Collaborative Attention Mechanism for Multi-View Action Recognition

Yue Bai *1, Zhiqiang Tao†2, Lichen Wang‡1, Sheng Li§3, Yu Yin¶1, and Yun Fu∥1

1Department of Electrical and Computer Engineering, Northeastern University, USA
2Department of Computer Science and Engineering, Santa Clara University, USA
3Department of Computer Science, University of Georgia, USA

Abstract

Multi-view action recognition (MVAR) leverages complementary temporal information from different views to enhance the learning process. Attention is an effective mechanism which has been extensively adopted for modeling temporal data. However, most existing MVAR methods only utilize attention to extract view-specific patterns. They ignore the potential to dig latent mutual-support information in attention space. To fully take the advantage of the multi-view cooperation, we propose a collaborative attention mechanism (CAM). It detects the attention differences among multi-view inputs, and adaptively integrates complementary frame-level information to benefit each other. Specifically, we utilize recurrent neural network (RNN) by expanding long short-term memory (LSTM) as a Mutual-Aid RNN (MAR). CAM takes advantages of view-specific attention pattern to guide another view and unlock potential information which is hard to explore by itself. Extensive experiments on three action datasets illustrate our CAM achieves better result for each single view, and also boosts the multi-view performance.

1 Introduction

Multi-view action recognition (MVAR) has drawn more attention Cai et al. [2014], Cheng et al. [2012], Holte et al. [2011] as the increasing usage of multi-modality sensors to boost the learning performance. It is a challenging task due to the difficulties of extracting complicated temporal patterns and fully utilizing multi-view information. Most MVAR scenarios are roughly grouped into two categories. The first category captures multi-view action videos using multiple sensors belonging to the same modality (e.g. RGB cameras). These sensors are deployed in different viewpoints (e.g. front, top, and side). Corresponding methods Cai et al. [2014], Holte et al. [2011], Ji et al. [2014] are designed to make use of distinctive information obtained from each of them. The second category simultaneously collects multi-view action sequences via using different modality sensors (e.g., RGB, depth, and skeleton Lin et al. [2012], Rahmani et al. [2016], Wang et al. [2019c]). For example, Kinect sensors collect RGB, depth, and skeleton data simultaneously Pagliari and Pinto [2015], Zhang [2012]. Many modalities containing physical characteristics (e.g., electromyography (EMG), acoustical, and acceleration) are also used to explore multi-view data for action recognition Bu et al. [2009], Chen et al. [2015], De la Torre et al. [2009]. Various modalities provide complementary information to achieve better performance Wang et al. [2019b].

In this paper, we mainly focus on the second category MVAR task with RGB and depth views. RGB-D action recognition is one of the most popular and important MVAR research directions due to the increasing usage of depth sensor and many attractive applications Horaud et al. [2016], Keselman et al. [2017], Zhang...
Figure 1: Illustration of multi-view difference in attention space. RGB view easily captures the visible changes shown in images; while the depth view is more sensitive to the changes of distance in depth dimension. We take a "kicking" action as example. In RGB view, the visible changes are obvious during the lifting leg and drawing back the leg. These frames are easily captured by RGB view as patterns. On the other hand, in the middle of example sequence, the changes of RGB view are tiny and hard to be discovered; while the depth view illustrates significant changes during this period (when the leg is at the top position, the changes are mainly in depth dimension).

Its main challenge is how to efficiently represent view-specific information and employ them for better multi-view performance. Subspace learning is a widely used strategy to derive a common subspace Jia and Fu [2016], Jia et al. [2014]. It aims to find consistent characteristics among multi-view to derive effective representations for recognition. However, only focusing on the synchronous patterns ignores the distinctive information of each single view. Fusion mechanism is another popular way for multi-view learning Wang et al. [2019b], Zadeh et al. [2018], Wang et al. [2019a]. Effective fusion takes advantage of distinctive information of each view and combines them appropriately for higher performance, while trivial fusions (e.g. average, concatenation, and summation) may hurt the final performance. Late fusion algorithms fully explore distinctive feature from each view individually and focus on fusion in label space Wang et al. [2019a], Bruno and Marchand-Maillet [2009]. The usage of mutual-support information is reflected via wisely fusing the predicted scores. On the other hand, early fusion methods pay more attention on augmenting the capacity of each view by borrowing information from the other view in Zadeh et al. [2018], Liang et al. [2018]. It enhances the representation learning process for each view at the same time in feature space. However, both late and early fusion strategies benefit multi-view learning only by exploiting the readily available mutual-support information directly. For example, the late fusion uses the predicted scores obtained from view-specific classifier; early fusion mutually borrows information for each other. These information have already been there. The information, which provides help, has limitation itself and sets the boundary of potential learning capacity. Hence, existing fusion methods cannot further discover valuable clues to raise the ceiling of model.

To improve the learning capacity from the root, we propose a collaborative attention mechanism (CAM) to discover additional clues which are obscure and difficult to explore. CAM fully excavates the latent information to improve the performance. Attention has been proved as an effective mechanism to boost learning accuracy in a variety of tasks. Meanwhile, it is also capable of interpreting model and provides many intuitions of data. According to the interpretability of attention mechanism, we start our work from the observations of multi-view action videos. Different views have different attention distributions (see Fig. 1). Specifically, RGB view pays attention to certain frames, while depth view values more contributions from some other frames. Based on the different observations in attention space, we propose Mutual-Aid RNN (MAR) to collaboratively guide pattern mining process for MVAR task (see Fig. 2). One view, according to the attention distribution as the instructive information, selectively direct the other view to attach importance to certain frames which are not easily noticed by itself. Overall, our contributions are summarized as belows:
Figure 2: The illustration of our whole framework. The view-specific attention is the first-stage. Two LSTM encoders make view-specific recognition, respectively. The attention score $z$ of two views are obtained and fed into the second stage. Mutual-Aid RNN, as the second stage, achieves the collaborative attention mechanism. Mutual-Aid block, $MA$, collaborate the two views frame-by-frame. We deploy attention module for each view and make a fusion for multi-view results.

- We propose a collaborative attention mechanism (CAM) framework to improve the MVAR learning performance. It efficiently utilizes the attention information to mutually enhance multi-view learning process, which boosts the single-view and multi-view learning performance simultaneously.

- We propose a Mutual-Aid RNN (MAR) structure to enhance the multi-view sequential learning. It relies on the attention information to capture the latent patterns and adaptively enhances the frame-level representations for each single view. The multi-view performance is also boosted.

- To the best of our knowledge, we open up a novel perspective to reacquaint multi-view learning by fully utilizing the interpretability of attention mechanism to guide learning process. Extensive experiments on three action datasets illustrate the effectiveness of our proposed model.

We organize the rest of this paper as follows: we first introduce related works in section 2; next, we propose our collaborative attention mechanism (CAM) framework in section 3; then we show the experimental results to illustrate the effectiveness of our model in section 4; finally, we conclude the whole paper in section 5.

2 Related Work

Numerous multi-view learning algorithms have been proposed for MVAR task Junejo et al. [2008], Wang et al. [2014, 2019b]. Attention mechanism is popular in modeling sequential data, especially for human action videos. In this section, we mainly focus on introducing multi-view learning algorithms and a variety of attention based methods.

2.1 Multi-View Learning

Multi-view learning technique has been adopted in multiple applications (e.g., image classification, emotion recognition, and face detection) Jones and Viola [2003], Farfade et al. [2015], Zhang et al. [2014], Nie et al. [2017a], Bai et al. [2019], Liu et al. [2015]. Action recognition task is also expressively developed by multi-view learning Cai et al. [2014], Ahmad and Lee [2006], Shao et al. [2016]. Wang et al. Wang et al. [2019b] proposed a generative model for MVAR task which expands model to handle missing view scenario. Wang et al. designed generative feature fusion strategy for human action recognition Wang et al. [2018]. Cai et al. fused the action feature descriptors fusion for recognition using a multi-view super vector Cai et al. [2014]. However, these method only pay attention to making use of complementary information directly without discovering additional latent knowledge.

2.2 Attention Mechanism

The pioneering work Mnih et al. [2014] introduced the attention mechanism into computer vision community to make image classification using recurrent model. The natural language processing (NLP) field was
significantly explored by firstly employing attention on machine language translation task Bahdanau et al. [2014]. After that, Attention based methods were widely adopted into many other applications (e.g., image caption Xu et al. [2015], text classification Yang et al. [2016], and image genreation Gregor et al. [2015]). Benefit from the inherent advantage of attention for modeling sequential data, plenty works were proposed for action recognition. Sharma et al. Sharma et al. [2015] proposed a visual attention framework for action recognition with a soft attention-based model. Song et al. proposed a spatial-temporal attention model based on skeleton data for recognizing human action Song et al. [2017]. However, these works only use attention to boost performance instead of taking advantage of its instructive function of interpretability to achieve further improvement.

3 Methodology

We first introduce the multi-view problem setting and start from the temporal attention model for action recognition. Then, we describe the whole collaborative attention framework with our proposed MAR model. Specifically, our model follows a two-stage training strategy. The first stage obtains the view-specific attention-distribution among temporal data. The second stage benefits from the attention-distribution to achieve a multi-view collaboration. It mutually enhances the view-specific representations. Finally, an effective correlative late fusion is utilized to boost learning performance by fully exploring multi-view label correlations.

3.1 Preliminary

Let $X_{t\text{r}}^1$ and $X_{t\text{r}}^2$ are multi-view training data, where $X_{t\text{r}}^1 \in \mathbb{R}^{n_{tr} \times T \times d^1}$ and $X_{t\text{r}}^2 \in \mathbb{R}^{n_{tr} \times T \times d^2}$ are the feature matrices of two views. $n_{tr}$ is the number of training samples. $T$ represents the number of training samples and length of action clip. $d^1$ and $d^2$ are feature dimensions of two views. $Y_{t\text{r}} \in \mathbb{R}^{n_{tr} \times d^l}$ is the one-hot label matrix, where $d^l$ is the dimension of label space. Correspondingly, $X_{t\text{e}}^1 \in \mathbb{R}^{n_{te} \times T \times d^1}$, $X_{t\text{e}}^2 \in \mathbb{R}^{n_{te} \times T \times d^2}$ and $Y_{t\text{e}} \in \mathbb{R}^{n_{te} \times d^l}$ are the test features of two views and label matrix, respectively. We train our model on training set, including the training and validation steps, by leveraging the mutual-support information from different views and evaluate model on test set. We refer $X$ and $Y$ as input features and labels of $\forall i$-th sample by default for training and test processes.

Before introducing each model component, we briefly summarize the whole logic of our proposed framework. The first phase contains multiple view-specific encoders. We use LSTM network with self-attention to encode the sequence input and obtain the attention information. In the second phase, the multi-view collaboration module utilizes the view-specific attention to guide the other view explore more latent but useful patterns which is hard to be found by itself. The view-specific representation is enhanced to achieve higher single view performance. The correlative late fusion is deployed to obtain final multi-view result which is also boosted by the enhanced single-view representations.

3.2 Temporal Attention for Action Recognition

Given an action video and the corresponding label, the temporal attention model aims to optimize the following objective:

$$\theta^* = \arg\max_\theta \sum_{(X, y)} \log p(y|X; \theta),$$

where $\theta$ is the set of parameters of model. $X = \{x_1, ..., x_t\}$ is the multiple frames of a video sample, and $y$ is the corresponding label. The dynamic information is the key factor for recognition. Thus, wisely choosing temporal encoder is decisive for temporal feature extraction. In our work, we deploy recurrent neural network (RNN) as the sequential modeling method. Specifically, we adopt long short-term memory (LSTM) Hochreiter and Schmidhuber [1997], a widely used variant of the vanilla RNN. Each input frame $x_t$ is encoded as a hidden representation $h_t$, and the cell state $c_t$ is updated correspondingly. Specifically,
the update processes in each LSTM block are:

\[
\begin{align*}
    f_t &= \sigma_g(W_f x_t + U_f h_{t-1} + b_f), \\
    i_t &= \sigma_g(W_i x_t + U_i h_{t-1} + b_i), \\
    o_t &= \sigma_g(W_o x_t + U_o h_{t-1} + b_o), \\
    c_t &= f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c h_t + U_c h_{t-1} + b_c), \\
    h_t &= o_t \circ \sigma_h(c_t),
\end{align*}
\]

where \( f_t, i_t, o_t, c_t, \) and \( h_t \) represent forget gate, input gate, output gate, cell state and hidden state at current time \( t \). \( c_{t-1} \) and \( h_{t-1} \) are cell and hidden states at last time \( (t-1) \). \( \sigma_g, \sigma_c, \) and \( \sigma_h \) are activation functions. \( \circ \) represents the element-wise product. \( W, U \) and \( b \) are corresponding parameters. To introduce our proposed mechanism clearly, we go deeper here to discover more insights of LSTM block. The key factor is the \( c_t \). It reflects the memory states of the whole sequence. \( f_t \) and \( i_t \) update the \( c_t \) internally through the forget and input procedures. The contents of forget and input are derived from current input \( x_t \) and last hidden state \( h_{t-1} \). The content of current hidden state \( h_t \) is also extracted from \( x_t \) and \( h_{t-1} \), then filtered by \( c_t \). All information flows cross several control gates center on the \( c_t \). \( c_t \) only records the temporal dynamic characteristic instead of specific domain knowledge. Own to this insight, we conclude that fully exploiting the cell state is decisive for temporal encoding.

Original video sequence \( X \) is encoded as \( H = \{ h_1, ..., h_T \} \). The common way to choose the representation is picking the last hidden state \( h_T \) which contains information about all hidden sequence. However, it may lose information in the middle of video to some degree. A reasonable way is using the weighted summation of \( h_t \). The weights are calculated based on the importance of each frame by attention mechanism. Concretely, we adopt a self-attention variant Yang et al. [2016] which is proposed for document classification. It fits our task well and could be easily utilized for modeling temporal data. We formulate our attention network as follows:

\[
\begin{align*}
    u_t &= \tanh(W_u h_t + b_u), \\
    z_t &= \frac{\exp(u_t^T u_w)}{\sum_i \exp(u_i^T u_w)}, \\
    r &= \sum_t z_t h_t,
\end{align*}
\]

where \( u_t \) denotes the attention vector derived from \( h_t \). \( W_u \) and \( b_u \) are learnable parameters. \( u_w \) is the context vector, which is random initialized and updated through the optimization procedure. It depicts the global meaning of the video sequence itself. \( z_t \) means the degree of importance for each \( u_t \) among the whole video context \( u_w \) by using softmax activation. \( r \) is the weighted summation of \( h_t \). In the following section, we introduce our framework step-by-step based on our basic temporal attention.

## 3.3 View-Specific Attention Mechanism

Multi-view video data contain mutual-support information for each other, while each view has its own distinctive patterns. To fully exploit the distinctive information from each view, we propose the view-specific attention mechanism formulated as follows:

\[
\begin{align*}
    H^v &= E^v(X^v, \phi_E^v), \\
    r^v &= Q^v(H^v, \phi_Q^v), \\
    \hat{Y}_a^v &= C^v(r^v, \phi_C^v), \\
    L_a^v &= \ell(Y, \hat{Y}_a^v),
\end{align*}
\]

where superscript \( v \) represents the RGB or depth view. \( E \) is LSTM model (Eq. 2), encoding video sequence \( X \) into hidden sequence \( H \). \( Q \) is the attention model (Eq. 3), transferring \( H \) into weighted summation vector \( r \). \( C \) is the view-specific classifier implemented by linear mapping, resulting in the predicted label \( \hat{Y}_a \). \( \phi_E, \phi_Q, \) and \( \phi_C \) are learnable parameters. They are optimized by minimizing following objective:

\[
    L_a^v = \ell(Y, \hat{Y}_a^v),
\]

where \( \ell \) represents cross-entropy loss, \( Y \) is the ground truth.
Figure 3: The illustration of Mutual-Aid block. $x_t$ and $h_{t-1}$ of two views are set as input to integrate information from cross view. They update the $c_t$ of two views via point multiplication, respectively. Attention distributions $z_t$ of two views are involved in weighted summation to adaptively collaborate the multi-view knowledge.

The goal of view-specific attention aims to derive the $Z^v = \{z^v_1, ..., z^v_T\}$ which is the intermediate product of Eq. 3. They preserve the view-specific dynamic patterns. The learning process supervised by $Y$ is conducted for obtaining better attention distribution. We regard the view-specific attention as our first-stage model. Multi-view attention distribution $Z^v$ is reserved for the second-stage model.

3.4 Multi-View Collaboration

To substantially take advantage of multi-view data, in our second stage, we propose the multi-view collaboration mechanism. It mutually supports the multi-view learning. Please note that, unlike some data augmentation strategies cross multi-view (e.g., representation mapping, feature fusion, and conditional GAN), our goal is borrowing the knowledge from other view to help the target view discovering more clues by itself, instead of transferring information directly. To convey our insight clearly, we briefly discuss it here. View-specific attention mechanism provides the distinctive temporal pattern by attention-distribution. It is extracted through optimizing single view classifier individually and reflect the view-specific characteristics. Particularly, for the same video sample, the attention distribution of on view focuses on certain frames, while that of the other view focuses on different frames. This is caused by the inherent attribute of each view. Specifically, the RGB view always takes advantages of color changes to recognize human action, while the depth view focuses more on the distance changes. As a result, the effective frames for two views could be different. However, the differences are not opposite but complementary for each other. Some frames are ignored by certain view, due to its inherent attribute, still restore valuable information. This information may be easily discovered by the other view. To this end, we propose the multi-view collaboration mechanism. It encourages multi-view data to help with each other by guiding other view to focus on implicit but effective information.

We first encode the multi-view video sequence $X^v$ with LSTM (Eq. 2). We formulate it briefly as follows:

$$
\begin{align*}
    c_t &= f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c h_t + U_c h_{t-1} + b_c), \\
    h_t &= o_t \circ \sigma_h(c_t),
\end{align*}
$$

where $c_t$ and $h_t$ are the cell state and hidden state for time $t$, deriving information from $c_{t-1}$, $h_{t-1}$, and $x_t$. We expand the LSTM as Mutual-Aid RNN (MAR) by designing a novel recurrent block (see Fig. 3). Instead of setting the $c_t$ as the cell state for next time step directly, MAR leverages the knowledge from the other view to update the $c_t$ of target view, and then guides the target view for information mining. We formulate our proposed MAR step-by-step.
Cross-View Collaborator. It is proposed to integrated multi-view information and formulated as follow:
\[
G_{r \rightarrow d} = \sigma(W_{rd}x_t^r + W_{d}h_{t-1}^d),
\]
\[
G_{d \rightarrow r} = \sigma(W_{dr}x_t^d + W_{r}h_{t-1}^r),
\]
where $W$ is the learnable parameters. $G$ extract information from current input $x_t$ and collaborate with last hidden state $h_{t-1}$ from the other view with sigmoid activation. The frame-level knowledge from the other view is reserved in $G$.

Mutual Filtering. Cell state $c_t$ has already been updated internally in LSTM (Eq. 6). It only contains the memory information from single view. We update the $c_t$ using cross-view collaborator to derive knowledge from the other view which is given by
\[
c_t' = G_{d \rightarrow r} \circ c_t^r,
\]
\[
c_t'' = G_{r \rightarrow d} \circ c_t^d,
\]
where $\circ$ is the point-wise product. $c_t'$ is the updated cell state containing mutual-support information from the other view.

Mutual Collaboration. Attention distributions $z_t$ reflect the importance of each frame for multi-view. It also decides the importance of information which is used for updating $c_t$. We first normalize the attention score by
\[
z_t' = \frac{z_t'}{z_t' + z_t^d},
\]
\[
z_t'' = \frac{z_t''}{z_t'' + z_t^r}.
\]
The original cell state $c_t$ is updated by single view information, while $c_t'$ is updated by the cross-view collaborator $G$. $z_t'$ represents the dynamic knowledge from different views. We use it to integrate the multi-view information for updating cell state via the weighted summation:
\[
c_t''' = z_t''c_t'' + z_t'd_t,c_t',
\]
\[
c_t''' = z_t'c_t' + z_t''d_t,c_t'',
\]
where $c_t''$ is the final cell state containing the dynamic knowledge from multi-view data. Through being the input for next time step, it brings the knowledge from the other view to overcome the inherent drawback of specific view. Some implicit information could be discovered by each single view via the guidance from mutual collaboration.

So far, we introduce the proposed multi-view collaboration via the MAR encoder. As an integrated model, the input and output are multi-view sample data and representation sequences, respectively. In order to fully utilize the discovered information via our collaboration mechanism, we reuse the self-attention (Eq. 3) to obtain the final representation and make the view-specific recognition again similar to Eq. 4. We briefly formulate these steps by
\[
H_M^v = E_M^v(X^v, \phi_{E_M}^v),
\]
\[
r_M^v = Q_M^v(H_M^v, \phi_{Q_M}^v),
\]
\[
\hat{Y}_M^v = C_M^v(r_M^v, \phi_{C_M}^v),
\]
where all the terms with subscript $M$ represents the similar meanings for multi-view collaboration compared with Eq. 4. We obtain another attention distribution $Z_M^v$ and the predicted label $\hat{Y}_M^v$ for two views. The learnable parameters are optimized by minimizing following loss:
\[
L_M^v = \ell(Y, \hat{Y}_M^v),
\]
where $v$ denotes RGB or depth view. $\ell$ represents cross-entropy loss, and $Y$ is the ground truth. The view-specific attention (first-stage) and the multi-view collaboration (second-stage) form our whole framework Collaborative Attention Mechanism (CAM). It exploits the knowledge from multi-view attention distributions to guide the multi-view information discovering and enhance the learning process. After obtaining the $\hat{Y}_M^v$ from two views, we use an effective correlative late fusion to evaluate multi-view recognition results.
3.5 Correlative Late Fusion

Our CAM discovers more clues for single view representation and boost its recognition accuracy. Efficient late fusion strategy benefits from it to further obtain multi-view results. We deploy a concise yet effective late fusion model via building a correlative graph structure Wang et al. [2019a]. It explores the label correlation among two views. We formulate the late fusion strategy as follows:

\[ D = \hat{Y}_M \cdot \hat{Y}_M^T, \]

where \( \hat{Y}_M \in \mathbb{R}^{d \times 1} \) and \( \hat{Y}_M^T \in \mathbb{R}^{1 \times d} \) are the predicted label from two views. \( D \in \mathbb{R}^{d \times d} \) is the correlative matrix constructed by the multiplication of predicted labels from two views. It extracts pair-wise label correlations of multi-view data. \( D \) is flatten into a \( d \times d \) dimension vector and set as input of the final classifier \( C_f : \mathbb{R}^{d \times d} \rightarrow \mathbb{R}^{d} \). \( C_f \) is parameterized by \( \phi_{C_f} \) and updated by minimizing following loss:

\[ L_f = \ell(Y, C_f(D, \phi_{C_f})), \]

where \( Y \) is the ground truth, \( \ell \) is the cross-entropy loss. \( L_f \) represents the final multi-view loss. Our model mainly consists of the view-specific attention and the multi-view collaboration, followed by a late fusion model for multi-view learning performance. As a summary, view-specific attention aims to capture the differences between two views, especially focusing on the attention distribution. These differences are set as guidance information for multi-view collaboration. A novel MAR structure is proposed for extracting cross-view knowledge and updating memory cell effectively. More implicit yet valuable information could be discovered. A concise late fusion is deployed for multi-view performance.

4 Experiments

We make extensive experiments on three challenging multi-view human action datasets to evaluate our proposed framework. Recognition performances for both single-view and multi-view demonstrate the effectiveness of our model. Detailed ablation study proves the necessity of each model component.

4.1 Multi-View Action Datasets

We use three multi-view human action dataset. EV-Action dataset is a new and large-scale dataset. The other two, UWA3DII and DHA, are relatively small. We test our model on both large and small datasets for a comprehensive evaluation.

- **EV-Action Dataset** Wang et al. [2019c] is a large-scale multi-view human action dataset. It contains RGB, depth, skeleton, and electromyography (EMG) views. RGB and depth are used in our multi-view action recognition experiments. Ev-Action contains 20 human common actions. 10 actions are finished by single subject like walking, sitting, jumping, etc, and the other 10 actions finished by the same subjects interacting with other objects like moving table, drinking, reading book, etc. It includes 53 subjects (the first 53 out of 70 subjects in original EV-Action dataset) performing each action 5 times so that we have 5300 samples in total. Each subject performs 100 action clips for 20 classes. We choose the action clips collected from first 40 (10 out of 40 for vallidation) subjects as training set and the rest 13 subjects as test set.

- **UWA3D Multiview Activity II Database (UWA3DII)** Rahmani et al. [2016, 2014] contains 30 human actions performed by 10 subjects. There are RGB, depth, and skeleton views recorded from 4 different viewpoints. We use the RGB and depth recorded from front view for evaluation. There are totally 270 samples available and we randomly choose 150 for training (50 for validation) and 120 for test.

- **Depth-included Human Action Dataset (DHA)** Lin et al. [2012] is a multi-view dataset containing RGB and depth views. It contains 23 actions performed by 21 subjects. There are 483 video samples in total. We randomly choose 240 for training (80 for validation) and the rest 243 for test.
Table 1: Recognition Performance on EV-Action Dataset

| Method       | RGB    | Depth  | RGB-D  |
|--------------|--------|--------|--------|
| RC Classifier| 0.5992 | 0.5790 | 0.6213 |
| MFN          | 0.5743 | 0.4082 | 0.6423 |
| MLSTM-FCN    | 0.6804 | 0.6926 | 0.7014 |
| GMVAR        | 0.5902 | 0.5661 | 0.6653 |
| Ours         | 0.7022 | 0.7123 | 0.7359 |

4.2 Baseline Methods

For EV-Action dataset, we choose several recently proposed methods for comparison. They include advanced temporal modeling algorithms and the state-of-the-art multi-view learning methods. For UWA3DII and DHA datasets, we choose some dataset specific baselines for them.

- **MLSTM-FCN** Karim et al. [2019] is a novel deep framework proposed for handling multivariate temporal data. It contains a two-pathway structure (CNN and LSTM) to encode temporal data. Fully exploited patterns are captured for classification.

- **RC framework** Bianchi et al. [2018] proposes a reservoir computing (RC) approach to model temporal data as vectorial representations in an unsupervised fashion.

- **MFN** Zadeh et al. [2018] designs a memory fusion mechanism for multi-view learning based on temporal data. It proposes an early fusion strategy to integrate multi-view information in feature space.

- **TSN** Wang et al. [2016] is an effective benchmark model for action recognition. It utilizes a two-stream structure to collect more valuable features from RGB and flow modality. TSN proposes a efficient sampling method to extract video information.

- **AMGL** Nie et al. [2016] is a novel multi-view classification method based on graph learning. It aims to optimize weights for each graph automatically in a parameter-free fashion.

- **MLAN** Nie et al. [2017b] proposes an adaptive graph-based algorithm. It achieves the local structure and semi-supervised learning at the same time for multi-view learning.

- **GMVAR** Wang et al. [2019b] utilizes the generative strategy to mutually augment the multi-view representations. It boosts the multi-view learning performance significantly and improves the model robustness simultaneously.

We use the first three baselines for EV-Action dataset, and the next four baselines for UWA3DII and DHA datasets. The GMVAR, as a state-of-the-art method for multi-view action recognition, is used for evaluation on all three datasets.

4.3 Implementation

We use the same strategy to preprocess the raw videos for three datasets. We use TSN Wang et al. [2016] to extract frame-level features for RGB view using the BNInception network as backbone. Each RGB frame is extracted into 1024 dimension feature vector. The depth view is transferred into RGB format first using HHA encoding algorithm Gupta et al. [2014]. Then, we use the exactly the same TSN framework to extract depth features. We arrange the length of video with a unified number for each dataset via the cutting and repeating strategies. Specifically, for longer video, we pick the first certain frames and cut the rest video off; while for shorter video, we repeat the whole video sequence several times until it reaches the target number. The lengths of the sequence we set are 60, 40, and 60 for EC-Action, DHA, and UWA3DII, respectively. Since the EV-Action is a relatively large-scale dataset, we extract features using TSN without fine-tune. The other two are smaller, we fine-tune the TSN to extract features for higher recognition accuracy. As a summary, each video sample is preprocessed into a $60 \times 1024$, $40 \times 1024$, and $60 \times 1024$ feature matrix for
Table 2: Recognition Performance on UWA/DHA Dataset

| Dataset  | DHA       | UWA3DII   |
|----------|-----------|-----------|
| Method   | RGB       | Depth     | RGB-D     | RGB       | Depth     | RGB-D     |
| TSN      | 0.6785    | 0.8324    | -         | 0.4833    | 0.5936    | -         |
| AMGL     | 0.6461    | 0.7284    | 0.7489    | 0.3067    | 0.3667    | 0.3933    |
| MLAN     | 0.6791    | 0.7296    | 0.7613    | 0.2933    | 0.2867    | 0.3800    |
| GMVAR    | 0.6972    | 0.8348    | 0.8872    | 0.4917    | 0.5846    | 0.6035    |
| Ours     | **0.7407**| **0.8642**| **0.8724**| **0.5083**| **0.6073**| **0.6314**|

EV-Action, DHA, and UWA3DII, respectively. We concatenate the multi-view data as multivariate time series to implement the MLSTM-FCN and RC classifier baselines. The MFN and GMVAR are designed for multi-view learning which fit our input data appropriately. TSN can be conducted for RGB and depth (transferred into RGB) individually without multi-view learning scenario. We adopt the AMGL and MLAN to fit our multi-view learning scenario.

As shown on Fig. 2, the view-specific attention is first trained individually. The input is multi-view sequence data. The attention distributions $Z^v$ are derived through optimizing Eq. 5 during first-stage model. The same input sequence is set as input for the multi-view collaboration (second-stage). The MAR model is conducted with the additional input $Z^v$. Single view results from MAR model are fed into the final fusion model for multi-view performance. We set 128 batch size for EV-Action, and 32 for the other two datasets. The hidden dimensions for both temporal encoders (first and second stages) and attention are 128. The learning rates are 0.0005 and 0.001 for first-stage and second-stage. Our model is implemented by PyTorch with GPU acceleration.

4.4 Performance Analysis

The recognition performance of EV-Action is shown in Tab. 1. Our method outperforms all other baselines on both single-view and multi-view settings. MLSTM-FCN is an effective model for single view temporal data. It achieves competitive results. However, the fusion result is lower than ours. More importantly, our single-view performances are also higher which demonstrates our MAR model works well for discovering more valuable information. MFN and GMVAR are multi-view learning methods. However, they may suffer from the complicated high-dimensional temporal data which hinders the early fusion process and increases the difficulty of training generative model.

In Tab. 2, we report the performances on DHA and UWA3DII datasets. GMVAR is the state-of-the-art method for DHA. It achieves high accuracy on single-view and boosts the multi-view performance significantly. Our results cannot outperform GMVAR in multi-view scenario. However, we obtain the comparable result. Moreover, for single-view scenario, we make the substantial improvement which illustrates the effectiveness of our MAR encoding model. For UWA3DII dataset, our model always achieves the best accuracy. The TSN and GMVAR are efficient. We improve the RGB in relatively small scale, while obtain more improvement on depth and RGB-D scenarios. Overall, our model performs well on three datasets for both single-view and multi-view scenarios. The evaluation results prove the effectiveness of our proposed framework.

4.5 Ablation Study

We provide detailed ablation study to prove the necessity of each model component. We use the EV-Action dataset and results are shown in Tab. 3. We use single-view data as input in LSTM model and report RGB and Depth, respectively. Feature extracted from TSN lays a good foundation for our temporal modeling process. No Collaboration denotes the we train the multi-view data synchronous and add late fusion without any collaborative procedures. The RGB and Depth views show almost the same accuracy to LSTM with sight improvement. It denotes that the multi-view training makes effects on the single-view encoding to some degree. Late fusion boosts the RGB-D performance. RGB Collaboration only utilizes the multi-view
Table 3: Ablation Study

| Method               | RGB    | Depth  | RGB-D  |
|----------------------|--------|--------|--------|
| LSTM                 | 0.6878 | 0.6772 | -      |
| No Collaboration     | 0.6894 | 0.6796 | 0.7154 |
| RGB Collaboration    | 0.6978 | 0.6802 | 0.7285 |
| Depth Collaboration  | 0.6874 | 0.7084 | 0.7255 |
| Ours                 | 0.7022 | 0.7123 | 0.7359 |

Figure 4: Visualization results of the changes of attention distribution for two views. X-axis is the time dimension. y-axis is the index of selected samples.

collaboration mechanism to update the cell state of RGB. It show more improvement on RGB and depth only obtains limited enhancement. RGB-D obtains higher accuracy. RGB benefits from the collaboration process to discover more information. Similarly, Depth Collaboration shows the better results on depth, while RGB is relatively lower. Our complete model deploys the multi-view collaboration achieving high performance on each single view. RGB-D leverages the enhanced single-view representations and obtains the best multi-view performance.

4.6 Attention Visualization

To substantially provide the intuition about our model and its insight, we visualize the $Z^v$ and $Z^v_M$, which are the attention distributions before and after our multi-view collaboration model, respectively. Based on the changes among them, we expect to provide more clues for better understanding to our model. All the visualization results are picked out from EV-Action dataset.

First, we randomly choose 10 samples from test samples to compare the differences between $Z^v_M$ and $Z^v$ on two views (see Fig. 4). Comparing two columns, the attention scores are distributed more comprehensively, which indicates more information is collected for the final representation. For a specific view, some frames with low scores are gained attention after our multi-view collaboration procedure. Specifically, we visualize one of the selected sample with its raw RGB and depth frames to illustrate the changes with more details. We pick a sample from Moving Table action to illustrate our insights (see Fig. 5). For both two views, the green box indicates the frames have been noticed by single view itself. The red box denotes the frames gained attention after our collaboration model, which is hard to discovered by single view itself. Through the frames shown on figure, we can notice the new-found frames also provide valuable clues for action recognition.
Figure 5: The colorbars represent the attention distributions. The frames and the fragment in the same color box indicate the temporal index and its corresponding frames.

5 Conclusions

In this paper, we propose a Collaborative Attention Mechanism (CAM) for multi-view action recognition (MVAR) task. A view-specific attention is utilized for capturing multi-view attention distributions in first-stage. The multi-view collaboration is achieved via the novel Mutual-Aid RNN (MAR). By this way, each view is guided by knowledge from the other view and enhanced to discover more latent information. Our model provides a novel perspective to utilize the attention mechanism. The interpretability of attention is appropriately exploited to guide the learning process. Due to the collaboration strategy, our model outperforms other state-of-the-art methods on three action datasets on both single and multi-view scenarios. Evaluation results and ablation study prove the effectiveness of our proposed framework and the necessity of each model component.

References

Mohiuddin Ahmad and Seong-Whan Lee. Hmm-based human action recognition using multiview image sequences. In 18th International Conference on Pattern Recognition (ICPR’06), volume 1, pages 263–266. IEEE, 2006.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.

Yue Bai, Lichen Wang, Zhiqiang Tao, Sheng Li, and Yun Fu. Multi-view time series classification via global-local correlative channel-aware fusion mechanism. arXiv preprint arXiv:1911.11561, 2019.

Filippo Maria Bianchi, Simone Scardapane, Sigurd Løkse, and Robert Jenssen. Reservoir computing approaches for representation and classification of multivariate time series. arXiv preprint arXiv:1803.07870, 2018.

Eric Bruno and Stephane Marchand-Maillet. Multiview clustering: a late fusion approach using latent models. In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pages 736–737, 2009.

Nan Bu, Masaru Okamoto, and Toshio Tsuji. A hybrid motion classification approach for emg-based human–robot interfaces using bayesian and neural networks. IEEE Transactions on Robotics, 25(3):502–511, 2009.

Zhuowei Cai, Linlin Wang, Xiaojiang Peng, and Yu Qiao. Multi-view super vector for action recognition. In Proceedings of CVPR, pages 596–603, 2014.
Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In 2015 IEEE International conference on image processing (ICIP), pages 168–172. IEEE, 2015.

Zhongwei Cheng, Lei Qin, Yituo Ye, Qingming Huang, and Qi Tian. Human daily action analysis with multi-view and color-depth data. In Proceedings of ECCV, pages 52–61. Springer, 2012.

Fernando De la Torre, Jessica Hodgins, Adam Bargteil, Xavier Martin, Justin Macey, Alex Collado, and Pep Beltran. Guide to the carnegie mellon university multimodal activity (cmu-mmac) database. Carnegie Mellon University, 2009.

Sachin Sudhakar Farfade, Mohammad J Saberian, and Li-Jia Li. Multi-view face detection using deep convolutional neural networks. In Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, pages 643–650, 2015.

Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. arXiv preprint arXiv:1502.04623, 2015.

Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from rgb-d images for object detection and segmentation. In Proceedings of ECCV, pages 345–360. Springer, 2014.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.

Michael B Holte, Thomas B Moeslund, Nikos Nikolaidis, and Ioannis Pitas. 3d human action recognition for multi-view camera systems. In 2011 International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission, pages 342–349. IEEE, 2011.

Radu Horaud, Miles Hansard, Georgios Evangelidis, and Clément Ménier. An overview of depth cameras and range scanners based on time-of-flight technologies. Machine vision and applications, 27(7):1005–1020, 2016.

Xiaofei Ji, Ce Wang, and Yibo Li. A view-invariant action recognition based on multi-view space hidden markov models. International Journal of Humanoid Robotics, 11(01):1450011, 2014.

Chengcheng Jia and Yun Fu. Low-rank tensor subspace learning for rgb-d action recognition. IEEE Transactions on Image Processing, 25(10):4641–4652, 2016.

Chengcheng Jia, Yu Kong, Zhengming Ding, and Yun Raymond Fu. Latent tensor transfer learning for rgb-d action recognition. In Proceedings of the ACM international conference on Multimedia, pages 87–96, 2014.

Michael Jones and Paul Viola. Fast multi-view face detection. Mitsubishi Electric Research Lab TR-20003-96, 3(14):2, 2003.

Imran N Junejo, Emilie Dexter, Ivan Laptev, and Patrick PÚrez. Cross-view action recognition from temporal self-similarities. In Proceedings of ECCV, pages 293–306. Springer, 2008.

Fazle Karim, Somshubra Majumdar, Houshang Darabi, and Samuel Harford. Multivariate lstm-fcns for time series classification. Neural Networks, 116:237–245, 2019.

Leonid Keselman, John Iselin Woodfill, Anders Grunnet-Jepsen, and Achintya Bhowmik. Intel realsense stereoscopic depth cameras. In Proceedings of CVPR, pages 1–10, 2017.

Paul Pu Liang, Ziyin Liu, Amir Zadeh, and Louis-Philippe Morency. Multimodal language analysis with recurrent multistage fusion. arXiv preprint arXiv:1808.03920, 2018.

Yan-Ching Lin, Min-Chun Hu, Wen-Huang Cheng, Yung-Huan Hsieh, and Hong-Ming Chen. Human action recognition and retrieval using sole depth information. In Proceedings of the ACM international conference on Multimedia, pages 1053–1056, 2012.
Meng Liu, Yong Luo, Dacheng Tao, Chao Xu, and Yonggang Wen. Low-rank multi-view learning in matrix completion for multi-label image classification. In Proceedings of AAAI, 2015.

Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In Advances in neural information processing systems, pages 2204–2212, 2014.

Feiping Nie, Jing Li, Xuelong Li, et al. Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification. In Proceedings of IJCAI, pages 1881–1887, 2016.

Feiping Nie, Guohao Cai, Jing Li, and Xuelong Li. Auto-weighted multi-view learning for image clustering and semi-supervised classification. IEEE Transactions on Image Processing, 27(3):1501–1511, 2017a.

Feiping Nie, Guohao Cai, and Xuelong Li. Multi-view clustering and semi-supervised classification with adaptive neighbours. In Proceedings of AAAI, 2017b.

Diana Pagliari and Livio Pinto. Calibration of kinect for xbox one and comparison between the two generations of microsoft sensors. Sensors, 15(11):27569–27589, 2015.

Hossein Rahmani, Arif Mahmood, Du Q Huynh, and Ajmal Mian. Hope: Histogram of oriented principal components of 3d pointclouds for action recognition. In Proceedings of ECCV, pages 742–757. Springer, 2014.

Hossein Rahmani, Arif Mahmood, Du Huynh, and Ajmal Mian. Histogram of oriented principal components for cross-view action recognition. IEEE transactions on pattern analysis and machine intelligence, 38(12):2430–2443, 2016.

Ling Shao, Li Liu, and Mengyang Yu. Kernelized multiview projection for robust action recognition. International Journal of Computer Vision, 118(2):115–129, 2016.

Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov. Action recognition using visual attention. arXiv preprint arXiv:1511.04119, 2015.

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 Proceedings of AAAI, 2017.

Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-view action modeling, learning and recognition. In Proceedings of CVPR, pages 2649–2656, 2014.

Lan Wang, Chenqiang Gao, Luyu Yang, Yue Zhao, Wangmeng Zuo, and Deyu Meng. Pm-gans: Discriminative representation learning for action recognition using partial-modalities. In Proceedings of ECCV, pages 384–401, 2018.

Lichen Wang, Zhengming Ding, Seungju Han, Jae-Joon Han, Changkyu Choi, and Yun Fu. Generative discovery correlation network for multi-label learning. In Proceedings of ICDM, pages 588–597. IEEE, 2019a.

Lichen Wang, Zhengming Ding, Zhiqiang Tao, Yunyu Liu, and Yun Fu. Generative multi-view human action recognition. In Proceedings of ICCV, pages 6212–6221, 2019b.

Lichen Wang, Bin Sun, Joseph Robinson, Taotao Jing, and Yun Fu. Ev-action: Electromyography-vision multi-modal action dataset. arXiv preprint arXiv:1904.12602, 2019c.

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 Proceedings of ECCV, pages 20–36. Springer, 2016.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of ICML, pages 2048–2057, 2015.
Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, pages 1480–1489, 2016.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In Proceedings of AAAI, 2018.

Zhengyou Zhang. Microsoft kinect sensor and its effect. IEEE multimedia, 19(2):4–10, 2012.

Zixing Zhang, Eduardo Coutinho, Jun Deng, and Björn Schuller. Cooperative learning and its application to emotion recognition from speech. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(1):115–126, 2014.