MTGA: Multi-View Temporal Granularity Aligned Aggregation for Event-Based Lip-Reading

Wenhao Zhang[1*], Jun Wang[1*], Yong Luo[1], Lei Yu[2], Wei Yu[1], Zheng He[1†], Jialie Shen[3]

[1]School of Computer Science, National Engineering Research Center for Multimedia Software and Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, China
[2]School of Electronics and Information, Wuhan University, China
[3]Department of Computer Science, City, University of London, The United Kingdom
{wenhaozhang, junwang_ai, luoyong, ly.wd, yuwei, hezheng}@whu.edu.cn, jialie@gmail.com

Abstract

Lip-reading is to utilize the visual information of the speaker's lip movements to recognize words and sentences. Existing event-based lip-reading solutions integrate different frame rate branches to learn spatio-temporal features of varying granularities. However, aggregating events into event frames inevitably leads to the loss of fine-grained temporal information within frames. To remedy this drawback, we propose a novel framework termed Multi-view Temporal Granularity aligned Aggregation (MTGA). Specifically, we first present a novel event representation method, namely time-segmented voxel graph list, where the most significant local voxels are temporally connected into a graph list. Then we design a spatio-temporal fusion module based on temporal granularity alignment, where the global spatial features extracted from event frames, together with the local relative spatial and temporal features contained in voxel graph list are effectively aligned and integrated. Finally, we design a temporal aggregation module that incorporates positional encoding, which enables the capture of local absolute spatial and global temporal information. Experiments demonstrate that our method outperforms both the event-based and video-based lip-reading counterparts.

Code — https://github.com/whu125/MTGA

Introduction

With the continuous advancement of human-computer interaction technology, lip reading as a silent form of communication has attracted widespread attention. Existing lip-reading approaches are typically based on video inputs (Sheng et al. 2024), which are required to be high resolution. Poor performance may be obtained when the video is blurry or involves rapid movement.

Event cameras (Gallego et al. 2020), as innovative visual sensors, can capture changes in pixel brightness at the microsecond level, producing an event stream output. Their sensitivity to motion and ability to reduce redundancy make them well-suited for lip-reading tasks that require capturing fine-grained features and lip movements.

There exist many event-based recognition approaches. For example, the event streams are represented as point clouds in (Wang et al. 2019) to preserve maximal information, but the computational cost may be very high for a large number of events. Li *et al.* (Li et al. 2021a) employ neighboring graph structures to capture of local correlations, yet lip-reading task may introduce an excessive number of nodes and edges, and thus make the feature extraction difficult. Some other approaches assemble asynchronous events into group tokens (Peng et al. 2023), based on timestamps and polarity. This can efficiently reduce feature dimension but do not adequately exploit the temporal information, rendering them less effective for lip-reading task that demands a focus on subtle temporal variations.

A recent lip-reading approach based on event camera is presented in (Tan et al. 2022). In this approach, the event points within a set of time are aggregated into a single frame, so that the video processing strategies can be employed. However, during the talking, lips often change subtly and rapidly, and using image frame for representation may obscure these detailed variations. Tan *et al.* (Tan et al. 2022) alleviate this issue by developing a Multi-grained Spatio-Temporal Features Perceived (MSTP) Network that merges features learned from two branches with different frame rates through a Message Flow Module (MFM). However, MSTP only aggregate the events into frames, and hence inevitably leads to the loss of intra-frame local information.

To remedy this drawback, we propose a novel multi-view learning method termed multi-view temporal granularity aligned aggregation (MTGA) for event-based lip-reading. In particular, we first represent the events from two different viewpoints, as shown in Figure 1. One view is to aggregate the events into event frames at an appropriate temporal segment following (Tan et al. 2022), and convolution is applied to extract global spatial features. The other view is to partition the event flow (at the same temporal segment as the first view) into a voxel grid within a three-dimensional space. Within each time segment, we construct a three-dimensional geometric neighboring graph using the most informative voxels. The resulting local graphs are temporally connected to form a graph list, which are utilized to extract critical local details and contextual correlations using Gaussian Mixture Model convolutions (Li et al. 2021a).

[*]These authors contributed equally.
[†]Corresponding author.

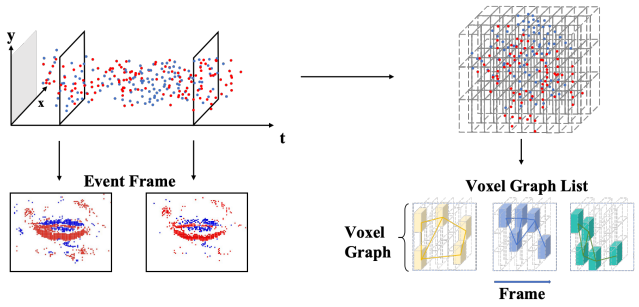


Figure 1: The two event representations we adopt. (Left) An event frame is obtained by integrating event points within a certain temporal range in the event stream. (Right) After dividing the event stream into a voxel grid in three-dimensional space, a three-dimensional geomeric neighboring graph is constructed at regular time intervals, resulting in a voxel graph list used for event representation.

Our designed voxel graph list are temporally aligned with the event frames, and then we devise a fusion module that combines features extracted from both branches through convolution within each temporal segment. Due to the stronger spatio-temporal connectivity, the voxel graphs can serve as supplement to the global spatial information extracted from the event frame, and thereby result in more discriminative spatio-temporal features.

Finally, we propose a temporal aggregation module to combine the obtained features of different temporal segments. This is achieved by introducing the idea of positional encoding adopted in Transformer (Vaswani et al. 2017). When constructing the voxel graph, the absolute position information of the voxel in the events is discarded. Therefore, to capture the local absolute spatial information, the voxel coordinates are utilized to obtain absolute positional features through positional encoding. The result embeddings are appended to the temporal segment features and fed into the Bi-GRU and Self-Attention layers to exploit the global temporal information. We verify the effectiveness of our proposed method on the DVS-Lip (Tan et al. 2022) dataset, and the experiments demonstrate that our model can significantly outperform the state-of-the-arts in the field of event-based lip-reading recognition. For example, we obtain a 4.1% relative improvement in terms of overall accuracy compared the most competitive counterpart. In addition, we conducted experiments on the DVS128-Gait-Day dataset, and the experimental results proved that our model has good generalization performance.

In summary, the main contributions of our work are as follows:

- We propose a novel representation method for event streams termed temporal segmented voxel graph list, which can effectively exploit the local spatio-temporal correlations.
- We design a temporal granularity aligned fusion module, which combines the voxel graph list features and frame features within each temporal segment, thereby obtaining more discriminative spatio-temporal features.
- We design a temporal aggregation module, which can capture the local position and global temporal information.
- To the best of our knowledge, MTGA is the first work in the field of event-based lip reading that combines features from multiple views.

Related Work

Lip-reading

Lip-reading aims to understand what the people say according to the lip and facial movements. Traditional lip-reading approaches extract handcrafted features and adopt classic classifiers for words recognition (Anderson et al. 2013; Kim et al. 2015). Recently, some deep learning-based recognition networks are proposed for lip-reading. Front-end of the network is used for feature extraction and some commonly utilized backbones are CNNs (Li et al. 2021b; Sheng et al. 2021a; Feng et al. 2020), GCNs (Sheng et al. 2021b; Liu, Chen, and Yang 2020; Zhang and Zhao 2021), and Visual Transformers (Prajwal, Afouras, and Zisserman 2022; Wang 2019; Han et al. 2022). In the back-end, RNNs (Luo et al. 2020; Zhao et al. 2020; Xiao et al. 2020), Transformer (Afouras et al. 2018; Zhang, Cheng, and Wang 2019; Vaswani et al. 2017), and TCN (Martinez et al. 2020; Afouras et al. 2018) are usually adopted to aggregate temporal information. For instance, Feng (Feng et al. 2020) *et al.* chose the typical RNN-based Bi-GRU as the back-end network for lip-reading, where Bi-GRU's strong context learning and sequence modeling capabilities are effectively leveraged for word recognition. Martinez (Martinez et al. 2020) *et al.* proposed a Multi-Scale Temporal Convolutional Network (MS-TCN) structure, which is able to capture long-term dependency. Most of the existing lip recognition approaches are video-based. These approaches are limited in that the video resolution may be not enough to capture some subtle movements and the videos may contain motion blur. Inspired by (Tan et al. 2022), we proposed a method based on event streams, which often contains more rich information than videos.

Event-based Recognition

Event camera is a groundbreaking visual sensor, where each pixel independently detects changes in luminance instead of capturing fixed-interval full-image frames (Gallego et al. 2020). When the brightness change of a pixel exceeds a preset threshold, the event camera generates an event point that includes the pixel's location, timestamp, and the polarity indicating the change in brightness. This allows for the capture of visual information at a high temporal resolution and provides a broad dynamic range. In the lip-reading application, high temporal resolution of event cameras is crucial for detecting subtle motion, and the sensitivity to motion makes them to particularly suitable to capture lip movements.

In event-based recognition, a representative work is presented in (Zhu et al. 2019), where the event stream is segmented by timestamps, and the event points within a certain timestamp range are unified to induce a single time

point, resulting in 'event frames' similar to image frames. The event streams can also be regarded as point clouds, and a representative work is conducted by Wang (Wang et al. 2019) *et al.*, which utilized PointNet for gesture recognition. Another solution for event-based recognition is to integrate time and polarity information with tokens to represent events (Peng et al. 2023), and this facilitates effective feature communication and integration in the spatial and temporal-polarity domains. In (Jiang et al. 2023), the events are represented as both point clouds and voxels, which are then constructed as graphs for feature extraction and combination using a graph network. These event-based recognition approaches are used for the recognition of static objects or actions of short duration, and the temporal information are not exploited and thus cannot achieve satisfactory performance in the lip-reading task during a long time range. As far as we know, the first lip-reading work based on event data is conducted by Tan *et al.* (Tan et al. 2022), where a multi-granularity spatio-temporal feature perception network is designed to exploit the temporal information to some extent. However, the events are only represented as event frames in (Tan et al. 2022), and thus the fine-grained information within frames cannot be exploited. This issue is addressed in this paper by integrating information from mulitple views.

Methodology

Framework Overview

In this paper, we present a multi-view learning method for event-based lip-reading, and an overview of the proposed framework is shown in Figure 2. In the following, we will depict each module.

Multi-view Event Representation

Event Frame. Converting the event stream into frames requires an intermediary, and a suitable choice is to encode the event stream into a spatio-temporal voxel grid (Zhu et al. 2019). For the event camera output, the event stream $\varepsilon = \{(x_n, y_n, t_n, p_n)\}_{n=1}^{N}$ includes x, y direction coordinate, time and polarity. we discrete the timestamps of the event stream into T time intervals through normalization. Each event point weights its polarity according to temporal proximity and is contributed to the two closest time intervals. The event frames $V(t, x, y)$ are generated using the method described in (Tan et al. 2022).In our work, we divide the event stream into 60 event frames to represent the event stream.

Voxel Graph List. Another representation method initially maps the event data onto discretized voxels in a three-dimensional space. This approach enhances the capture of the events' spatial positioning and distribution information. To aggregate the event points while preserving polarity features, we convert the voxels into graph nodes, with the event points' polarities serving as the feature matrix of the nodes.

To align with event frames in the temporal dimension and supplement the temporal information within each frame, we first normalize time over $n * T$ (n=3 in our work), dividing the three-dimensional space sized $(n * T, H, W)$ into voxels with each voxel sized $(1, h, w)$, where H and W denote the spatial capture range of the event camera. . The three-dimensional space is respectively partitioned into $(n*T, \frac{H}{h}, \frac{W}{w})$ segments. After ensuring temporality aligned, each event frame corresponds to $n * \frac{H}{h} * \frac{W}{w}$ voxels.

To reduce computational load, we select the k voxels with the highest number of event points per frame as graph nodes, denoting $\mathrm{O}_i \in (\mathrm{o}_{i1}...\mathrm{o}_{ik})$ as the chosen voxels. We select K event points within each voxel to obtain their polarity matrix as node features $\mathbf{a} \in \mathrm{R}^K$, thus each graph node is described by $\mathrm{o}_{ij} \in (\mathrm{t}_{ij}, \mathrm{x}_{ij}, \mathrm{j}_{ij}, \mathbf{a}_{ij})$. An edge exists between nodes when the Euclidean distance between their three-dimensional coordinates $[t_i, x_i, y_i]$ is less than threshold R, using this Euclidean distance as the edge feature, and the edge set is defined as E_i. The edge feature set $\mathrm{W}_i \in \mathrm{R}^{num\,edges}$ is obtained by taking the Euclidean distance as the feature of the edges. Represent the $i-$th frame corresponding to the geometric neighboring graph with k nodes as $\mathrm{g}_i \in (\mathrm{O}_i, \mathbf{E}_i, \mathbf{W}_i)$.Then the voxel graph list is represented as $g = [\mathrm{g}_1...\mathrm{g}_T]$. In Figure 2, the voxel graph list maintains consistency with the event frames in the temporal dimension, and the three-dimensional geometric neighboring graph also compensates for the loss of intra-frame temporal information. Selecting voxels with with a higher count of points as nodes also emphasizes key local features.

Feature Extraction

Event frames can reflect the overall spatio-temporal information of the event stream but overlook the timestamp differences within a single frame, and they lack advantages in capturing local changes in the image. In contrast, voxel graph list maintain temporal associations within a frame, and the selection of graph nodes highlights areas with significant local changes, effectively representing the temporal information and local features within a frame. Consequently, we design a dual-branch network that extracts features from both image frames and voxel graph list. These features are then input into a feature fusion module, which concurrently perceives the global spatial and local spatio-temporal features.

Event Frame Feature Extraction. In the branch of the event frame, we select ResBlock (He et al. 2016) as the core feature extractor to capture characteristics. ResBlock, centered around CNN, employs residual connections to circumvent gradient issues and is extensively utilized in feature extraction of image frames. The event frame f is initially subjected to convolution to acquire feature $F_{in}^{f} \in R^{T*C_{in}^{f}*H_{in}^{f}*W_{in}^{f}}$, which is then processed through ResBlock to yield feature $F_{out}^{f} \in R^{T*C_{out}^{f}*H_{out}^{f}*W_{out}^{f}}$, as depicted below:

$$F_{out}^{f} = ResBlock(F_{in}^{f}). \tag{1}$$

Voxel Graph List Feature Extraction. In the branch of the voxel graph list, for the graph g_i within the list, con-

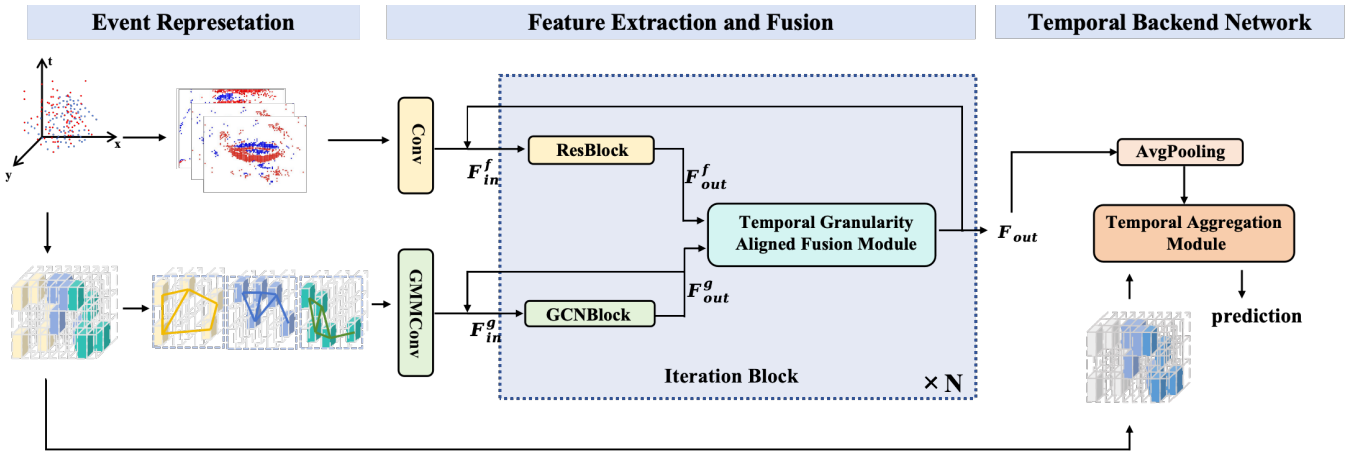


Figure 2: The architecture of our proposed network. The model is divided into three components: (1) Event representation, which describes the events from two different viewpoints, i.e., event frames and voxel graph list; (2) Feature extraction and fusion, which extracts features for the two views separately, and combines them through a temporal granularity aligned fusion module; (3) Temporal backend network, which aggregates the global temporal information.

sidering its node $o_i = ((x_i, y_i, t_i), \mathbf{a_i})$ where $\mathbf{a_i}$ represents the feature vector, we utilize GMMConv (Monti et al. 2017) and a GCNBlock with a residual structure to extract the spatio-temporal features of the collective neighboring graph. GMMConv employs a set of Gaussian kernels to perform convolution operations on each node in the graph and its neighbors. The GCNBlock is expressed as follows:

$$F^g_{res} = GMMConv(F^g_{in}), \tag{2}$$

$$F^g_{out} = ELU(F^g_{res} + GMMConv(ELU(BN(F^g_{res})))), \tag{3}$$

where $F^g_{in} \in R^{T*N^g_{in}*C^g_{in}}$ represents the initial features obtained from the voxel graph list g after the convolution with the Gaussian Mixture Model, and $F^g_{out} \in R^{T*N^g_{out}*C^g_{out}}$ is the output feature after passing through the GCNBlock. $GMMConv()$ denotes the Gaussian Mixture Model convolution operation, which computes new node features based on node features $\mathbf{A_i}$, edge indices $\mathbf{E_i}$, and edge attributes $\mathbf{W_i}$. $BN()$ stands for batch normalization, and $ELU()$ is the activation function applied to the output of the convolution.

Feature Fusion

After obtaining the event frame and voxel graph list features, we design a fusion module centered around an iterative fusion mechanism in the fusion model for deeply integrating the event frame features and voxel grid features. This fusion model fully utilizes the complementary nature of these two types of features. Within the fusion module, based on the consistency of the features in the time dimension, we adopted temporal granularity alignment fusion approach. As illustrated in the Figure 3, for each event frame, a voxel grid from the same time segment is selected. The voxel graph list features are dimensionally expanded to adapt to the size of the event frame features, and then convolutional and residual modules are added to maintain the integrity of the information during the feature fusion process. The algorithm of the fusion module is as follows:

$$F_{res} = Concat(F^f_{out}, Conv(F^g_{out})), \tag{4}$$

$$F^{fus}_{out} = ELU(BN(Conv(F_{res}))) + F_{res}, \tag{5}$$

where F^f_{out} and F^g_{out} represent the output features of the previous module, serving as the input features of the fusion module. F^{fus}_{out} represents the output of the fusion module.

To fully explore the contextual associations and deep fine-grained information of the features while retaining the coarse-grained features presented by the frame features, we utilized an iterative strategy. In our configured N-layer (N=4 in our work) fusion model, each layer contains a fusion module. As shown in the feature extraction and fusion section of the Figure 2, the output features of the previous layer's fusion module are processed through a ResBlock for feature extraction, and re-entered into the fusion module together with the voxel grid features processed by the GMMBlock. Through multiple iterations, the features of the two branches are continuously fused, thereby realizing the role of voxel grid contextual features and local features in guiding the overall feature learning of the event frame. The complete iterative fusion block is represented as follows:

$$F^{fus}_i = FM(ResBlock(F^{fus}_{i-1}), GCNBlock(F^g_{i-1})) \tag{6}$$

$$F_{out} = F^{fus}_N, \tag{7}$$

where F^{fus}_i denotes the output of the fusion module after the i-th iteration, F_{out} denotes the fused features of the final output and FM refers to FusionModule.

Temporal Backend Network

In this section, we propose a temporal back-end network based on Bi-GRU and Self-Attention with positional encoding. We first incorporate positional encoding using graph convolution to capture the absolute position information of

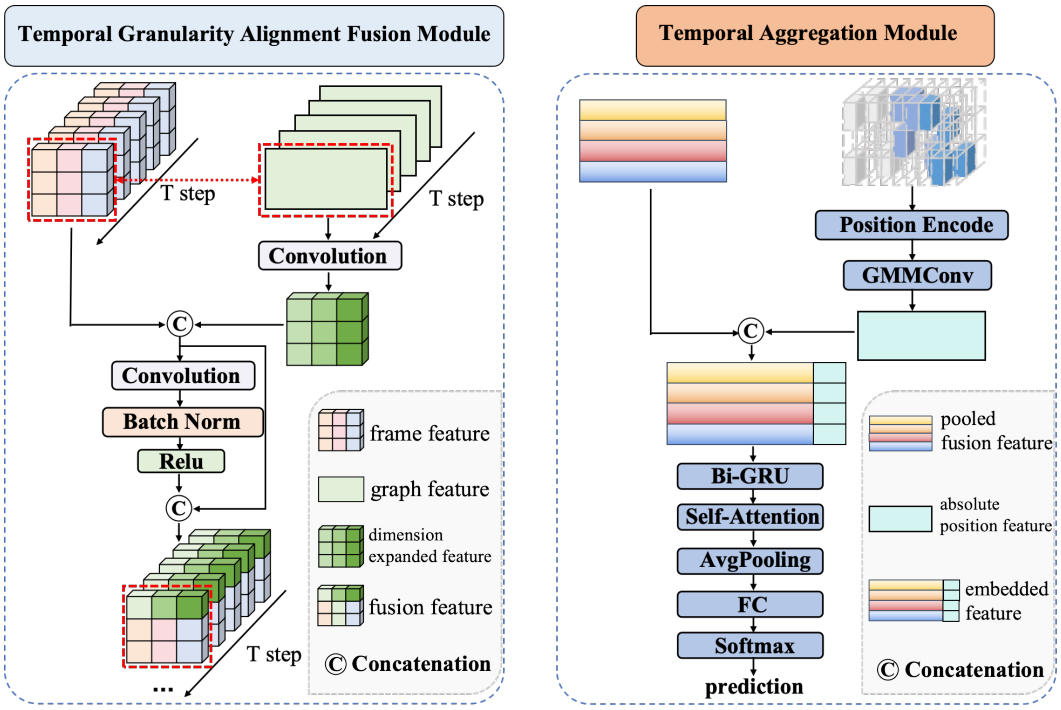


Figure 3: Illustrations of our designed two modules. (Left) The temporal granularity aligned fusion module. At each time step, the voxel graph node features are convolved to the same shape and then merged. The merged features achieve the fusion of spatio-temporal features through convolution and residual block. (Right) The temporal aggregation module. Extract features from the original voxel according to encoding, and then concatenate them with the sequence model, aggregating temporal information through Bi-GRU and Self-Attention layers.

nodes. This information is then concatenated with the spatio-temporal features obtained in the previous section. Subsequently, the Bi-GRU is employed to learn the temporal features of the feature sequences and contextual associations. Finally, the Self-Attention module is utilized to emphasize important temporal divisions.

Due to the properties of graph structures, graph convolution can only capture the relative positional relationships between nodes, corresponding to local spatio-temporal features in three-dimensional coordinates. As nodes' voxels are crucial local features, in order to capture the absolute position information of nodes, we apply positional encoding to the nodes before the temporal module. Specifically, we replace the polarity in voxel graph with the coordinates of nodes in the voxel grid as node features. This incorporation of node positions within the voxel grid allows the model to comprehend the spatial arrangement of nodes in the voxel grid, thus enhancing the discrimination ability of similar graph structures.

In the selection of temporal aggregation modules, similar to previous work (Tan et al. 2022), we employ Bidirectional Gate Recurrent Units (Bi-GRU) to further aggregate temporal information, obtaining feature sequences containing temporal information. Through visual analysis, we observed variations in the number of event points across different temporal divisions. To ensure the model focuses on crucial sequences, we integrate a Self-Attention module to learn attention weights for each temporal division. These weights are then used to weight the feature sequences accordingly. The computation of the complete temporal back-end network is as follows:

$$Encode_{pos} = GMMConv(VoxelGraphList_{coor}), \quad (8)$$

$$W_{att} = Softmax(\frac{QK^T}{\sqrt{d_k}})V, \quad (9)$$

$$P = Softmax(FC(GAP_t(Bi(Cat(Encode_{pos}, F_{out}) * W_{att})))), \quad (10)$$

where $VoxelGraphList_{coor}$ represents voxel graph list with coordinates used as node features. $Encode_{pos}$ represents the positional encoding obtained. Q, K, and V respectively represent the queries, keys, and values in Self-Attention. d_k denotes the dimensionality of the feature vectors, W_{att} stands for the Self-Attention weights, and P signifies the output probabilities for each word in the vocabulary.

Experimental Evaluation

Evaluation Datasets

DVS-Lip. Acknowledging the significant contribution of Tan (Tan et al. 2022) *et al.* to event-based lip-reading tasks, we utilize the DVS-Lip dataset they compiled, which is focused on lip-reading of words through an event camera. This dataset, gathered using event camera DAVIS346, is based

on selecting words from the conventional camera-based lip-reading dataset LRW, including words that are often confused. It comprises event streams and intensity images for 100 words, with the lip region extracted within a 128*128 range. The validation of our model will be conducted using the DVS-Lip dataset, which will facilitate a comparison between our method and other approaches, highlighting our method's superiority.

DVS128-Gait-Day. To demonstrate the generalization capability of the models, we also conducted experiments on this dataset. The DVS128-Gait-Day(Wang et al. 2021) dataset contains 4,000 gait event samples from 20 volunteers,who are asked to walk normally 100 times in front of a DVS128 sensor mounted on a tripod at approximately 90 degrees to the walking direction.

Experimental Results

Since we utilized the same dataset and evaluation method as MSTP(Tan et al. 2022), we selected MSTP for comparison. Additionally, we compared our method with some of the most advanced video or event-based action recognition and object recognition methods to demonstrate the superiority of our approach in the lip-reading task.

Our experimental results, along with those of other methods, are presented in Table 1. The table illustrates that our Multi-view Temporal Granularity Aligned Aggregation significantly outperforms other action and object recognition methods in lip-reading tasks. This is primarily because the key to lip-reading is recognizing the syllable changes caused by lip movements. Recognition methods for objects generally lack the ability to capture local temporal features and struggle with lip-reading tasks that have a long time axis. Our method also surpasses the MSTP structure on the same dataset, indicating that our voxel graph list effectively compensates for the intra-frame temporal information lost during the temporal normalization of event frames. The structure of the voxel graph emphasizes the areas with a higher density of event points, adding local details to the overall features. Furthermore, the feature vectors representing the graph nodes effectively replenish some of the polarity information lost due to polarity weighted encoding.

In addition, we conducted experiments on the DVS128-Gait-Day dataset. As this is an event action stream dataset, we made a fair comparison with MSTP. The experimental results showed that under the same number of training epochs, MTGA achieved better performance. This proves that MTGA has broad applicability.

Ablation Studies

Effects of Branch. We compared our fused model with individual branches in Table 3:(1) Only Event Frame Branch: this branch solely involves feature extraction for each frame of the event image. (2) Only Voxel Graph List Branch: this branch exclusively focuses on feature extraction from the temporally segmented voxel graph list. According to Table 3, both branches perform less effectively compared to the fused model. The Event Frame Branch, using CNN, loses temporal detail due to frame aggregation. The Voxel Graph List Branch, capturing temporal features with GMMConv, is less efficient and generalized than CNN. The data show that the accuracy of the Voxel Graph List Branch is only 70.03%, which is 1.68% lower than the Event Frame Branch, and the higher accuracy among the two branches is 3.37% lower than the accuracy of the fused branch.

Effects of Fusion Methods. We further explored the impact of different fusion methods. We set up a comparison among three fusion approaches: (1) Overall feature fusion based on given weights, (2) Feature fusion based on the attention mechanism, (3) Fusion based on temporal granularity alignment. The first method combined features linearly with fixed weights, a simple approach that lacks adaptability. The second employed an attention mechanism to dynamically allocate feature importance, enhancing adaptability. The third is the fusion based on temporal granularity alignment we introduced earlier, where the features of each frame are considered and fused individually, maintaining temporal details and dynamic changes. Table 4 displays our experimental results. It shows that the accuracy of the first fusion method is only 71.26%. The accuracy of the second fusion method is 72.51%. Although the attention mechanism can adaptively focus on the most useful features according to the context, in the lip-reading task where the front-end and back-end tasks are separated, the front-end feature fusion considers information within asynchronous time, which adversely affects the feature extraction at each timestep, leading to a decrease in accuracy. The experimental results further illustrate that our fusion method is the most effective in handling such time-series data.

Effects of Temperal Aggregation Module. Our approach integrates a Bi-GRU and Self-Attention model with position encoding, enhancing the back-end network beyond the Bi-GRU used in (Tan et al. 2022). Position encoding enriches the node positions absent in the front-end network. The Bi-GRU captures temporal dynamics and contextual links, followed by Self-Attention to weigh each time step.We conducted ablation experiments to verify the effectiveness by controlling whether to use position encoding and the Self-Attention module. The experimental results in Table 5 demonstrate that compared to using only Bi-GRU (M1) as the back-end network, the accuracy is improved by 0.59% with position encoding (M2), 1.31% with the Self-Attention module (M3), and 1.52% with both modules combined. It can be concluded that position encoding can indeed provide additional information, and Self-Attention can help our model focus on more important parts.

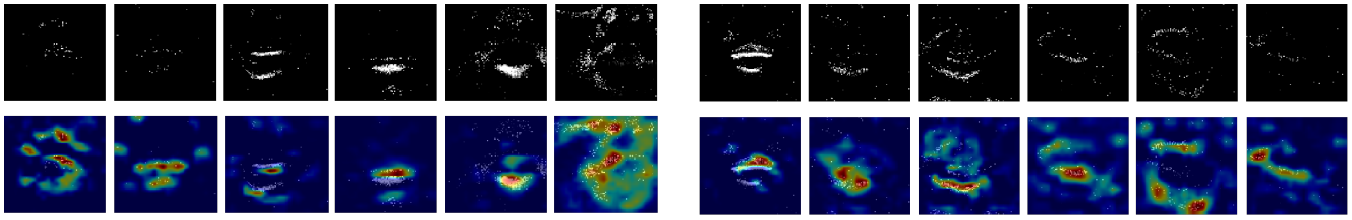
Qualitative Analysis

Visualization of Frontend Network. To showcase the excellent effects of front-end feature extraction and fusion networks, we applied Grad-CAM (Selvaraju et al. 2017) to our model using samples from the DVS-Lip test dataset. The results from Grad-CAM vividly illuminate the visual saliency areas by computing gradients with respect to a specific class.

As shown in Figure 4, we present two examples to demonstrate the effectiveness of our model. With 60 bins, we selected one event frame image every 10 bins, where the top

Model	Representation	Acc1(%)	Acc2(%)	Acc(%)
Event Clouds (Wang et al. 2019)	point clouds	35.82	48.51	42.15
EST (Gehrig et al. 2019)	event spike tensor	40.91	56.45	48.66
ACTION-Net (Wang, She, and Smolic 2021)	video frame	58.32	79.41	68.84
Martinez et al. (Martinez et al. 2020)	video frame	55.60	75.46	65.51
AGCN (Jiang et al. 2023)	(point,voxel)	55.52	80.47	67.74
GET (Peng et al. 2023)	group token	58.96	80.82	69.80
MSTP (Tan et al. 2022)	event frame	62.17	82.07	72.10
MTGA	**(event frame,voxel graph)**	**63.90**	**86.38**	**75.08**

Table 1: Comparisons with existing event-based models and the state-of-the-art video-based models on the DVS-Lip test set. Acc1 represents the accuracy achieved on the first subset of the test data, Acc2 corresponds to the accuracy on the second subset, and Acc indicates the overall accuracy across the entire test dataset.



(a) Ground Truth:EDUCATION Prediction:EDUCATION **(b) Ground Truth:TOMORROW Prediction:TOMORROW**

Figure 4: Visualization of the saliency maps for words (a) “education” and (b) “tomorrow”.

Model	Epoch	Acc(%)
MSTP	10	90.28
MTGA	**10**	**95.42**

Table 2: Comparison results of different models on the DVS128-Gait-Day dataset.

Branch	Acc1(%)	Acc2(%)	Acc(%)
Event Frame	60.05	83.81	71.71
Voxel Graph List	57.45	83.12	70.03
Ours	**63.90**	**86.38**	**75.08**

Table 3: Ablation study of adopting different branches.

Fusion Method	Acc1(%)	Acc2(%)	Acc(%)
Weight	59.39	83.63	71.26
Attention	60.28	85.22	72.51
Ours	**63.90**	**86.38**	**75.08**

Table 4: Comparison of different fusion model.

Model	PEB	AM	Acc(%)
M1	✗	✗	73.56
M2	✓	✗	74.15
M3	✗	✓	74.87
MTGA	✓	✓	**75.08**

Table 5: Ablation study of adopting temperal aggregaion moudle.PEB refers to Position Embedding, and AM refers to Self-Attention Mechanism.

row shows the input original images, and the bottom row displays the overlaid images with heatmaps generated by applying Grad-CAM to the last network layer. It is observable that, after the fusion of the two branches, our model focuses on the most crucial local information in the images, while also attributing certain weight to the scattered points around the edges, representing the global information.

Conclusion

In this paper, we propose a model named MTGA, wherein the feature of a large number of event data are extracted from multiple views. Through our designed Temporal Granularity aligned Fusion Module, the fused features simultaneously possess global spatial and local spatio-temporal information, addressing the issue of feature loss caused by previous single-frame aggregating methods. Experiments on the DVS-Lip dataset validate the superiority of our model. From the results, we mainly conclude that: (1) Each branch of the model has the ability to capture different features, and our fusion module effectively integrates them.; (2) Our fusion approach has a competitive advantage over other methods; (3) Our supplementary position encoding and Self-Attention effectively improve the aggregation effect of the back-end network. In the future, we intend to integrate other event views and further explore fusion strategies to enable the model to comprehensively reflect the features. Additionally, we also aims to enhance the model’s generalization performance to effectively handle other event recognition tasks.

Acknowledgments

This work is supported in part by the National National Key Research and Development Program of China (No. 2022YFF0712300), the National Natural Science Foundation of China (Grant No. U23A20318, 62276195 and 62376200), the Science and Technology Major Project of Hubei Province (Grant No. 2024BAB046) and the Innovative Research Group Project of Hubei Province (Grant No. 2024AFA017). The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

Afouras, T.; Chung, J. S.; Senior, A.; Vinyals, O.; and Zisserman, A. 2018. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 44(12): 8717–8727.

Anderson, R.; Stenger, B.; Wan, V.; and Cipolla, R. 2013. Expressive visual text-to-speech using active appearance models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3382–3389.

Feng, D.; Yang, S.; Shan, S.; and Chen, X. 2020. Learn an effective lip reading model without pains. *arXiv preprint arXiv:2011.07557*.

Gallego, G.; Delbrück, T.; Orchard, G.; Bartolozzi, C.; Taba, B.; Censi, A.; Leutenegger, S.; Davison, A. J.; Conradt, J.; Daniilidis, K.; et al. 2020. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1): 154–180.

Gehrig, D.; Loquercio, A.; Derpanis, K. G.; and Scaramuzza, D. 2019. End-to-end learning of representations for asynchronous event-based data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5633–5643.

Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. 2022. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1): 87–110.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Jiang, B.; Yuan, C.; Wang, X.; Bao, Z.; Zhu, L.; and Luo, B. 2023. Point-voxel absorbing graph representation learning for event stream based recognition. *arXiv preprint arXiv:2306.05239*.

Kim, T.; Yue, Y.; Taylor, S.; and Matthews, I. 2015. A decision tree framework for spatiotemporal sequence prediction. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 577–586.

Li, Y.; Zhou, H.; Yang, B.; Zhang, Y.; Cui, Z.; Bao, H.; and Zhang, G. 2021a. Graph-based asynchronous event processing for rapid object recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 934–943.

Li, Z.; Liu, F.; Yang, W.; Peng, S.; and Zhou, J. 2021b. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 33(12): 6999–7019.

Liu, H.; Chen, Z.; and Yang, B. 2020. Lip Graph Assisted Audio-Visual Speech Recognition Using Bidirectional Synchronous Fusion. In *INTERSPEECH*, 3520–3524.

Luo, M.; Yang, S.; Shan, S.; and Chen, X. 2020. Pseudo-convolutional policy gradient for sequence-to-sequence lip-reading. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, 273–280. IEEE.

Martinez, B.; Ma, P.; Petridis, S.; and Pantic, M. 2020. Lipreading using temporal convolutional networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6319–6323. IEEE.

Monti, F.; Boscaini, D.; Masci, J.; Rodola, E.; Svoboda, J.; and Bronstein, M. M. 2017. Geometric deep learning on graphs and manifolds using mixture model cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5115–5124.

Peng, Y.; Zhang, Y.; Xiong, Z.; Sun, X.; and Wu, F. 2023. GET: group event transformer for event-based vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6038–6048.

Prajwal, K.; Afouras, T.; and Zisserman, A. 2022. Sub-word level lip reading with visual attention. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 5162–5172.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Sheng, C.; Kuang, G.; Bai, L.; Hou, C.; Guo, Y.; Xu, X.; Pietikäinen, M.; and Liu, L. 2024. Deep learning for visual speech analysis: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Sheng, C.; Pietikäinen, M.; Tian, Q.; and Liu, L. 2021a. Cross-modal self-supervised learning for lip reading: When contrastive learning meets adversarial training. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2456–2464.

Sheng, C.; Zhu, X.; Xu, H.; Pietikäinen, M.; and Liu, L. 2021b. Adaptive semantic-spatio-temporal graph convolutional network for lip reading. *IEEE Transactions on Multimedia*, 24: 3545–3557.

Tan, G.; Wang, Y.; Han, H.; Cao, Y.; Wu, F.; and Zha, Z.-J. 2022. Multi-grained spatio-temporal features perceived network for event-based lip-reading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20094–20103.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, C. 2019. Multi-grained spatio-temporal modeling for lip-reading. *arXiv preprint arXiv:1908.11618*.

Wang, Q.; Zhang, Y.; Yuan, J.; and Lu, Y. 2019. Space-time event clouds for gesture recognition: From RGB cameras to event cameras. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1826–1835. IEEE.

Wang, Y.; Zhang, X.; Shen, Y.; Du, B.; Zhao, G.; Cui, L.; and Wen, H. 2021. Event-stream representation for human gaits identification using deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7): 3436–3449.

Wang, Z.; She, Q.; and Smolic, A. 2021. Action-net: Multipath excitation for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13214–13223.

Xiao, J.; Yang, S.; Zhang, Y.; Shan, S.; and Chen, X. 2020. Deformation flow based two-stream network for lip reading. In *2020 15th IEEE international conference on automatic face and gesture recognition (FG 2020)*, 364–370. IEEE.

Zhang, C.; and Zhao, H. 2021. Lip reading using local-adjacent feature extractor and multi-level feature fusion. In *Journal of Physics: Conference Series*, volume 1883, 012083. IOP Publishing.

Zhang, X.; Cheng, F.; and Wang, S. 2019. Spatio-temporal fusion based convolutional sequence learning for lip reading. In *Proceedings of the IEEE/CVF International conference on Computer Vision*, 713–722.

Zhao, X.; Yang, S.; Shan, S.; and Chen, X. 2020. Mutual information maximization for effective lip reading. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, 420–427. IEEE.

Zhu, A. Z.; Yuan, L.; Chaney, K.; and Daniilidis, K. 2019. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 989–997.