Analysis of Linguistically Significant Eyebrow Events in American Sign Language

Anonymous FG2020 submission
Paper ID 78

***Abstract*— To do**

I. INTRODUCTION

The non-manual channel provides critical linguistic information in signed languages, including American Sign Language (ASL), the focus of this paper. Facial expressions and movements of the head and upper body convey information essential to recognition of lexical and, especially, syntactic information. The eyebrows are involved in almost all of the non-manual grammatical markings in ASL. However, achieving accuracy in the computer-based detection and analysis of patterns of eyebrow movements from 2D video (e.g., identifying start and end points of the onsets, offsets, and core portions of these events, as well as the overall degree to which the eyebrows are raised and lowered in a given eyebrow event) poses a variety of challenges, which our current work addresses.

Previous research [1], [2], [3], [4] has generally analyzed eyebrow events with respect to (1) estimating the intensity of eyebrow movements, and (2) recognizing and temporally locating eyebrow events. Most eyebrow movement intensity estimation approaches first extract face landmarks using face trackers, and then compute eyebrow movement intensity from the obtained landmarks. For example, Liu *et al* [3] first detect facial landmarks using a 3D face tracker, and then compute and normalize the Euclidean Distance between landmarks of the eyes and eyebrows to estimate eyebrow movement intensity or height. However, these face-tracker based methods are not accurate and robust enough for linguistic use, given the difficulty of extracting accurate 3D facial landmarks in sign language videos, where obstructions caused by hand occlusions, large head pose changes and video-focusing issues are prevalent. As for previous approaches to eyebrow event detection and recognition, previous researchers have generally formulated and solved the task as a frame-level prediction problem. For example, Liu *et al* [3] first assign a label to each video frame based on whether or not the frame is inside an event, and if so, what type of event the frame is in; they then train a CRF-based model to predict these labels. These methods are not good because xxx.

To address previous limitations and create an eyebrow event estimation system accurate enough to be used for linguistic purposes, we propose a new machine learning approach based on deep neural networks to accurately estimate 3D eyebrow movement intensity from ASL video input. Our approach takes advantage of the fact that there are multiple facial visual cues that can signal intensity of eyebrow deformations, such as wrinkling of the forehead; deep neural networks have particularly well-suited to learning and leveraging such visual cues. The proposed method uses spatial-temporal information from video sequences, as shown in Figure 1. It uses a sequence of frames, instead of a single frame, as input, and then predicts the eyebrow movement intensity of the middle frame. The proposed network is composed of a ResNet [5] and a Bidirectional Long short-term memory (LSTM) layer [6]. The ResNet extracts spatial features from each frame, and then the LSTM layer further extracts temporal information. The network is trained end-to-end using data with labeled eyebrow movement intensity. Add? This is another improvement of our method over previous approaches, which have not analyzed the data end-to-end.

In order to train and evaluate our proposed method, we used the American Sign Language Linguistic Research Project (ASLLRP) SignStream® 3 Corpus, a large-scale, linguistically annotated, ASL video dataset of native ASL signing shared publicly on the Web through the Data Access Interface (DAI) 2 (a collaboration between Boston and Rutgers Universities): <http://dai.cs.rutgers.edu/dai/s/dai> [cite Neidle, et al., 2018]. The data we analyzed included a total of 1810 utterances, and over 3700 annotated eyebrow events (instances of raised or lowered eyebrows).

It should be noted that labeling each frame with a continuous intensity value for eyebrow deformation would be non-trivial and extremely laborious. To tackle the difficulty, we use a discrete-to-continuous annotation strategy. In ASL, signers employ eyebrow movements as important components for marking a wide variety of types of grammatical information. However, eyebrows are constantly in motion in the course of a sentence. Thus it is a challenge to identify the regions of the video in which the eyebrows are participating in a linguistically significant grammatical marking of some kind, as well as detecting theonsets, offsets, and transitions between such eyebrow events. We have exploited the linguistic labels from the ASLLRP SignStream® 3 Corpus, which include annotations of raised and lowered eyebrow onsets, offsets, and transitions bewteen eyebrow events. We have assigned a discrete value of the corresponding eyebrow height based on the face tracking. We then interpolate the eyebrow level between those events and obtain a continuous intensity value for eyebrow deformation. This strategy significantly cuts the difficulty and workload of data annotation while allowing us to produce superior results compared to traditional face trackers.

We train and test the proposed method on the collected dataset. The results demonstrate the effectiveness of our method. We further visualize the activation maps of the learned model and show that our model focuses ? and leverages meaningful visual cues. Finally, we extend our method to estimatation of linguistically important head pose events (head position in 3 dimensions) and eye aperture.

We also compare to the state of the art 3D trackers...

The main contributions of this paper are summarized as follows:

- We exploit a large-scale linguistically annotated ASL video dataset with annotations of eyebrow movement intensity for this research.
- We propose a CNN-RNN based network which can

estimate eyebrow movement intensity directly from image inputs.

• We further extend the approach to linguistically important head movement and eye aperture.

• We validate our approach using state-of-the-art 3D face trackers.

Next we present details of our approach, followed by a report of our experiments and discussion of the results.

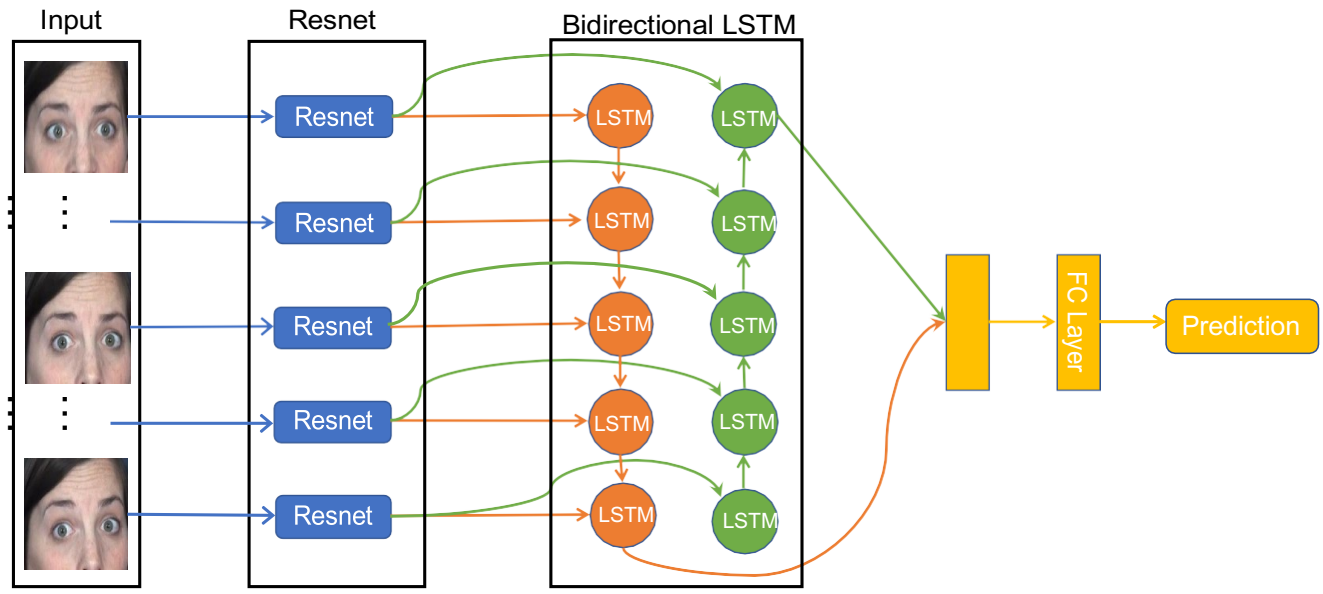

Fig. 1. Illustration of our method.

II. Related Work

III. Methodology

The overview of our approach is shown in Figure 1. Our model takes a short video segment as input and estimates the intensity of eyebrow deformation of the middle frame in the segment. First, ResNet is applied on input frames to extract and embed visual clues, as introduced in Section III-A. The extracted sequence of spatial features is fed to the bidirectional LSTM to further learn temporal information. The final hidden states of each direction are concatenated and used as the feature representation of the input video for the prediction of the intensity of the eyebrow deformation, as presented in Section III-B. The eyebrow events are temporally located and predicted from the predicted deformation intensity in Section III-C

A. ResNet

Each input frame contains multiple facial visual cues of the intensity of eyebrow deformation. For example, wrinkling of the forehead signals a raised eyebrow event (although wrinkling may not always be visible, especially in the event of blurring in the video). The larger extent of any visible wrinkling, the larger intensity of eyebrow raising.

B. Bidirectional LSTM

ResNet [5], whose superior performance in feature learning has been demonstrated for many vision tasks [5], [7], [8], is used to extract and embed these visual cues from each frame. Specifically, given a input video I = {Xt|t = 1, . . . , T }, where T represents the length of input video, Xt denotes the frame at t-th time step, the ResNet G extracts a sequence of deep feature representations F = {xt|t = 1, . . . , T } from I, where xt denotes the spatial feature representation of Xt, i.e xt = G(Xt). The F is fed to Bidirectional LSTM to further extract temporal information.

Temporal information is useful for correcting for occlusion caused by the hands. To be more specific, when a frame is occluded, other nearby frames that are not occluded can provide complementary information for determining the eyebrow deformation intensity of the occluded frames.

To extract such temporal information, we use a Bidirectional LSTM layer. The Bidirectional LSTM layer consists of two LSTM units [6], one of which extracts temporal information in chronological order, *i.e.,* from beginning to the end, and the other one does this in reverse chronological order, *i.e.,* from the end to the beginning. Compared to traditional one-directional LSTM, the Bidirectional LSTM layer can more effectively extract temporal information [9], [10], [11], [12].

Specifically, the LSTM unit which extracts temporal infor- mation in chronological order takes $\boldsymbol{F}$ as input, and outputs a sequence of T vectors $\boldsymbol{H} = \{h_t \ t = 1, \ldots, T\}$, where $\mathbf{h}_t$ is the t-th hidden state, which embeds learned information from the first time step to the t-th time step. The LSTM unit implements this by the following operations. At each time step, it first uses gates to control the information flows (current input, previous and current state) in the units, and then updates cell state $\mathbf{c}_t$ to memorize these information flows. Finally, the $\mathbf{h}_t$ is generated to reflect the $\mathbf{c}_t$ [6]. These operations are formulated in Equation 1:

states, $\mathbf{x}_t$ denotes the spatial feature representation of t-th frame, and W and U denote the weight matrices, σ is the Sigmoid activation function, $\odot$ denotes element-wise dot between two vectors. $\mathbf{i}_t$, $\mathbf{f}_t$ and $\mathbf{o}_t$ is the activation vector of update gate, forget gate, output gate respectively, which measures the importance of previous cell states, current input information $\mathbf{g}_t$ and output [6].

The LSTM unit which extracts temporal information in reverse chronological order works the same way as shown in Equation 1. The difference is that it takes the reversed $\boldsymbol{F}$, *i.e* $x_t\ t = T, T-1, \ldots, 1$ as input. We denote the last hidden state of this LSTM unit as $\mathbf{h}_T$. $\mathbf{h}_T$ is then concatenated with $\mathbf{h}_T$, and the resulting vector is used as the feature representation of input video $\boldsymbol{I}$ for eyebrow deformation intensity estimation.

C. Eyebrow Event Detection

As shown in Figure XX. The temporal location of a raised (or lowered) eyebrow event can be determined by four points:

- the start point of onset: the time when eyebrow starts to raise (or lower);
- the end point of the onset (which is also the start point of the core event): the time when the eyebrows essentially stop raising (or lowering) and reach a positive (or negative) maximum intensity value;
- the start point of the offset (which is also the end point of the core event): the time when eyebrows start to lower (or raise) to return to neutral or to transition to the next eyebrow event;
- the end point of offset: the time when the eyebrow movement has ended.

Therefore, we can temporally locate and recognize eyebrow movement events by detecting these four points. We implement this by computing and checking the left and right side scope, and the deformation intensity of each frame. To be more specific, given a video with its predicted eyebrow deformation intensity $\boldsymbol{P} = \{p_i | i = 1, \ldots, N\}$, where p_i is the predicted eyebrow intensity of i-th and N denotes the length of video, the left side slope l_i and the right side slope r_i of i-th frame can be obtained by:

$$l_i = p_i - p_{i-1}, \qquad r_i = p_i - p_{i-1}, \tag{2}$$

For a specific frame, it is obvious that if its left and right side slopes are not equal, and one of the slopes is zero, that frame may correspond to one of the four points just defined. We find and chronologically sort all such frames, and the result as $\boldsymbol{B} = \{b_j | j = 1, \ldots, M, M \leq N\}$,

where $\mathbf{b}_j$ is j-th candidate frame. In addition, we have a function M that can map the index of obtained frames back to $\boldsymbol{P}$. In other words, the $M(j)$ is the corresponding index of j-th candidate frame in $\boldsymbol{P}$.

For each of the four continuous points (b_j, b_{j+1}, b_{j+2}, b_{j+3}) from $\boldsymbol{B}$, if they meet following conditions, we predict them to be the temporal boundaries of a raised eyebrow event:

- $l_{M(j)} = 0$ and $r_{M(j)} > 0$;
- $l_{M(j+1)} > 0$ and $r_{M(j+1)} = 0$ and $p_{M(j+1)} > 0$;
- $l_{M(j+2)} = 0$ and $r_{M(j+2)} < 0$ and $p_{M(j+2)} = p_{M(j+1)}$;
- $l_{M(j)} < 0$ and $r_{M(j)} = 0$;

If they meet following conditions, we predict them to be the temporal boundaries of a lowered eyebrow event:

- $l_{M(j)} = 0$ and $r_{M(j)} < 0$;
- $l_{M(j+1)} < 0$ and $r_{M(j+1)} = 0$ and $p_{M(j+1)} < 0$;
- $l_{M(j+2)} = 0$ and $r_{M(j+2)} > 0$ and $p_{M(j+2)} = p_{M(j+1)}$;
- $l_{M(j)} > 0$ and $r_{M(j)} = 0$;

IV. Experiments and Discussion

V. Conclusion

References

[1] J. Liu, B. Liu, S. Zhang, F. Yang, P. Yang, D. N. Metaxas, and C. Neidle, "Non-manual grammatical marker recognition based on multi-scale, spatio-temporal analysis of head pose and facial expressions," *Image and Vision Computing*, vol. 32, no. 10, pp. 671–681, 2014.

[2] C. Neidle, J. Liu, B. Liu, X. Peng, C. Vogler, and D. Metaxas, "Computer-based tracking, analysis, and visualization of linguistically significant nonmanual events in American Sign Language (ASL)," in *LREC Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel*, vol. 5, Citeseer, 2014.

[3] B. Liu, J. Liu, X. Yu, D. N. Metaxas, and C. Neidle, "3D face tracking and multi-scale, spatio-temporal analysis of linguistically significant facial expressions and head positions in ASL," in *LREC*, pp. 4512–4518, Citeseer, 2014.

[4] J. Liu, B. Liu, S. Zhang, F. Yang, P. Yang, D. N. Metaxas, and C. Neidle, "Recognizing eyebrow and periodic head gestures using CRFs for non-manual grammatical marker detection in ASL," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1–6, IEEE, 2013.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, pp. 91–99, 2015.

[8] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-CNN," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

[9] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional LSTM with CNN features," *IEEE Access*, vol. 6, pp. 1155–1166, 2017.

[10] Y. Bin, Y. Yang, F. Shen, X. Xu, and H. T. Shen, "Bidirectional long-short term memory for video description," in *Proceedings of the 24th ACM international conference on Multimedia*, pp. 436–440, ACM, 2016.

[11] Á. Peris, M. Bolaños, P. Radeva, and F. Casacuberta, "Video description using bidirectional recurrent neural networks," in *International Conference on Artificial Neural Networks*, pp. 3–11, Springer, 2016.

[12] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *European conference on computer vision*, pp. 766–782, Springer, 2016.

ADD: C. Neidle, A.Opoku, G. Dimitriadis, and D. Metaxas, NEW Shared & Interconnected ASL Resources: SignStream® **3 Software; DAI 2 for Web Access to Linguistically Annotated Video Corpora; and a Sign Bank.** 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community (pp. 147-154). LREC 2018, Miyagawa, Japan. May 2018.