An End-to-End Spatio-Temporal Attention Model for Human Action Recognition from Skeleton Data

Sijie Song,[1] **Cuiling Lan,**[2*] **Junliang Xing,**[3] **Wenjun Zeng,**[2] **Jiaying Liu**[1*]
[1] Institute of Computer Science and Technology, Peking University, Beijing, China
[2] Microsoft Research Asia, Beijing, China
[3] Institute of Automation, Chinese Academy of Sciences, Beijing, China
{ssj940920, liujiaying}@pku.edu.cn, {culan,wezeng}@microsoft.com, jlxing@nlpr.ia.ac.cn

Abstract

Human action recognition is an important task in computer vision. Extracting discriminative spatial and temporal features to model the spatial and temporal evolutions of different actions plays a key role in accomplishing this task. In this work, we propose an end-to-end spatial and temporal attention model for human action recognition from skeleton data. We build our model on top of the Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM), which learns to selectively focus on discriminative joints of skeleton within each frame of the inputs and pays different levels of attention to the outputs of different frames. Furthermore, to ensure effective training of the network, we propose a regularized cross-entropy loss to drive the model learning process and develop a joint training strategy accordingly. Experimental results demonstrate the effectiveness of the proposed model, both on the small human action recognition dataset of SBU and the currently largest NTU dataset.

1 Introduction

Recognition of human action is a fundamental yet challenging task in computer vision. It facilitates many applications such as intelligent video surveillance, human-computer interaction, video summary and understanding (Poppe 2010; Weinland, Ronfard, and Boyerc 2011). The key to the success of this task is how to extract discriminative spatial temporal features to effectively model the spatial and temporal evolutions of different actions.

One general approach focuses on the recognition from RGB videos (Weinland, Ronfard, and Boyerc 2011). Since each frame is a capture of the highly articulated human in a two-dimensional space, it loses some information of the three-dimensional (3D) space and then loses the flexibility of achieving human location and scale invariance. The other general approach leverages the high level information of skeleton data, which represents a person by the 3D coordinate positions of key joints (i.e., head, neck,$\cdots$, foot). Such representation is robust to variations of locations and viewpoints. Without combining RGB information, there is a lack of appearance information. Fortunately, biological observations from the early seminal work of Johansson suggest that the positions of a small number of joints can effectively represent human behavior even without appearance information (Johansson 1973). Skeleton-based human representation has attracted increasing attention for recognizing human actions thanks to its high level representation and robustness to variations of locations and appearances (Han et al. 2016). The prevalence of cost-effective depth cameras such as Microsoft Kinect (Zhang 2012) and the advance of a powerful human pose estimation technique from depth (Shotton et al. 2011) make 3D skeleton data easily accessible. This boosts research on skeleton-based human action recognition. In this work, we focus on recognition from skeleton data.

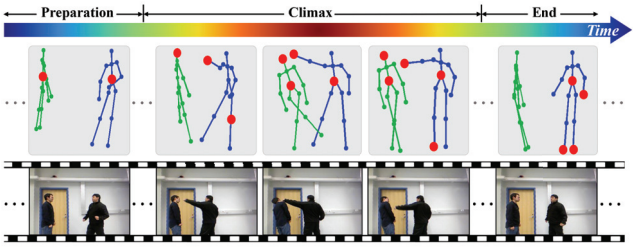


Figure 1: Illustration of the procedure for an action "punching". An action may experience different stages, and involve different discriminative subsets of joints (as the red circles).

Fig. 1 shows an example of a series of skeleton frames (and RGB images) for the action "punching". Each human body is represented by several key joints in terms of the coordinate positions in the 3D space. The articulated configurations of joints constitute various postures and a series of postures in a certain time order identifies an action. With the skeleton as an explicit high level representation of human pose, many works design algorithms taking the positions of joints as inputs. There are two basic components. One is the design and mining of discriminative features from the skeleton, such as the histograms of 3D joint locations (HOJ3D) (Xia, Chen, and Aggarwal 2012), pairwise relative position features (Wang, Liu, and Yuan 2012), relative 3D geometry features (Vemulapalli, Arrate, and Chellappa 2016). The other is the modeling of temporal dynamics, such

*Corresponding author. This work was done at Microsoft Research Asia. This work was supported by National Natural Science Foundation of China under contract No. 61472011 and No. 61303178.

as Hidden Markov Model (Xia, Chen, and Aggarwal 2012), Conditional Random Fields (Sminchisescu, Kanaujia, and Metaxas 2006), and Recurrent Neural Networks (Du, Wang, and Wang 2015). In this work, we present a spatio-temporal attention model to effectively incorporate the two components into an end-to-end deep learning architecture.

For spatial joints of skeleton, we propose a spatial attention module which conducts automatic mining of discriminative joints. A certain type of action is usually only associated with and characterized by the combinations of a subset of kinematic joints (Wang, Liu, and Yuan 2012). As the action proceeds, the associated joints may also change accordingly. For example, the joints "hand", "elbow", and "head" are discriminative for the action "drinking" while the joints from legs can be considered as noise. For an action "approaching and shaking hands", at the beginning, the legs may be paid attention to; at the middle stage, the arms attract more attention. In contrast to actionlet (Wang, Liu, and Yuan 2012), the attentions to joints are allowed to vary over time, being content-dependent.

Furthermore, for a sequence of frames, we propose a temporal attention module explicitly learning and allocating the content-dependent attentions to the output of each frame. For a sequence of some action, the flow of the action may experience different stages, e.g., the preparation, climax, and the end (Fig. 1). Taking the action "punching" as an example, the two persons approach each other, stretch out the hands, and kick out the legs. The frames for identifying stretching out the hands and kicking out the legs are a part of the key sub-stages. Different sub-stages/frames have different degrees of importance and robustness to variations. In this paper, in contrast to the ideas of extracting key frames (Carlsson and Sullivan 2001; Zhao and Elgammal 2008), our proposed scheme pays different attentions to different frames instead of simply skipping frames.

In summary, we have made the following four main contributions in this work.

- An end-to-end framework with two types of attention modules is designed based on the LSTM networks for skeleton based human action recognition.
- A spatial attention module with joint-selection gates is designed to adaptively allocate different attentions to different joints of the input skeleton within each frame. A temporal attention module with frame-selection gate is designed to allocate different attentions to different frames.
- Spatio-temporal regularizations are proposed to enable the better learning of the networks.
- A joint training strategy is designed to efficiently train the entire end-to-end network.

2 Related Work

2.1 Spatial Co-Occurrence Exploration

An action is usually associated with and characterized by the interactions and combinations of a subset of skeleton joints. An actionlet ensemble model is proposed to mine such discriminative joints (Wang, Liu, and Yuan 2012), where an actionlet is a particular conjunction of the features for a subset of the joints and an action is represented as a linear combination of the actionlets. For example, for the action "drinking", the subset of joints including "hand", "elbow", and "head" composes an actionlet. Orderlet makes an extension of actionlet by including the feature of pairwise joint distance and allowing various sizes of a subset (Yu, Liu, and Yuan 2014). Actionlets or orderlets are mined from training samples for robust performance. With RNN, a group sparsity constraint is introduced to the connection matrix to encourage the network to explore the co-occurrence of joints (Zhu et al. 2016).

In the above methods, once the mining is done, the degrees of importance of joints/features are fixed and there will be no change for different temporal frames and sequences. In contrast, our spatial attention module determines the degrees of importance of joints on the fly based on the contents.

2.2 Temporal Key Frame Exploration

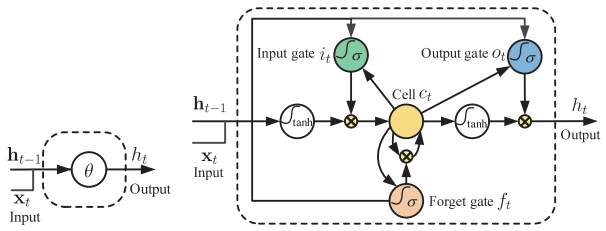
For identifying an action, not all frames in a sequence have the same importance. Some frames capture less meaningful information, or even carry misleading information associated with other types of actions, while some other frames carry more discriminative information (Liu, Shao, and Rockett 2013). A number of approaches have proposed using key frames as a representation for action recognition. One is to utilize the conditional entropy of visual words to measure the discriminative power of a given frame and the classification results from the top 25% most discriminative frames are employed to make a majority vote for recognition (Zhao and Elgammal 2008). Another one employs AdaBoost to select the most discriminative key frames for action recognition (Liu, Shao, and Rockett 2013). The learning of key frames can also be cast in a max-margin discriminative framework by treating them as latent variables (Raptis and Sigal 2013).

Leveraging key frames can help exclude noise frames, e.g., frames which are less relevant to the underlying actions. However, in comparisons to the holistic based approaches (Simonyan and Zisserman 2014; Wu et al. 2015; Zhu et al. 2016) which use all the frames, it loses some information. In this paper, our temporal attention module determines the degree of importance for each frame. Instead of skipping frames, it allocates different attention weights to different frames to automatically exploit their respective discriminative power and focus more on the important frames.

2.3 Attention-Based Models

When observing the real-world, a human usually focuses on some fixation points at the first glance of the scene, i.e., paying different attentions to different regions (Goferman, Zelnik-Manor, and Tal 2012). Many applications leverage predicted saliency maps for performance enhancement (Yu, Mann, and Gosine 2010; Jiang, Xu, and Zhao 2014; Bazzani, Larochelle, and Torresani 2016), which explicitly learn the saliency maps guided by human labeled groundtruths.

The human labeled groundtruths for the explicit attention are generally unavailable and might not be consistent with real attention related to the specific tasks. Recently, the exploitation of an attention model implicitly learning attention



(a) (b)

Figure 2: Structures of the neurons. (a) RNN, (b) LSTM.

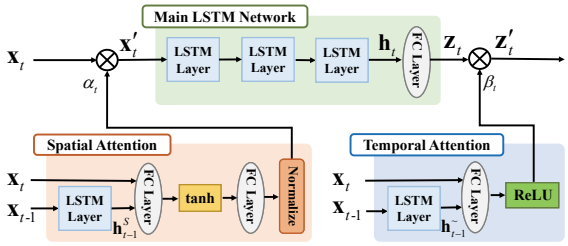


Figure 3: Overall architecture of our proposed network, which consists of the main LSTM network, the spatial attention subnetwork, and the temporal attention subnetwork.

has attracted increasing interest in various fields, such as machine translation (Bahdanau, Cho, and Bengio 2014), image caption generation (Xu et al. 2015), and image recognition (Ba, Mnih, and Kavukcuoglu 2014). Selective focus on different spatial regions is proposed for action recognition on RGB videos (Sharma, Kiros, and Salakhutdinov 2015). Ramanathan et al. propose an attention model to detect events in RGB videos while attending to the people responsible for the event (Ramanathan et al. 2015). The fusion of neighboring frames within a sliding window with learned attention weights is proposed to enhance the performance of dense labeling of actions in RGB videos (Yeung et al. 2015). However, all the attention models mentioned above for action recognition are based on RGB videos. There is a lack of investigation of skeleton sequences, which exhibit different characteristics from RGB videos.

3 Overview of RNN and LSTM

In this section, we briefly review the Recurrent Neural Network (RNN), and Long Short-Term Memory (LSTM) to make the paper self-contained.

RNN is a popular model for sequential data modeling and feature extraction (Graves 2012). Fig. 2(a) shows an RNN neuron. The output response h_t at time step t is determined by the input $\mathbf{x}_t$ and the hidden outputs from RNN themselves at the last time step $\mathbf{h}_{t-1}$

$$h_t = \theta\left(\mathbf{w}_{xh}^{\mathrm{T}}\mathbf{x}_t + \mathbf{w}_{hh}^{\mathrm{T}}\mathbf{h}_{t-1} + b_h\right), \tag{1}$$

where θ represents a non-linear activation function, $\mathbf{w}_{xh}$ and $\mathbf{w}_{hh}$ denote the learnable connection vectors, and b_h is the bias value. The recurrent structure and the internal memory of RNN facilitate its modeling of the long-term temporal dynamics of the sequential data.

LSTM is an advanced RNN architecture which mitigates the vanishing gradient effect of RNN (Hochreiter and Schmidhuber 1997; Hochreiter et al. 2001; Graves 2012). As illustrated in Fig. 2(b), an LSTM neuron contains a memory cell c_t which has a self-connected recurrent edge of weight 1. At each time step t, the neuron can choose to write, reset, and read the memory cell governed by the input gate i_t, forget gate f_t and output gate o_t.

4 Deep LSTM with Spatio-Temporal Attention Model

We propose an end-to-end multi-layered LSTM network with spatial and temporal attention mechanisms for action recognition. The network is designed to automatically select dominant joints within each frame through the spatial attention module, and assign different degrees of importance to different frames through the temporal attention module. Fig. 3 shows its overall architecture, which consists of a main LSTM network, a spatial attention subnetwork, and a temporal attention subnetwork. Because of the inter-play among the three subnetworks, it is challenging to train the network.

In the following, we discuss the proposed spatial attention module and temporal attention module, respectively, both built based on the LSTM networks. We then introduce a regularized learning objective of our model and a joint training strategy to help overcome the difficulty of model learning for the highly coupled network.

4.1 Spatial Attention with Joint-Selection Gates

The action of persons can be described by the evolution of a series of human poses represented by the 3D coordinates of joints. In general, different actions involve different subsets of joints as discussed in Section 2.1.

We propose a spatial attention model to automatically explore and exploit the different degrees of importance of joints. With a soft attention mechanism, each joint within a frame is assigned a spatial attention weight based on the joint-selection gates. This enables our model to adaptively focus more on those discriminative joints.

At each time step t, given the full set of K joints $\mathbf{x}_t = (\mathbf{x}_{t,1}, ..., \mathbf{x}_{t,K})^{\mathrm{T}}$, with $\mathbf{x}_{t,k} \in \mathbb{R}^3$, the scores $\mathbf{s}_t = (s_{t,1}, \cdots, s_{t,K})^{\mathrm{T}}$ for indicating the importance of the K joints are jointly obtained as

$$\mathbf{s}_t = U_s \tanh(W_{xs}\mathbf{x}_t + W_{hs}\mathbf{h}_{t-1}^s + \mathbf{b}_s) + \mathbf{b}_{us}, \tag{2}$$

where U_s, W_{xs}, W_{hs} are the learnable parameter matrixes, $\mathbf{b}_s$, $\mathbf{b}_{us}$ are the bias vectors. $\mathbf{h}_{t-1}^s$ is the hidden variable from an LSTM layer as illustrated in Fig. 3. For the k^{th} joint, the activation as the joint-selection gate is computed as

$$\alpha_{t,k} = \frac{\exp(s_{t,k})}{\sum_{i=1}^{K}\exp(s_{t,i})}, \tag{3}$$

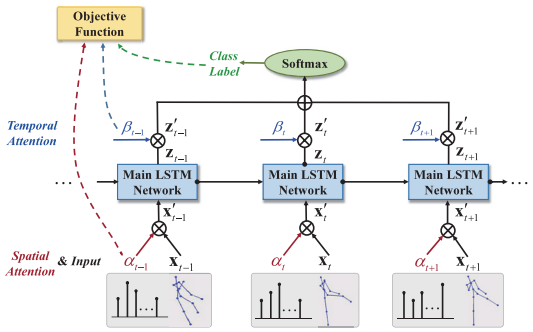


Figure 4: Illustration of how spatial attention output α and temporal attention output β influence the LSTM network.

which is a normalization of the scores. The set of gates controls the amount of information of each joint to flow to the main LSTM network. Among the joints, the larger the activation, the more important this joint is for determining the type of action. We also refer to the activation values as attention weights. Instead of assigning equal degrees of importance to all the joints $\mathbf{x}_t$, as illustrated in Fig. 4, the input to the main LSTM network is modulated to $\mathbf{x'}_t = (\mathbf{x'}_{t,1}, ..., \mathbf{x'}_{t,K})^{\mathrm{T}}$, with $\mathbf{x'}_{t,k} = \alpha_{t,k} \cdot \mathbf{x}_{t,k}$.

Note that the proposed spatial attention model determines the importance of joints based on all the joints of the current time step and the hidden variables from an LSTM layer. On one hand, the hidden variables $\mathbf{h}_{t-1}$ contain information of past frames, benefiting from the merit of LSTM which is capable of exploring temporal long range dynamics. In this paper, the spatial attention subnetwork composes of an LSTM layer, two fully connected layers and a normalization unit as illustrated in Fig. 3. On the other hand, leveraging all joints within the current frame provides necessary ingredient for determining their importance.

Bridged by the joint-selection gate, the main LSTM network and the spatial attention subnetwork can be jointly trained to implicitly learn the spatial attention model.

4.2 Temporal Attention with Frame-Selection Gate

For a sequence, the amount of valuable information provided by different frames is in general not equal. Only some of the frames (key frames) contain the most discriminative information while the other frames provide context information. For example, for the action "shaking hands", the sub-stage "approaching" should have lower importance than the sub-stage of "hands together". Based on such insight, we design a temporal attention module to automatically pay different levels of attention β to different frames.

For the sequence level classification, based on the output $\mathbf{z}_t$ of the main LSTM network and the temporal attention value β_t at each time step t, the scores for C classes are the weighted summation of the scores at all time steps

$$\mathbf{o} = \sum_{t=1}^{T} \beta_t \cdot \mathbf{z}_t, \tag{4}$$

where $\mathbf{o} = (o_1, o_2, \cdots, o_C)^{\mathrm{T}}$, T denotes the length of the sequence. Fig. 4 illustrates how the temporal attention output β is incorporated to the main LSTM network. The predicted probability being the i^{th} class given a sequence X is

$$p\left(C_i|X\right) = \frac{e^{o_i}}{\sum_{j=1}^{C} e^{o_j}}, \quad k = 1, ..., C. \tag{5}$$

As illustrated in Fig. 3, the attention module is composed of an LSTM layer, a fully connected layer, and a ReLU non-linear unit. It plays the role of soft frame selection. The activation as the frame-selection gate can be computed as

$$\beta_t = \mathrm{ReLU}(\mathbf{w}_{x\sim}\mathbf{x}_t + \mathbf{w}_{h\sim}\tilde{\mathbf{h}_{t-1}} + b_{\sim}), \tag{6}$$

which depends on the current input $\mathbf{x}_t$, and the hidden variables $\tilde{\mathbf{h}_{t-1}}$ of time step $t-1$ from an LSTM layer. We use the non-linear function of ReLU due to its good convergence performance. The gate controls the amount of information of each frame to be used for making the final classification decision. The works (Du, Wang, and Wang 2015; Zhu et al. 2016) are our special cases where the attention weights on each frame are equal.

Bridged by the frame-selection gate, the main LSTM network and the temporal attention subnetwork can be jointly trained to implicitly learn the temporal attention model.

4.3 Joint Spatial and Temporal Attention

To enable the network to pay different levels of attention to different joints and assign different degrees of importance to different frames as an action proceeds, we integrate spatial and temporal attention in the same network as illustrated in Fig. 3. How the spatial attention model acts on the input and how the temporal attention model acts on the output of the main LSTM network are illustrated in Fig. 4.

Regularized Objective Function We formulate the final objective function of the spatio-temporal attention network with a regularized cross-entropy loss for a sequence as,

$$\begin{aligned} L = &- \sum_{i=1}^{C} y_i \log \hat{y}_i + \lambda_1 \sum_{k=1}^{K} \left(1 - \frac{\sum_{t=1}^{T} \alpha_{t,k}}{T}\right)^2 \\ &+ \frac{\lambda_2}{T} \sum_{t=1}^{T} \|\beta_t\|_2 + \lambda_3 \|W_{uv}\|_1, \end{aligned} \tag{7}$$

where $\mathbf{y} = (y_1, \cdots, y_C)^{\mathrm{T}}$ denotes the groundtruth label. If it belongs to the i^{th} class, then $y_i = 1$ and $y_j = 0$ for $j \neq i$. $\hat{y}_i$ indicates the probability that the sequence is predicted as the i^{th} class, where $\hat{y}_i = p(C_i|X)$. The scalars λ_1, λ_2, and λ_3 balance the contribution of the three regularization terms. We discuss the regularization designs in the following.

The first regularization item aims to encourage the spatial attention model to dynamically focus on more spatial joints in a sequence. We found the spatial attention model is prone to consistently ignoring many joints along time even though these joints are also valuable for determining the type of action, i.e., trapped to a local optimum. We introduce this regularization item to avoid such ill-posed solutions. For clarity, we re-describe it as $\sum_{t=1}^{T} \alpha_{t,k} \approx T$, with $k = 1, \cdots, K$. This encourages paying equal attentions to different joints.

The second regularization item is to regularize the learned temporal attention values under control with l_2 norm rather than to increase them unboundedly. This alleviates gradient vanishing in the back propagation, where the back-propagated gradient is proportional to $1/\beta_t$.

The third regularization item with l_1 norm is to reduce overfitting of the networks. W_{uv} denotes the connection matrix (merged to one matrix here) in the networks.

Joint Training of the Networks Due to the mutual influence of the three networks, the optimization is rather difficult. We propose a joint training strategy to efficiently train the spatial and temporal attention modules, as well as the main LSTM network. The separate pre-training of the attention modules ensures the convergence of the networks. The training procedure is described in Algorithm 1.

Algorithm 1 Joint Training of the LSTM Network with Spatio-Temporal Attention Model.

Input: model training parameters N_1, N_2 (e.g., $N_1 = 1000$, $N_2 = 500$).

1: Initialize the network parameters using Gaussian.
// Pre-train Temporal Attention Model.
2: With spatial attention weights being fixed as ones, jointly train the main LSTM network with only one LSTM layer and the temporal attention subnetwork to obtain the temporal attention model.
3: Fix this learned temporal attention subnetwork. Train the main LSTM network after increasing its number of LSTM layers to three by N_1 iterations.
4: Fine-tune this temporal attention subnetwork and the main LSTM network by N_2 iteration.
// Pre-train Spatial Attention Model.
5: With temporal attention weights being fixed as ones, jointly train the main LSTM network with only one LSTM layer and the spatial attention subnetwork to obtain the spatial attention model.
6: Fix this learned spatial attention subnetwork. Train the main LSTM network after increasing its number of LSTM layers to three by N_1 iterations.
7: Fine-tune this spatial attention subnetwork and the main LSTM network for N_2 iterations.
// Train the Main LSTM Network.
8: Fix both the temporal and spatial attention subnetworks learned in Step-4 and Step-7. Fine-tune the main LSTM network by N_1 iterations.
// Jointly Train the Whole Network.
9: Jointly fine-tune the whole network (main LSTM network, the spatial attention subnetwork, and the temporal attention subnetwork) by N_2 iterations.

Output: the final converged whole model.

5. Experimental Results

5.1 Datasets and Settings

We perform our experiments on the following two datasets: the SBU Kinect interaction dataset (Yun et al. 2012), and the largest RGB+D dataset of NTU (Shahroudy et al. 2016).

SBU Kinect Interaction Dataset (SBU). The SBU dataset is an interaction dataset with two subjects. It contains 230 sequences of 8 classes (6614 frames) with subject independent 5-fold cross validation. Each person has 15 joints and the dimension of the input vector is $15 \times 3 \times 2 = 90$. Note that we smooth each joint's position of the skeleton in the temporal domain to reduce the influence of noise (Du, Wang, and Wang 2015; Zhu et al. 2016).

NTU RGB+D Dataset (NTU). The NTU dataset is currently the largest action recognition dataset with high quality skeleton (Shahroudy et al. 2016). It contains 56880 sequences (with 4 million frames) of 60 classes, including Cross-Subject (CS) and Cross-View (CV) settings. Each person has 25 joints. We apply the similar normalization preprocessing step to have position and view invariance (Shahroudy et al. 2016). To avoid destroying the continuity of a sequence, no temporal down-sampling is performed.

Implementation Details. For the network and parameter settings, we use three LSTM layers for the main LSTM network, and one LSTM layer for each attention network. Each LSTM layer composes of 100 LSTM neurons. We set λ_1, λ_2, and λ_3 to 0.001, 0.0001, and 0.0005 for the SBU dataset, and 0.01, 0.001 and 0.00005 for the NTU dataset experimentally. Adam (Kingma and Ba 2014) is adopted to automatically adjust the learning rate during optimization. The batch sizes for the SBU dataset and the NTU dataset are 8 and 256 respectively. Dropout is utilized to mitigate overfitting (Zaremba, Sutskever, and Vinyals 2014).

5.2 Visualization of the Learned Attentions

We analyze where the learned spatial and temporal attention attend to by visualizing the attention weights in the test.

Spatial Attention. For a sequence of action "kicking", Fig. 5(a) shows the amplitude of the spatial attention weights on joints by the sizes of the red circles. We also present concrete attention values in Fig. 6. The attention weights on the left foot, right elbow and left hand of the right person are large. Meanwhile, the weights on the torso and right foot of the left person are large. Being content-dependent, the attentions vary across frames. The learned important types of joints are consistent with what human perceives.

Temporal Attention. Fig. 5(b) shows the temporal attention weights β. Fig. 5(c) shows the differentiated attention weights (i.e., $\triangle\beta_t = \beta_t - \beta_{t-1}$) for "kicking". Since the LSTM network usually accumulates more information as time goes, the attention weight usually increases correspondingly. The increased amplitude of the attention weight, i.e., $\triangle\beta_t$, can indicate the importance of the frame t. We can see the differentiated attention weight goes up to a climax as the person on the right lifts his foot to the highest point, which human also considers as more discriminative.

5.3 Effectiveness of the Proposed Attention Models

To validate the effectiveness of our designs, we conduct experiments with different configurations as follows.

- **LSTM**: main LSTM network without attention designs.
- **SA-LSTM(w/o reg.)**: LSTM + spatial attention without regularization (only includes 1^{st} and 4^{th} items in (7)).

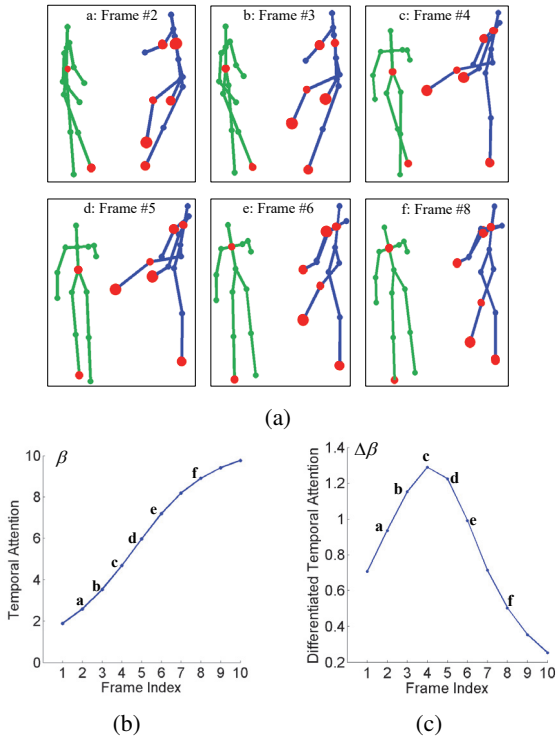


Figure 5: Visualization of the spatial and temporal attention weights from our model for the action "kicking". (a) Spatial attention weights. The larger of the red circle, the higher of the attention on that joint. We only mark on the 8 joints with the largest attentions. (b) Temporal attention weights β on each frames. (c) Differentiated temporal attention weights (i.e., $\triangle\ \beta_t = \beta_t - \beta_{t-1}$). Best viewed in color.

- **SA-LSTM**: LSTM + spatial attention network.
- **TA-LSTM(w/o reg.)**: LSTM + temporal attention without regularization(only includes 1^{st} and 4^{th} items in (7)).
- **TA-LSTM**: LSTM + temporal attention network.
- **STA-LSTM**: LSTM+spatio-temporal attention network.

Fig. 7 shows the performance comparisons on the SBU, NTU (Cross-Subject), NTU (Cross-View) datasets. With the baseline scheme LSTM, the introduction of the spatial attention module (SA-LSTM) and the temporal attention module (TA-LSTM) brings up to 5.1% and 6.4% accuracy improvement, respectively. The best performance is achieved by combining both modules (STA-LSTM). In the objective function as defined in (7), the second and the third items for regularizations are designed for the spatial and temporal attention model, respectively. We can see they improve the performance of both spatial attention model and temporal attention model.

5.5 Comparisons to Other State-of-the-Art

We show performance comparisons of our final scheme with the other state-of-the-art methods in Table 1 and Table 2 for the SBU and NTU datasets, respectively. Thanks to the introduction of the spatio-temporal attention models with efficient regularizations and the training strategy, our model is capable of extracting discriminative spatio-temporal features. We can see that our scheme achieves about 10% accuracy gain on the NTU dataset for the Cross-Subject and Cross-View settings, respectively.

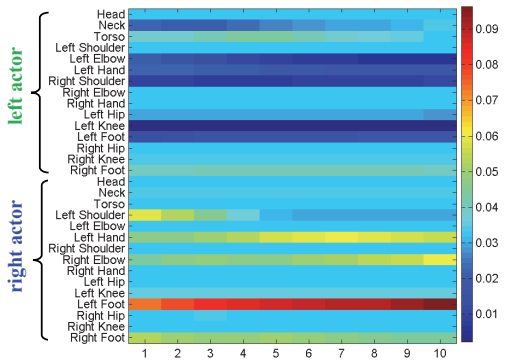


Figure 6: Visualization of spatial attention on the two actors of the action "kicking" for a sequence. Vertical axis denotes the joint indexes. Horizontal axis denotes the frame indexes (time). Color values indicate the spatial attention weights.

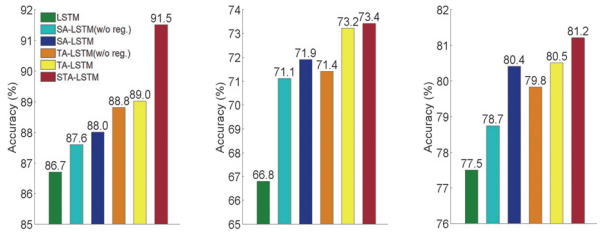


Figure 7: Performance evaluation of our attention models, and the regularization items on two datasets in accuracy (%).

Table 1: Comparisons on the SBU dataset in accuracy (%).

Methods	Acc. (%)
Raw skeleton (Yun et al. 2012)	49.7
Joint feature (Yun et al. 2012)	80.3
Raw skeleton (Ji, Ye, and Cheng 2014)	79.4
Joint feature (Ji, Ye, and Cheng 2014)	86.9
Hierarchical RNN (Du, Wang, and Wang 2015)	80.35
Co-occurrence RNN (Zhu et al. 2016)	90.41
STA-LSTM	**91.51**

6. Conclusion

We present an end-to-end spatio-temporal attention model for action recognition from skeleton data. To select dominant joints automatically and adaptively, we propose a spatial attention module with joint-selection gates to assign different importance to each joint. To automatically exploit the different levels of importance for each frame, we propose a temporal attention module to allocate different attention

Table 2: Comparisons on the NTU dataset with Cross-Subject and Cross-View settings in accuracy (%).

Methods	CS	CV
Lie Group (Vemulapalli et al. 2014)	50.1	52.8
Skeleton Quads (Evangelidis et al. 2014)	38.6	41.4
Dynamic Skeletons (Hu et al. 2015)	60.2	65.2
HBRNN (Du, Wang, and Wang 2015)	59.1	64.0
Deep LSTM (Shahroudy et al. 2016)	60.7	67.3
Part-aware LSTM (Shahroudy et al. 2016)	62.9	70.3
STA-LSTM	**73.4**	**81.2**

weights to each frame within a sequence. Finally, we design a joint training procedure to efficiently combine spatial and temporal attention with a regularized cross-entropy loss. Experimental results show the effectiveness of our proposed model which achieves remarkable performance in comparison with other state-of-the-art methods.

References

Ba, J.; Mnih, V.; and Kavukcuoglu, K. 2014. Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bazzani, L.; Larochelle, H.; and Torresani, L. 2016. Recurrent mixture density network for spatiotemporal visual attention. *arXiv preprint arXiv:1603.08199*.

Carlsson, S., and Sullivan, J. 2001. Action recognition by shape matching to key frames. In *Workshop on Models versus Exemplars in Computer Vision*.

Du, Y.; Wang, W.; and Wang, L. 2015. Hierarchical recurrent neural network for skeleton based action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1110–1118.

Goferman, S.; Zelnik-Manor, L.; and Tal, A. 2012. Context-aware saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34(10):1915–1926.

Graves, A. 2012. Supervised sequence labelling with recurrent neural networks. *Volume 385 of Studies in Computational Intelligence*.

Han, F.; Reily, B.; Hoff, W.; and Zhang, H. 2016. Space-time representation of people based on 3D skeletal data: A review. *arXiv preprint arXiv:1601.01006*.

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

Hochreiter, S.; Bengio, Y.; Frasconi, P.; and Schmidhuber, J. 2001. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. A field guide to dynamical recurrent neural networks. IEEE Press.

Hu, J.-F.; Zheng, W.-S.; Lai, J.; and Zhang, J. 2015. Jointly learning heterogeneous features for RGB-D activity recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Ji, Y.; Ye, G.; and Cheng, H. 2014. Interactive body part contrast mining for human interaction recognition. In *IEEE International Conference on Multimedia and Expo Workshops*.

Jiang, M.; Xu, J.; and Zhao, Q. 2014. Saliency in crowd. In *European Conference on Computer Vision*.

Johansson, G. 1973. Visual perception of biological motion and a model for it is analysis. *Perception and Psychophysics* 14(2):201–211.

Kingma, D., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Liu, L.; Shao, L.; and Rockett, P. 2013. Boosted key-frame selection and correlated pyramidal motion-feature representation for human action recognition. *Pattern Recognition* 46(7):1810–1818.

Poppe, R. 2010. A survey on vision-based human action recognition. *Image and Vision Computing* 28(6):976–990.

Ramanathan, V.; Huang, J.; Abu-El-Haija, S.; Gorban, A.; Murphy, K.; and Li, F.-F. 2015. Detecting events and key actors in multi-person videos. *arXiv preprint arXiv:1511.02917*.

Raptis, M., and Sigal, L. 2013. Poselet key-framing: A model for human activity recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*.

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

Sharma, S.; Kiros, R.; and Salakhutdinov, R. 2015. Action recognition using visual attention. *arXiv preprint arXiv:1511.04119*.

Shotton, J.; Fitzgibbon, A.; Cook, M.; Sharp, T.; Finocchio, M.; Moore, R.; Kipman, A.; and Blake, A. 2011. Real-time human pose recognition in parts from a single depth image. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Simonyan, K., and Zisserman, A. 2014. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*.

Sminchisescu, C.; Kanaujia, A.; and Metaxas, D. 2006. Conditional models for contextual human motion recognition. *Computer Vision and Image Understanding* 104(2):210–220.

Vemulapalli, R.; Arrate, F.; and Chellappa, R. 2016. R3DG features: Relative 3D geometry-based skeletal representations for human action recognition. *Computer Vision and Image Understanding*.

Wang, J.; Liu, Z.; and Yuan, J. 2012. Mining actionlet ensemble for action recognition with depth cameras. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Weinland, D.; Ronfard, R.; and Boyerc, E. 2011. A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding* 115(2):224–241.

Wu, Z.; Wang, X.; Jiang, Y.-G.; Ye, H.; and Xue, X. 2015. Modeling spatial-temporal clues in a hybrid deep learning

framework for video classification. In *ACM International Conference on Multimedia*.

Xia, L.; Chen, C.-C.; and Aggarwal, J. K. 2012. View invariant human action recognition using histograms of 3D joints. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

Xu, K.; Ba, J.; Kiros, R.; Courville, A.; Salakhutdinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference for Machine Learning*.

Yeung, S.; Russakovsky, O.; Jin, N.; Andriluka, M.; Mori, G.; and Li, F.-F. 2015. Every moment counts: Dense detailed labeling of actions in complex videos. *arXiv preprint arXiv:1507.05738*.

Yu, G.; Liu, Z.; and Yuan, J. 2014. Discriminative orderlet mining for real-time recognition of human-object interaction. In *Asian Conference on Computer Vision*. Springer.

Yu, Y.; Mann, G. K.; and Gosine, R. G. 2010. An object-based visual attention model for robotic applications. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 40(5):1398–1412.

Yun, K.; Honorio, J.; Chattopadhyay, D.; Berg, T. L.; and Samaras, D. 2012. Two-person interaction detection using body-pose features and multiple instance learning. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

Zaremba, W.; Sutskever, I.; and Vinyals, O. 2014. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*.

Zhang, Z. 2012. Microsoft kinect sensor and its effect. *IEEE MultiMedia* 19(2):4–10.

Zhao, Z., and Elgammal, A. 2008. Information theoretic key frame selection for action recognition. In *British Machine Vision Conference*.

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 *AAAI Conference on Artificial Intelligence*.