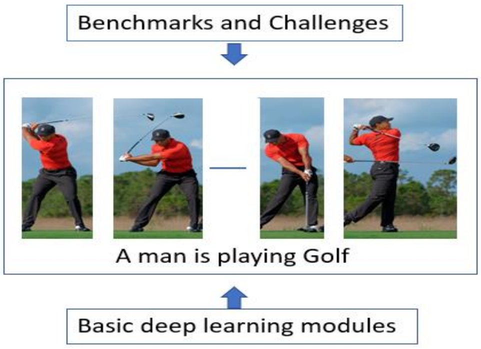

Figure 1. Example of Video Captioning, depicting an action of playing golf game in different frames of one video clip

Captioning is a text form substituting to audio information in video and animation clips. Words, phrases are part of captioning statement which deliver a meaning. Captioning goes hand in hand with visual content. Captions aim to describe a significant meaning to the deaf and hard of hearing people and it also can be considered as application of captioning a video.

Classification of Captioning: There are two types to caption the video Open Captioning and Closed Captioning. Open Captioning states to captioning the video movement to user viewable form.

1.2 What is Gesture Recognition?

Here, there is a difference in captioning and recognition. Captioning is nothing but labelling to an action in video or to label an image. And Recognition has a meaning of identifying an accurate action in video or picture in image.
Gesture itself has a meaning as motion or movement of different human body parts, such as Hand gesture, Eye Gesture etc. If gesture recognition is being performed on hand, it is called as Hand Gesture Recognition. Hand Gesture Recognition states particular

movement performed by hand. As we have seen that every movement of hand has meaning, for example if a person is showing his/her thumb in upward direction to someone then it can have a meaning as All the Best. This can be the example of Image Recognition because here action is static.

Therefore, in Recognition there are two parts that are Image Recognition and Video Recognition. Example of image recognition is given in above paragraph. In case of Video Recognition, for example if a person is shaking his hand side by side in air can have meaning of saying Hello. This can be small action depicting a video with a meaning.

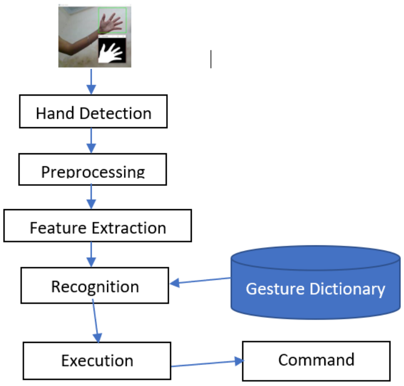

Figure 2. Block diagram of Gesture Recognition, and the steps involved in gesture recognition

II. RESEARCH STATUS

Gesture Recognition with video captioning is a challenging task in deep learning era of computer science and human computer language interaction. And deriving strategies as well as approaches for gesture recognition with video captioning is also a thoughtful challenge.

Video captioning approaches are derived in two ways from previous research status: template-based language model [1, 2, 3] and CNN/RNN fusion sequence learning model [4, 5, 6, 7]. Both of the strategies mentioned above generates caption result for video when already a set of templates are available. Combining to those templates with ruling syntax caption is generated for video.

2.1 Template-based model

It is majorly depending on the predefined templates and thus the resulting sentences are always with a constant syntactical structure.

Template based models contains alignment of words detected from video frames with sentence ruling fragments such as subject, verb, object and then try to generate caption with language patterns that are predefined.

Actually, as depicted in the below figure 3 different objects are derived through single frames and are in pipe. At very first faster RNN is applied on frames to detect object in a frame. Then those objects are linked to each other to depict the action. Each object gives a meaning then those meaningful words are correlated with each other to generate a ruling caption for a video clip. After pipes correlated the encoder-decoder scheme is applied and encoder is a bidirectional LSTM, which is used to gain dynamic information of objects to generate caption.

It is majorly depending on the predefined templates and thus the resulting sentences are always with a constant syntactical structure.

Template based models contains alignment of words detected from video frames with sentence ruling fragments such as subject, verb, object and then try to generate caption with language patterns that are predefined.

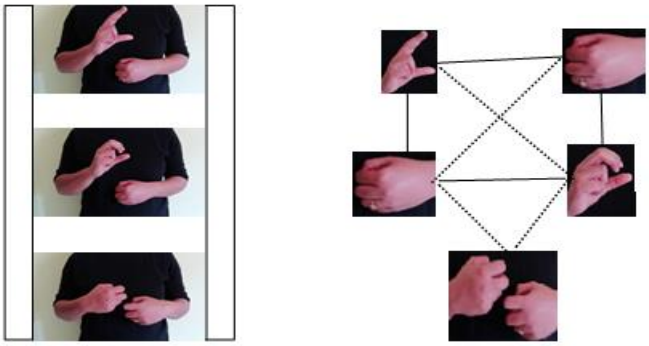

Figure 3. The video similarity in three different frames depicting ‘not sure’ action, to identify an action frames features extracted are shown

2.2 CNN/RNN fusion Sequence learning Model

In divergence Recurrent neural network (RNN) are well known for more accuracy because of their repeated nature to extract the features. Sequence Learning Model is also inspired by RNN in Deep Neural Networks (DNN). The concept of RNN/CNN convolutional neural network is totally depending on the concept of encoder and decoder work for translation. Encoder and decoder concept are derived from working of CNN and RNN fusion. Encoder of convolutional neural network (CNN) produces vector representation of features from frames extracted from scanned videos. And then result of encoder CNN are given as a input to decoder RNN which intern predict a caption for scanned video frames. Fusion of CNN and RNN is considered very effective for generating the natural language sentence for frames in video.

2.3 2D/3D CNN on Frames

2D/3D networks have been applied to study features containing some movement**.** The features studied on image set which are extracted through small duration video, the duration which is few seconds to some minutes. The network better performs on set of layers, because single baseline features does not provide marginal information. Then collectively the features provide a caption.

2.4 CNN/RNN on Frames

CNN and RNN is deep fusion network, specially evaluated for gesture recognition. Previous models have some followed procedures, as in Template-based model which is having set of templates identified from frames in its last convolutional layer are given as a input decoder of LSTM. In CNN/RNN fusion Sequence Learning Model the layers of convolutional model are fully connected for better accuracy in recognition of the gesture, and that output is given as input to decoder LSTM. As 2D/3D CNN is a bit different, where output of decoder LSTM are combined and result for gesture recognition is generated. But the fourth model is fusion of all where result of convolutional layer patterns and result of fusion sequence learning model is combined then the result will be generated. The result of the last model has higher accuracy[10]

According to [20] study for research, the CNN model contains well known Convolutional and Deconvolutional blocks. And each block consists of convolutional layer with filter of 3x3, ReLU (Rectified Linear Units), max pooling layer with filter of 2x2 or deconvolutional layer with filter of 4x4.

III. APPLICATIONS AND CHALLENGES

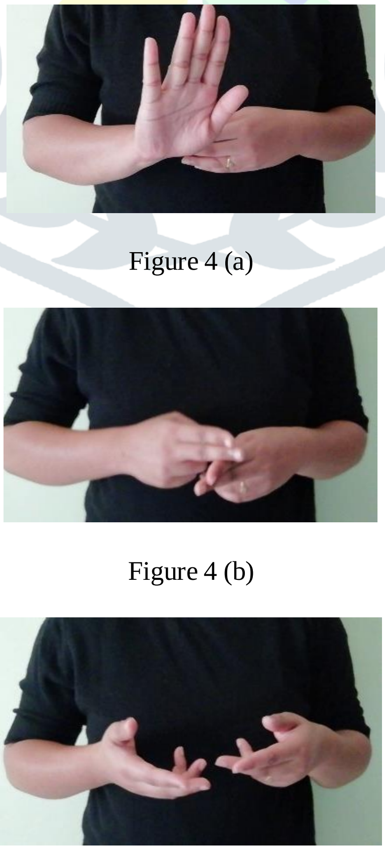

Figure 4 (a)

Figure 4 (b)

Figure 4 (c)

Figure 4 (a, b, c). Example of Application and Challenges in Gesture Recognition with Captioning

Above Figure 4 depicts an action from 3 different frames as Figure 4(a, b, c). Above frames also can be considered as Application and Challenge in Gesture Recognition and Video Captioning. Above frames are derived from a video of sign language of action ‘what is your name’. Frame 4(a) tells a sign ‘your’, frame 4(b) tells a sign ‘name’, frame 4(c) tells a sign ‘what’. Therefore, combining the model generates a caption ‘what is your name’.

3.1 Applications of Gesture Recognition

1. Improvement of social life of elderly and deaf people. Human Computer Interaction is beneficial for communicating with kids, deaf-mutes and the computers. Hand Gesture Recognition can simplify the way of communication with deaf mutes.

2. Control on Virtual reality and e-Environmental things. Hand Gesture Recognition is beneficial for completing some tasks like surfing, selection or handling of the virtual environment. Hand Gestures can govern the task in it can be a reality.

3. Smart Appliances in home are also in queue of applications for gesture recognition. Nowadays, home appliances also have control on home or one can have control on anything through smart devices using smart assistance by intelligence. When there is a control by hand gesture one can thought as controlling tv or laptop through hand gestures. Smartphones that are touchscreen devices also can be controlled through hands.

4. Robotics is a huge are where use offhand gestures can be seen on wide range. New devised robots can be assisted through hand gestures and the recognizing ability of the robot on human actions or reactions can be studied.

3.2 Challenges of Hand Gesture Recognition with Video Captioning

Gesture Recognition is quite difficult because detecting a static object and naming it is considered easy but if a clip contains movable action then detecting an object as well as its related action it a bit difficult and challenging. Hand gesture recognition with video captioning is with lot of work done but the area still facing some challenges, such as:

Different feature mining, moving gestures, sign language recognition, recognition approach with RGB values and scalability features, and recognition of multiple gestures [11].

1. Object detection difficulty

   Object Detection from a frame that is nothing but image extracted from video clip. Object is any entity, person, living or non-living thing that has particular features. In an image object can be static, therefore detecting that object is easy but if compared with frames which are sequenced depicts an action. So, detecting object as well as its features from frames of video is little bit challenging.

2. Object recognition difficulty

   As object detection is difficult, object recognition is also important as well as difficult task. Need of object recognition is easy when recognition is from image because there no action present in image. If compared with case of videos object recognition is little bit challenging. In video action is present, after video is separated into frames, recognizing object as well as action of an object is difficult.

IV. CONCLUSION

In this paper, Video Captioning techniques and gesture recognition techniques are reviewed include template-based model, sequence learning model, 2D/3D CNN on Frames, CNN/RNN on Frames are often used gesture recognition techniques due to their accuracy, both of their applications and challenges are discussed. Also, difficulty in detection and recognizing object for captioning is important concern, it can be thought as challenge in Gesture Recognition. Application and challenges of video captioning. As we know deaf and dump people have disability to hear and talk, so hand gestures are considered as their sign language to communicate. Studying sign language is also a thoughtful and challenging task. Because of the variety and complex nature in gesture recognition and captioning is facing many challenges due to background environment factors.

REFERENCES

[1] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In ICCV, 2013.

[2] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele. Translating video content to natural language descriptions. In ICCV, 2013.

[3] R. Xu, C. Xiong, W. Chen, and J. J. Corso. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In AAAI, 2015.

[4] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui. Jointly modeling embedding and translation to bridge video and language. In CVPR, 2016.

[5] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko. Sequence to sequence - video to text. In ICCV, 2015.

[6] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. In ICCV, 2015.

[7] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu. Video paragraph captioning using hierarchical recurrent neural networks. In CVPR, 2016.

[8] I.Sutskever,O.Vinyals,andQ.V.Le. Sequence to sequence learning with neural networks. In NIPS, 2014.

[9] Q. Wu, C. Shen, L. Liu, A. Dick, and A. v. d. Hengel. What value do explicit high level concepts have in vision to language problems? In CVPR, 2016.

[10] Human Action Recognition by Learning Spatio- Temporal Features With Deep Neural Networks, Lei Wang , Yangyang Xu, Jun Cheng, Haiying Xia, Jianqin Yin, And Jiaji Wu, (Member, IEEE).

[11] Yanan Xu and Yunhai Dai Review of Hand Gesture Recognition Study and Application. In CES, 2017.

[12] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In ACL workshop, 2005.

[13] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach,S.Venugopalan,K.Saenko,andT.Darrell. Long-termrecurrent convolutional networks for visual recognition and description. In CVPR, 2015.

[14] C. Gan, T. Yao, K. Yang, Y. Yang, and T. Mei. You lead, we exceed: Labor-free video concept learning by jointly exploiting web videos and images. In CVPR, 2016.

[15] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei. Boosting image captioning with attributes. arXivpreprintarXiv:1611.01646, 2016.

[16] X. Ji, J. Cheng, D. Tao, X. Wu, and W. Feng, ‘‘The spatial Laplacian and temporal energy pyramid representation for human action recognition using depth sequences,’’ Knowl.-Based Syst., vol. 122, pp. 64–74, Apr. 2017.

[17] A. Kojima, T. Tamura, and K. Fukunaga. Natural language description of human activities from video images based on concept hierarchy of actions. IJCV, 2002.

[18] Q. Wu, C. Shen, L. Liu, A. Dick, and A. v. d. Hengel. What value do explicit high level concepts have in vision to language problems? In CVPR, 2016.

[19] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, et al. From captions to visual concepts and back. In CVPR, 2015.

[20] Haitam Ben Yahia. “Frame Interpolation using Convolutional Neural Networks on 2D animation”. MA thesis. University of Amsterdam, Aug. 2016.