Predicting Human Interaction via Relative Attention Model

Yichao Yan, Bingbing Ni, Xiaokang Yang
Shanghai Jiao Tong University, Shanghai, China
{yanyichao, nibingbing, xkyang}@sjtu.edu.cn

Abstract

Predicting human interaction is challenging as the on-going activity has to be inferred based on a partially observed video. Essentially, a good algorithm should effectively model the mutual influence between the two interacting subjects. Also, only a small region in the scene is discriminative for identifying the on-going interaction. In this work, we propose a relative attention model to explicitly address these difficulties. Built on a tri-coupled deep recurrent structure representing both interacting subjects and global interaction status, the proposed network collects spatio-temporal information from each subject, rectified with global interaction information, yielding effective interaction representation. Moreover, the proposed network also unifies an attention module to assign higher importance to the regions which are relevant to the on-going action. Extensive experiments have been conducted on two public datasets, and the results demonstrate that the proposed relative attention network successfully predicts informative regions between interacting subjects, which in turn yields superior human interaction prediction accuracy.

1 Introduction

Action prediction is defined as the problem of recognizing on-going activities based on temporally incomplete observations. It is a challenging task as only a part of the video is available for observation. Compared to individual action prediction, human interaction prediction is even harder, because the activities are more complex and involve more actors in the scene. More importantly, the incoming action of a subject might depend on the intention of the other subject, and this intention has to be inferred based on certain movement of this subject. In other words, to predict interaction, a good model should understand one subject's current action and how it will affect the other's response to this action in the near future.

Despite significant progress in the past few years, human interaction prediction is still challenging mainly due to the following two unanswered questions. The first one is how to model interaction or relative information. Second is how to discover the most discriminative regions and make use of them to make prediction. Solving these two difficulties will always bring performance gain over holistic or global feature learning methods.

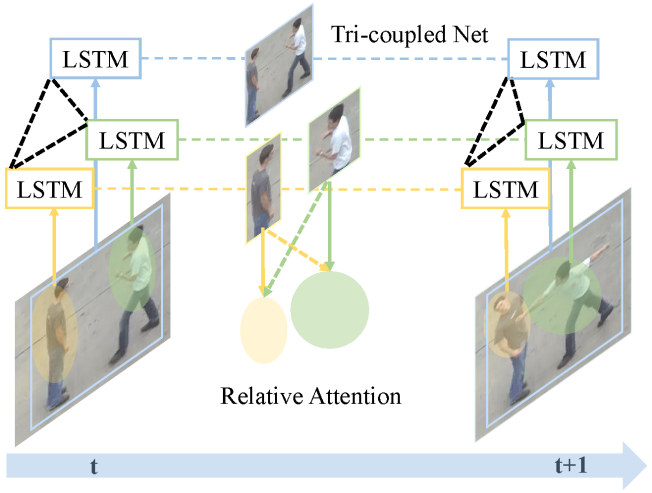


Figure 1: Overview of our framework. We design a tri-coupled deep recurrent structure representing both interacting subjects and global interaction status, and embed an attention module to predict the discriminative regions of each subject.

However, previous methods do not address these questions in a proper way. Recent methods mainly resort to: (1) holistic representation [Cao *et al.*, 2013; Ryoo, 2011; Kong *et al.*, 2014]; (2) individual representation [Lan *et al.*, 2014] and (3) discriminative part based representation [Xu *et al.*, 2015; Chen, 2015]. Despite their favorable performance on recent benchmark datasets [Ryoo and Aggarwal, 2010; Kong *et al.*, 2012], we have the following observations on their limitations. First, holistic feature based methods [Cao *et al.*, 2013; Ryoo, 2011; Kong *et al.*, 2014] usually encode the whole scene into a global feature vector, the richer information contained in individual subjects is ignored. Second, although individual representation based method [Lan *et al.*, 2014] models both interactive subjects, they are usually modeled separately. How to effectively model their relationship is not well exploded. Third, discriminative part based methods [Xu *et al.*, 2015; Chen, 2015] try to select discriminative patches/parts to represent the actions. Such discriminative patch/part detectors usually apply to the video frame-by-frame, thus the detected patches/parts are not temporally consistent. Moreover, such

methods are hard to distinguish similar movements, as the generated patches are also similar.

To explicitly address the above issues, we propose a tri-coupled relative attention framework. On one hand, a **tri-coupled interaction fusion network** is proposed to model mutual influence between subjects involved in the interaction. This network is composed of three recurrent substructures, which accept three streams of information representing both interacting subjects and the global interaction region enclosing both subjects. To capture the dependency between subjects, at each time-step, information flows from all three streams are aggregated to the hidden node of the current time-step, and then output the new status information for both interacting subjects. We make two remarks. First, we denote it by *coupled recurrent network* because status information of one subject is linked to the other stream, in order to assist the prediction of the next status of the other subject. Second, information extracted from the global scene (which encloses both subjects) is also utilized to predict the interaction status of both subjects. In this way, both local motion information and global motion information are fully utilized, which are complements to each other. On the other hand, built on this tri-coupled recurrent infrastructure, we introduce a **relative attention network**. The motivation is that some local motion (attended small regions) might give very useful information to predict the other subject's response in the future. For example, if a person extends his arm or leg, another person is likely to dodge, a punching/kicking is more likely to happen. In this situation, the arm/leg region is crucial for predicting another person's response. Motivated by this observation, a visual attention module is embedded to the recurrent structure to predict the discriminative regions of each subject. At each time-step, the attention module receives information from both interactive subjects, as well as their hidden states of previous time-step, and then output the attended regions of both subjects. In this way, only the attended regions are input into the recurrent networks, providing discriminative local information.

The proposed network is extensively compared with some popular methods for encoding human interaction on two popular datasets, the results of the proposed method show favorable performance against the state-of-the-art methods.

2 Related Work

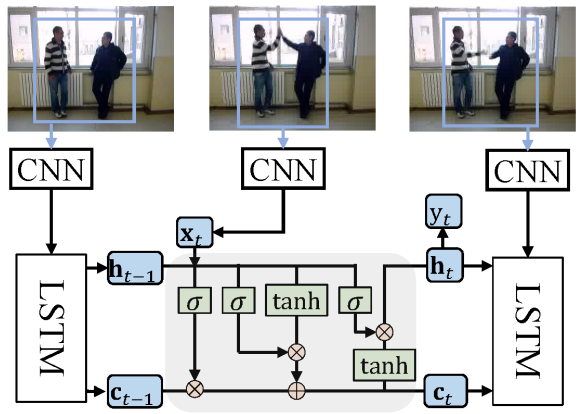
Traditional methods. For action prediction, many previous works focus on finding good feature representation (usually bag-of-words features or sparse coding) and training SVM-like classifiers. For example, Ryoo [2011] proposes two BoW based representation, i.e., the integral bag-of-words (IBoW) and dynamic bag-of-words (DBoW). Cao et al. [2013] apply sparse coding to derive the activity bases, and use the reconstruction error in the likelihood computation. Lan et al. [2014] propose a hierarchical representation and combine it with a max-margin learning framework for action prediction. Another two max-margin frameworks [Kong *et al.*, 2014; Nguyen and la Torre, 2012] are built upon structured SVM model, but extend it to accommodate sequential data. Kong and Fu [2016] further extend this framework using compositional kernels to model the relationship of partial observations. Xu et al. [2015] consider action prediction as a query auto-completion problem. These methods use hand-crafted features and encoding methods to represent the video. The difference of our work lies in the using CNN/LSTM features rather than hand-crafted features, which enables our model to be trained end-to-end.

CNN based methods. Many CNN based methods have been focused on activity recognition and video classification. In [Ji *et al.*, 2013], a 3D CNN model is proposed for action recognition. Karpathy et al. [2014] explore several approaches for fusing information over temporal dimension trough the CNN, but only achieving marginal improvement than the single frame baseline, which indicates that learning motion information is difficult for CNN. To address this issue, Simonyan and Zisserman [2014] propose a two-stream architecture which directly incorporate motion information from optical flows, achieving significant improvements compared to previous CNN based methods. However, such approaches are based on single frames, not able to represent long-term temporal clue.

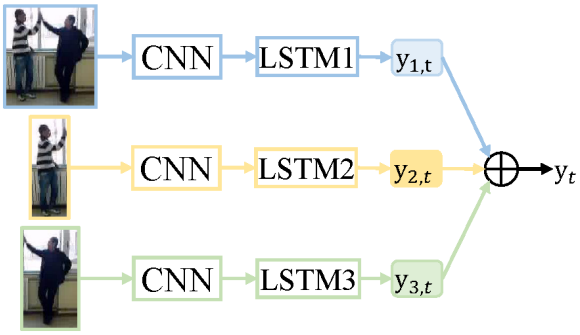
RNN based methods. Recurrent neural network (RNN) and Long Short Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997] are powerful tools to model sequential data. LSTMs have been applied to action classification in [Baccouche *et al.*, 2010; Donahue *et al.*, 2015]. The work of [Wu *et al.*, 2015; Ng *et al.*, 2015] further improves the performance by building a hybrid model incorporating both spatial and temporal clue. Ibrahim et al. [2016] build a 2-stage deep temporal model for group activity recognition. Ma et al. [2016] design novel ranking losses for training LSTM which enforce the score margin between the correct and incorrect categories to be monotonically non-decreasing. Visual attention model is also investigated for action recognition in [Sharma *et al.*, 2015]. Song et al [2016] build a spatio-temporal attention model from skeleton data. These works mainly focus on recognizing action of a single object or group activity, they achieve promising result when the complete video is observed. In contrast, our framework is explicitly designed for person interaction involving a pair of persons in the scene, and it still achieves satisfactory results when only a small part of the video is observed.

3 Methodology

The problem is formulated as follows. We denote a complete video of duration T as $V[1:T]$, the task is to predict the action y with only partial observation $V[1:t], t \in \{1, ..., T\}$, the observation ratio is $\frac{t}{T}$. The complete videos are only accessible for training, and the performance is evaluated by calculating the prediction accuracy with a fixed observation ratio for all the test videos. In this work, we assume the bounding box of each person and the global scene enclosing the two actors are located in each frame. In the rest of this section, we use $\mathbf{X}$ to denote the CNN feature extracted from raw frames. $\mathbf{L}$ denotes the attention weights corresponding to the attended region. The inputs and hidden states of LSTM network are denoted as $\mathbf{x}$ and $\mathbf{h}$ respectively. The weights and bias terms in our networks are denoted as $\mathbf{W}, \mathbf{U}, \mathbf{V}$ and $\mathbf{b}$.



(a) Global LSTM network.



(b) Naive fusion method.

Figure 2: Two baseline methods for action prediction.

3.1 Tri-coupled Interaction Fusion Network

With an LSTM network, information could be propagated from the first node to the last one, and the good nature of LSTM is very useful for our given task, i.e., to make full use of the observed information and make a prediction. Motivated by the success of recurrent neural networks in temporal sequence analysis, we employ LSTM network as our network prototype. The frame-level features are input into LSTMs to model the spatio-temporal information.

In particular, each LSTM node includes three gates, (i.e., the input gate $\mathbf{i}$, the output gate $\mathbf{o}$ and the forget gate $\mathbf{f}$) as well as a memory cell. At each time-step t, the input feature $\mathbf{x}_t$ and the previous hidden state $\mathbf{h}_{t-1}$ are input into the LSTM, as illustrated in Figure 2(a). The LSTM network updates as follows:

$$\mathbf{i}_t = \sigma(\mathbf{W}_i\mathbf{x}_t + \mathbf{U}_i\mathbf{h}_{t-1} + \mathbf{V}_i\mathbf{c}_{t-1} + \mathbf{b}_i) \quad (1)$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_f\mathbf{x}_t + \mathbf{U}_f\mathbf{h}_{t-1} + \mathbf{V}_f\mathbf{c}_{t-1} + \mathbf{b}_f) \quad (2)$$

$$\mathbf{c}_t = \mathbf{f}_t \cdot \mathbf{c}_{t-1} + \mathbf{i}_t \cdot \tanh(\mathbf{W}_c\mathbf{x}_t + \mathbf{U}_c\mathbf{h}_{t-1} + \mathbf{b}_c) \quad (3)$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_o\mathbf{x}_t + \mathbf{U}_o\mathbf{h}_{t-1} + \mathbf{V}_o\mathbf{c}_t + \mathbf{b}_o) \quad (4)$$

$$\mathbf{h}_t = \mathbf{o}_t \cdot \tanh(\mathbf{c}_t) \quad (5)$$

where σ is the sigmoid function and $\cdot$ denotes the element-wise multiplication operator. $\mathbf{W}_*$, $\mathbf{U}_*$ and $\mathbf{V}_*$ are the weight matrices, and $\mathbf{b}_*$ are the bias vectors. The memory cell $\mathbf{c}_t$ is a weighted sum of the previous memory cell $\mathbf{c}_{t-1}$ and a function of the current input. The weights are the activations of forget gate and input gate respectively.

For the task of interaction prediction, the most straightforward idea is to model the global interaction regions enclosing both subjects with a single LSTM network, as other activity recognition system [Donahue *et al.*, 2015]. We denote it a **global LSTM network**, which takes the complete region of action as input and models the global information of the observed video. As shown in Figure 2(a), the frame-level features are extracted by a CNN extractor, and then input into the LSTM network for classification. Here, we use Alexnet [Krizhevsky *et al.*, 2012] as CNN feature extractor. The good nature of this structure is that all the information is modeled by a global LSTM, which is simple and effective for action recognition. However, the interaction of individual subjects is not explicitly modeled in the structure, the performance might be limited for the task of interaction prediction.

There are multiple options to model the mutual interactions of the interactive subjects. A naive approach is to model each subject with an individual LSTM model and then combine their predictions, which can be further enhanced by employing the prediction of the global LSTM network. We denote this structure as **naive fusion network**, see Figure 2(b). This structure employs both global and local interactive information, but it also suffers from a major limitation. Some subjects are likely to have very similar behaviours in different interactions, e.g., the *dodge* action in both *kick* and *box*. The prediction scores of these subjects can be very confusing, directly summing up their prediction scores may bring side effects to the overall results.

To address this issue, we design a joint training scheme that simultaneously models the interactive state of the two subjects. In particular, each subject is also represented by an LSTM network, but the hidden states of the two LSTMs are shared at each time-step. In this case, the terms $\mathbf{U}_*\mathbf{h}_{t-1}$ in Equation 1 to Equation 5 are further represented by:

$$\mathbf{U}_*\mathbf{h}_{t-1} = \mathbf{U}_{*,s_1}\mathbf{h}_{t-1,s_1} + \mathbf{U}_{*,s_2}\mathbf{h}_{t-1,s_2}, \quad (6)$$

where $\mathbf{h}_{t-1,s_1}$ is the previous hidden state of the network and $\mathbf{h}_{t-1,s_2}$ is the previous hidden state of the other subject. This enables the information communication between the subjects, i.e., the statues of one subject can be used to help predict the action of the other subject. Moreover, the outputs of the LSTMs are concatenated as a union feature for prediction, which is in contrast of combing the prediction scores of individual subject level LSTMs. This structure allows the two the LSTMs to be trained together, i.e., there is a single loss for the networks. We denote it a **coupled network**, see the top part of Figure 3.

Although the coupled network explicitly models the spatio-temporal correlations of the two subjects, the global interactive information is not used in the structure. To integrate the global information into the network, we design a **Tri-coupled interaction fusion network**, as shown in the middle part of Figure 3. For the tri-couple structure, the LSTM representing the global interaction status is pre-trained as a vanilla LSTM, the other two LSTMs modeling the mutual interactions are modeled as:

$$\mathbf{U}_*\mathbf{h}_{t-1} = \mathbf{U}_{*,s_1}\mathbf{h}_{t-1,s_1} + \mathbf{U}_{*,s_2}\mathbf{h}_{t-1,s_2} + \mathbf{U}_{*,g}\mathbf{h}_{t,g}, \quad (7)$$

where $\mathbf{h}_{t,g}$ is the hidden state of global LSTM.

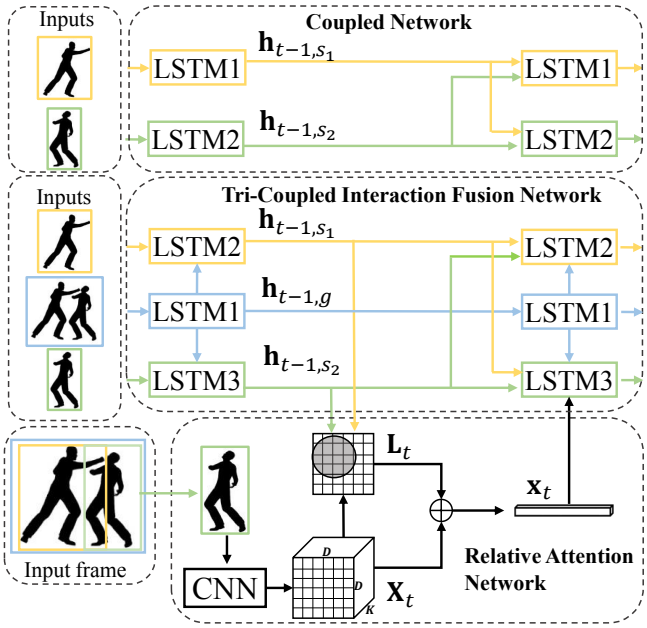


Figure 3: Illustration of the Coupled Network, Tri-coupled Recurrent Network and the Relative Attention Network.

3.2 Relative Attention Network

For the task of action prediction, usually only a certain region is crucial for identifying an action. Therefore, we would like our model to focus on these regions and to model the fine-grained details. Here, we embed our tri-coupled network with a relative attention module.

Two kinds of attention model have been used to address this issue. Hard attention [Mnih *et al.*, 2014; Ba *et al.*, 2014] samples attention location at each time stamp, which causes the system not differentiable. In contrast, soft attention [Bahdanau *et al.*, 2014; Sharma *et al.*, 2015] aims to learn a set of weights corresponding to each region, the model is differentiable and can be trained end-to-end using standard back-propagation. Therefore, we adopt the soft attention model in our work. Instead of extracting feature from the last fully connected layer, the soft attention model employs the last convolutional layer, resulting to K convolutional maps of size $D * D$, which can be denoted as:

$$\mathbf{X}_t = \{\mathbf{X}_{t,1}, ..., \mathbf{X}_{t,D^2}\}, \qquad \mathbf{X}_{t,i} \in \mathbb{R}^K. \tag{8}$$

Specially, each vector $\mathbf{X}_{t,i}$ corresponds to a specific receptive field in the original image.

At each time-step t, we would like to assign weights to each location in the $D * D$ feature map. The attended region should have higher weights compared to less important regions. As each location in the feature maps corresponds to a certain receptive field in the original image, attending to the feature map plays the same role as attending to the original image. The attention weights $\mathbf{L}_t = \{l_{t,1}, ..., l_{t,D^2}\}$ at time-step t is usually calculated using the following two features: the hidden state of the previous time-step $\mathbf{h}_{t-1}$ and the CNN feature map of the current time-step $\mathbf{X}_t$. See Figure 3. The weights are normalized after a softmax layer:

$$l_{t,i} = \frac{\exp\left(\mathbf{W}_{h,i}\mathbf{h}_{t-1} + \mathbf{W}_{X,i}\mathbf{X}_t\right)}{\sum_{j=1}^{D*D} \exp\left(\mathbf{W}_{h,j}\mathbf{h}_{t-1} + \mathbf{W}_{X,j}\mathbf{X}_t\right)}, \tag{9}$$

where $i \in 1, ..., D^2$ and $\mathbf{W}_{*,i}$ are the weights for the inputs. $l_{t,i}$ can be viewed as the probability of the i-th region to be important. For the tri-coupled network, we can also take advantage of the mutual information to help locate the interesting region, i.e., to use the hidden states of neighboring LSTMs. The $\mathbf{W}_{h,i}\mathbf{h}_{t-1}$ term in Equation 9 can be further decomposed into hidden state information from both subjects:

$$\mathbf{W}_{h,i}\mathbf{h}_{t-1} = \mathbf{W}_{h,i,s_1}\mathbf{h}_{t-1,s_1} + \mathbf{W}_{h,i,s_2}\mathbf{h}_{t-1,s_2}. \tag{10}$$

The final inputs for LSTM is a weighted summation of the attention vector $\mathbf{L}_t$ and the CNN features $\mathbf{X}_t$:

$$\mathbf{x}_t = \sum_{i=1}^{D^2} l_{t,i}\mathbf{X}_{t,i}.$$

3.3 Training the Network

The proposed tri-coupled network and relative attention network can be jointly trained as a classification problem of N classes (N is the number of human interaction category). At each time-step, the hidden state $\mathbf{h}_t$ of each sub-LSTM is concatenated as the feature representation vector, which is further connected to a softmax layer. The output of the N-way softmax is the prediction of the probability distribution over N different actions:

$$\mathrm{y_i} = \frac{\exp(\mathrm{y'_i})}{\sum_{k=1}^{N} \exp(\mathrm{y'_k})}, \tag{11}$$

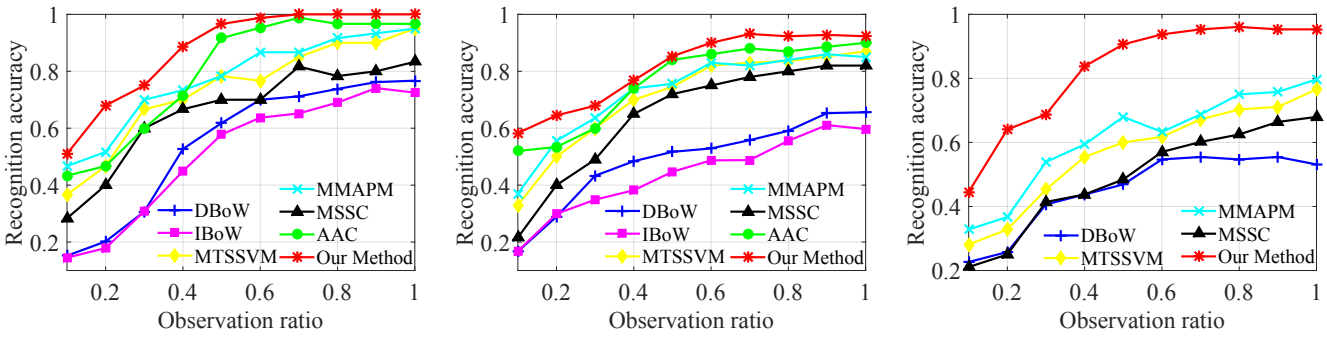
where $\mathrm{y'_j} = \mathbf{w_j} \cdot \mathbf{h_t} + \mathrm{b_j}$ linearly combines the LSTM outputs, and $\mathbf{w}$ and b are the weight matrix and bias term of the softmax layer. The network is learned by minimizing $-\log \mathrm{y_k}$, where k is the index of the true label for a given input. Stochastic gradient descent is used with gradients calculated by back-propagation.

4 Experiments

In this section, we present extensive experimental evaluations and in-depth analysis of the proposed method on the following two human interaction prediction benchmarks:

UT dataset. The UT-Interaction dataset (UTI) [Ryoo and Aggarwal, 2010] contains videos of 6 classes of human-human interactions: shake-hands, point, hug, push, kick and punch. Except that point is a single action, all other activities are performed by a pair of actors. This dataset contains two subsets: UTI #1 and UTI #2. The backgrounds of UTI #1 are mostly static with little camera jitter, while the backgrounds of UTI #2 are moving slightly and containing more camera jitters. Both of the two subsets contain 10 videos of each interaction class. We adopt 10-folder leave-one-out cross validation setting to measure the performance of the two subsets.

BIT dataset. The BIT dataset [Kong *et al.*, 2012] contains 8 types of interactions: bend, box, handshake, hifive, hug, kick, pat and push, all the activities are performed by a pair of actors. Each activity contains 50 video sequences, i.e., totally 400 videos in the dataset. Following [Kong *et al.*, 2012], a random subset containing 272 videos is used for training, and the remaining 128 videos are used for testing.



(a) UTI # 1 (b) UTI # 2 (c) BIT

Figure 4: Prediction results on UTI #1, #2, and BIT dataset. Our method means the result achieved by our tri-coupled relative attention network on top of optical flows.

Table 1: Activity prediction performance on UTI #1 dataset

Methods	OR=0.5	OR=1
Our method	**96.7%**	**100%**
DBoW [Ryoo, 2011]	70.0%	85.0%
IBoW [Ryoo, 2011]	65.0%	81.7%
MTSSVM [Kong *et al.*, 2014]	78.33%	95.00%
MMAPM [Kong and Fu, 2016]	78.33%	95.00%
MSSC [Cao *et al.*, 2013]	70.0%	83.3%
AAC [Xu *et al.*, 2015]	91.67%	96.67%

4.1 Implementation Details

The implementation of the proposed networks are based on Caffe [Jia *et al.*, 2014]. The LSTM layer contains 512 hidden units, and a dropout layer is placed after it to avoid overfitting. To increase training instances and to make our model applicable for sequences of variable length, we randomly extract subsequences of fixed length L ($L = 10$ in our experiments) for training. To train the LSTM networks, the original learning rate is initialized as 0.001, and the learning rate is decreased to $\frac{1}{10}$ of the original value after each 10 epochs. The whole training phase includes 30 epochs. The complete duration of training time is about 12 hours on a Titan X GPU. During testing, we extract the subsequences in the testing video with a stride of 5, and averaging their classification score as prediction. We test our network on top of both RGB frames and optical flows. The optical flow is computed using the approach of [Brox *et al.*, 2004]. As *point* action in the UTI dataset is a single action, we duplicate the image as input for the networks that require both subjects, i.e., the naive fusion network, the coupled network and the tri-coupled network.

4.2 Results on UTI Dataset

The proposed tri-coupled relative attention network is compared with some leading approaches on interaction prediction. (1) Bag-of-words based methods: DBow and IBoW [Ryoo, 2011]; (2) Sparse coding based method: MSSC [Cao *et al.*, 2013]; (3) Max margin structure SVM based methods: MTSSVM [Kong *et al.*, 2014] and MMAPM [Kong and Fu, 2016]; and (4) discriminative patch based method: AAC [Xu *et al.*, 2015]. The comparative results on UTI #1 is shown in Figure 4(a), and the quantitative results with observation ratio 0.5 and 1 are shown in Table 1. We report our best performance with tri-coupled relative attention network on top of optical flow inputs. Our method achieves favorable performance compared to other methods. It's remarkable that our tri-coupled structure achieves 100% recognition accuracy when the observation ratio is larger than 0.6. This is better than the previous state-of-the-art method [Xu *et al.*, 2015], which also achieves remarkable performance on this dataset, i.e., 91.67% and 96.67% for half video and full video. We further notice that our results are significantly higher than DBow, IboW [Ryoo, 2011] and other encoding based models. This is mainly because that tri-coupled network explicitly employs the interactive information, while most other methods only rely on the global information.

Comparative results on UTI #2 are displayed in Figure 4(b). We notice that other methods have significant lower prediction accuracies compared to the results on UTI #1, due to more complex backgrounds and more camera jitter. Even the discriminative patch based method AAC [Xu *et al.*, 2015] suffers from about 10% decrease. Compared to these methods, our tri-coupled relative attention model achieves better performance, more than 90% prediction accuracies when observation ratio is larger than 0.5, which is higher than other methods. This well demonstrates the robustness of the proposed method in existence of noise, and it is mainly due to our relative attention network, which is able to attend to discriminative regions on each interactive subject.

Component analysis. Our framework consists of two major components: the tri-coupled network and the relative attention network. To evaluate the effectiveness of each component, we compare our network with some baseline structures introduced in Section 3.1: (1) global LSTM network; (2) naive fusion network; (3) coupled network; and (4) tri-coupled network without relative attention. The results on UTI #2 dataset with both RGB inputs and optical flow inputs are shown in Table 2, we have three observations. First, our baseline networks with optical flows achieves much better performance than the baseline methods using RGB frames. This is mainly because that the motion information contained in optical flows is crucial for identifying the actions. Second, we note that the naive fusion method only achieves marginally increase to the performance compared to global LSTM network, for both optical flows and RGB frames. This is because that some motion patterns of individual subjects can be

Table 2: Interaction prediction accuracies on UTI #2 dataset with different observation ratio (OR).

Methods	RGB inputs				Optical flows			
	OR=0.1	OR=0.2	OR=0.5	OR=1	OR=0.1	OR=0.3	OR=0.5	OR=1
Global LSTM	16.7%	33.3%	51.7%	63.3%	16.7%	58.3%	70.0%	76.7%
Naive fusion	23.3%	38.3%	58.3%	68.3%	30.0%	61.7%	71.7%	80.0%
Coupled network	16.7%	33.3%	50.0%	58.3%	21.7%	58.7%	68.3%	75.0%
Tri-coupled network	35.0%	48.3%	65.0%	73.3%	53.3%	65.0%	81.7%	90.0%
Our method	**36.7%**	**53.3%**	**71.7%**	**76.7%**	**58.3%**	**68.3%**	**85.0%**	**91.3%**

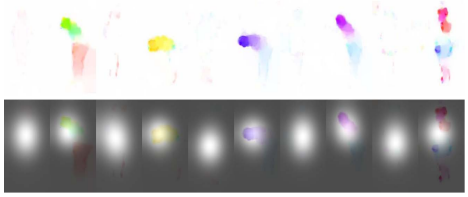
very similar though different interactions, which may even provide negative information for prediction. e.g., the dodge motion occurs in both kick and punch, it will be difficult to make a prediction when observing such pattern. Last but not least, the tri-coupled network brings significant performance gain to the above baseline methods, especially in the case of high observation ratios. When embedded with the relative attention network, the performance is further improved. This demonstrates the effectiveness of the proposed tri-coupled network as well as the relative attention network.

4.3 Result on BIT Dataset

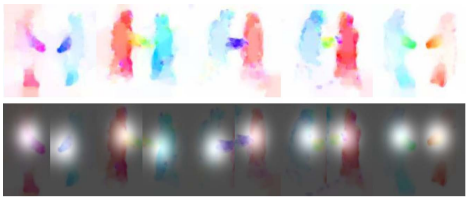
The results on BIT dataset are shown in Figure 4(c). All the other methods get worse results compared to the results on UTI datasets, due to the fact that BIT dataset contains more category of interactions, and the videos in this dataset are with more complex backgrounds and sometimes with heavy occlusion. Therefore, the compared methods [Ryoo, 2011; Kong and Fu, 2016; Cao *et al.*, 2013; Kong *et al.*, 2014] achieve less than 80% prediction accuracies even with full observation, which is far away from real-world applications. While our network achieves more than 90% accuracy when only only half the video is observed, which outperforms the compared methods by a large margin (more than 10% with any observation ratio). This is because of the effectiveness of the relative attention module, which is able to attend to the discriminative regions in the scene, thus make the proposed method more robust to occlusions.

4.4 Qualitative Results

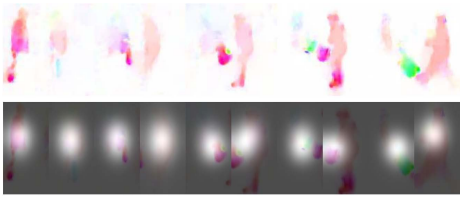
Figure 5 visualizes the attended regions generated by our relative attention model. As the best performance of our method is achieved upon optical flows, the visualization is based on optical flow images. The first example illustrates the interaction of *bend*. It's easy to notice that the major subject is on the right side, and the subject on the left nearly has no movement during the interaction. For the subject on the right, we can find very strong correlation between the attended regions and the movements of the upper part of the body. The second example depicts two people shaking their hands. Both subjects are involved during the interaction, and they share similar behaviours: stepping forward and reaching out their hands. Our model consistently attends to the arms of both subjects, which shares similar intuition of human cognition. In the last example, the subject on the left is kicking the right subject. The attended regions are focused on the extended leg of the left subject, and the upper body of the right subject is attended to as he/she falls down.



(a) Bow



(b) Handshake



(c) Kick

Figure 5: Examples of attended regions on optical flows.

5 Conclusion

In this paper, we propose a tri-coupled relative attention network for human interaction prediction. Experimental results convincingly demonstrate that the proposed relative attention network successfully predicts informative regions between interacting subjects, which in turn yields superior human interaction prediction accuracy. Although this paper is explicitly designed to model two subject interaction, our method is easily extendable to model group people interaction. Here is a brief illustration for this generalization. For each subject, the relative attention could be calculated with his/her nearest neighbors. The computational complexity only increases linearly w.r.t. the number of neighboring subjects. Finally, we can aggregate all groups via a LSTM structure to achieve global group activity label prediction.

Acknowledgments

The work was supported by State Key Research and Development Program (2016YFB1001003). This work was partly supported by National Natural Science Foundation of China (NSFC61502301, NSFC61521062), China's Thousand Youth Talents Plan, the 111 Project (B07022) and the Shanghai Key Laboratory of Digital Media Processing and Transmissions.

References

[Ba *et al.*, 2014] Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. Multiple object recognition with visual attention. *CoRR*, abs/1412.7755, 2014.

[Baccouche *et al.*, 2010] Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt. Action classification in soccer videos with long short-term memory recurrent neural networks. In *ICANN*, 2010.

[Bahdanau *et al.*, 2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2014.

[Brox *et al.*, 2004] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow estimation based on a theory for warping. In *ECCV*, 2004.

[Cao *et al.*, 2013] Yu Cao, Daniel Paul Barrett, Andrei Barbu, Siddharth Narayanaswamy, Haonan Yu, Aaron Michaux, Yuewei Lin, Sven J. Dickinson, Jeffrey Mark Siskind, and Song Wang. Recognize human activities from partially observed videos. In *CVPR*, 2013.

[Chen, 2015] Jia-Lin Chen. Weakly supervised learning of part-based models for interaction prediction via LDA. In *ACM MM*, 2015.

[Donahue *et al.*, 2015] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Trevor Darrell, and Kate Saenko. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015.

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

[Ibrahim *et al.*, 2016] Mostafa S. Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *CVPR*, 2016.

[Ji *et al.*, 2013] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *TPAMI*, 35(1):221–31, 2013.

[Jia *et al.*, 2014] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[Karpathy *et al.*, 2014] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Fei-Fei Li. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[Kong and Fu, 2016] Yu Kong and Yun Fu. Max-margin action prediction machine. *TPAMI*, 38(9):1844–1858, 2016.

[Kong *et al.*, 2012] Yu Kong, Yunde Jia, and Yun Fu. Learning human interaction by interactive phrases. In *ECCV*, 2012.

[Kong *et al.*, 2014] Yu Kong, Dmitry Kit, and Yun Fu. A discriminative model with multiple temporal scales for action prediction. In *ECCV*, 2014.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[Lan *et al.*, 2014] Tian Lan, Tsung-Chuan Chen, and Silvio Savarese. A hierarchical representation for future action prediction. In *ECCV*, 2014.

[Ma *et al.*, 2016] Shugao Ma, Leonid Sigal, and Stan Sclaroff. Learning activity progression in lstms for activity detection and early detection. In *CVPR*, 2016.

[Mnih *et al.*, 2014] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention. In *NIPS*, 2014.

[Ng *et al.*, 2015] Joe Yue-Hei Ng, Matthew J. Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *CVPR*, 2015.

[Nguyen and la Torre, 2012] Minh Hoai Nguyen and Fernando De la Torre. Max-margin early event detectors. In *CVPR*, 2012.

[Ryoo and Aggarwal, 2010] M. S. Ryoo and J. K. Aggarwal. UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA), 2010.

[Ryoo, 2011] M. S. Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *ICCV*, 2011.

[Sharma *et al.*, 2015] Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov. Action recognition using visual attention. *CoRR*, abs/1511.04119, 2015.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014.

[Song *et al.*, 2016] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. *CoRR*, abs/1611.06067, 2016.

[Wu *et al.*, 2015] Zuxuan Wu, Xi Wang, Yu-Gang Jiang, Hao Ye, and Xiangyang Xue. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In *ACM MM*, 2015.

[Xu *et al.*, 2015] Zhen Xu, Laiyun Qing, and Jun Miao. Activity auto-completion: Predicting human activities from partial videos. In *ICCV*, 2015.