Joint Dynamic Pose Image and Space Time Reversal for Human Action Recognition from Videos

Mengyuan Liu[1,2], Fanyang Meng[3], Chen Chen[4], Songtao Wu[5]
[1]Tencent Research [2]School of Electrical and Electronic Engineering, Nanyang Technological University
[3]Key Laboratory of Machine Perception, Peking University, Shenzhen Graduate School
[4]University of North Carolina at Charlotte
[5]College of Computer Science and Software Engineering, Shenzhen University
nkliuyifang@gmail.com fymeng@pkusz.edu.cn chenchen870713@gmail.com csstwu@szu.edu.cn

Abstract

Human action recognition aims to classify a given video according to which type of action it contains. Disturbance brought by clutter background and unrelated motions makes the task challenging for video frame-based methods. To solve this problem, this paper takes advantage of pose estimation to enhance the performances of video frame features. First, we present a pose feature called dynamic pose image (DPI), which describes human action as the aggregation of a sequence of joint estimation maps. Different from traditional pose features using sole joints, DPI suffers less from disturbance and provides richer information about human body shape and movements. Second, we present attention-based dynamic texture images (att-DTIs) as pose-guided video frame feature. Specifically, a video is treated as a space-time volume, and DTIs are obtained by observing the volume from different views. To alleviate the effect of disturbance on DTIs, we accumulate joint estimation maps as attention map, and extend DTIs to attention-based DTIs (att-DTIs). Finally, we fuse DPI and att-DTIs with multi-stream deep neural networks and late fusion scheme for action recognition. Experiments on NTU RGB+D, UTD-MHAD, and Penn-Action datasets show the effectiveness of DPI and att-DTIs, as well as the complementary property between them.

Introduction

Human action recognition (HAR) is an active topic in the field of artificial intelligence (Wang et al. 2018). This task has a wide range of applications in human-robot interaction and intelligent video surveillance.

HAR can be divided into image-based HAR and video-based HAR. Single image can barely distinguish similar human actions, e.g., “sitting down” and “standing up”. It is more common and natural to analyze human action using videos, as human action is actually a sequence of submovements. According to the type of human action, HAR can also be categorized into “human body action”, “hand gesture”, and “group action”. We focus on “human body action”, and simplify this term as “action”.

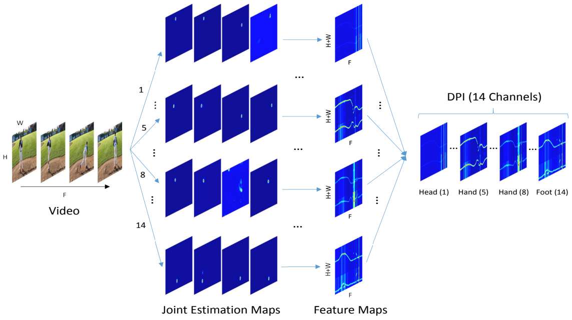

Figure 1: Dynamic pose image (DPI) for action description from a video. All generated images are colored to facilitate observation.

Action recognition from videos remains challenging for two reasons. First, each video frame concurs traditional problems in image analysis, such as clutter background, illumination variation, and low resolution. Second, motions among video frames bring new problems, including unrelated motions and motion blur. We focus on decreasing the effect from clutter background and unrelated motions. In other words, we intend to recognize actions from complex scenes using unfixed cameras. Without assuming the camera is stable, common background modeling (Monnet et al. 2003) can barely work against camera movements. Besides, background modeling cannot distinguish human movements from unrelated motions, e.g., “waving tree” and “moving car”. An alternative way is to use human detection (Dalal and Triggs 2005) to extract human region from the scenes. While, the detected human regions still contain background. What’s worse, human detection methods may concur detection failure and mismatch whole human body region.

Recent development of human pose estimation enables the usage of human pose for solving above mentioned problem. The general pipeline of human pose estimation includes the inference of joint estimation maps and the extraction of joint locations. The location of a joint is defined as the position with maximum value on the joint estimation map. Compared with joint location, the joint estimation map contains richer information about the distribution of body parts. Therefore, we use joint estimation map instead of joint location for further robust feature extraction.

We present dynamic pose image (DPI) to describe human action as the aggregation of joint estimation maps, which is shown in Fig. 1. Suppose we use pose estimation method to estimate human body as 14 joint estimation maps. Given a video, we apply pose estimation on each video frame, leading to 14 channels of new sequences. Each new sequence denotes the movements of one joint. As the data in each new sequence is sparse, we remove the redundancy by aggregating each new sequence as one compact feature map. The DPI is defined as the concatenation of all feature maps. As can be seen, DPI is a 14-channel image, which can simultaneously capture human body shape and movements. Specifically, human body shapes are explicitly expressed in separate feature map, and joint movements are implicitly expressed across multiple channels. Considering the semantic meaning of joint estimation maps, DPI mainly focuses on human body and thus suffers less from disturbance, including clutter background and unrelated motions.

Moreover, DPI shows complementary property to video frame-based features. DPI can be viewed as feature extracted from a new modal called joint estimation maps, whose pixel values represent probabilities. Compared with RGB frames, joint estimation maps can directly reflect the locations of human body parts, but fail to capture textural information. Combining video frame-based features and DPI benefits the description of both textural and human body information. Especially for the description of a special action type – "human-object interaction", the object is ignored by DPI, but can be captured by video frame-based features.

We present dynamic texture image (DTI) as a new type of video frame-based feature. The redundancy in video motivates us to compress the video as a multi-channel image. We introduce space time reversal method to reduce spatial and temporal redundancy in video, respectively. A video is treated as a space time volume. The space time reversal means to reverse the space and time by observing the volume from front, side, and top views. From these three views, we develop three DTIs, i.e., f-DTI, t-DTI, and s-DTI. Specifically, f-DTI reduces temporal redundancy, meanwhile t-DTI and s-DTI reduce spatial redundancy. Since original video frames may contain clutter background and un-related motions, we accumulate joint estimation maps as attention map, which is combined with video frame for designing DTIs. These attention-based DTIs, named as att-DTIs, show more distinctive power for representing human action.

Generally, our contribution is three-fold:

- We propose dynamic pose image (DPI) as a compact pose feature for human action recognition. Based on joint estimation maps, DPI captures richer information about human body parts, compared with pose-based methods using joint locations. What's more, DPI suffers less from clutter background and unrelated motions.
- We present a new type of video frame-based feature, called dynamic texture image (DTI). Based on space time reversal, we develop three DTIs, i.e., f-DTI, s-DTI, and t-DTI to capture more spatial and temporal cues. Attention-based DTIs (att-DTIs) are further proposed to suppress the effect of clutter background and unrelated motions.
- With multi-stream CNNs and late fusion scheme, our method achieves state-of-the-art performances on three benchmark datasets. Experimental results consistently verify the effectiveness of DPI, att-DTIs, and the complementary property between them.

Related Work

Video Frame-Based Methods

Previous methods use Convolutional neural network (CNN) (Simonyan and Zisserman 2014), 3D CNN (Du et al. 2016; Baradel et al. 2018; Carreira and Zisserman 2017; Qiu, Yao, and Mei 2017), and recurrent neural network (RNN) (Luo et al. 2017) for video description. Two-stream convolutional network (Simonyan and Zisserman 2014) learns spatial-temporal features by fusing convolutional networks spatially and temporally. Compared with traditional 2D CNN, 3D CNN (Du et al. 2016), is more suitable for learning spatiotemporal features. Based on 3D CNN (Du et al. 2016), Two-Stream Inflated 3D ConvNet (I3D) (Carreira and Zisserman 2017) is proposed to enlarge perception field along the temporal direction, and Pseudo-3D Residual Net (P3D ResNet) (Qiu, Yao, and Mei 2017) is proposed to build deeper 3D CNN model. With 3D CNN as backbone, the spatial and temporal relationships among visual features can be further explored (Qiu, Yao, and Mei 2017). Compared with CNN, RNN is born to model temporal relationships among video frames. In (Luo et al. 2017), a RNN based encoder-decoder framework is proposed to effectively learn a representation that predicts the sequence of basic motions. These methods follow a common pipeline, i.e., extracting spatial feature from each frame (or several consecutive frames) and then modeling temporal relationships among frames. Different from such pipeline, our DTI feature describes a video as a multi-channel image, which simultaneously capture both spatial and temporal information of video frames in a compact manner. In addition, the s-DTI (side view) and t-DTI (top view) are formulated from novel views of the original video volume. These DTIs provide richer spatial and temporal cues for describing action from videos.

Human Pose-Based Methods

Aforementioned methods ignore the semantic meaning of human actions which are inherently structured patterns of body movements. Recent studies (Zhu et al. 2016; Ma, Sigal, and Sclaroff 2015; Gkioxari and Malik 2015; Singh, Arora, and Jawahar 2016; Ma, Fan, and Kitani 2016) extract whole human body or body parts instead of whole video for analysis. Further, human action recognition and pose estimation tasks have been integrated to extract pose guided features for recognition. Wang *et al.* (Wang, Wang, and Yuille 2013) improve an existing pose estimation method, and then design pose features to represent both spatial and temporal configurations of body parts. Nie *et al.* (Xiaohan Nie, Xiong, and Zhu 2015) propose a framework to integrate training and testing of action recognition and pose estimation. They decompose action into poses which are further divided to mid-level ST-parts and then part. In (Chron, Laptev, and Schmid 2016), joint locations are used to guide the sampling of patches for extracting CNN features. In (Iqbal, Garbade, and Gall 2017), action recognition and pose estimation are conducted in an iterative manner. After iteration, the pose and video frame features are fused for action recognition. In (Zolfaghari et al. 2017), the estimated poses and video frames are directly encoded and fused by multi-stream 3D CNN model. Above pose features (Wang, Wang, and Yuille 2013; Iqbal, Garbade, and Gall 2017; Zolfaghari et al. 2017) or pose-guided video frame features (Xiaohan Nie, Xiong, and Zhu 2015; Chron, Laptev, and Schmid 2016) do not use the relative position of multiple human joints over time. Different from these methods, our DPI representation naturally incorporates this information. Moreover, existing methods usually rely on joint locations. Our DPI is built upon joint estimation maps, which contain richer spatial information than joint locations. Besides, we show the complementary property between DPI and att-DTIs, which achieves the state-of-the-art performances on three benchmark datasets.

Proposed Model

In following, we first present DPI based on joint estimation maps, and then present att-DTIs based on space time reversal and attention map. Finally, we propose multi-stream fusion method to combine both type of features.

DPI

Human pose estimation from a single image is actually a structure prediction problem. Recent progress in CNN-based methods boost the accuracy of estimated poses. In (Ramakrishna et al. 2014), a pose machine is proposed to sequentially predict pose estimation maps for body parts, where previous predicted pose estimation maps iteratively improve the estimates in following stages.

Let $\mathcal{Y}_k \in \{x, y\}$ denote the set of coordinates from body part k. The structural output can be formulated as $\mathcal{Y} = \{\mathcal{Y}_1, ..., \mathcal{Y}_k, ..., \mathcal{Y}_K\}$, where K is the total number of body parts. Multi-class classifier g_t^k is trained to predict the k-th body part in the m-th stage. For a position $\mathbf{z}$, the joint estimation map for the k-th body part is formulated as:

$$\mathbf{J}_m^k(\mathcal{Y}_k = \mathbf{z}) = g_m^k\left(\mathbf{f_z}; \bigcup_{i=1,...,K} \psi(\mathbf{z}, \mathbf{J}_{m-1}^i)\right), \tag{1}$$

where $\mathbf{f_z}$ is the color feature at position $\mathbf{z}$, $\mathbf{J}_{m-1}^i$ is the joint estimation map predicted by g_{m-1}^i, $\cup$ is the operator for vector concatenation, ψ is the feature function for computing contextual features from previous joint estimation maps. After M stages, the generated joint estimation maps are used to predict locations of body parts.

The pose machine (Ramakrishna et al. 2014) uses boosted classifier with random forests for the weak learners. Instead, this paper applies the convolutional pose machine (Wei et al. 2016; Cao et al. 2017) to combine pose machine with convolutional architectures, which does not need graphical-model style inference and boosts the performances of pose machine.

Let $\mathbf{J}_f^k$ denote the k-th joint estimation map on the f-th video frame of a video $\mathbf{V}_c \in \mathbb{R}^{H\times W\times 3\times F}$, where H is the height, W is the width, and F is the number of frames. We use pose estimation method (Cao et al. 2017) to generate 18 joint estimation maps. Four joints on the head are not used since they are redundant for denoting human body movements. Generally, a set of joint estimation maps $\{\mathbf{J}_f^k\}_{f=1,...,F}^{k=1,...K}$ are used for further feature extraction, where K equals to 14. Each joint estimation map has the same size of the original video frame, we organize the set of joint estimation maps in the matrix form: $\mathbf{E} \in \mathbb{R}^{H\times W\times K\times F}$.

Compared with the original video $\mathbf{V}_c$, the scale of $\mathbf{E}$ is $K/3$ times larger, which brings extra computation burden. Observing the sparse property of joint estimation maps, we present DPI as a compact representation of $\mathbf{E}$, which is shown in Fig. 1. Let $\mathbf{E}^k \in \mathbb{R}^{H\times W\times F}$ be the sequence of joint estimation maps for the k-th joint. We reduce the spatial redundancy of $\mathbf{E}^k$ by horizontal and vertical projection methods. Using vertical projection, the reduced data called $\mathbf{H}^k \in \mathbb{R}^{W\times F}$ is formulated as:

$$\mathbf{H}^k[w, f] = \frac{1}{H}\sum_{h=1}^{H} \mathbf{E}^k[h, w, f]. \tag{2}$$

Using horizontal projection, the reduced data called $\mathbf{W}^k \in \mathbb{R}^{H\times F}$ is formulated as:

$$\mathbf{W}^k[h, f] = \frac{1}{W}\sum_{w=1}^{W} \mathbf{E}^k[h, w, f]. \tag{3}$$

We combine both reduced data to formulate $\mathbf{P}^k$ as $[\mathbf{H}^k, \mathbf{W}^k]$, which belongs to $\mathbb{R}^{(H+W)\times F}$. Our proposed DPI, $\mathbf{P} \in \mathbb{R}^{(H+W)\times F\times K}$, can be formulated by taking $\mathbf{P}^k$ as the feature map on the k-th channel. The data scale of $\mathbf{P}$ is only $(H + W)/(H * W)$ times of $\mathbf{E}$. Suppose $H \approx W$, $(H + W)/(H * W)$ equals to $2/H$. Usually, the height H of video frame is larger than 100 pixels. In other words, we compress $\mathbf{E}$ by at least 50 times.

DPI is a compact description of joint estimation maps. It contains less spatial redundancy, and can reflect both shape and movements of human body parts. We consider DPI as a multiple channel image. In this way, we can take advantage of pre-trained CNNs for extracting deep features. To this end, we normalize pixels in DPI to the scope of 0 to 255. The normalized $\hat{\mathbf{P}}$ is formulated as:

$$\hat{\mathbf{P}} = 255 \times \frac{\mathbf{P} - min\{\mathbf{P}\}}{max\{\mathbf{P}\} - min\{\mathbf{P}\}}, \tag{4}$$

where function $max\{\cdot\}$ calculates the maximum value of a given matrix, and $min\{\cdot\}$ calculates the minimum value. To facilitate the usage of pre-trained CNNs, we further normalize the size of $\hat{\mathbf{P}}$ to fixed size. In this work, we use pre-trained ResNet model on ImageNet dataset. Correspondingly, the processed DPI is resized to $\hat{\mathbf{P}} \in \mathbb{R}^{224\times 224\times K}$.

att-DTIs

To describe the texture of a color video, we first transform it to a gray scale video $\mathbf{V}_g \in \mathbb{R}^{H\times W\times F}$. Previous deep learning methods commonly process $\mathbf{V}_g$ in three ways. First, CNN-based methods treat the video as a bag of frames. Each frame of the video is processed by CNN to predict the action label, and all predictions are fused to obtain final prediction. Second, RNN-based methods treat the video as a sequence of frames. Each frame is described as a feature vector. The

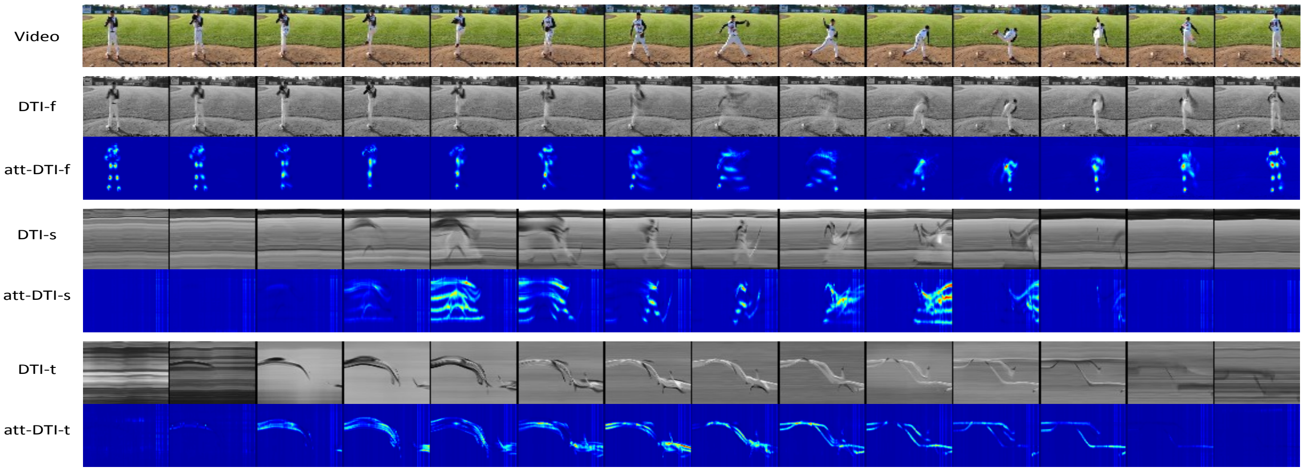

Figure 2: Comparison between DTIs and att-DTIs. Three att-DTIs are colored to facilitate observation.

RNN is applied to model temporal relationship among feature vectors. Third, 3DCNN-based methods treat the video as a 3D volume. They use 3D convolution to simultaneously fuse spatial temporal information.

Different from above methods, we treat the video as a multi-channel image, which is named as dynamic texture image (DTI). In this way, we are able to take advantage of pre-trained CNN models on large-scale image classification task for transfer learning. Compared with CNN-based methods, our method captures the temporal relationship among video frames. Compared with RNN-based and 3DCNN-based methods, our method is able to capture long term temporal information through frame sampling.

To implement DTI, a simple way is treating $\mathbf{V}_g \in \mathbb{R}^{H\times W\times F}$ as a multi-channel image, where the frame number F is treated as the number of channels. However, this definition has two drawbacks. First, the size of DTI varies from different videos according to the number of frames of a video. Second, to handle a long video, we have to use CNN with a significant number of parameters, which cost more storage and computation. What's worse, CNN with deep channels is difficult to train and converge. To solve these problems, we resize the matrix $\mathbf{V}_g$ to $\hat{\mathbf{V}}_g \in \mathbb{R}^{H\times W\times \hat{F}}$, where $\hat{F} \ll F$. To keep the continuity among frames, bicubic interpolation method is used to resize the matrix. Similar to the processing of DPI, we finally obtain a normalized form of DTI, $\hat{\mathbf{V}}_g \in \mathbb{R}^{224\times 224\times K}$. Noted that we set $\hat{F}$ to K to keep the uniformity of our following CNN model.

As DTI is sampled from $\mathbf{V}_g$, it inevitably ignores a portion of spatial and temporal information. We introduce space time reversal to alleviate this problem. Given $\mathbf{V}_g$, we transform the channel order to form three new matrices, namely, $\mathbf{V}_g^f \in \mathbb{R}^{H\times W\times F}$, $\mathbf{V}_g^t \in \mathbb{R}^{F\times W\times H}$, and $\mathbf{V}_g^s \in \mathbb{R}^{H\times F\times W}$. Note that $\mathbf{V}_g^f$ equals to $\mathbf{V}_g$, that is to keep the original channel order unchanged. Based on these matrices, we generate three types of DTIs, namely, DTI-f, DTI-t, and DTI-s, by treating F, H, and W as the temporal channel, respectively. Mathematically, we denote DTIs as $\mathbf{T}_f$, $\mathbf{T}_t$ and $\mathbf{T}_s$. The normalized versions are denoted as $\hat{\mathbf{T}}_f$, $\hat{\mathbf{T}}_t$ and $\hat{\mathbf{T}}_s$, which belong to $\mathbb{R}^{224\times 224\times K}$. The merit of our method is that the three DTIs are complementary to each other. Specifically, DTI-f is able to capture the main portion of spatial data, and DTI-t and DTI-s can characterize the main portion of temporal data. Jointly using these DTIs can effectively model the spatial and temporal information in a video.

The original video usually contains clutter background and unrelated motions, these disturbances decrease the distinctive power of DTIs. To solve this problem, we use attention maps to weight video frames as a pre-processing step before building DTIs. For the f-th frame, we define A_f as $A_f = \sum_{k=1}^{K} \mathbf{J}_f^k$, which is the accumulation of joint estimation maps. We normalize pixel values of A_f to the range of zero to one. The normalized A_f is defined as the attention map. We weight the f-th frame by multiplying each pixel value on the frame with the corresponding pixel value on the attention map. Based on the weighted video frames, we present the attention-based DTIs, termed att-DTIs for short. Fig. 2 shows the comparison between DTIs and att-DTIs. We sample 14 key frames to represent the video. For DTIs and att-DTIs, we unfold them along the channel. Compared with DTIs, att-DTIs contain less background information and are more related to the human action.

Fusion

We use multi-stream CNN model to fuse DPI and DTIs. A simplified version of our model is shown in Fig. 3, where one DTI is fused with DPI. Naturally, three types of DTI can be fused with DPI in a similar way. Considering the appearance gap between DTI and DPI, we use CNN to separately process each stream, and use late fusion for final prediction.

For each stream of our input data, we use the pre-trained ResNet152 model due to its impressive performance and strong generalization ability. Since ResNet152 is originally designed to process three-channel images, we modify the first convolutional layer to process our proposed K-channel images, i.e., DPI and DTIs. For the first convolutional layer, the number of input channel is K; the number of output feature maps is 64; its kernel size is 7 with a stride of 2; the padding size is 3. In order to classify N types of actions,

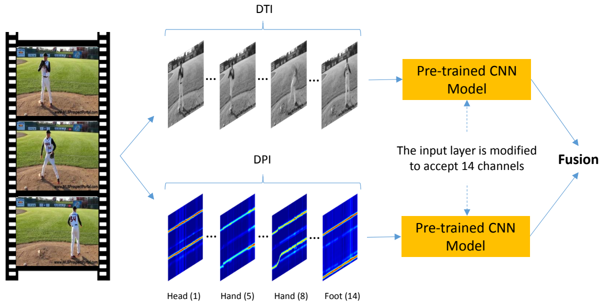

Figure 3: Our proposed model for fusing DTI and DPI

we remove the final full connection layer from the basic network, and add a new one with N neurons as output. For the first and last layers, their parameters are initialized using Xavier initialization. For other layers, pre-trained parameters are used for initialization.

The network weights are learned using the mini-batch stochastic gradient descent with the momentum value set to 0.9 and weight decay set to 0.0004. The learning rate is set to 0.001 and the maximum training epoch is set to 60. After every 20 epochs, the learning rate is multiplied by 0.1. In each epoch, a mini-batch of B samples is constructed by randomly sampling B images from the training set. For NTU RGB+D dataset, UTD-MHAD dataset and Pen-Action dataset, the batch size B is set to 32, 16 and 16, considering the scale of training set. We do not use any data augmentation method to ensure the running speed of our model. When the accuracy on the training set is higher than 0.98, the training procedure is early stopped. To reduce the effect of random parameter initialization and random sampling, we repeat the training of CNN model for five times and report the average results. We implement our method using PyTorch with four Tesla K80 GPUs.

Experiments

To compare with both RGB-based and pose-based methods, we conduct experiments on NTU RGB+D (Shahroudy et al. 2016), UTD-MHAD (Chen, Jafari, and Kehtarnavaz 2015), and Penn-Action (Zhang, Zhu, and Derpanis 2013) datasets, which contain both RGB and 3D/2D pose data.

Datasets and Protocols

NTU RGB+D dataset (NTU) contains 60 actions performed by 40 subjects from various views, generating more than 56K videos and 4 million frames. Following the cross subject protocol in (Shahroudy et al. 2016), we split the 40 subjects into training and testing groups. Each group contains samples captured from different views performed by 20 subjects. This is currently the largest dataset for 3D pose-based human action recognition. Despite that estimating poses from depth is much easier than from RGB, the estimated poses from depth are still noisy, which reflects the difficulty of using estimated poses from RGB for action recognition task. Besides, the view point and large intra-class variations bring new challenges to this dataset. For the evaluation, the training and testing sets have 40320 and 16560 samples, respectively.

Table 1: Evaluation of our method with different settings

Type	Feature	NTU	UTD	Penn
Human Pose	DPI	74.41%	74.05%	90.32%
Video Frame	DTI-f	70.60%	68.28%	83.50%
	DTI-s	77.49%	65.16%	84.66%
	DTI-t	70.13%	57.95%	77.23%
	DTIs	85.39%	79.81%	92.51%
Early Fusion	att-DTI-f	73.83%	71.72%	85.77%
	att-DTI-s	83.05%	74.00%	89.61%
	att-DTI-t	79.86%	65.77%	83.82%
	att-DTIs	88.02%	85.81%	94.25%
Late Fusion	DPI+att-DTI-f	84.12%	82.14%	93.90%
	DPI+att-DTI-s	88.76%	84.19%	94.53%
	DPI+att-DTI-t	86.57%	80.60%	93.28%
	DPI+att-DTIs	**90.23%**	**88.37%**	**95.86%**

UTD-MHAD dataset (UTD) was collected using a Microsoft Kinect sensor and a wearable inertial sensor in an indoor environment. It contains 27 actions performed by 8 subjects. Each subject repeated each action 4 times, generating 861 sequences. We use this dataset to compare the performances of methods using different data modalities. Cross subject protocol (Chen, Jafari, and Kehtarnavaz 2015) is used for evaluation.

Penn-Action dataset (Penn) contains 15 action categories and 2326 sequences in total. Since all sequences are collected from internet, complex body occlusions, large appearance and motion variations make it challenging for pose-related action recognition (Xiaohan Nie, Xiong, and Zhu 2015; Du, Wang, and Qiao 2017). We follow (Xiaohan Nie, Xiong, and Zhu 2015) to split the data into half and half for training and testing.

Ablation Study

Table 1 shows the ablation study of our method on three datasets. DTIs means jointly using DTI-f, DTI-s, and DTI-t as features. The predictions of these three features are multiplied as the final prediction of DTIs. Similarly, att-DTIs means the fusion result of att-DTI-f, att-DTI-s, and att-DTI-t. “Early Fusion”denotes that we fuse human pose information (organized as attention map) and video frame at feature level. “Late Fusion” means that we fuse human pose feature and video frame feature at decision level. The symbol “+” means multiplying posterior probability matrices.

DTIs: We evaluate the performances of DTI-f, DTI-s, DTI-t, and their combined form called DTIs on describing video frames. On NTU, UTD and Penn datasets, DTI-f achieves the accuracy of 70.60%, 68.28% and 83.50%, respectively. This shows that DTI feature is able to describe common video frames. By representing video volume from side view, we find that DTI-s achieves comparable performance with DTI-f. Especially on NTU dataset, DTI-s achieves an accuracy of 77.49%, which is even 6.89% higher than DTI-f. The possible reason is that DTI-s facilities the neural networks to capture space and time information. As shown in Fig. 2, several feature maps in DTI-s capture the spatial and temporal information at the same time. We can even guess the action with single feature map from DTI-s.

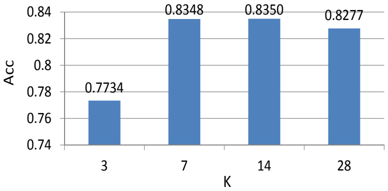

Figure 4: Evaluation of DTI-f with channel K on Penn dataset

Another finding is that DTI-t also works for action recognition, but it performs worse than DTI-f and DTI-s. The reason is that human is only a small portion in each image, when observing the video volume from the top view. Therefore, clutter background and unrelated motions may overpower the human body information. On all datasets, the combined form, DTIs, performs much better than sole DTI feature, i.e., DTI-f, DTI-s, and DTI-t. For example, DTIs outperforms sole DTI feature by more than 7%. This result verifies that three DTI features are complementary to each other.

Early fusion: We evaluate the function of attention maps on DTI features. Generally, attention maps improve the performances of DTI-f, DTI-s, DTI-t, and DTIs. For example, att-DTI-f achieves accuracy of 73.83% on NTU dataset, which is 3.23% higher than DTI-f. Among DTI features, attention maps significantly boost the performance of DTI-t. On NTU dataset, att-DTI-t outperforms DTI-t by 9.73%. The reason is that attention maps can effectively alleviate the effect of disturbances, from which DTI-t suffers the most.

Late fusion: We evaluate the performances of combining both human pose-based feature and video frame-based feature. Specifically, we use DPI as human pose-based feature and use attention-based DTI features to describe video frames. We claim that the combination of both types of features boost the performances of either feature. On NTU dataset, DPI+att-DTI-f achieves accuracy of 84.12%, which is 9.71% higher than DPI and 10.29% higher than att-DTI-f. Highest accuracy of 90.23% is achieved by DPI+att-DTIs, as it has the richest information including DPI and DTI features from three views.

Channel K: As shown in Fig. 4, we take DTI-f as an example to show the selection of channel number K for designing DTI-f. We set K to 3, 7, 14, and 28. When K equals to 14, DTI-f achieves the best accuracy of 83.50%. When K equal to 3, DTI-f only achieves accuracy of 77.34%, since many video frames which may contain significant action cues are discarded. When K changes from 14 to 28, the accuracy treads to decrease. The reason is that DTI-f with more channels increases the the number of parameters of CNN model, making the learning difficult from limited training data. It is more reasonable to set K to 7, taking both performance and the scale of CNN model into account. We actually set K to 14 to ensure that only one CNN structure is needed to take either DPI or DTI features as input.

Interpolation vs. Sampling: Instead of resizing a video to DTI, another choice is to uniformly sample K frames from the video to construct DTI. We take DTI-f with $K = 3$ as an example, our DTI-f achieves an accuracy of 77.34%, which outperforms the uniformly sampling-based DTI-f by 1.29%. This is because DTI-f is based on frame interpolation, which fuses information from multiple frames. While, the uniformly sampling method only captures information from K frames.

Table 2: Comparison between our method and state-of-the-art approaches on NTU dataset using cross subject protocol. JEM is short for joint estimation map, which is a byproduct of estimating 2D pose from RGB data.

Method	Modal	Acc
HON4D (Oreifej and Liu 2013)	Depth	30.56%
Super Normal Vector (Yang and Tian 2014)	Depth	31.82%
Lie Group (Vemulapalli, Arrate, and Chellappa 2014)	3D Pose	50.10%
HBRNN-L (Du, Wang, and Wang 2015)	3D Pose	59.07%
FTP Dynamic Skeletons (Hu et al. 2015)	3D Pose	60.23%
2 Layer P-LSTM (Shahroudy et al. 2016)	3D Pose	62.93%
ST-LSTM + Trust Gate (Liu et al. 2016)	3D Pose	69.20%
Unsupervised Learning (Luo et al. 2017)	RGB	56.00%
LieNet-3Blocks (Huang et al. 2017)	3D Pose	61.37%
GCA-LSTM network (Liu et al. 2017)	3D Pose	74.40%
Clips + CNN + MTLN (Ke et al. 2017)	3D Pose	79.57%
Chained Network (Zolfaghari et al. 2017)	RGB+2D Pose	80.80%
ST-GCN (Yan, Xiong, and Lin 2018)	3D Pose	80.70%
Ind-RNN (Li et al. 2018)	3D Pose	81.80%
RGB + 2D Pose (Luvizon, Picard, and Tabia 2018)	RGB+2D Pose	85.50%
Glimpse Clouds (Baradel et al. 2018)	RGB	86.60%
Proposed DPI	JEM	74.41%
Proposed DTIs	RGB	85.39%
Proposed att-DTIs	RGB+JEM	88.02%
Proposed DPI+att-DTIs	RGB+JEM	**90.23%**

Table 3: Comparison between our method and state-of-the-art methods on UTD dataset using cross subject protocol

Method	Modal	Acc
Cov3DJ (Hussein et al. 2013)	3D Pose	85.58%
Kinect (Chen, Jafari, and Kehtarnavaz 2015)	3D Pose	66.10%
Inertial (Chen, Jafari, and Kehtarnavaz 2015)	Inertial	67.20%
Fusion (Chen, Jafari, and Kehtarnavaz 2015)	3D Pose+Inertial	79.10%
JTM (Wang et al. 2016)	3D Pose	85.81%
Optical Spectra (Hou et al. 2016)	3D Pose	86.97%
3DHOT-MBC (Zhang et al. 2017)	Depth	84.40%
JDM (Li et al. 2017)	3D Pose	88.10%
Proposed DPI	JEM	74.05%
Proposed DTIs	RGB	79.81%
Proposed att-DTIs	RGB+JEM	85.81%
Proposed DPI+att-DTIs	RGB+JEM	**88.37%**

Comparisons with State-of-the-Art

State-of-the-art methods can be roughly divided into video frame-based methods and human pose-based methods, where human pose can be estimated from depth data or RGB data. As we use sole RGB data, our method can be fairly compared with video frame-based methods and human pose-based methods using RGB data. The performances of human pose-based methods using depth data are listed to show the superior performance of our method, even compared with methods using depth data.

Ours vs. 2D pose-based methods: We evaluate the performance of DPI verses estimated poses on describing human poses. According to (Luvizon, Picard, and Tabia 2018), sole estimated poses can achieve an accuracy of 71.70% on NTU dataset. DPI outperforms estimated poses by 2.71%. The reason is that DPI is built on joint estimation maps, which provide richer information than estimated poses. In (Luvizon, Picard, and Tabia 2018), they crop multiple clips from a video, and the final score on multi-clip is computed by the average result on all clips from one video. Combining this method with estimated poses, the performance is boosted to 74.30%. Without applying any temporal information

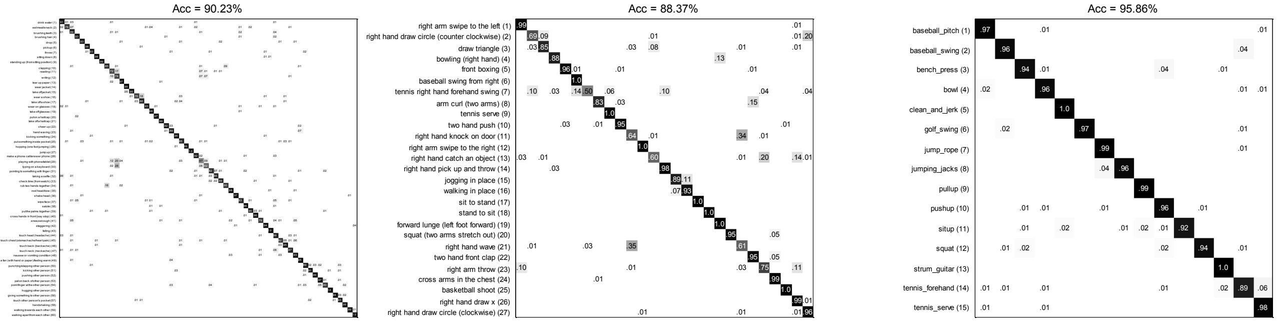

Figure 5: Confusion matrices of our method on NTU, UTD, and Penn datasets

enhancement methods, our DPI still performs slightly better than sole estimated poses.

Ours vs. 3D pose-based methods: As shown in Table 2, many deep learning based methods (Ke et al. 2017; Yan, Xiong, and Lin 2018; Li et al. 2018) have achieved high performances by applying CNN, GCN, or RNN model on 3D pose. Ind-RNN (Li et al. 2018) achieves 81.80%, which is the best performance for 3D pose based methods. Our proposed DPI yields 74.41%, which is 7.39% lower than Ind-RNN. This indicates that 3D pose from depth data is more accurate than 2D pose from RGB data. Furthermore 3D pose contains depth information, which facilitates the description of human motions along the depth direction.

Ours vs. RGB-based methods: Unsupervised Learning (Luo et al. 2017) obtains an accuracy of 56.00% on NTU dataset using video frame features. Our proposed DTIs achieves 85.39%, which outperforms (Luo et al. 2017) by a large margin. With attention maps, the proposed att-DTIs further boosts the accuracy by 2.63% as compared to DTIs. The accuracy of att-DTIs is also 1.42% higher than that of Glimpse Clouds (Baradel et al. 2018), which is the most recent video frame-based method on NTU dataset.

Ours vs. Depth/Inertial-based methods: In Table 2, the performances of HON4D (Oreifej and Liu 2013) and Super Normal Vector (Yang and Tian 2014) are far lower than our method. Although they use more advanced depth data, our method using deep neural networks can still effectively learn distinctive features from RGB features. Table 3 compares our method with related methods on UTD dataset. Our DPI+att-DTIs has an accuracy of 88.37%, which outperforms methods using depth or inertial data.

Ours vs. RGB+2D pose-based methods: On NTU dataset, our method is most related to (Zolfaghari et al. 2017) and (Luvizon, Picard, and Tabia 2018), which jointly use video frame and 2D pose features for action recognition. Our DPI+att-DTIs reaches 90.23%, which is 9.43% higher than (Zolfaghari et al. 2017) and 4.73% higher than (Luvizon, Picard, and Tabia 2018). Our method fuses video frame and pose information at different stages, which effectively leverages the complementary property between different modals. Table 4 compares our method with the state-of-the-art approaches on Penn dataset, which is collected in the wild. Pose + IDT-FV method (Iqbal, Garbade, and Gall 2017) uses both pose and hand-crafted video-frame based features, and achieves 92.00% of recognition accuracy. Our proposed DPI+att-DTIs has 3.86% better accuracy than (Iqbal, Garbade, and Gall 2017), which again verifies the superiority of our method. Confusion matrices of DPI+att-DTIs are shown in Fig. 5. It is evident that most ambiguities among similar actions are suppressed.

Table 4: Comparison between our method and state-of-the-art methods on Penn dataset using half/half protocol

Method	Modal	Acc
Action Bank (Zhang, Zhu, and Derpanis 2013)	RGB	83.90%
AOG (Xiaohan Nie, Xiong, and Zhu 2015)	RGB+2D Pose	85.50%
C3D (Du et al. 2016)	RGB	86.00%
JDD (Cao et al. 2016)	RGB+2D Pose	87.40%
Pose + IDT-FV (Iqbal, Garbade, and Gall 2017)	RGB+2D Pose	92.00%
Proposed DPI	JEM	90.32%
Proposed DTIs	RGB	92.51%
Proposed att-DTIs	RGB+JEM	94.25%
Proposed DPI+att-DTIs	RGB+JEM	**95.86%**

Conclusion and Future Work

This paper jointly uses human pose and video frame features for action recognition. The proposed dynamic pose image (DPI) and attention-based dynamic texture images (att-DTIs) can effectively capture spatial and temporal information of action. Moreover, they are robust to clutter background and unrelated motions. The DPI aggregates joint estimation maps and provides richer human body cues than traditional estimated 2D poses. The att-DTIs are built by observing video volume from three views, which follows the proposed space time reversal rule. Experiments are conducted on three benchmark datasets, where the combination of DPI and att-DTIs outperforms RGB-based methods and even some depth-based methods. In our future work, we will focus on inserting space time reversal rule to more video volume description methods, such as I3D model (Carreira and Zisserman 2017) and P3D model (Qiu, Yao, and Mei 2017). Instead of using 2D CNN model, modifying 3D CNN model to process DPI and att-DTIs is a new direction.

Acknowledgement

This work is supported in part by start-up funds of University of North Carolina at Charlotte. Special thanks to Prof. Junsong Yuan for his useful suggestions.

References

Baradel, F.; Wolf, C.; Mille, J.; and Taylor, G. W. 2018. Glimpse clouds: Human activity recognition from unstructured feature points. In *CVPR*.

Cao, C.; Zhang, Y.; Zhang, C.; and Lu, H. 2016. Action recognition with joints-pooled 3D deep convolutional descriptors. In *IJCAI*, 3324–3330.

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Real-time multi-person 2D pose estimation using part affinity fields. In *CVPR*, 1302–1310.

Carreira, J., and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 4724–4733.

Chen, C.; Jafari, R.; and Kehtarnavaz, N. 2015. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *ICIP*, 168–172.

Chron, G.; Laptev, I.; and Schmid, C. 2016. P-CNN: Pose-Based CNN Features for Action Recognition. In *ICCV*, 3218–3226.

Dalal, N., and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *CVPR*, 886–893.

Du, T.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2016. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 4489–4497.

Du, W.; Wang, Y.; and Qiao, Y. 2017. RPAN: An end-to-end recurrent pose-attention network for action recognition in videos. In *ICCV*, 3745–3754.

Du, Y.; Wang, W.; and Wang, L. 2015. Hierarchical recurrent neural network for skeleton based action recognition. In *CVPR*, 1110–1118.

Gkioxari, G., and Malik, J. 2015. Finding action tubes. In *CVPR*, 759–768.

Hou, Y.; Li, Z.; Wang, P.; and Li, W. 2016. Skeleton optical spectra based action recognition using convolutional neural networks. *TCSVT*.

Hu, J.-F.; Zheng, W.-S.; Lai, J.; and Zhang, J. 2015. Jointly learning heterogeneous features for RGB-D activity recognition. In *CVPR*, 5344–5352.

Huang, Z.; Wan, C.; Probst, T.; and Van Gool, L. 2017. Deep learning on lie groups for skeleton-based action recognition. In *CVPR*, 6099–6108.

Hussein, M. E.; Torki, M.; Gowayyed, M. A.; and El-Saban, M. 2013. Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations. In *IJCAI*, 2466–2472.

Iqbal, U.; Garbade, M.; and Gall, J. 2017. Pose for action-action for pose. In *FG*, 438–445.

Ke, Q.; Bennamoun, M.; An, S.; Sohel, F.; and Boussaid, F. 2017. A new representation of skeleton sequences for 3D action recognition. In *CVPR*, 1063–6919.

Li, C.; Hou, Y.; Wang, P.; and Li, W. 2017. Joint distance maps based action recognition with convolutional neural networks. *SPL* 24(5):624–628.

Li, S.; Li, W.; Cook, C.; Zhu, C.; and Gao, Y. 2018. Independently recurrent neural network (indrnn): Building a longer and deeper rnn. In *CVPR*.

Liu, J.; Shahroudy, A.; Xu, D.; and Wang, G. 2016. Spatio-temporal LSTM with trust gates for 3D human action recognition. In *ECCV*, 816–833.

Liu, J.; Wang, G.; Hu, P.; Duan, L.-Y.; and Kot, A. C. 2017. Global context-aware attention LSTM networks for 3D action recognition. In *CVPR*, 3671–3680.

Luo, Z.; Peng, B.; Huang, D.-A.; Alahi, A.; and Fei-Fei, L. 2017. Unsupervised learning of long-term motion dynamics for videos. In *CVPR*, 7101–7110.

Luvizon, D. C.; Picard, D.; and Tabia, H. 2018. 2d/3d pose estimation and action recognition using multitask deep learning. In *CVPR*.

Ma, M.; Fan, H.; and Kitani, K. M. 2016. Going deeper into first-person activity recognition. In *CVPR*, 1894–1903.

Ma, S.; Sigal, L.; and Sclaroff, S. 2015. Space-time tree ensemble for action recognition. In *CVPR*, 5024–5032.

Monnet, A.; Mittal, A.; Paragios, N.; and Ramesh, V. 2003. Background modeling and subtraction of dynamic scenes. In *ICCV*, 1305–1312.

Oreifej, O., and Liu, Z. 2013. Hon4D: Histogram of oriented 4D normals for activity recognition from depth sequences. In *CVPR*, 716–723.

Qiu, Z.; Yao, T.; and Mei, T. 2017. Learning spatio-temporal representation with pseudo-3d residual networks. 5534–5542.

Ramakrishna, V.; Munoz, D.; Hebert, M.; Bagnell, J. A.; and Sheikh, Y. 2014. Pose machines: Articulated pose estimation via inference machines. In *ECCV*, 33–47.

Shahroudy, A.; Liu, J.; Ng, T.-T.; and Wang, G. 2016. NTU RGB+D: A large scale dataset for 3D human activity analysis. In *CVPR*, 1010–1019.

Simonyan, K., and Zisserman, A. 2014. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 568–576.

Singh, S.; Arora, C.; and Jawahar, C. 2016. First person action recognition using deep learned descriptors. In *CVPR*, 2620–2628.

Vemulapalli, R.; Arrate, F.; and Chellappa, R. 2014. Human action recognition by representing 3d skeletons as points in a lie group. In *CVPR*, 588–595.

Wang, P.; Li, Z.; Hou, Y.; and Li, W. 2016. Action recognition based on joint trajectory maps using convolutional neural networks. In *ACM MM*, 102–106.

Wang, P.; Li, W.; Ogunbona, P.; Wan, J.; and Escalera, S. 2018. RGB-D-based Human Motion Recognition with Deep Learning: A Survey. *CVIU*.

Wang, C.; Wang, Y.; and Yuille, A. L. 2013. An approach to pose-based action recognition. In *CVPR*, 915–922.

Wei, S.-E.; Ramakrishna, V.; Kanade, T.; and Sheikh, Y. 2016. Convolutional pose machines. In *CVPR*, 4724–4732.

Xiaohan Nie, B.; Xiong, C.; and Zhu, S.-C. 2015. Joint action recognition and pose estimation from video. In *CVPR*, 1293–1301.

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

Yang, X., and Tian, Y. 2014. Super normal vector for activity recognition using depth sequences. In *CVPR*, 804–811.

Zhang, B.; Yang, Y.; Chen, C.; Yang, L.; Han, J.; and Shao, L. 2017. Action recognition using 3D histograms of texture and a multi-class boosting classifier. *TIP* 26:4648–4660.

Zhang, W.; Zhu, M.; and Derpanis, K. G. 2013. From actemes to action: A strongly-supervised representation for detailed action understanding. In *ICCV*, 2248–2255.

Zhu, W.; Hu, J.; Sun, G.; Cao, X.; and Qiao, Y. 2016. A key volume mining deep framework for action recognition. In *CVPR*, 1991–1999.

Zolfaghari, M.; Oliveira, G. L.; Sedaghat, N.; and Brox, T. 2017. Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection. In *ICCV*, 2923–2932.