Instance-based metric learning models for 2D RGB sign language identification from multiple cameras

Ch.Raghava Prasad

Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation (KLEF), Deemed to be University, Vaddeswaram, Green fields, Guntur, Andhra Pradesh, India chrp@kluniversity.in.

DOI : 10.48047/IJFANS/11/S6/005

Abstract

The authors of this study developed a multi-view motion modeled deep attention (MV2MDA) method to address the limitations of existing approaches. Most significantly, this network can be trained from the ground up using any conceivable backbone structure. The data on motion is calculated by Throughout all the layers, attention features derived from Lukas kande optical flow are combined with spatial convolutional features. The Four-view spatial-temporal fusion features are used in the suggested V2PN, or the Viewer-to-Viewer Pooling Network. V2PN generates a view-invariant feature matrix of user-specific attention features. The model was a fully retrainable one. To this point uses mF1 score and mRA to quantify performance. Despite the fact that the outcomes on our 5 views 200 class sign were excellent, datasets in languages.

1.Introduction

Human biomechanics place limitations on the range of motion that can be achieved in the hands and fingers. Therefore, it is inevitable that identical signs produced by various signers or by the same signer on separate occasions will display some degree of difference. The visual features of joints and their orientations provided by the hand and fingers to a machine interpreter are crucial to a full SLR. Multi-view gesture [1], face [2], and action recognition [3] were the primary areas of study for earlier approaches. Very little research has been done on multi-view sign language recognition to investigate how different viewpoints affect the effectiveness of deep learning systems. Traditional SLR models rely on feature extraction techniques developed by hand and automated learning algorithms [4]. Regrettably, computer vision algorithms used in hand-crafted features rarely model all the characteristics of a sign language in a variety of perspectives. In order to increase the recognition algorithm's real-time performance, multi-view learning expands the camera's operational freedom in real-time. To effectively recognize signs, it is crucial to use multi view learning implemented in deep networks. The primary results of this section are: 1. a video dataset of 200 signs in five different camera views, filmed with five different signers against a variety of backgrounds. Our data collection, dubbed KL2DMVSL, contains 5 observations per symbol. To generate a maximally discriminative feature vector for recognition, this chapter offers an 8-stream convolutional neural network (CNN) with a motion attention model to extract features in various perspectives.

2. MV2MDA – Net Architecture

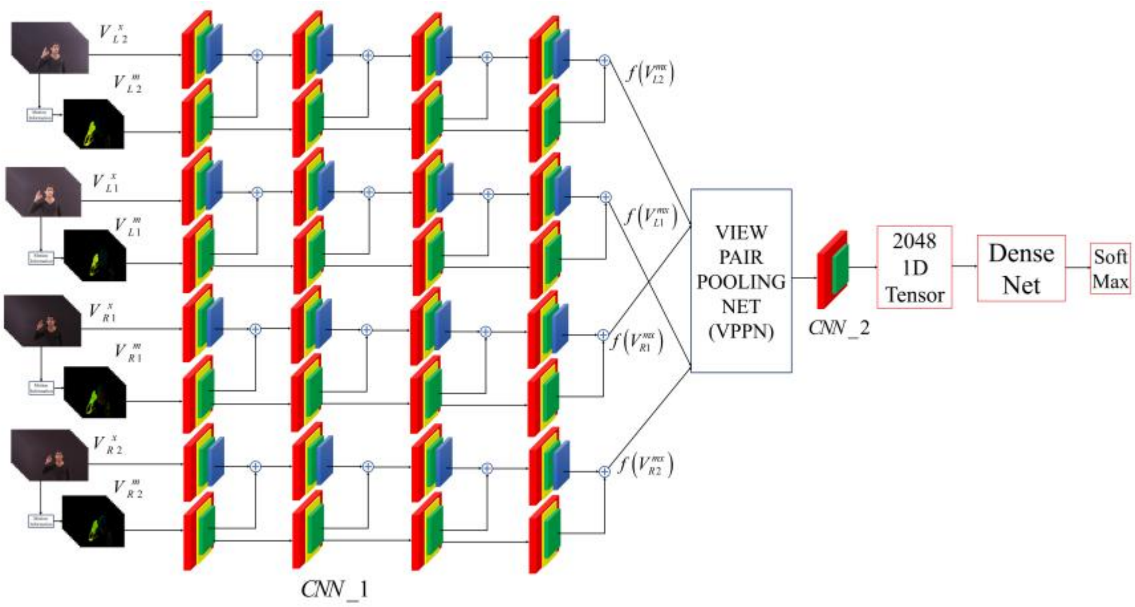

Fig. 1: Proposed MV2MDA – Net Architecture for RGB video based sign language recognition

Figure 1 depicts the suggested deep learning architecture for sign language recognition across many viewpoints. Only four of the eight feature extraction streams in the motion-modelled deep attention network (MV2MDA - Net) are spatial, whereas the remaining six are motion attention. The probabilities calculated in the SoftMax layer are used to make the final call on the assigned class labels. Each incoming video frame has an input size of 256 by 256 by 3.The issue with recordings depicting human motion (action or sign) is that these motions can be interpreted in both subjective and objective ways.

3.View Pair Pooling Network (VPPN)

To create an ensemble feature matrix, the VPPN uses the network depicted in fig.3.2 to combine the various view-oriented features. Figure 3 depicts a network that was conceptualized using GoogleNet's inception layers. To combat the vanishing gradients problem, GoogleNet's inception layers expand the network horizontally rather than vertically. This led to higher recognition accuracy at the expense of simplified computation.

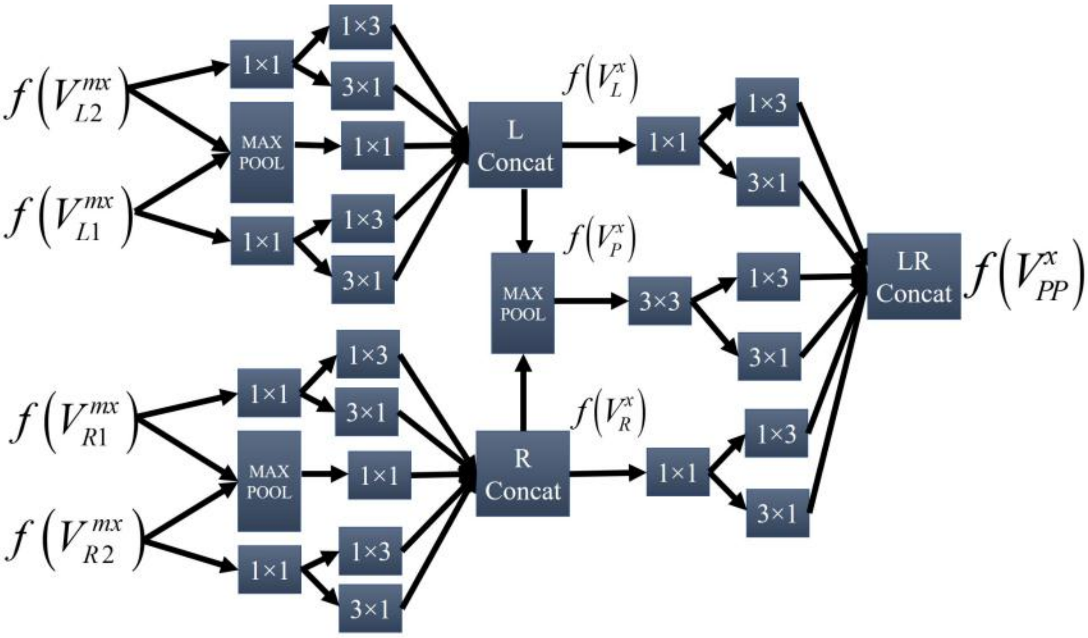

Fig. 2: View Pair Pooling Network (VPPN)

4.conclusion

For multi-view sign language recognition on the KL2DMVSL 2D video dataset, this chapter presents a deep learning approach called MV2MDA. The 8-channel convolutional neural network processes data from 4 spatial streams and 4 motion streams across 4 perspectives. Every one of the layers is forced to pay more attention to the spatial details thanks to the motion streams. To produce features that are independent of the viewpoint, this chapter introduces VPPN, a multi-view feature learning network that acquires the skill of pooling. Recognizability could be improved by using the learnt pooling procedure to extract view-specific information.

References

1. W. Wei, Y. Wong, Y. Du, Y. Hu, M. Kankanhalli, and W. Geng, “A multi-stream convolutional neural network for sEMG-based gesture recognition in muscle-computer interface,” Pattern Recognition Letters, vol. 119, pp. 131–138, mar 2019. [Online]. Available: https://doi.org/10.1016%2Fj.patrec.2017.12.005

2. S. S. Farfade, M. J. Saberian, and L.-J. Li, “Multi-view face detection using deep convolutional neural networks,” in Proceedings of the 5th ACM on International Conference on Multimedia Retrieval - ICMR '15. ACM Press, 2015, pp. 643–650. [Online]. Available: https://doi.org/10.1145%2F2671188.2749408

3. A. Fathi and G. Mori, “Action recognition by learning mid-level motion features,” in 2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2008, pp. 1–8.

4. M. J. Cheok, Z. Omar, and M. H. Jaward, “A review of hand gesture and sign language recognition techniques,” International Journal of Machine Learning and Cybernetics, vol. 10, no. 1, pp. 131–153, aug 2017. [Online]. Available: https://doi.org/10.1007%2Fs13042-017-0705-5

5. Q. Wang, X. Chen, L.-G. Zhang, C. Wang, and W. Gao, “Viewpoint invariant sign language recognition,” Computer Vision and Image Understanding, vol. 108, no. 1-2, pp. 87–97, 2007.

6. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.

7. A. Plyer, G. Le Besnerais, and F. Champagnat, “Massively parallel lucas kanade optical flow for real-time video processing applications,” Journal of Real-Time Image Processing, vol. 11, no. 4, pp. 713–730, 2016.

8. A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, “Ntu rgb+ d: A large scale dataset for 3d human activity analysis,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1010–1019.

9. S. Singh, S. A. Velastin, and H. Ragheb, “Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods,” in 2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance. IEEE, 2010, pp. 48–55.

10. M. Nageswara Rao, K. Narayana Rao and G. Ranga Janardhana,. "Machines and AGVs Scheduling in Flexible Manufacturing System with Mean Tardiness Criterion," International journal of Advanced Materials Manufacturing and Characterization, vol. 4, pp. 100-105, 2014.

11. M. Nageswara Rao, K. Narayana Rao and G. Ranga Janardhana, " Integrated Scheduling of Machines and AGVs in FMS by Using Dispatching Rules," Journal of Production Engineering, vol. 20(1), pp. 75-84, 2017.

12. M. Nageswara Rao and K. Narayana Rao and G. Ranga Janardhana,. "Machines and AGVs Scheduling in Flexible Manufacturing System with Mean Tardiness Criterion,"

International journal of Advanced Materials Manufacturing and Characterization, vol. 4, pp. 100-105, 2014.

13. M. Nageswara Rao, K. Narayana Rao and G. Ranga Janardhana, "Machines and AGVs Scheduling in Flexible Manufacturing System with Mean Tardiness Criterion," International journal of Advanced Materials Manufacturing and Characterization, vol. 4, pp. 100-105, 2014.