Multimodal Fusion via Teacher-Student Network for Indoor Action Recognition

Bruce X.B. Yu, Yan Liu,* Keith C.C. Chan
Department of Computing, The Hong Kong Polytechnic University
{csxbyu, csyliu}@comp.polyu.edu.hk, keithccchan@gmail.com

Abstract

Indoor action recognition plays an important role in modern society, such as intelligent healthcare in large mobile cabin hospitals. With the wide usage of depth sensors like Kinect, multimodal information including skeleton and RGB modalities brings a promising way to improve the performance. However, existing methods are either focusing on a single data modality or failed to take the advantage of multiple data modalities. In this paper, we propose a Teacher-Student Multimodal Fusion (TSMF) model[1] that fuses the skeleton and RGB modalities at the model level for indoor action recognition. In our TSMF, we utilize a teacher network to transfer the structural knowledge of the skeleton modality to a student network for the RGB modality. With extensive experiments on two benchmarking datasets: NTU RGB+D and PKU-MMD, results show that the proposed TSMF consistently performs better than state-of-the-art single modal and multimodal methods. It also indicates that our TSMF could not only improve the accuracy of the student network but also significantly improve the ensemble accuracy.

Introduction

Since the release of the depth sensor called Kinect, vision-based action recognition has been attracting increasing attention. Kinect can provide multiple data modalities like skeleton and RGB. Plenty of methods have been proposed to learn neural representations from different data modalities captured by the sensor. For example, neural representation approaches have been proposed to learn human actions from skeleton and RGB data (Wei et al. 2017; Baradel, Wolf, and Mille 2017, 2018). A key motivation of these multimodal approaches is to improve the action recognition accuracy by learning mutually independent or complementary features from different data modalities. However, prior work on action recognition in the Kinect sensor usually focus on handling a single data modality, which is either the skeleton modality (Yan, Xiong, and Lin 2018; Si et al. 2019) or the RGB modality (Carreira and Zisserman 2017; Xie et al. 2018; Tran et al. 2015).

Methods that make use of the skeleton modality for action recognition usually focus on learning spatial-temporal features so as to infer the particular type of an action (Yan, Xiong, and Lin 2018; Si et al. 2019). To understand the limitations of the current skeleton-based action recognition methods, we illustrate four challenging actions (i.e., "reading", "writing", "playing with phone" and "typing") from the NTU-RGB+D dataset (Shahroudy et al. 2016) in Figure 1. The four actions have similar movements in the skeleton modality, which makes them hard to be recognized by even more advanced graph-based models like (Shi et al. 2019b; Liu et al. 2020). The major limitation of these skeleton-based approaches is that they do not consider the appearance feature of the RGB modality that contains different object information. On the other hand, video-based methods are mainly developed to model representations from the optical flow and RGB video data modalities (Carreira and Zisserman 2017; Xie et al. 2018; Tran et al. 2015; Feichtenhofer, Pinz, and Zisserman 2016; Feichtenhofer et al. 2019). For these video-based approaches, they fail to consider the 3D structure information in the skeleton modality. Moreover, existing video-based models (Carreira and Zisserman 2017; Xie et al. 2018; Tran et al. 2015) seem to perform well for outdoor actions with different background scenes but not for indoor actions collected by Kinect. Specifically, (Choi et al. 2019) conducted experiments with human subjects masked from video data so that only information about the background scenes was retained. Despite the loss of human subject information, video-based methods in (Carreira and Zisserman 2017; Xie et al. 2018; Tran et al. 2015) still could

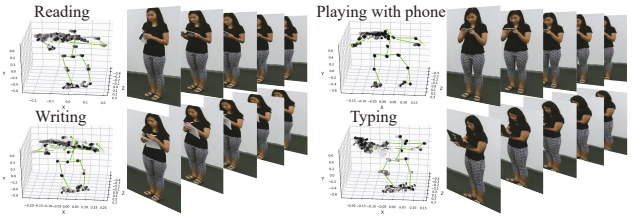

Figure 1: Four sample actions that are challenging for the skeleton modality in NTU-RGB+D (their left parts represent the skeleton modality, while right parts represent the RGB modality). Our motivation is to build effective representation for the RGB modality that will compensate the lack of appearance features in the skeleton modality.

*denotes the corresponding author.

[1]Code is available: https://github.com/bruceyo/TSMF

classify outdoor actions in datasets like UCF-101 (Soomro, Zamir, and Shah 2012) and Kinetics (Kay et al. 2017). This is due the fact that existing video-based methods actually capture features in the background scenes. However, for indoor actions with a consistent background as shown in Figure 1, such video-based methods work poorly as reported in (Tran et al. 2015) and (Luo et al. 2018), which indicates these methods are not well-suited to capture descriminative appearance features from objects like book, paper, mobile phone and laptop as shown in Figure 1.

Although there are some attempts to deal with multimodal action recognition, how effective representations can be learned from multiple data modalities so as to improve the recognition accuracy remains an open problem. For example, multimodal methods with object recognition was proposed in (Wei et al. 2017) and (Zhang et al. 2019). These methods require to perform objection detection on the whole video and handle the human-object relationship. Similarly, with the motivation of involving object appearance features, multimodal methods that put attention on the body areas around the two hands of human bodies are proposed in (Baradel et al. 2018; Baradel, Wolf, and Mille 2017, 2018). These attention-based methods seem to improve the accuracy of the whole model when it aggregates the results of the skeleton and RGB modalities. However, they neglect other body areas, e.g., feet and head, that also provide discriminative information for actions like "wearing a shoe", "taking off a hat", "shaking head" and "touching head".

To overcome the above limitations in existing multimodal methods, we propose a model-based multimodal fusion method called Teacher-Student Multimodal Fusion (TSMF) that includes two subnetworks (i.e., a teacher network and a student network). The student network borrows knowledge from the teacher network to build a fused representation of the RGB modality. In such a way, we construct an effective representation of the RGB modality that can complement the inadequacy of the skeleton modality. In our TSMF, we capture features in the RGB video that can represent both object and body movements. Comparing with video-based methods, this approach alleviates the problem of overfitting to the features in the background scenes. While compared with existing multimodal methods that focus on the features around hand areas (Baradel et al. 2018; Baradel, Wolf, and Mille 2017, 2018), our method also focuses on extra body areas including head and feet.

In our TSMF, modality-specific sub-models (i.e., the student and teacher networks) are utilized to learn representations from different data modalities of the Kinect sensor. We extensively evaluate our model on two large popular datasets: NTU RGB+D (Shahroudy et al. 2016) and PKU-MMD (Liu et al. 2017a). Our novel representation scheme not only improves modality-specific and ensemble accuracies but also consistently outperforms state-of-the-art single modal and multimodal methods. In the following, we describe our TSMF that successfully makes use of the mutual complementary information of different data modalities of Kinect for action recognition.

Related Work

Representation methods of action recognition with Kinect could be classified into three categories: skeleton, video, and multimodal. In this section, we introduce these representation methods that relate to our TSMF.

Skeleton Representation. Various representations that focus on the spatial and temporal features of the skeleton modality like recurrent neural network (Liu et al. 2017b), and graph convolutional network (Yan, Xiong, and Lin 2018; Li et al. 2019; Shi et al. 2019a) have been proposed. Another representation is to lean co-occurrence features referring to the interactions and combinations of some subsets of skeleton joints by using the cooccurrence learning method (Zhu et al. 2016). (Si et al. 2019) considered both graph and co-occurrence representation methods. Except modeling novel representations of skeleton data, data preprocessing or data cleaning methods that learn a model to reconstruct more accurate skeleton data have been proposed (Liu, Liu, and Chen 2017; Zhang et al. 2017). Although these techniques are successful in modeling features of the skeleton modality, they usually overfit to the training data as there is no larger action recognition dataset that has millions of training samples like ImageNet. Moreover, they neglect to represent the appearance features from the RGB modality.

Video Representation. Two modalities (i.e., RGB and optical flow) of video data are widely used to learn representations for action recognition. The optical flow data is usually extracted from the RGB modality by using the TV-L^1 algorithm (Zach, Pock, and Bischof 2007). To learn representations from the RGB and optical modalities, methods using 3D Convolutional Neural Networks (CNNs) like 3D CNN (C3D) (Tran et al. 2015), Inflated 3D CNN (I3D) (Carreira and Zisserman 2017), and separable 3D CNN (S3D) (Xie et al. 2018) are proposed for outdoor action recognition. As an advanced version of C3D (Tran et al. 2015), the 3D CNN in I3D (Carreira and Zisserman 2017) is based on 2D CNN which is a pre-trained Inception-V^1 (Ioffe and Szegedy 2015). While S3D (Xie et al. 2018) is a modified version of I3D (Carreira and Zisserman 2017) that utilized both 2D and 3D CNNs to further improve the accuracy and speed of such video-based methods. However, S3D (Xie et al. 2018) is trained on 56 GPUs with a batch size set to six per GPU, which reflects its huge computational cost. Moreover, these video-based methods could not perform well for indoor actions (Tran et al. 2015; Luo et al. 2018).

Multimodal Representation. Representations for multimodal learning could be categorized to joint and coordinated representations (Baltrušaitis, Ahuja, and Morency 2019). Joint representations are related to model-agnostic approaches that concatenate representations at the feature level or decision level (Wei et al. 2017; Wu et al. 2016; Pan et al. 2019), which are also known as early fusion and late fusion, respectively. Coordinated representations could focus on enforcing either similarity between modality-specific representations (Luo et al. 2018; Garcia, Morerio, and Murino 2018) or more structure on the resulting space like correlation-independence analysis (Shahroudy et al.

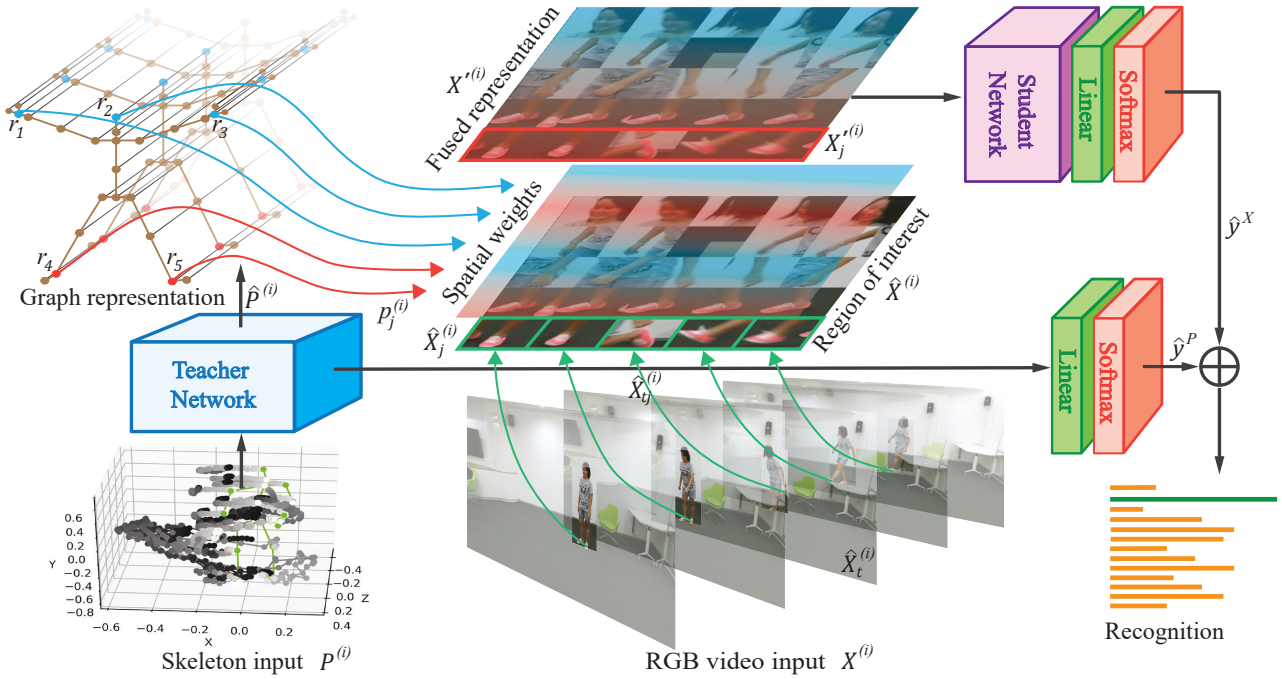

Figure 2: Illustration of model-based multimodal fusion approach in our TSMF for indoor action recognition. $p_j^{(i)}$ represents the spatial attention weights derived from the graph representation of the teacher network, we use red color to indicate relatively larger weight values that will put more attention on the corresponding regions of interest $\hat{X}_j^{(i)}$ cropped from the RGB video input. After feature fusion, the fused representation $X'^{(i)}_j$ will be fed to the student network.

2017). However, whether enforcing the similarity between the probability distribution before the SoftMax could improve the ensemble result was not reported in (Luo et al. 2018; Garcia, Morerio, and Murino 2018). On the other hand, the methods in (Luo et al. 2018; Garcia, Morerio, and Murino 2018) still rely on the performance of their cumbersome models to regularize their modality-specific networks. While the correlation analysis in (Shahroudy et al. 2017) also failed to disentangle which data modality is good for the recognition of which actions. Our method is different from these joint and coordinated representations. Instead, our TSMF fuses different data modality at the model level.

Model-based Multimodal Fusion. Multimodal fusion is one of the multimodal learning settings where all the used data modalities are used for both training and testing phases (Ngiam et al. 2011). Most existing multimodal fusion methods, which are model-agnostic, concatenate their high level features of their fully connected layers or add their results of the final SoftMax layers (Wei et al. 2017; Wu et al. 2016; Pan et al. 2019). These model-agnostic fusion methods could achieve very limited improvement and the relationship between their modality-specific networks remains closed. Different with model-agnostic fusion methods that depend on modality-specific representations, model-based fusion methods address fusion in their model level construction. Model-based fusion methods have been attempted by using the skeleton and RGB modalities to capture their complementary information (Baradel et al. 2018; Baradel, Wolf, and Mille 2017, 2018). Our model-based fusion differs from existing work with simpler structure and less number of loss items but performs better.

Proposed Method

An overview of our proposed TSMF is illustrated in Figure 2. Given a dataset with M samples, the ith training sample could be symbolized as $(P^{(i)}, X^{(i)}, y^{(i)})$ where $P^{(i)} \in \mathbb{R}^{T\times J\times 3\times 2}$ represents a skeleton sequence with two human subjects, $X^{(i)} \in \mathbb{R}^{T\times H\times W\times 3}$ represents an RGB video sequence, while $y \in \{0, 1, ..., N-1\}$ is the label of the action that has N possible action classes. Here T is the number of temporal frames, J is the number of skeleton joints, H and W denote the height and width of an RGB frame, respectively. The goal is to learn two feature extractors including a teacher network G_T with parameters Θ_T and a student network G_S with parameters Θ_S for inferring the action class by aggregating their predictions, which could be represented as

$$\hat{y} = G_T(\Theta_T, P) + \lambda G_S(\Theta_S, X) \quad (1)$$

The central part of our proposed model is constructed with two separate neural representations for the skeleton and RGB modalities with a teacher network and a student network, respectively. While the model-based data fusion happens in between the two networks. The teacher network is a Graph Convolutional Network (GCN) that learns a feature

representation from the skeleton data. This representation could not only deliver modality specific prediction but also provide spatial weights that function as an attention mechanism on the region of interest of the RGB modality. In such a way, the teacher network transfers its structural knowledge to the student network to construct a fused representation of the RGB modality. Whereas the student network is a basic CNN model that learns features from the fused representation to deliver a prediction for the RGB modality, which will be aggregated with the prediction of the teacher network to make an overall prediction.

Spatial Attention Weights

Given a skeleton sequence of an action with a set of structured joints $P^{(i)} = \{P_{tj} | P_{tj} \in \mathbb{R}^{3\times 2},\ t = (1, \ldots, T),\ j = (1, \ldots, J)\}$, where t denotes the temporal position of the skeleton frame, j denotes the spatial index of the skeleton joint, 3 and 2 are the numbers of skeleton joint attributes and skeleton bodies, respectively. We adopt a graph structure to model the spatial and temporal characteristics of the skeleton modality. Precisely, a skeleton frame at time t could be represented as a graph $\boldsymbol{\vartheta}_t = \{\boldsymbol{\upsilon}_t, \boldsymbol{\varepsilon}_t\}$, where the graph nodes $\boldsymbol{\upsilon}_t$ denote the skeleton joints and the graph edges $\boldsymbol{\varepsilon}_t$ denote the skeleton bones. In this skeleton graph, $\boldsymbol{\upsilon} = \{\upsilon_{tj} | \upsilon_{tj} = P_{tj}\}$ is a node set that contains all joints of the skeleton sequence. The convolutional operation of GCNs is similar with that of general 2D CNNs except that the sampling area of a node υ_{tj} is defined as a neighbor set $N(\upsilon_{tj}) = \{\upsilon_{tj} \mid d(\upsilon_{ti}, \upsilon_{tj}) \leq D\}$, where D is the minimum path length of $d(\upsilon_{ti}, \upsilon_{tj})$. Our convolutional sampling strategy on the skeleton graph follows the spatial partitioning strategy in (Yan, Xiong, and Lin 2018). Suppose there is a fixed number of K subsets in the $N(\upsilon_{tj})$, every neighbor set will be labeled numerically with a mapping $l_{tj} : N(\upsilon_{tj}) \rightarrow \{0, \ldots, K-1\}$. The neighborhood concept could be extended to temporally connected joints as $N(\upsilon_{tj}) = \{\upsilon_{q,j} | d(\upsilon_{tj}, \upsilon_{ti}) \leq K, | q - t| \leq \Gamma/2\}$, where Γ is the temporal kernel size that controls the temporal range of the neighbor set. Then the graph convolution for node υ_{tj} could be calculated as

$$\hat{\upsilon}_{tj} = \sum_{\upsilon_{tj} \in N \upsilon_{ti}} \frac{1}{Z_{ti}(\upsilon_{tj})} f_{in}(\upsilon_{tj}) W(l(\upsilon_{tj})) \qquad (2)$$

where $f_{in}(\upsilon_{tj})$ is the feature map that gets the attribute vector of υ_{tj}, $W(l(\upsilon_{tj}))$ is a weight function $W(\upsilon_{ti}, \upsilon_{tj}) : N(\upsilon_{ti}) \rightarrow \mathbb{R}^c$ that could be implemented by indexing a tensor of (c, K) dimension. $Z_{ti}(\upsilon_{tj}) = |\{\upsilon_{tk} | l_{ti}(\upsilon_{tk}) = l_{ti}(\upsilon_{tj})\}|$ is a normalization term that equals the cardinality of the corresponding subset.

The feature map of the skeleton sequence could be represented by a tensor of $(C,\ T,\ J)$ dimensions, where J denotes the number of vertices, T denotes the temporal length and C denotes the number of attributes of the joint vertex. With a specific partitioning strategy determined, it could be represented by an adjacent matrix $\mathbf{A}$ with its elements indicating if a vertex υ_{tj} belongs to a subset of $N(\upsilon_{tj})$. The spatial graph convolution is implemented by performing a 1×1 classical 2D convolution and multiplies the resulting tensor with the normalized adjacency matrix $\mathbf{\Lambda}^{-\frac{1}{2}} \mathbf{A} \mathbf{\Lambda}^{-\frac{1}{2}}$ on the second dimension. Then it is followed by a $1 \times \Gamma$ convolutional layer as the implemtation of the temporal graph convolution. With K partitioning strategies $\sum_{k=1}^{K} \mathbf{A}_k$, Equation 2 could be transformed into

$$\hat{P} = \sum\nolimits_{k=1}^{K} \mathbf{\Lambda}^{-\frac{1}{2}} \mathbf{A} \mathbf{\Lambda}^{-\frac{1}{2}} f_{in} W_k \odot M_k \qquad (3)$$

where $\mathbf{\Lambda}_k^{ii} = \sum_j (\mathbf{A}_k^{ij}) + \alpha$ is a diagonal matrix with α set to 0.001 to avoid empty rows. W_k is a weight tensor of the 1×1 convolutional operation with $(C_{in},\ C_{out},\ 1,\ 1)$ dimensions, which represents the weighting function of Equation 2. M_k is an attention map with the same size of A_k, which indicates the importance of each vertex. $\odot$ denotes the element-wise product between two matrixes.

In our method, we utilize features from the activation tensor $\hat{P} \in \mathbb{R}^{c \times t \times j}$ of the graph representation, which is a tensor sized $(c,\ t,\ j)$ with c denotes the number of output channels, t denotes the temporal length and j denotes the number of vertices. We use an arithmetic mean of the individual activations over the corresponding sequential locations of a skeleton joint as its knowledge that will be transferred as spatial attention weights for training the student network. Let $\hat{P}_{ctj} \in \hat{P}$ denote one neuron activation of the activation tensor, the spatial attention weight p_j for the feature fusion of the jth skeleton joint and its corresponding appearance features from the RGB video input could be calculated as

$$p_j = \frac{1}{tc} \sum\nolimits_1^t \sum\nolimits_1^c \sqrt{\hat{P}_{ctj}^2} \qquad (4)$$

Fused Representation

Given $X^{(i)} \in \mathbb{R}^{T \times H \times W \times 3}$ as the RGB video of an action, an ordered image sequence in the time interval $[1, T]$ could be represented as $X^{(i)} = \left(X_1^{(i)}, \ldots, X_t^{(i)}, \ldots, X_T^{(i)}\right)$, where $X_t^{(i)} \in \mathbb{R}^{H \times W \times 3}$ is the image frame at time t. We propose to use the Region of Interest (ROI) from the video frames to represent the appearance feature of the RGB modality. Unlike the approaches introduced in (Baradel et al. 2018; Baradel, Wolf, and Mille 2017, 2018) that focus on two hand areas of human subjects, we focus on more body areas including head, hands and feet in a temporal manner. To crop the ROI from an action video, we use joints of the 2D skeleton retrieved with the OpenPose tool (Cao et al. 2017). Given an RGB frame $X_t^{(i)}$, we define this cropping process as a feature transformation function g. This process can be written as

$$\hat{X}_{tj}^{(i)} = g\left(X_t^{(i)}, o_{tj}^{(i)}\right),\ j \in \{r_1,\ \ldots,\ r_m\},\ r_m \leq J \qquad (5)$$

where $\hat{X}_{tj}^{(i)}$ is the ROI cropped by using the jth joint $o_{tj}^{(i)}$ of the 2D OpenPose skeleton at time t. r_m is the last index of the skeleton joint corresponding to the body part that we are focusing on (see Figure 2), which is not larger than the total number of skeleton joints. Given $X^{(i)} = \left(X_1^{(i)}, \ldots, X_t^{(i)}, \ldots, X_T^{(i)}\right)$, we randomly select τ sample frames at temporal positions and horizontally concatenate their transformed features to a feature map $\hat{X}_j^{(i)}$ as illustrated in Figure 2.

Provided the spatial attention weight $p_j^{(i)}$ in Equation 4 for the jth joint derived from the teacher network, we can construct the fused representation $X'^{(i)}$, which holds the appearance feature of the RGB modality, by multiplying $p_j^{(i)}$ with its corresponding ROI feature map. This process can be formulated as

$$X'^{(i)}_j = p_j^{(i)} \times \hat{X}_j^{(i)},\ j \in \{r_1,\ \ldots,\ r_m\} \tag{6}$$

As Figure 2 shows, the size of $X'^{(i)}_j$ is related with the number of ROIs along the sequential video frames, which is horizontally concatenated into a rectangular shaped feature map. As we focus on multiple body parts, the input of the student network could be denoted as $X'^{(i)}$. $X'^{(i)}$ is a vertical concatenation of different fused representations of body areas like head, hands, and feet.

Optimization

We build an end-to-end format of our objective function as a sum of loss terms from the teacher and student networks that are both supervised by the action labels as

$$\mathcal{L} = \mathcal{L}_P\left(\hat{y}^P, y\right) + \lambda \mathcal{L}_X\left(\hat{y}^X, y\right) \tag{7}$$

The loss term $\mathcal{L}_P$ is from the teacher network that is a graph convolutional model fueled with the skeleton data. The prediction of the teacher network $\hat{y}^P$ could be represented as

$$\hat{y}^P = \sigma(G_T(P^{(i)}, \Theta_T)) \tag{8}$$

where G_T is the GCN model that will deliver attention feature $\hat{P}$ as defined in Equation 3, Θ_T is the learnable parameters of the GCN model. σ represents the linear layer that transforms the $\hat{P}$ to a one-hot representation.

For the loss term of the student network $\mathcal{L}_X$, recall that we have proposed the fused representation of the RGB modality, which is intrinsically a 2D feature map, hence we adopt the ResNet proposed in (He et al. 2016). The one-hot representation of student network prediction $\hat{y}^X$ could be formulated as

$$\hat{y}^X = \sigma(G_S(X'^{(i)},\ \Theta_S) + X'^{(i)}) \tag{9}$$

where $G_S(X'^{(i)}, \Theta_S)$ represents the residual mapping to be learned, Θ_S denotes the learnable parameters of ResNet (He et al. 2016).

For two submodels of TSMF (i.e., G_T and G_S), we formulate the optimization problem as two independent objectives as the following

$$\arg\min_{\Theta_T} -\sum\nolimits_{c=1}^{N} \underbrace{y_c log(\hat{y}_c^P)}_{\mathcal{L}_P} \tag{10}$$

$$\arg\min_{\Theta_S} -\sum\nolimits_{c=1}^{N} \underbrace{y_c log(\hat{y}_c^X)}_{\mathcal{L}_X} \tag{11}$$

where $\mathcal{L}_P$ and $\mathcal{L}_X$ are cross-entropy losses enforcing the prediction abilities of the teacher and student networks, respectively. To train the whole pipeline of TSMF, the teacher network with the parameters Θ_T needs to be trained first. Then two training strategies could be adopted for optimizing the student network. The first one is tuning Θ_T together with Θ_S. While the second one is fixing Θ_T (i.e., setting the teacher network to the evaluation mode) when Θ_S is being updated in the training mode.

Experiments

In this section, we introduce the datasets used in our experiments and a comparison of our TSMF with state-of-the-art methods. We performed our experiments on the following two human action recognition datasets: NTU RGB+D (Shahroudy et al. 2016) and PKU-MMD (Liu et al. 2017a).

Datasets

NTU RGB+D. The NTU RGB+D dataset (Shahroudy et al. 2016) was collected with Kinect v2 sensors, which contains over 56K samples of 60 different actions including individual actions, interactions between multiple people, and health-related actions. The actions were performed by 40 subjects and recorded from 80 viewpoints. Multiple data modalities like depth, RGB video, and skeleton are available from the datasets. In our method, the RGB video and skeleton channels are utilized. We followed the Cross-Subject (CS) and Cross-View (CV) split settings from (Shahroudy et al. 2016) for evaluating our method.

PKU-MMD. The PKU-MMD dataset (Liu et al. 2017a) is another popular large dataset collected with Kinect v2. It contains 1076 long untrimmed video and skeleton sequences. The dataset was performed by 66 subjects in three camera views. With 51 activity categories annotated, we retrieved 21,545 valid action sequences and 6 invalid samples that has no skeleton frames are not used. Similar with NTU RGB+D, we adopt the two evaluation protocols (i.e., cross-subject and cross-view) recommended in (Liu et al. 2017a). For action samples that have more than 300 frames, we evenly select 300 frames (details are in the source code).

Implementation Details

For the RGB modality, the sizes of the feature map cropped from the videos of NTU-RGB+D and PKU-MMD are 96×96. We set both m and τ to 5 for the two datasets. Hence, the size of the fused representations of both datasets are 480×480, which are resized to 225×225 and normalized before being fed to our student network. For our teacher network, we utilize the graph convolutional model introduced in (Yan, Xiong, and Lin 2018) for both datasets. Both implementations are trained with the stochastic gradient descent optimizer. The initial learning rate is set as 0.1, which is decayed by 0.1 at epochs 10 and 50 and ended at the epoch 80. The minibatch size is set to 64. All experiments are conducted on a workstation with 4 GTX 1080 Ti GPUs.

Comparison with State-of-the-art

We show the performance comparison with previous single modal and multimodal methods in Table 1 for both the NTU-RGB+D and PKU-MMD datasets. Our method outperforms existing skeleton-based, video-based and multimodal methods on both datasets. We could observe that existing skeleton-based methods could hardly achieve further improvement due to the lack of appearance features and the inadequacy of larger training data. While, on the NTU-RGB+D dataset, our method significantly improves the average accuracies of state-of-the-art skeleton-based (MS-G3D),

Method	Modality		NTU RGB+D			PKU-MMD		
	P	X	CS	CV	Average	CS	CV	Average
Lie Group (Vemulapalli, Arrate, and Chellappa 2014)	√	-	50.1	52.8	51.5	-	-	-
Dynamic Skeletons (Hu et al. 2015)	√	-	60.2	65.2	62.7	-	-	-
Part-aware LSTM (Shahroudy et al. 2016)	√	-	62.9	70.3	66.6	-	-	-
GCA-LSTM (Liu et al. 2017b)	√	-	74.4	82.8	78.6	-	-	-
STA-LSTM (Song et al. 2018)	√	-	73.4	81.2	77.3	86.9	92.6	89.8
View-invariant (Liu, Liu, and Chen 2017)	√	-	80.0	87.2	83.6	-	-	-
CNN-Based (Li et al. 2017)	√	-	83.2	89.3	86.3	90.4	93.7	92.1
ST-GCN (Yan, Xiong, and Lin 2018)	√	-	81.5	88.3	84.9	-	-	-
DPRL+GCNN (Tang et al. 2018)	√	-	83.5	89.8	86.7	-	-	-
HCN (Li et al. 2018)	√	-	86.5	91.1	88.8	92.6	94.2	93.4
2s-AGCN (Shi et al. 2019b)	√	-	88.5	95.1	91.8	-	-	-
AGC-LSTM (Si et al. 2019)	√	-	89.2	95.0	92.1	-	-	-
MS-G3D (Liu et al. 2020)	√	-	91.5	96.2	93.9	-	-	-
C3D (Tran et al. 2015)	-	√	63.5	70.3	66.9	-	-	-
Glimpse Clouds (Baradel et al. 2018)	-	√	86.6	93.2	89.9	-	-	-
RGB distillation (Garcia, Morerio, and Murino 2018)	√	√	79.7	81.4	80.6	-	-	-
DSSCA - SSLM (Shahroudy et al. 2017)	√	√	74.9	-	-	-	-	-
STA-Hands (Baradel, Wolf, and Mille 2017)	√	√	82.5	88.6	85.6	-	-	-
Hands Attention (Baradel, Wolf, and Mille 2018)	√	√	84.8	90.6	87.7	-	-	-
Our multimodal method	√	√	**92.5**	**97.4**	**95.0**	**95.8**	**97.8**	**96.8**

Table 1: Comparison with state-of-the-art methods on NTU RGB+D and PKU-MMD with Cross-Subject (CS) and Cross-View (CV) settings (accuracy in %). P denotes the skeleton modality. X denotes the RGB modality. √ means used. - means not used.

video-based (Glimpse Clouds), and multimodal (Hands Attention) methods by 1.1%, 5.1%, and 7.3%, respectively. For the PKU-MMD dataset, our method also achieves a significant improvement of 3.4% average accuracy comparing with the current best result achieved by HCN (Li et al. 2018). The results indicate that our TSMF successfully compensates the lack of appearance features in the skeleton modality with its fused representation of the RGB modality. Moreover, existing mutimodal methods usually utilize much more complex CNN models like ResNet50 in Glimpse Clouds (Baradel et al. 2018) or ResNet101 in SlowFast Networks (Feichtenhofer et al. 2019), which indicates the effectiveness and further potential of our proposed TSMF for action recognition. However, it remains open regarding if Glimpse Clouds could contribute back to the ensemble results and whether SlowFast Networks could perform well for indoor actions. It is also worth mentioning that, by using the ResNet101 and I3D as the backbone, SGFB (Ji et al. 2020) could perform better than SlowFast Networks on outdoor actions. However, it requires huge computational resources and a perfect scene graph prediction method, which may not hold for indoor actions in this study.

Discussion

To validate the effectiveness of our method, we analyze the design choices of our TSMF with three questions.

Which training strategy is good for the student network? As illustrated in rows #2, #3 and #4 of Table 2, training the student model without updating the parameter of the teacher model (i.e., in evaluation mode) could be more effective than updating the teacher and student models together. Although the student model could not achieve as good performance as the teacher model, it could compensate the lack of appearance features in the skeleton modality no matter with which training strategy is used. However, training the student model without the spatial weights in our proposed TSMF could not contribute as much as the fused representation of the RGB modality. Figure 3 shows some visualized views of the fused representation of the RGB modality during the training process. We could observe that the unimportant body areas are masked by the spatial weights, which improves the ability of the student network to compensate the lack of appearance features in the skeleton modality.

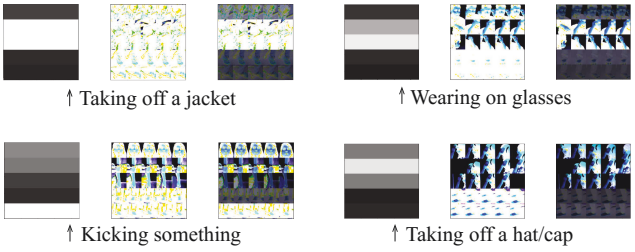

Figure 3: Visualized views of the data during the training process for NTU-RGB+D (left, middle and right parts are spatial weights, normalized ROI and fused representation, respectively). The unimportant body areas are masked.

#	Method	Submodel		NTU RGB+D			PKU-MMD		
		G_T	G_S	CS	CV	Average	CS	CV	Average
1	ST-GCN (Yan, Xiong, and Lin 2018)	√	-	81.6	88.8	85.2	91.5	92.4	92.0
2	ResNet18 (He et al. 2016)	-	√	72.7	81.3	77.0	75.3	75.1	75.2
3	ResNet18+Spatial Weights	√	√	73.8	85.2	79.5	76.8	75.8	76.4
4	ResNet18+Spatial Weights	○	√	76.8	86.2	81.5	82.8	82.2	82.5
5	Ensemble (1+2)	○	○	88.9	94.9	91.9	93.5	95.2	94.4
6	Ensemble (1+3)	○	○	89.4	95.0	92.2	93.5	95.0	94.3
7	Ensemble (1+4)	○	○	89.7	95.4	92.6	94.3	95.7	95.0
8	AGCN (Shi et al. 2019b)	√	-	83.4	90.0	86.7	93.3	96.4	94.9
9	MS-G3D (Liu et al. 2020)	√	-	89.6	95.0	92.3	95.0	96.2	95.6
10	Ensemble (4+8)	○	○	90.4	95.5	93.0	95.4	97.7	96.6
11	Ensemble (4+9)	○	○	**92.5**	**97.4**	**95.0**	**95.8**	**97.8**	**96.8**

Table 2: Ablation study for NTU RGB+D and PKU-MMD (accuracy in %). G_T and G_S are the teacher and student network models, respectively. √ means in training mode. ○ means in evaluation mode. - means the submodel is not used.

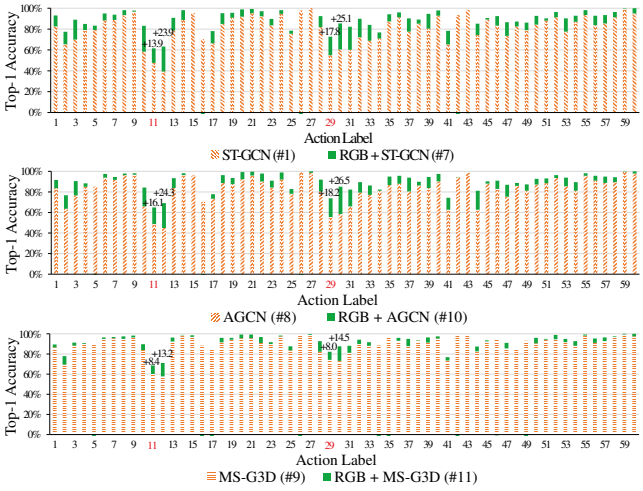

Figure 4: Visualization of improvements on skeleton-based models trained with the skeleton joint stream (the cross-subject setting of NTU-RGB+D).

Could the fused representation effectively contribute to the skeleton modality? Existing skeleton-based methods like AGCN (Shi et al. 2019b) and MS-G3D (Liu et al. 2020) have achieved encouraging improvements by considering two streams of the skeleton modality (i.e., skeleton joints and bones). To investigate if the RGB modality trained with our TSMF could further contribute the recognition accuracy, we consider the skeleton joint stream in this work since our teacher network is based on it. As shown in rows #10 and #11 of Table 2, the results of the RGB modality trained with our TSMF could bring significant improvements to the results of AGCN and MS-G3D trained with the skeleton joint stream, which is more significant than the contribution of the skeleton bone stream as compared to their corresponding two-stream results shown in Table 1. Figure 4 shows the improvements gained for each action of NTU-RGB+D with our TSMF when aggregate the results of our student network with that of various GCN models trained with the skeleton joint stream (i.e., rows #1, #8 and #9 in Table 2).

Are the challenging actions improved significantly? From Figure 4, we could observe that challenging actions labeled 11, 12, 29 and 30 (i.e., "reading", "writing", "playing with phone" and "typing" in the NTU-RGB+D dataset as mentioned in Figure 1) could not be well recognized by skeleton-based models like ST-GCN (Yan, Xiong, and Lin 2018), AGCN (Shi et al. 2019b) and MS-G3D (Liu et al. 2020) due to the lack of appearance features. By aggregating the results of the student network to that of various GCN models, our proposed TSMF achieves consistent and significant improvements for these challenging actions. Except these challenging actions, we could also observe that the recognition accuracies of almost all of the actions are actually improved with our student network (for ST-GCN, AGCN and MS-G3D, the accuracies of 59, 58 and 52 out of 60 actions are improved, respectively). From the above analysis, it is obvious that our TSMF effectively makes use of the multimodal information.

Conclusion

This paper proposes a model-based multimodal fusion approach called TSMF for indoor action recognition with heterogeneous skeleton and RGB data modalities. Comparing with previous single modal and multimodal methods, the proposed TSMF model has achieved superior performance on NTU-RGB+D and PKU-MMD. Based on the analysis of the experimental results, the fused representation of our TSMF successfully takes the advantage of multimodal information as it complements the inadequacy of appearance features in the skeleton modality. In the future, we will extend our work to the task of outdoor action recognition and explore more effective action recognition techniques (Feichtenhofer et al. 2019; Ji et al. 2020) under complex and various environment like in UCF-101 (Soomro, Zamir, and Shah 2012) and Kinetics (Kay et al. 2017).

Acknowledgements

This work is supported by The Hong Kong Polytechnic University for the Project: Music Therapy for Dementia Prevention Using Algorithmic Composition under Grant No.: P0008740. The author also would like to thank group members who helped with proofreading the manuscript.

References

Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41(2): 423–443. ISSN 0162-8828.

Baradel, F.; Wolf, C.; and Mille, J. 2017. Human action recognition: Pose-based attention draws focus to hands. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 604–613.

Baradel, F.; Wolf, C.; and Mille, J. 2018. Human activity recognition with pose-driven attention to rgb. In *BMVC 2018 - 29th British Machine Vision Conference*, pp.1–14.

Baradel, F.; Wolf, C.; Mille, J.; and Taylor, G. W. 2018. Glimpse clouds: Human activity recognition from unstructured feature points. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 469–478.

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7291–7299.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Choi, J.; Gao, C.; Messou, J. C.; and Huang, J.-B. 2019. Why Can't I Dance in the Mall? Learning to Mitigate Scene Bias in Action Recognition. In *Advances in Neural Information Processing Systems*, 851–863.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE international conference on computer vision*, 6202–6211.

Feichtenhofer, C.; Pinz, A.; and Zisserman, A. 2016. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1933–1941.

Garcia, N. C.; Morerio, P.; and Murino, V. 2018. Modality distillation with multiple stream networks for action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 103–118.

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.

Hu, J.-F.; Zheng, W.-S.; Lai, J.; and Zhang, J. 2015. Jointly learning heterogeneous features for RGB-D activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5344–5352.

Ioffe, S.; and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* .

Ji, J.; Krishna, R.; Fei-Fei, L.; and Niebles, J. C. 2020. Action Genome: Actions As Compositions of Spatio-Temporal Scene Graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10236–10247.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; and Natsev, P. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* .

Li, C.; Zhong, Q.; Xie, D.; and Pu, S. 2017. Skeleton-based action recognition with convolutional neural networks. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 597–600. IEEE.

Li, C.; Zhong, Q.; Xie, D.; and Pu, S. 2018. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. *arXiv preprint arXiv:1804.06055* .

Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; and Tian, Q. 2019. Actional-Structural Graph Convolutional Networks for Skeleton-based Action Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3595–3603.

Liu, C.; Hu, Y.; Li, Y.; Song, S.; and Liu, J. 2017a. PKU-MMD: A large scale benchmark for skeleton-based human action understanding. In *Proceedings of the Workshop on Visual Analysis in Smart and Connected Communities*, 1–8.

Liu, J.; Wang, G.; Hu, P.; Duan, L.-Y.; and Kot, A. C. 2017b. Global context-aware attention LSTM networks for 3D action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1647–1656.

Liu, M.; Liu, H.; and Chen, C. 2017. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition* 68: 346–362. ISSN 0031-3203.

Liu, Z.; Zhang, H.; Chen, Z.; Wang, Z.; and Ouyang, W. 2020. Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 143–152.

Luo, Z.; Hsieh, J.-T.; Jiang, L.; Carlos Niebles, J.; and Fei-Fei, L. 2018. Graph distillation for action detection with privileged modalities. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 166–183.

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y. 2011. Multimodal deep learning. In *The 28th International Conference on Machine Learning*.

Pan, B.; Sun, J.; Lin, W.; Wang, L.; and Lin, W. 2019. Cross-Stream Selective Networks for Action Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 0–0.

Shahroudy, A.; Liu, J.; Ng, T.-T.; and Wang, G. 2016. NTU RGB+ D: A large scale dataset for 3D human activity analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1010–1019.

Shahroudy, A.; Ng, T.-T.; Gong, Y.; and Wang, G. 2017. Deep multimodal feature analysis for action recognition in rgb+ d videos. *IEEE transactions on pattern analysis and machine intelligence* ISSN 0162-8828.

Shi, L.; Zhang, Y.; Cheng, J.; and Lu, H. 2019a. Skeleton-Based Action Recognition With Directed Graph Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7912–7921.

Shi, L.; Zhang, Y.; Cheng, J.; and Lu, H. 2019b. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12026–12035.

Si, C.; Chen, W.; Wang, W.; Wang, L.; and Tan, T. 2019. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1227–1236.

Song, S.; Lan, C.; Xing, J.; Zeng, W.; and Liu, J. 2018. Spatio-temporal attention-based LSTM networks for 3D action recognition and detection. *IEEE Transactions on image processing* 27(7): 3459–3471.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* .

Tang, Y.; Tian, Y.; Lu, J.; Li, P.; and Zhou, J. 2018. Deep progressive reinforcement learning for skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5323–5332.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 4489–4497.

Vemulapalli, R.; Arrate, F.; and Chellappa, R. 2014. Human action recognition by representing 3d skeletons as points in a lie group. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 588–595.

Wei, P.; Zhao, Y.; Zheng, N.; and Zhu, S.-C. 2017. Modeling 4D human-object interactions for joint event segmentation, recognition, and object localization. *IEEE transactions on pattern analysis and machine intelligence* 39(6): 1165–1179. ISSN 0162-8828.

Wu, D.; Pigou, L.; Kindermans, P.-J.; Le, N. D.-H.; Shao, L.; Dambre, J.; and Odobez, J.-M. 2016. Deep dynamic neural networks for multimodal gesture segmentation and recognition. *IEEE transactions on pattern analysis and machine intelligence* 38(8): 1583–1597. ISSN 0162-8828.

Xie, S.; Sun, C.; Huang, J.; Tu, Z.; and Murphy, K. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 305–321.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *32nd AAAI conference on artificial intelligence*.

Zach, C.; Pock, T.; and Bischof, H. 2007. A duality based approach for realtime TV-L 1 optical flow. In *Joint pattern recognition symposium*, 214–223. Springer.

Zhang, P.; Lan, C.; Xing, J.; Zeng, W.; Xue, J.; and Zheng, N. 2017. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In *Proceedings of the IEEE International Conference on Computer Vision*, 2117–2126.

Zhang, Y.; Tokmakov, P.; Hebert, M.; and Schmid, C. 2019. A structured model for action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9975–9984.

Zhu, W.; Lan, C.; Xing, J.; Zeng, W.; Li, Y.; Shen, L.; and Xie, X. 2016. Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. In *Thirtieth AAAI Conference on Artificial Intelligence*.