xIlkogretim Online - Elementary Education Online,2021; Vol 20 (Issue 5): pp.2602-2609
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.05.282

Integrated Application Model for Hand Gesture Recognition

Sijjad Ali Khuhro, School of Computer Science and Technology, University of Science and Technology of China
Iftikhar Ahmed Koondhar, School of Computer Science and Technology, Beijing Institute of Technology, China
Adeel Abro, School of Electronic Engineering, Beijing University of Posts and Telecommunications, China
Khurram Hussain, Department Control Science and Engineering, Beijing Institute of Technology, China
Saleem Raza, Quaid-e-Awam University of Engineering, Science and Technology, Larkana, Pakistan
Zulfiqar Ali Bhutto, Dawood University of Engineering and Technology, Karachi, Pakistan

Abstract: Hand gesture-based human-computer interaction is both intuitive and versatile, with diverse applications such as in smart homes, games, operating theaters, and vehicle infotainment systems. An effective human-computer interaction system is required a good accuracy rate of recognition and speed. In our work, we have proposed a system model for static hand gesture recognition by using multiple common features. There are three contributions in this model:(1) A multiple features classification based on the Non-Dominated Sorting Genetic Algorithm II (NSGAII). The use of NSGAII can be reduced redundant features and minimizing feature value which effective on the execution cost of the system. (2) Proposed a new methodology of multiple features convolutional neural network (MFCNN) model to recognize both common and real-time hand gestures. (3) The generation of sequence sentences based on the Beam Search (BS) algorithm. Data of image labels that were received from the recognition stage combine with the CNN/Dailymail dataset is used to generate sentences.

Keywords: Hand gesture, Non-Dominated Sorting Genetic Algorithm II (NSGAII), multiple features convolutional neural network (MFCNN)

I. INTRODUCTION

a. Object features extraction

Commonly, the objectof hand gesturefeatures extraction canbecharacterizedby the special identification of an interesting object in the images. The proper selectionof these characteristics is the main ideato increasethe qualityof the system recognition and tasks analysis [1]. It is an essential workthat we needto periodically calibrate to serveamodel of hand gesturesrecognition. Generally, the naivety approach of extractionmethodswillfind as the features that expect to extract from the images,thosefeaturescan beutilizedfor training and testing. Some methods exist for predictingthesimilarityof images based on the multiple featurematching.Zhang et al. presented that applyingexternalfeaturescanachievea high accuracyrate and handle some partialchangesin the image such as objects moving in a scene [2].In addition; some methods are robust to extract the image featuresin different environments such as scaling and rotations, size, light condition changes, etc. For example,theyapplied ageometry technique to verify the matching of image gestures [3].

More recently, there are many successesin the application of various methods to extract image features. Adapted an extraction method to extract and classifyfeatures from anobject. Their worksweretested and compared with a few methods such as the principal component analysis (PCA), Fourier descriptor (FD), and the standard Zernikemoment (ZM). The result shows that the best performancewasobtained from the usesofboth Zernike Moment and discriminative Zernike Moment[4]. Ng and Ranganath have proposed a system to represent hand blobs[5]. The study considers a vision-based system that can interpret hand gestures in real-time. Overwhelmingly, the hand segmentation procedure extracts binary of the sequence image gestures from each frame by using Zernike moments and Hu moment. The resultsshowthat it canachieve a good result inpresentingeach hand blobs. Also, we noted thatfeaturesextractfrom Zernike moment can perform better than Hu-moment [6].Wang C. and Wang K. applied theHu-moment method to extract features along with valley circle features for real-time recognition of static hand gestures from a2D image [7].In this work, they designed the robot movement control system and used the NTMR algorithm for hand gesturesrecognition. The experimentalresultsof this work demonstrated that the combination of NTMR and Hu-moment can gainan accuracy recognition rate of4% and reduce the time of execution by112ms.Similarly; Guo et al. [8] presented a solution to improve the version of the Zernike moment method. This method can be used for measuring accuracy and cost on a general static of hand gestures. The experiment results showthat their method canreduce the amount of complexity of an image, and increasethe speed of execution.Significantly, compared to the original Zernike moment, this effect improvement becomes obvious as the order increases. Pichao W et al. [9]appliedan RGB-D method to extract the features from abinary image and to combine them with the recognition model. From this work, we observed that the use of RGB-D can present comprehensively depending on objects. Based on the hypothesisof this research, the useful features of the RGB-D method can achieve a good result as listedin Table 3 of [9].However, the RGB-D method has faced the risk of annoyancefrom color, light condition, and different environments. These problems can be anobstacletoincreasingspeed, reducingresources, andcost, etc. To challenge these problems, some studies have been used depth data to improve the quality of the image features. Palacios et al. presentedhand gesture recognition by using the featuresextracted from thosedatato serve in the system [10]. The experimentalresults of this work shownthat the different static hand gestures can berecognized fluently on different combinations of spread fingers, open hand movement, and 6 dynamic gestures. As mentionedabove, many methods can be usedforextractingthefeatures of animage. However, some methods cannot extract agood feature set because of the problem of color changes, density of image, rotation, scaling, and light condition. So, the recognition model will be faced with problems suchasthe confusioncase, timely execution, and high-cost expense. Our research will apply the special feature extraction methods that can extract the good features and those that deservetobecombinedwith the recognition model.

b. Inspiration for using NSGAII to classify object features

The main purpose of the objectfeatures classification is to define the fitness feature set. All features will be evaluatedand selectedthrough an evolutionary algorithm called theNon-Dominated Sorting Genetic algorithm II (NSGAII). Additionally, NSGAII can perform to reduce noise and minimize the feature value. The fitness feature set will select to feed in the recognition model. In this section, feature classification is one of the special tasks in our research. However, the use of numerous methods to extract features from the object should be revised clearly before being combined with the recognition model because those features can be redundant and add little value.Especially,those are the core problems that the recognition model. The enforcement of a good feature set can gain the capability of a recognition model as well as reduce thenoise of features so cost, time, speed, and accuracy rate will be increased also[1]. The most related work, F. Chevtchenko et al. [11] presented about the feature evaluation and selection. To evaluate and select features set they use an evolutionary algorithm called Non-Dominated Sorting Algorithm II (NSGAII). This approach also usedthe recurrent neural network (RNN) combined with multiple features to classify the hand gestures from Benchmark's dataset. More significantly, to gain the rate of accuracy, firstly theyextractedfeatures byusing somemethods such as Gabor filter, Zernike moment, and Pseudo Zernike moment, and then those features were evaluated andselected by NSGAII algorithm. Finally, the fitness features wereutilizedtofeed into the recognition model. The experiment results demonstrated that the rate of recognition is higher than someresearches especially; it can save more time execution and resources. Generally, the NSGAIIalgorithm can organize in some works to select the finest solution and resolve multiple problems also. Delgado, M. et al. [12] presentedmultipleobjectives optimization and training in RNN. The research shows thatthe use of the RNN model still lacked a capable training algorithm and sometimes the processing due to problems of vanishing gradient and addressed the training and topology optimization of using RNN multi-objective hybrid procedures. In this context, to measure the hybrid of objectives, theyhaveusedthe NSGAII algorithm. The research results illustrated that it can achieve a good result. Agarwal, A. et al. [13] applied the NSGAII algorithm for generating optimal heat exchanger networks. The study involves randomly generating the number of intermediate temperatures in each of their value and stream, bothhot and cold. In each solution, the chromosomes of NSGAII responded to store the decision variable, and this variable can also reveal the length of the chromosomes. The experiment results have shownthat the use of NSGAII can help to improve the speed of the convergence, and all the solutions can be analyzedalso like the above mentions, NSGAII can help usto definethe fitness value of the multiple objectives in numerous works. The motivation of using the NSGAII from thoseworks,so we also apply NSGAII algorithm to classifymultiple features that were extracted from all methods. The fitness features set will be usedto feed the recognition model.

c. Handgesturesrecognitionby using CNN

The CNNis the most popularmodelthat has been applied by numerous researchers. This model can recognize the small and big data of human acquisitions. Moreover, it canuse the common features to train and test data inside processes. We believe that the combination of common features with theCNN model can reduce the costexpending, boost speed execution, and especially produce a high system accuracy rate also.CNN can recognize both common and real-time gestures. The recognition of an objectgesturecan expressacharacter and the meaning of its gestures. In this section,theuse of the CNN model to recognize objectgestures will brieflybe reviewed. More recently, many researchersaboutobjectgesturerecognitionhave useddifferent techniques. Barros et al. [14]have adopted imagegesturerecognition by utilizingaCNN model todecomposea layer of CNN into three layers. The architecture demands to convert theoriginalimage as grayscale and rescale image size into 32-dimensional pixels. Based on the experiment results of the recognition shown that the system accuracy rate isabout 90% more than some methods. Inspiredby the aforementioned studies, the multi-channel of the CNN model is usedin the present article. Sergio et al. [15]proposed RNN multiple channels by using some external features to recognize hand gestures in real-time. This research demanded to covert the image as grayscale and convolves by a Gabor filterbefore combiningwith the system.The feature of the Gabor filterwasarrangedby usinga hyper-parameter selection algorithm. Besidesthat,a set of pairs of convolution and max-pooling layers are enabled to extract and motivate thespecific features of an image'sgestures. In max-pooling layers, a region of the previous layer is connected to a unit in the current layer and then the dimension of feature map valuesis reduced. For each layer, only the maximum value is assigned. The parameters of this model can be trained either by both a supervised approach tuning the filters in a training database [15] and an unsupervised approach [16].This study used a supervised approach. The experiment results illustrate that the system accuracy rate is higher than some researches, theaccurate rate rangesfrom 95% to 98%. In this context, to improve the rateof accuracy, some studies used a depth camera or sensor to filter the high-quality gestures before combiningwithaCNN model. Sha, L. et al. [17] applied a KinectTM depth sensor toincreasethe robustness of image capture and features extraction forCNN single channel. We observed thatacross preliminary training of random gestures classifier, the systemcanfiltera characteristic with less influence, computation costs are also reduced. Similarly, Plouffe et al. presented a combination of aKinectTM sensor with a CNN model for hand gestures recognition [18, 19]. The area of the hand's fingertips and palm have been detected and extracted as feature descriptors to feed in the network layers. The experiment results show that this application can recognize gestures with an average delay of 100 msand the accuracy rate is higher than some researches that used a webcam camera. As mentioned above, the CNN model can achieve good results for object gesture recognition. The use of multiple objects features to combine with CNN multiple channels still doesn't haveenough research. To challenge this, we will propose a novel architecture of CNN two channels to recognize the human hand gestures. Additionally, this architecture will feed by multiple common features descriptor. The architecture is designed as the last fully-connected layer to take information from both convolutional and common image features. The objects can recognize by using an auxiliaryfeature at the activation layer, the highest activation as aresult of recognition. Especially, our modelcan recognize hand gestures on depth or 2D objects frame, common and real-time also.

d. Inspiration for using beam search to generate sequence sentences

The sequence sentence generation is a critical task in our research. The generation of sequence sentences will make peopleeasilyunderstand the meaning of hand gestures are performed by auser. To generate sentences, we have applieda beam search (BS) algorithm to generate an object label performed in the recognition model. Currently, many kinds of research have applied BSalgorithm to generate sentences. Some researchers jointly embed image labels and language into a multimodal embedding with a neural network-based language model to generate sentences [20], [21]. Kiros et al. [22] proposed a multimodal log-bilinear neural language model which is biased by an image's label to decode the sentences. Tan et al. [23] andKarpathy& Li [24] applied a BSalgorithm to decode an image label of varying lengths. Wu, Q. et al. [25]useda BSalgorithm to decode image descriptions from their respective context. The context for word generation can be any described in the next section or a combination of several types of words. Chen et al. [26] applied a BSalgorithm to develop a dynamic visual representation of words generated to aid the next words predicted during caption generation. Tillmann, C., and Ney, H. proposed the machine translation by utilizingaBSalgorithm also [27].This study presented a novel technique tocombine the possible words reordering between resources and direction language to gain the ability of the searching algorithm. The experiment results showed that the use BSalgorithm is successfully tested on the Verbmobil tasks, and can translate some languages such as the German language to the English language with 8,000 words and the French language to the English language with 100,000 words. Significantly, this algorithm can translate in a short time, and only a minimal number of errors occurred. Concurrently, the BSalgorithm has alsobeenappliedto interpret and describe the image description, video caption, and movie description [28]. Barbu, A., et al. and Kojima, A. et al. [29, 30] presented themotionsof an objectfrom sequence video images to textual descriptions through Beam Search. Das, P. et al. and Guadarrama, S., et al. [31, 32] proposed theobject activities recognition of the videos with a small-scale and natural language descriptions generation by using BS algorithm. Similarly, Donahue, J., et al. [33] and Rohrbach, A., et al. [34] applied a BSalgorithm to provide video descriptions. Pan, Y.W., et al. [35] implemented a beam search algorithm joined with a visual-semantic embedding technique to produce sentences from YouTube videos. Xu, R., etal.[36] presented video captures and sentence generation by using BSand a deep neural network model. Rohrbach, A., et al. [37] presented movielinguisticdescriptionsgenerationbyutilizingBS.Yao, L., et al.[38]applied BS algorithm togenerate text descriptionsfromthe most relatedtemporal segments in a videowhich is selected by an attention-based model. As mentioned above, the applicable sentence generation canbe a convenient tool for humansto easily understand. However, generatingimage labels to the sequence sentences by using BSstill doesn't have enough researches. The use BSalgorithm bynumerousresearchershas beenmotivated us to apply it in our research also. In this context, we utilize a BSalgorithm to generate image labels as the sequence sentences to express the full meaning of each human gesture that is actedintherecognition model. We hope that the sentence generation can make users easily understand the meaning. Although sentence generation byusinga BSalgorithm is not a new application in this work, we hopethat itcan be a pioneer to help other related researches. In conclusion, many significant methodsare used to extract the features from the images. Commonly, features should be evaluated and selected carefully before using them because they may be affected by noise such as color changes, light condition changes, scaling, and rotation. In this context, an evolutionary algorithm called NSGAII is utilized in our research to classify the multiple features and select the fitness features to feed in the recognition model. Also, drawing inspiration fromthese works, we proposed a novel CNNtwo-channel combinedwith common multiple features fusion. In this network architecture, we designed the last fully-connected layer to receive information from both convolutional and common features descriptors. The objects will be evaluated by auxiliary features anactivationlayer. Finally, the BSalgorithm will perform at the last processes of the system to generate the sequence sentences from eachimagelabel.

II. CHALLENGES

In our proposed model there are three main challenges: 1.Reducingthe expense of the resourcesfor hand gesture recognition, the system can execute smoothly by utilizingthesmall computer cost.For Example, CPU and Webcam camera. 2. Good recognition,utilizinga small cost of machine capacity, butthe system can produce a high rate of accuracy and predictcorrectlyforhand gestures recognition in both common and real-time 3.The highsystemperforming,thissystem can perform faster than somesystemsin the same level (CPU, Webcam) of static hand gesturesrecognition also.

There are the main questions that are addressed in our proposed model: 1.What are the main methods that use to extract and classify the multiple features set from datasets? 2. What is the special key for a novel CNN model in this research? 3. What are the main challenge methods that use to generate the sequence sentences?

III. RESEARCH GOALS AND OBJECTIVES

The main goals of ourmodel are derived from the purpose of developing a method to use static hand gestures to interact with a computer system. These goals include the development of a high-quality and useful application that can be executed on a computer platform and perform with multiple functions such as features classification and selection, hand gestures recognition (both image and real-time recognition), and sequence sentence generation. Significantly, the goal presupposes using small computation costs and time while achieving good results. It is expected that the method can help humans in their real-life work and interactions with computers. In addition, there is the goal of providing a strong foundation for further research. The main objectives of this model are followings;

- To compromise various objects of hand gesture to be a standard, and produce a good feature set.
- To create a novel multiple feature convolutional neural network (MFCNN) model for image gesture recognition.
- To use image gestures labels to generate sequence sentences that can illustrate and contextualize the gestures and increase humans' understanding of them.
- To assess a novel recognizer that compares with the state-of-the-art in terms of recognition accuracy and speed.
- To determine an HCI system by utilizing a low computation cost, but increasing accuracy rate, and boosting system performance speeds.

IV. PROPOSED MODEL FOR HAND GESTURE APPLICATION

In the first process of the proposed model for hand gesture recombination application, the hand gestures are demanded to converts as grayscale and resize to32 dimensional. The features will extract from figure 1.

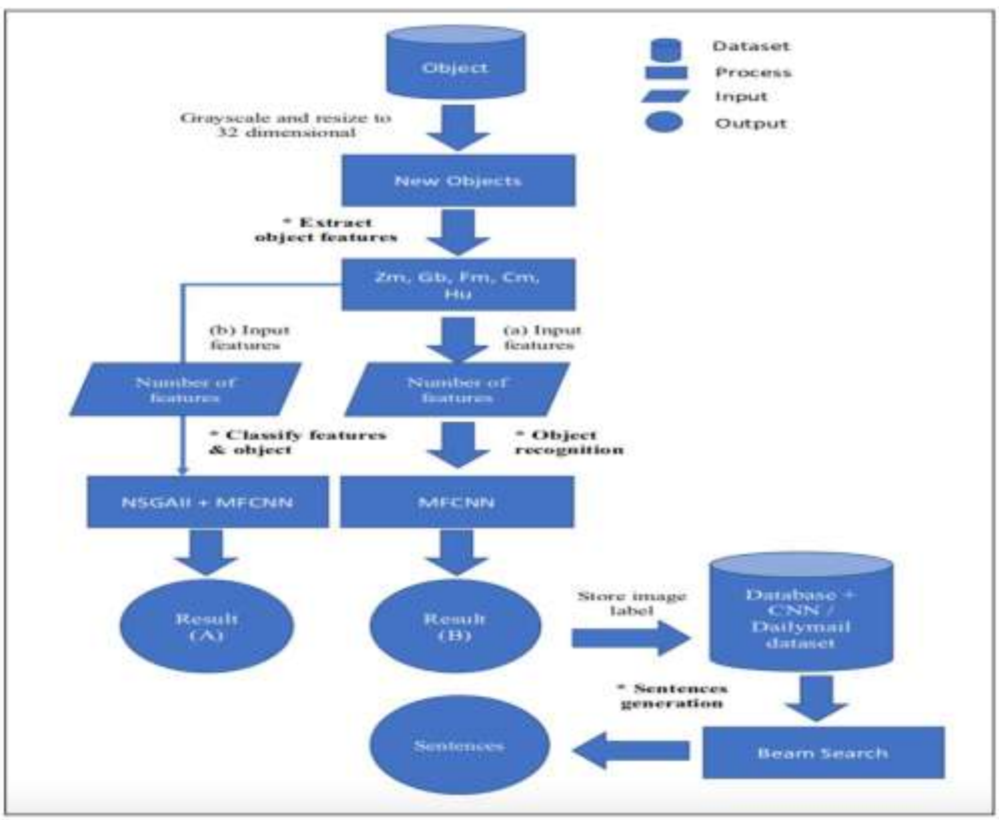

Figure: 1 Models for hand gesture application

Objects by using five methods such as Hu-moment, Gabor filter, Complex moment, Fourier moment, and Zernike moment also. The hand gestures can recognize in two ways. First, single or multiple features can be selected to recognize the common gestures (image) and real-time gestures; second, the multiple features will classify by utilizing an NSGAII algorithm and combined with Multiple Feature Convolution Neural Network MFCNN. Moreover, the image's label received from the recognition stage (Result B) combine with CNN/Dailymail datasetwilluse to generate the sentences. The sentences can be generated based on the Beam Search algorithm.The proposed model is given below in figure 1.

In this context,theresearch made threemain contributionsto knowledge:

- A multiple featuresclassification based ontheNon-Dominated Sorting Genetic Algorithm II (NSGAII). The use of NSGAII canbereducedredundant features and minimizing feature value which effective on the execution cost of the system.
- Proposed a new methodology of multiple featuresconvolutional neural network (MFCNN) model to recognize both common and real-time hand gestures.
- The generation of sequence sentences based on the Beam Search (BS) algorithm. Data of image labels that were received from the recognitionstagecombine with the CNN/Dailymail dataset is used to generate sentences.

V. CONCLUSION

Our research has succeeded to develop a system for interacting between humans and computers to predict static hand gestures. The model has proposed a novel architecture by using common features to feed in the channel of the networks. Meanwhile, the features extraction and classification through an evolutionary algorithm are also proposed. The system can be executed smoothly without demanding more resources but just use less time computation, cost, and can achieve a good result. Additionally, our system can execute in both CUP and GPU, common and real-time recognition according to our analysis. In future work, we will use these features in our proposed hand gesture recognition application.

REFERENCES

1. Ahmed, W.M., et al., Classification of bacterial contamination using image processing and distributed computing. IEEE Journal of Biomedical and Health Informatics, 2013. 17(1): p. 232-239.
2. Zhang, H., B. Li, and D. Yang, Key frame detection for appearance-based visual SLAM. IEEE/RSJ 2010 International Conference on Intelligent Robots and Systems (IROS 2010), 2010: p. 2071-2076.
3. Shahbazi, H. and H. Zhang, Application of locality sensitive hashing to realtime loop closure detection. 2011 IEEE /Rsj International Conference on Intelligent Robots and Systems, 2011: p. 1228-1233.
4. Chatterjee, S., D.K. Ghosh, and S. Ari, Static Hand Gesture Recognition Based on Fusion of Moments. Intelligent Computing, Communication and Devices, 2015. 309: p. 429-434.
5. Ng, C.W. and S. Ranganath, Real-time gesture recognition system and application, Image and Vision computing. Image and Vision Computing, 2002: p. 993-1007.
6. Zhou, Y.J. and M. Celenk, Color scene classification by Zernike moment invariants. Visual Information Processing X, 2001. 4388: p. 46-55.
7. Wang, M., W.-Y. Chen, and X.D. Li, Hand gesture recognition using valley circle feature and Hu's moment's technique for robot movement control. Measurement, 2016: p. 734–744.
8. Guo, Y., C.P. Liu, and S.R. Gong, Improved algorithm for Zernike Moments. Fourth International Conference on Control, Automation and Information Sciences (Ccais 2015), 2015: p. 307-312.
9. Wang, P., et al., RGB-D-based human motion recognition with deep learning: A survey. Computer Vision and Image Understanding, 2018: p. 118-139.
10. Badi, H., S.H. Hussein, and S.A. Kareem, Feature extraction and ML techniques for static gesture recognition. Neural Computing and Applications, 2014: p. 733–741.

11. Chevtchenko, S.F., R.F. Vale, and V. Macario, Multi-objective optimization for hand posture recognition. Expert Systems with Applications 92, 2018: p. 170–181.
12. Delgado, M., M.P. Cuellar, and M.C. Pegalajar, Multiobjective hybrid optimization and training of recurrent neural networks. IEEE Transactions on Systems Man and Cybernetics Part B-Cybernetics, 2008. 38(2): p. 381-403.
13. Agarwal, A. and S.K. Gupta, Multiobjective optimal design of heat exchanger networks using new adaptations of the elitist no dominated sorting genetic algorithm, NSGA-II. Industrial & Engineering Chemistry Research, 2008. 47(10): p. 3489-3501.
14. Barros, P., et al., A multichannel convolutional neural network for hand posture recognition. International Conference on Artificial Neural Networks, Springer, 2014: p. 403-410.
15. Bilal, S., et al., Vision-based hand posture detection and recognition for sign language. 4th International Conf. Mechatronics (ICOM), 2011: p. PP 1-6.
16. Ranzato, M., et al., Unsupervised learning of invariant feature hierarchies with applications to object recognition. 2007 IEEE Conference on Computer Vision and Pattern Recognition, Vols 1-8, 2007: p. 1429-+.
17. Sha, L., et al., A Framework of Real Time Hand Gesture Vision Based Human-Computer Interaction. Ieice Transactions on Fundamentals of Electronics Communications and Computer Sciences, 2011. E94a(3): p. 979-989.
18. Plouffe, G. and A.M. Cretu, Static and Dynamic Hand Gesture Recognition in Depth Data Using Dynamic Time Warping. IEEE Transactions on Instrumentation and Measurement, 2016. 65(2): p. 305-316.
19. Plouffe, G., A.M. Cretu, and P. Payeur, Natural human-computer interaction using static and dynamic hand gestures. 2015 IEEE International Workshop on Haptic Audio-Visual Environments and Games (HAVE), 2015: p. 57-62.
20. Vinyals, O., et al., Show and tell: A neural image caption generator. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015: p. 3156-3164.
21. Liu, A.A., et al., Hierarchical & multimodal video captioning: Discovering and transferring multimodal knowledge for vision to language. Computer Vision and Image Understanding, 2017. 163: p. 113-125.
22. Kiros, R., R. Salakhutdinov, and R.S. Zemel, Multimodal neural language models. Proceedings of the ICML, 2014: p. 595-603.
23. Tan, Y.H. and C.S. Chan, Phrase-based image caption generator with hierarchical LSTM network. Neurocomputing, 2019. 333: p. 86-100.
24. Karpathy, A. and L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015. 39(4): p. 3128-3137.
25. Wu, Q., et al., Image captioning and visual question answering based on attributes and external knowledge. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018. 40(6): p. 1367-1381.
26. Chen, X. and C.L. Zitnick, Mind's eye: a recurrent visual representation for image caption generation 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015: p. 2422-2431.
27. Tillmann, C. and H. Ney, Word reordering and a dynamic programming beam search algorithm for statistical machine translation. Computational Linguistics, 2003. 29(1): p. 97-133.
28. Rohrbach, A., M. Rohrbach, and B. Schiele, The long-short story of movie description. Pattern Recognition, Gcpr 2015, 2015. 9358: p. 209-221.
29. Barbu, A., et al., Video in sentences out. UAI (2012), 2012: p. 102-112.
30. Kojima, A., T. Tamura, and K. Fukunaga, Natural language description of human activities from video images based on concept hierarchy of actions. International Journal of Computer Vision, 2002. 50(2): p. 171-184.
31. Das, P., et al., Thousand frames in just a few words: lingual description of videos through latent topics and sparse object stitching. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013: p. 23-28.
32. Guadarrama, S., et al., YouTube2Text: Recognizing and Describing Arbitrary Activities Using Semantic Hierarchies and Zero-shot Recognition. 2013 IEEE International Conference on Computer Vision (Iccv), 2013: p. 2712-2719.
33. Donahue, J., et al., Long-term recurrent convolutional networks for visual recognition and description. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: p. 2625-2634.

34. Rohrbach, A., et al., Coherent multi-sentence video description with variable level of detail. Pattern Recognition, GCPR 2014, 2014. 8753: p. 184-196.
35. Pan, Y.W., et al., Jointly modeling embedding and translation to bridge video and language. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: p. 4594-4602.
36. Xu, R., et al., Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015: p. 2346-2352.
37. Rohrbach, A., et al., Movie description. International Journal of Computer Vision, 2017. 123(1): p. 94-120.
38. Yao, L., et al., Describing videos by exploiting temporal structure. IEEE International Conference on Computer Vision (ICCV), 2015: p. 4507-4515.