Temporal Grammatical Marker Localization for American Sign Language through Jointly Training Non-Manual Events

BMVC 2019 Submission # 37

Abstract

In American Sign Language (ASL), various types of important grammatical information are conveyed by combinations of different facial expressions and head gestures. It is challenging but of great linguistic importance to detect these non-manual events. Existing methods of grammatical marker localization are mainly built on hand-crafted features. However, such features can not well model complex spatial-temporal patterns of grammatical information inside ASL utterances. In this paper, we propose a two-stream multi-task framework for localizing grammatical markers. Rather than detecting non-manual events independently, we train them together from two-stream inputs. One stream utilizes 3D ResNets which extracts appearance features from the detected face bounding boxes while the other stream captures the geometry features from corresponding 3D facial landmarks. Experimental results on ASLLVD dataset demonstrate the effectiveness of multi-task training strategy and explainability of our proposed framework.

1 Introduction

Sign languages convey interpretation in the visual-manual modality, with signs (the equivalent of words) produced by the hands and arms in parallel with other types of information, such as critical grammatical information. In many sign languages such as American Sign Language, the essential grammatical information is conveyed through the patterns of head movements (such as periodic head nodes and shakes) and facial expressions that occur over phrasal domains, rather than over single signs. Therefore, recognition of non-manual information, which is known as non-manual grammatical marker (NMGM), is critical for modern sign language recognition in general, Besides, this non-manual information has also proved useful for recognition for manual signs [1, 18, 20]. In this paper, we consider recognizing the following 5 most popular NMGMs: yes-no question (*yes-no*), wh question (*whq*), negation (*neg*), topic/focus (*topic*) and conditional-when (*cond*). Different NMGM is able to recognize part of common facial appearance patterns. For example, lowered eyebrow usually occurs in both wh-question and negation with their difference can be identified by the head movement pattern (rapid head shake in wh-question but large and slow head shake in negation). This requires our proposing method gains the ability to accurately model head pose pattern and facial expression with head pose variations, which is a challenging task.

Early methods in sign language NMGM localization focus on designing various hand-crafted features, such as pyramid matching the SIFT features and bag of words [15] or combining different texture features [11, 12, 13, 14]. With the hand-crafted features introduced,

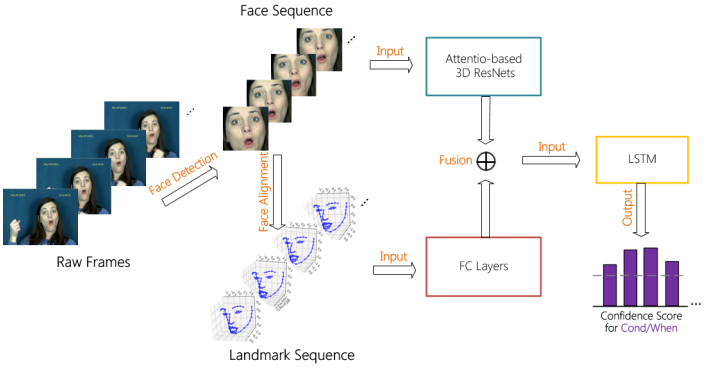

Figure 1: Overview of our two-stream multi-task framework for grammatical marker localization.

machine learning tools such as SVM or conditional random field (CRF) are utilized in detecting the non-manual grammatical information within sign language utterances. Though concise and explainable, such manually-designed features are not capable of handling complicated spatial-temporal linguistic patterns. In recent years, deep learning approaches achieve promising performances on many temporal visual detection tasks [19, 21, 25, 26]. The success of such techniques has also renovated the research in sign language recognition [5, 9, 18]. The above temporal detection methods treat all actions, signs or events independently from each other. However, the task of sign language NMGM localization has close relation with other non-manual events. Jointly training non-manual events and NMGM detection could enable the transition of useful information between these relevant tasks. Inspired by this, we propose an end-to-end multi-task framework on detecting grammatical markers.

In this work, we employ two-stream inputs to extract representative appearance and spatial features from given sign language utterances. The appearance branch takes in cropped face sequences and processes them by a 3D residual convolutional neural networks (3D ResNets) [8], which has demonstrated intuitively effectiveness in spatio-temporal operations. However, the appearance features directly acquired from raw videos are sensitive to background clutter and illumination conditions. As a supplementary to appearance features, we utilize 3D facial landmarks [3] to capture reliable spatial configurations. The two-stream features are then integrated and fed into a LSTM framework to make frame-level predictions of low-level non-manual sequences (such as eyebrow height), high-level non-manual events (such as raised eyebrow), and grammatical markers in turn. Figure 1 displays an overview of our framework. In summary, our key contributions are threefold:

- To the best of our knowledge, we are the first to propose a multi-task neural network approach for the task of temporal grammatical marker detection in sign language.
- We propose to combine appearance and landmark information for better feature representation. Experiments show it is capable of obviously imroving the mean average precision (mAP) on ASLLVD dataset.
- We explore a multi-task framework in our NMGM localization model. The ablation experiments and case study demonstrate that our approach indeed enables the transition of useful information between relevant tasks.

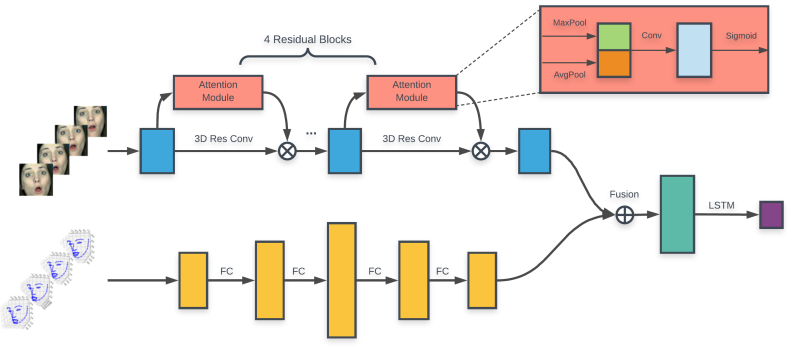

Figure 2: The architecture of the proposed multi-task approach.

2 Related Work

In this section, we review the works related to ours in the domain of non-manual event detection, temporal action localization, and landmark-based action recognition.

Non-Manual Event Detection. The research on detection of non-manual components has recently emerged as an auxiliary way to distinguish some signs from other similar signs. Several works [2, 11, 12, 13, 14, 17] have explored interpretation of the grammatical information from linguistically relevant non-manual events. Metaxas *et al.* [14] and Nguyen *et al.* [14] introduces eyebrow height, eye aperture, head pose and head movements as low-level features to assist detection of grammatical markers. Liu *et al.* [11, 12, 13] points out the necessity of high-level features to enhance non-manual grammatical marker localization. They combine low-level and high-level features and utilize HM-SVM for grammatical marker prediction. Their approach is indeed a two-stage framework. In this work, we design a deep multi-task framework to co-train NMGM and non-manual event detection.

Temporal Action Localization. The goal of temporal action localization is to identify the temporal boundaries of all specific action categories from untrimmed videos. Recently, many deep learning approaches [6, 10, 21, 22, 28] are proposed on localizing general actions in video segments. Segment-CNN (S-CNN) [21] relies on a multi-stage 3D ConvNets framework with multi-scale sliding windows to propose candidate segments, recognize actions, and localize their temporal boundaries. Structured Segment Networks (SSN) [28] builds a temporal stage structures (starting, course, ending) for each action instance via a structured temporal pyramid. A discriminative model comprising two classifiers is employed for classifying actions and determining completeness respectively. Boundary Sensitive Network (BSN) [10] follows a "local to global" fashion. It locally detects temporal boundaries with high probabilities and directly combines these onset and offset boundaries as proposals. A proposal evaluation module is adopted to select proposals with high confidence of containing an action within its region. Compared with general actions, grammatical markers are more closely distributed with each other in latent feature space. Therefore, it is critical to find representative features to distinguish between different grammatical markers. Inspired by linguistic knowledge, we consider learning low-level non-manual sequences and high-level non-manual events as separate representative features by intermediate supervisions. The

combination of the two features is able to effectively assist the localization of grammatical markers.

Landmark-based Action Recognition. Compared with ConvNets video features, human landmark sequences (body and face) conveys significant dynamic information for action recognition. Several recent works leverage human landmarks to capture the spatial and temporal evolutions of different action types. Two-stream RNN [24] incorporates both spatial and temporal networks to model human body kinematics and conducts skeleton based action recognition. STA-LSTM [23] designs spatial and temporal attention modules to allocate different attentions to divergent joints in human body and various video frames. ST-GCN [27] builds a spatial-temporal graph for human body landmark sequences and utilizes a spatial-temporal Graph ConvNets for modeling dynamic skeletons. All above works focus on recognizing actions with the help of human body landmarks. For non-manual detection in sign languages, facial landmarks are more suitable for spatial configuration representation. To this end, we build a landmark branch to extract representative spatial features from facial landmark sequences.

3 Methodology

In this section, we first introduce the layout of our two-stream module for extracting appearance features and spatial configurations. Then we presents the multi-task framework for jointly training temporal non-manual event detection and grammatical marker localization. Figure 2 shows the architecture of our proposed method.

3.1 Two-Stream Feature Extraction Module

Appearance Branch. The utterance in sign language is a sequence of continuous image frames performed by signers which contains various manual signs and non-manual gestures. Since non-manual information are mainly conveyed by facial regions, we first track, detect and crop the faces for each frame. Let $\mathbf{U} = (\mathbf{u}_1, \cdots, \mathbf{u}_T)$ denotes a cropped face sequence with T frames. We apply a sliding window with size L on $\mathbf{U}$ to construct a series of video clips $\mathbf{C} = (\mathbf{c}_1, \cdots, \mathbf{c}_L)$. According to previous work [4, 8], 3D Residual Networks are shown to achieve superior performance in representing temporal and spatial pattern of videos. In this work, we adopt 3D ResNets to extract appearance features for input face sequences. By passing each video clip $\{\mathbf{c}_i\}_{i=1}^{L}$ into the appearance branch, the output of 3D ResNets can be represented as:

$$\{\mathbf{f}_i\}_{i=1}^{L} = \mathcal{P}_\theta(\{\mathbf{c}_i\}_{i=1}^{L}), \tag{1}$$

Where $\mathcal{P}_\theta$ denotes the spatio-temporal modeling function of 3D ResNets, θ are the network weights to be learnt, and $\mathbf{f}_i$ is the representation of $\mathbf{c}_i$. Afterwards, $\{\mathbf{f}_i\}_{i=1}^{L}$ is fed into a bidirectional LSTM (Bi-LSTM) $\mathcal{R}_\psi$ to compute the hidden state sequence forward and backward:

$$\{\mathbf{h}_i^a\}_{i=1}^{L} = \mathcal{R}_\psi(\{\mathbf{f}_i\}_{i=1}^{L}), \tag{2}$$

Where ψ are the Bi-LSTM weights, $\mathbf{h}_i^a$ is the hidden state sequence for appearance branch. We employ $\mathbf{h}_i^a$ for make frame-level grammatical marker prediction in next step.

Spatial Branch. Compared with 2D facial landmarks, 3D facial landmarks offer richer and more discriminative spatial information. For this reason, we utilize the 3D facial landmarks as the input of our spatial branch. As shown in Figure 1, 3D facial landmark sequences are

detected by 3D-FAN [3] from cropped face sequence. Let $\mathbf{M} = (\mathbf{m}_1, \cdots, \mathbf{m}_L)$ denotes a 3D facial landmark sequence, where $\mathbf{m}_i \in \mathbf{R}^{d\times 3}$ is a tensor representing the absolute coordinate location of anchor landmarks, d is set to 68 in 3D-FAN. To eliminate the influence of absolute locations, we conduct normalization on $\mathbf{m}_i$. We first compute the origin $\mathbf{O} = (o_x, o_y, o_z)$ of the normalized coordinate: $o_k = \sum_{i\in\{34,37,46\}} \frac{1}{3} r_k^{(i)}$, where $k = \{x, y, z\}$ and $r_k^{(i)}$ is the raw k-axis coordinate for the i-th anchor(where the number 34, 37 and 46 represents the actual coordinate of two eyebrow corner anchors and nose center anchors in 3D-FAN). Thus, the normalization function $\mathcal{N}$ mapping the raw coordinates to normalized coordinates can be written as:

$$\mathcal{N}(r_k^{(i)}) = \frac{r_k^{(i)} - o_k}{\sqrt{\sum_{t\in\{x,y,z\}} (r_t^{(i)} - o_t)^2}} \tag{3}$$

After acquiring normalized 3D facial landmarks $\mathcal{N}(\mathbf{m}_i)$, we feed them into the Bi-LSTM $\mathcal{R}_\psi$ to compute the hidden state sequence $\mathbf{h}_i^s$ for spatial branch:

$$\{\mathbf{h}_i^s\}_{i=1}^L = \mathcal{R}_\psi(\{\mathcal{N}(\mathbf{m}_i)\}_{i=1}^L), \tag{4}$$

3.2 Non-Manual Event Detection

We divide non-manual event detection into two levels. The first level focuses on generating low-level non-manual sequences. In this work, we include five branches (eyebrow height, eye aperture, head yaw, head pitch, and head roll) for low-level sequence generation. The second level is detecting high-level non-manual events from generated low-level sequences. We are interested in four kinds of linguistically relevant non-manual events, namely raised eyebrow, lowered eyebrow, head nod and head shake. At the first step, the two-stream features from appearance and spatial branches are fused together through concatenation: $\mathbf{h}_i^t = [\mathbf{h}_i^a; \mathbf{h}_i^s]$. Then the concatenated features are fed into temporal convolutional networks $\mathcal{T}^v$ for value regression prediction:

$$\{\mathbf{v}_i\}_{i=1}^L = \mathcal{T}^v(\{\mathbf{h}_i^t\}_{i=1}^L), \tag{5}$$

Here $\{\mathbf{v}_i\}_{i=1}^L$ can be any low-level non-manual sequence. In the second level, we predict non-manual events with low-level non-manual sequences related to them. For instance, raised eyebrow and lowered eyebrow are predicted from eyebrow height curve and head shake is predicted from head yaw curve. For low-level non-manual sequence $\{\mathbf{v}_i\}_{i=1}^L$, we also utilize temporal convolution $\mathcal{T}^e$ to make frame-level categorical probability predictions:

$$\{\mathbf{e}_i\}_{i=1}^L = \sigma[\mathcal{T}^e(\{\mathbf{v}_i\}_{i=1}^L)], \tag{6}$$

Where σ stands for the Sigmoid function and $\{\mathbf{e}_i\}_{i=1}^L$ is the generated probability sequence for relevant non-manual events.

3.3 Temporal Grammatical Marker Localization

After non-manual event detection, we get five value and four event probability sequences. For related non-manual event $\mathbf{E}$ and value sequence $\mathbf{V}$, we employ cross-attention flow between them to generate an event-aware value representation. There are in total four eligible

value-event pairs: $\mathbf{W}:\{(\text{eb},\text{ebraise}),(\text{eb},\text{eblower}),(\text{yaw},\text{hshake}),(\text{pitch},\text{hnod})\}$.

$$\begin{aligned}\mathbf{Q} &= f(\mathbf{V},\mathbf{E}),\ \text{for all } (\mathbf{V},\mathbf{E})\in\mathbf{W}\\ \mathbf{V}^{Q} &= \mathbf{V}\odot\mathbf{Q}\end{aligned} \tag{7}$$

Here, $f(\mathbf{X},\mathbf{Y})=\text{softmax}(\mathbf{XY})$ defines a dot product attention function and $\mathbf{V}^{Q}$ is the generated event-aware value representation. Finally, we concatenate all intermediate representation from different input sequences (event-aware value $\mathbf{V}^{Q}$, value $\mathbf{V}$, event $\mathbf{E}$, hidden states for appearance $\mathbf{H}^{A}$ and spatial branches $\mathbf{H}^{S}$) and fed the combined sequences into a fully-connected layers with Sigmoid activation function to make categorical predictions for grammatical marker localization as Eq. 8. Here we use Sigmoid instead of softmax because multiple grammatical markers could occur at the same frame.

$$\mathbf{p} = \sigma([\mathbf{V};\mathbf{E};\mathbf{V}^{Q};\mathbf{H}^{A};\mathbf{H}^{S}]) \tag{8}$$

The prediction $\mathbf{p}$ includes six channels which corresponds to the five grammatical marker types and one other type. The training objective of our framework is to solve a multi-task optimization problem. The overall objective loss function is the weighted sum (α and β are used to balance each part of the overall loss) of low-level value regression loss (low), high-level event detection loss (high), and grammatical marker localization loss (gml):

$$\mathcal{L}_{overall} = \alpha\mathcal{L}_{low} + \beta\mathcal{L}_{high} + \mathcal{L}_{gml} \tag{9}$$

For low-level value regression, we apply a smooth L1 loss between the predicted values and ground truth coarse annotations. For high-level event detection, a binary cross-entropy loss is employed to distinguish between event and background. The grammatical marker localization loss is the mean binary cross-entropy loss over all six classes. Details of the above three losses are defined as:

$$\begin{aligned}\mathcal{L}_{low} &= \frac{1}{K_v N}\sum_{i}^{K_v}\sum_{j}^{N}\text{SmoothL1Loss}(g_{ij}^{v}-v_{ij})\\ \mathcal{L}_{high} &= -\frac{1}{K_e N}\sum_{i}^{K_e}\sum_{j}^{N}(g_{ij}^{e}\log(e_{ij})+(1-g_{ij}^{e})\log(1-e_{ij}))\\ \mathcal{L}_{gml} &= -\frac{1}{K_p N}\sum_{i}^{K_p}\sum_{j}^{N}(g_{ij}^{p}\log(p_{ij})+(1-g_{ij}^{p})\log(1-p_{ij}))\end{aligned} \tag{10}$$

Here N is the length of input sequence, K stands for the number of sequences for a specific task and g represents the annotations from ground truth. The mark v, e, p corresponds to value, event, and grammatical marker respectively.

4 Experiments

4.1 Experimental Setup

ASLLVD Dataset. The dataset we use in this work is the public American Sign Language Lexicon Video Dataset (ASLLVD) [16]. It includes videos of over 715 ASL utterances, each produced by 1-6 native ASL signers. For each utterance, its non-manual markers were

Split	#neg	#whq	#yes-no	#topic	#cond
Training	143	48	40	219	78
Validation	33	11	9	54	10
Testing	87	27	26	121	32

Table 1: Basic utterance statistics of five grammatical markers in ASLLVD dataset splits.

manually annotated using SignStream[1], with grammatical markers, facial expressions, head gestures and their corresponding temporal boundaries. If there is no NMGM detected, *others* will be marked. Each utterance video contains between 1 to 3 NMGMs. The dataset is split into training, validation, and testing sets by the ratio 6:1:3. More detailed statistics of the datasets are shown in Table 1.

Baselines. We compare our model with the following baselines, including [12] which utilizes HM-SVM and trained on hand-crafted features, [10] which constructs boundary-sensitive proposals for further classification, and [21] which adopts a multi-stage framework for action localization. We also conduct a series of ablation experiments as shown in Table 3, where *LM* and *VID* means landmark features and video features correspondingly, *ST* and *MT* stands for single-task training (i.e. NMGM localization only) and multi-task training respectively.

Metrics. For evaluation, F_1-score is applied to evaluate the performance of non-manual event detection, where $F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$. The evaluation of temporal grammatical marker localization can be regarded as a retrieval task and measured by mean average precision at different IoU thresholds (mAP@$tIoU$). A correct action instance prediction should satisfies correct category prediction plus IoU with ground truth instance larger than the evaluation threshold, where IoU is defined as: $IoU = \frac{|S \cap G|}{|S \cup G|}$, where S is detection result and G is ground truth.

Implementation Details. We track and crop the face by "face_recognition" library [7]. The length of sliding window L is to 30 raw video frames in our work. For appearance branch, we adopt 3D ResNets model [8] pretrained on Kinetics Dataset for feature extraction. In the multi-task loss, we set α and β both to 0.5 empirically. In order to balance the number of different NMGM classes, we repeat the instances of less frequent types and sample instances of frequent types. The distribution of non-manual events also become balanced since its distribution is highly related to grammatical markers'.

4.2 Quantitative Results

In this section, we quantitatively evaluate the performance of our proposed model compared to other baselines on public ASLLVD dataset with the setup described in the previous section.

Non-Manual Event Detection. Table 2 shows the performance on non-manual event detection by our best model VID+LM (MT). Our multi-task framework achieves high Precision and F1-score on all the four events (ebraise, eblower, hshake, and hnod), which enables the performance enhancement in grammatical marker localization. Fig 3 gives more examples of predicted low-level sequences. The performance

Grammatical Marker Localization. Apparently, as reported in Table 3 and Table 4, our proposed model VID+LM in multitask model outperforms all the other baselines. The reason of poor performance of HM-SVM is their hand-crafted features built in. While the weak

[1] http://www.bu.edu/asllrp/SignStream/3/

Metric	*ebraise*	eblower	*hshake*	*hnod*
Precision	98.1%	93.2%	89.2%	83.6%
Recall	95.6%	87.4%	87.9%	82.7%
F1-score	96.8%	90.2%	88.5%	83.1%

Table 2: Performance on non-manual event detection by our best model.

Method	mAP@0.3	mAP@0.4	mAP@0.5	mAP@0.6	mAP@0.7
HM-SVM [12]	27.56	25.18	23.16	18.72	13.17
S-CNN [21]	31.05	28.51	25.09	21.51	16.96
BSN [10]	34.44	31.45	28.95	23.40	16.47
LM (ST)	39.70	37.31	33.21	28.67	23.39
LM (MT)	39.76	37.44	33.53	28.76	23.00
VID (ST)	38.65	36.57	33.89	27.16	23.54
VID (MT)	37.43	36.24	34.27	26.71	23.21
VID+LM (ST)	40.31	38.15	35.35	28.32	23.51
VID+LM (MT)	41.65	39.42	36.53	29.27	24.29

Table 3: Grammatical marker localization performance comparison in terms of mAP@*tIoU* on testing split of ASLLVD.

performance of BSN comes from its inaccuracy detection of start point and end point due to the lack of corresponding training data. As for LM and VID individually, the LM is better than VID since LM contains more spatial information. We should point out that our proposed multi-task VID+LM method has significantly gains in terms of largest categories including *topic*, *whq*, *cond*, and *neg*.

5 Conclusion

In this work, we present a multi-task two-stream framework for temporal grammatical marker localization. We utilize 3D ResNets to extract appearance features from detected face ROIs. Additionally, 3D facial landmarks are adopt to capture the spatial configurations. With the

Method	*neg*	*whq*	*yes-no*	*topic*	*cond*	*others*	mAP@*0.5*
HM-SVM [12]	17.24	26.66	31.62	16.16	23.84	23.44	23.16
S-CNN [21]	18.70	28.91	34.28	17.52	25.85	28.84	25.09
BSN [10]	21.56	33.32	39.52	20.19	29.79	29.30	28.95
LM (ST)	12.21	42.94	48.09	21.48	31.92	37.63	33.21
LM (MT)	15.53	43.40	52.16	21.33	29.39	36.40	33.53
VID (ST)	43.30	45.99	22.91	52.23	34.75	23.91	33.89
VID (MT)	45.42	44.17	23.61	55.45	32.15	20.25	34.27
VID+LM (ST)	45.16	43.89	20.56	48.37	36.02	24.94	35.35
VID+LM (MT)	**46.67**	**47.41**	25.82	**56.29**	**37.20**	25.77	**36.53**

Table 4: Class-aware localization performance comparison in terms of mAP@0.5 for on testing split of ASLLVD.

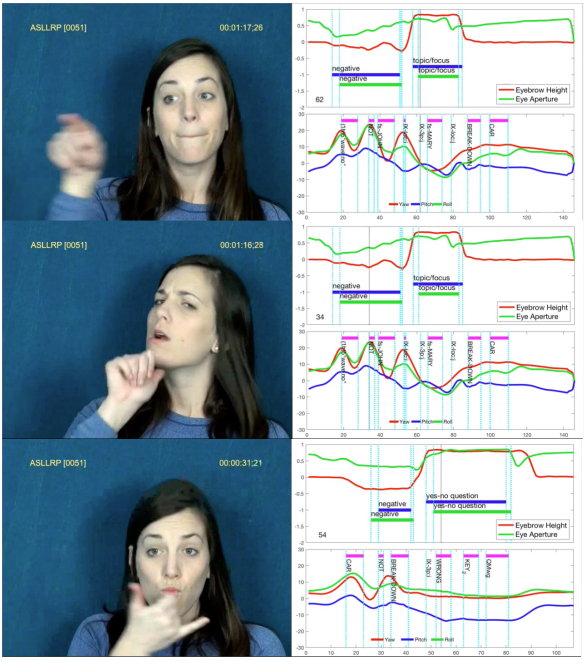

Figure 3: Qualitative visualization of the generated non-manual events graph.

two-stream features as input, we enable jointly training of non-manual events and grammatical marker localization, which encourages the information flow between these relevant tasks. Experimental and qualitative results on the ASLLVD dataset demonstrate the effectiveness and interpretability of our proposed framework.

References

[1] Oya Aran, Thomas Burger, Alice Caplier, and Lale Akarun. Sequential belief-based fusion of manual and non-manual information for recognizing isolated signs. In *International Gesture Workshop*, 2007.

[2] C Fabian Benitez-Quiroz, Kadir Gökgöz, Ronnie B Wilbur, and Aleix M Martinez. Discriminant features and temporal structure of nonmanuals in american sign language. *PloS one*, 2014.

[3] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *ICCV*, 2017.

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[5] Runpeng Cui, Hu Liu, and Changshui Zhang. Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In *CVPR*, 2017.

[6] Xiyang Dai, Bharat Singh, Guyue Zhang, Larry S Davis, and Yan Qiu Chen. Temporal context network for activity localization in videos. In *ICCV*, 2017.

[7] Adam Geitgey. face_recognition: the world's simplest facial recognition api for python and the command line. https://github.com/ageitgey/face_recognition, 2017.

[8] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet. In *CVPR*, 2018.

[9] Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. Video-based sign language recognition without temporal segmentation. *AAAI*, 2018.

[10] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. *ECCV*, 2018.

[11] Bo Liu, Jingjing Liu, Xiang Yu, Dimitris N Metaxas, and Carol Neidle. 3d face tracking and multi-scale, spatio-temporal analysis of linguistically significant facial expressions and head positions in asl. In *LREC*, 2014.

[12] Jingjing Liu, Bo Liu, Shaoting Zhang, Fei Yang, Peng Yang, Dimitris N Metaxas, and Carol Neidle. Recognizing eyebrow and periodic head gestures using crfs for non-manual grammatical marker detection in asl. In *FG*, 2013.

[13] Jingjing Liu, Bo Liu, Shaoting Zhang, Fei Yang, Peng Yang, Dimitris N Metaxas, and Carol Neidle. Non-manual grammatical marker recognition based on multi-scale, spatio-temporal analysis of head pose and facial expressions. *Image and Vision Computing*, 2014.

[14] Dimitris N Metaxas, Bo Liu, Fei Yang, Peng Yang, Nicholas Michael, and Carol Neidle. Recognition of nonmanual markers in american sign language (asl) using non-parametric adaptive 2d-3d face tracking. In *LREC*, 2012.

[15] Nicholas Michael, Dimitris Metaxas, and Carol Neidle. Spatial and temporal pyramids for grammatical expression recognition of american sign language. In *ACM SIGACCESS*, 2009.

[16] Carol Neidle, Jingjing Liu, Bo Liu, Xi Peng, Christian Vogler, and Dimitris Metaxas. Computer-based tracking, analysis, and visualization of linguistically significant non-manual events in american sign language (asl). In *LREC Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel*, 2014.

[17] Tan Dat Nguyen and Surendra Ranganath. Recognizing continuous grammatical marker facial gestures in sign language video. In *ACCV*, 2010.

[18] Junfu Pu, Wengang Zhou, and Houqiang Li. Dilated convolutional network with iterative optimization for continuous sign language recognition. In *IJCAI*, 2018.

[19] Colin Lea Michael D Flynn René and Vidal Austin Reiter Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *ICCV*, 2017.

[20] Sudeep Sarkar, Barbara Loeding, and Ayush S Parashar. Fusion of manual and non-manual information in american sign language recognition. In *Handbook of pattern recognition and computer vision*. 2010.

[21] Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *CVPR*, 2016.

[22] Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang. Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In *CVPR*, 2017.

[23] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *AAAI*, 2017.

[24] Hongsong Wang and Liang Wang. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In *CVPR*, 2017.

[25] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016.

[26] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *ECCV*, 2018.

[27] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018.

[28] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. *ICCV*, 2017.