Neurocomputing 322 (2018) 195–205

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A two-level attention-based interaction model for multi-person activity recognition

Lihua Lu, Huijun Di, Yao Lu*, Lin Zhang, Shunzhou Wang

Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing, China

ARTICLE INFO

Article history:
Received 19 February 2018
Revised 1 August 2018
Accepted 21 September 2018
Available online 27 September 2018

Communicated by Huaping Liu

Keywords:
Multi-person activity recognition
Individual level
Scene level
Attention mechanism

ABSTRACT

Multi-person activity recognition is a challenging task due to its elusive interactions in activities. We take into account these interactions at two levels. At the individual level, each person behaves depending on both its spatio-temporal features and interactions propagated from others in the scene. At the scene level, the multi-person activity is characterized by interactions between individuals' actions and the high-level activity. It is worth noting that interactions contribute unequally at both levels. To jointly explore these colorful interactions, we propose a two-level attention-based interaction model relying on two time-varying attention mechanisms. The individual-level attention mechanism conditioned on pose features, exploits various degrees of interactions among individuals in a scene while updating their states at each time step. The scene-level attention mechanism proposes an attention-based pooling strategy to explore various levels of interactions between individuals' actions and the high-level activity. We ground our model by a modified two-stage Gated Recurrent Units (GRUs) network to handle the long-range temporal variability and consistency. Our end-to-end trainable model takes as inputs a set of person detections in videos or image sequences and predicts labels of multi-person activities. Experimental results demonstrate comparable performance of our model and show the effectiveness of our attention mechanisms.

1. Introduction

Human social behavior can be characterized at numerous levels of details, such as individual action, human-object interaction, human–human interaction and so on. In this paper, we focus on multi-person activity recognition involving many persons behaving variously but confined to a global goal, a high-level detail such as "team right spiking activity" taken together by a group of people (see the example in Fig. 1).

Multi-person activity recognition is a valuable but challenging task in its own. Multiple people involved in a scene can behave uniformly at one time, but more generally they have varied actions and interactions that compose a high-level activity. We observe that the key cha llenge to multi-person activity recognition is to explicitly capture complicated spatio-temporal interactions at various levels. At the individual level, each person behaves relying on not only spatio-temporal features of himself but also interactive information provided by others in the scene. In addition, for the anchor one different persons share different levels of interactions with it. For example, in Fig. 1 for the person spiking the ball, persons blocking the ball have stronger effects than others. Therefore, when it comes to individual-level actions, a favorable model should attach different degrees of importance to interactions given by others. At the scene level, it is difficult to reason about the high-level activity from various individuals' actions because of the ambiguities on individuals' actions and the complexities of interactions between low-level individuals' actions and high-level activities. Moreover, different persons or subgroups active in a scene do different things that contribute differently to the actual high-level activity. Also in Fig. 1, for the actual activity "team right spiking activity", it is apparent that the person spiking the ball contributes more than others. Thus, a favorable model needs to bridge the gap between individuals' actions and the high-level activity, and identify individuals' various interactive effects on the actual group activity.

A volume of research efforts [1–8] have been contributed to infer spatio-temporal interactions in multi-person activity recognition. Commonly efficient approaches utilize kinds of graphical models to capture spatial relations and interactions. Lan et al. [7] propose a hierarchical graphical model that considers the interactions at the social role level. Amer et al. [2] adopt a HiRF model to perform recognition and detection simultaneously based on a

* Corresponding author.
E-mail addresses: lulihua@bit.edu.cn (L. Lu), ajon@bit.edu.cn (H. Di), vis_yl@bit.edu.cn (Y. Lu), zhanglin@bit.edu.cn, roryuna@163.com (L. Zhang).

https://doi.org/10.1016/j.neucom.2018.09.060
0925-2312/

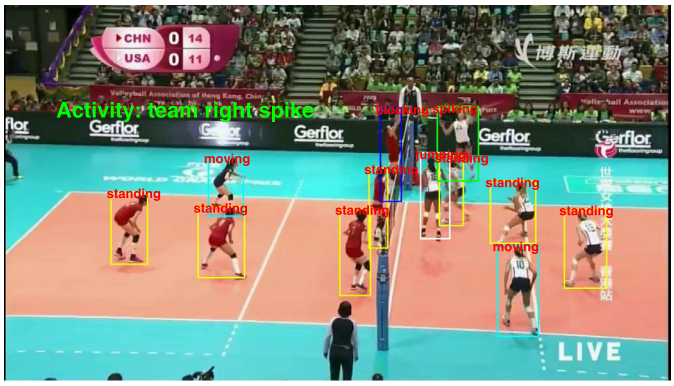


Fig. 1. We focus on multi-person activity recognition involving many persons perform variously but confined to a global goal, a high-level detail such as "team right spiking activity" taken together by a group of people (best viewed in color).

graphical model. However, since these approaches rely on handcrafted features, the expressive force is limited. Recently, inspired by recurrent neural networks, Deng et al. [3] propose structure inference machines that jointly integrate a graphical model with Recurrent Neural Networks. The model analyzes relations in group activity recognition and infers a relationship structure among persons on each iteration. But this model can't handle temporal information in videos or image sequences and is computationally complex.

Different from Deng et al. [3], at the individual level, we substitute an attention mechanism for complicated message passing and gating functions. Substantially, we ground our individual-level attention mechanism in pose feature space thanks to its powerful expression and robustness to variations of locations and appearances [9]. Some works [8,10] observe that context information can facilitate activity recognition. But context information may not work well under some situations, for example, in a volleyball game background information about the beach in a scene appears equally in all frames, which cannot be vivid semantic indicators for the volleyball activity recognition. In fact, we argue that pose features enable our individual-level attention mechanism to pay more importance to actions themselves than context information. Ultimately, our model mimic the interaction relationships among individuals and transmit information relevant to each other.

The attention mechanism, which is originally presented in language processing, has been applied in computer vision task [11–14]. In multi-person activity recognition domain, Ramanathan et al. [15] propose a model which learns to detect events in videos while automatically "attending" to the people responsible for the activity. However, automatically acquiring key actors is not always accurate and inevitably overlooks useful interaction information provided by other relevant persons in charge of the actual activity. In contrast, at the scene level, our scene-level attention-based interaction model roundly takes into consideration all individuals for the inference of the high-level multi-person activity. The simple but smart pooling strategy based on a time-varying attention mechanism can automatically eliminate ambiguities and redundancies of interactions flowing to the actual activity.

In brief, we observe that effective models need to jointly handle the colorful interactions at individual as well as scene levels. Accordingly, we propose a hierarchical attention-based model (see Fig. 2). Our model follows this outline: given a set of person detections for each video frame, we handle an individual-level attention-based interaction model to jointly analyze the evolutions of persons within a frame and reason about interaction relationships among them. For each anchor person, our individual-level attention mechanism captures different levels of interactions propagated from all others in the scene by computing pairwise attention weights and then updates the state of the anchor one conditioned on these interactions. A higher scene-level attention-based interaction model analyzes the evolutions of high-level activities and infers interaction relationships between low-level individuals' actions and high-level activities. For each high-level activity consisting of multiple persons, our scene-level attention mechanism seizes different degrees of interactions contributed by all individuals' actions in the scene by attentively pooling outputs of the individual-level model over all persons. This two-level attention mechanism enables our model to encode various interactions at both individual and scene levels.

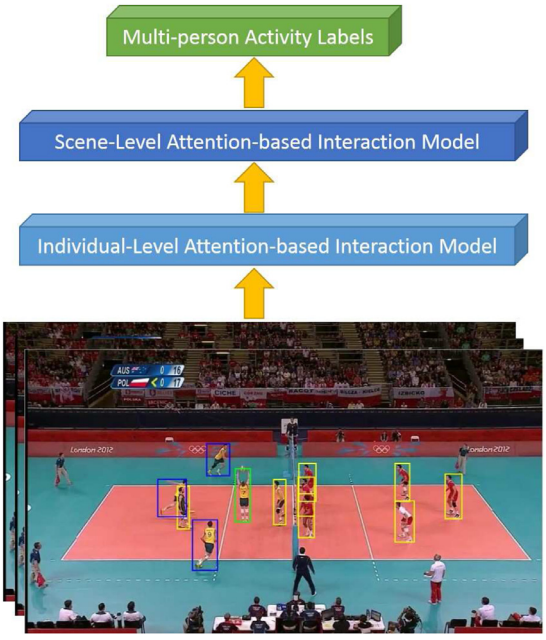


Fig. 2. General overview of our proposed model. Our model take as inputs a set of person detections for each video frame, and outputs the label of the high-level activity. The individual-level attention-based interaction model and scene-level attention-based interaction model respectively encode interaction relationships at both individual and scene levels (best viewed in color).

The main contributions of our work are fourfold: (1) Inspired by Bagautdinov et al. [4], we propose an individual-level attention-based interaction model to jointly map persons in temporal domain and infer interactions among persons. Our pose-dependent attention mechanism guarantees the individual-level interaction model's comprehensiveness and effectiveness. (2) We propose a scene-level attention-based interaction model to analyze the high-level activity. Our attention-based pooling strategy enables our model to capture interactions between individuals' actions and the high-level activity by weighting individuals' contributions to the activity, and then infer the actual activity from these interactions. (3) Our model is free of tracking in videos or image sequences. An exciting side-effect of our individual-level attention mechanism is that in temporal domain we can update each person at every time step by considering all persons' weighted states at last time step, averting from associating the same person across frames through tracking. (4) We ground our two-level attention-based interaction model by a two-stage Gated Recurrent Units (GRUs) network. These two GRUs networks respectively catch temporal variability and consistency at individual and scene levels by updating attention weights at every time step.

The remainder of this paper is organized as follows. Related work is discussed in Section 2. Our two-level attention-based interaction model is grounded in Section 3, especially, the individual-

level interaction model and the scene-level interaction model are respectively formulated in Section 3.2 and Section 3.3. Experiments and discussions on Volleyball dataset [16] are presented in Section 4. The conclusion and acknowledge are given in Section 5 and Section 6.

2. Related work

Multi-person activity recognition is an active area of research. Here, our work focuses on a hierarchical attention-based model which can jointly encodes interactions at both individual and scene levels. In what follows, we give a short overview of the existing work related to our work.

Multi-person activity recognition. Multi-person activity recognition is terrifically different from single-person action recognition owing to the fact that multi-person activities contain sophisticated interactions at multiple levels. In order to model interaction information between individuals in space and or time domain, on one hand, some works [17,18] use contexts as cues. However, these methods are restricted to smaller datasets. On the other hand, many previous works on multi-person activity recognition feed hand-crafted features to various graphical models. Among these models, hierarchical graphical models [2,7,19,20], AND-OR graphs [21], and dynamic Bayesian networks [22] are comparatively popular. Lan et al. [19] propose an adaptive latent structure learning that represents hierarchical relationships ranging from lower person-level information to higher group-level interactions. Shu et al. [21] detect group activities from aerial video using an AND-OR graph formalism.

However, human-engineered features used in these methods limit the ability for exploring intrinsic characteristics of multi-person activities. Thus, recent researches shift towards merging the power of neural networks with structured models [3,15,16]. Deng et al. [3] propose structure inference machines to refine individuals' estimates obtained from CNNs. They define a trainable graphical model with nodes for all the people and the scene, and pass messages between them to get the final scene-level estimate. Ibrahim et al. [16] propose a hierarchical temporal model consisting of a two-stage LSTMs. The first LSTM analyzes person-level actions, and then the second LSTM infers high-level activities on top of the first LSTM. Despite effectiveness, these methods arouse some issues such as deficiency of handling temporal information, terrible computations and so on.

Attention-based models. For humans, it is time-consuming, or even unavailable to explicitly label the ground truths for attentions. Therefore, researchers shift to attention models which can be implicitly learned from neural networks. Generally, attention mechanisms are categorized into two classes. On one hand, hard attention mechanism takes hard decisions that lead to stochastic algorithms, which cannot be easily learned through gradient descent and back-propagation [23]. For example, Yeung et al. [24] use a model based on hard attention for action detection. Their model can decide both which frame to observe next as well as when to emit an action prediction. On the other hand, soft attention mechanism takes entire input into account, weighting each part of the observations dynamically. Usually, models based on soft attention are differentiable, making gradient-based optimization possible. For example, Sharma et al. [25] propose a recurrent mechanism for action recognition from RGB data, which integrates convolution features from different parts of a space–time volume.

Recently, attention-based models excite various fields such as machine translation [26], image captioning [27], image classification [12] and detection [11]. Bahdanau et al. [26] propose an attention-based RNN model which effectively aligns input words to output words for machine translation. Following this, Xu et al. [27] use attention for image-captioning and video-captioning respectively. In all these methods, attention aligns a sequence of input features with words of an output sentence. Selective focus on different spatial regions is proposed for action recognition. Yeung et al. [28] propose a fusion of neighboring frames within a sliding window with learned attention weights to enhance the performance of dense labeling of actions. For group activity recognition, Ramanathan et al. [15] assume that only a small set of people are responsible for the actual activity, and propose a soft attention mechanism to identify key actors and infer activities base on these key actors actions. In our work, Instead of selecting key actors or subgroups, we propose a scene-level attention-based interaction model. To analyze the high-level activity, our model takes into account all individuals and these interactions adaptively using an attention mechanism.

Pose based action recognition. A large number of works have concentrated on action recognition in RGB or RGB-D data [29–31]. Liu et al. [29] investigate action recognition using an inexpressive RGB-D sensor. Here, we note that actions of persons can be described by the evolutions of a series of human poses. Based on pose features, many works have gained exciting performances [32–34]. Pose features are robust to some factors like illumination changes, occlusion and background clutter, and can explicitly model the dynamics of actions. Pose-based action recognition can be divided into two categories: joint-based approaches and body part-based approaches [35]. On one hand, joint-based approaches represent human poses by 3D or 2D coordinates of joints, and employ various coordinates based features such as joint positions [36] and pairwise relative joint positions [37] to characterize actions. On the other hand, body part-based approaches regard human poses as a connected set of segments, and then extract features from individual or connected pairs of body parts [38].

Recently, pose-based action recognition substitutes deep neural networks for human-engineered features to unleash the full potentials of data. Du et al. [39] design an end-to-end hierarchical RNN architecture for skeleton based action recognition. Veeriah et al. [40] propose differential RNN to temporally model the dynamics of states. Shahroudy et al. [41] present a part-ware LSTM to take into consideration the physical structure of the human body. More recently, wang et al. [35] offer a two-stream RNN architecture to respectively model temporal dynamics and spatial configures for actions. Furthermore, Fabien et al. [23] propose a pose-conditioned spatio-temporal attention for human action recognition. Inspired by previous works [23,35], we base our individual-level attention mechanism on pose features with spatial configurations in line of Baradel et al. [23].

3. Our method

3.1. Overview

The detailed flow of our model is illustrated in Fig. 3. Our model takes as inputs a set of person detections for each video frame, and predicts corresponding labels of multi-person activities. It dedicates to jointly capturing interaction relationships at individual and scene levels, reasoning about evolutions of individuals' actions as well as the high-level multi-person activities.

Our model can work with any effective human detection and feature extraction algorithms. In the line of Bagautdinov et al. [4], for every video frame $I_t \in \mathbb{R}^{H_0 \times W_0 \times 3}$, we obtain a set of reliable detections encoded as bounding boxes $b_t \in \mathbb{R}^{N \times 4}$, and extract fixed-size representations $f_t^i \in \mathbb{R}^{K \times K \times D}$, $i \in 1, \ldots, N$ for individuals, where N is the number of bounding boxes, K is the size of the fixed representation in pixels, and D is the number of features. On top of these representations $f_t^i \in \mathbb{R}^{K \times K \times D}$, a fully-connected layer produces more compact embedding representations $e_t^i \in \mathbb{R}^{D_e}$, $i \in 1, \ldots, N$, where D_e is the number of features in the embed-

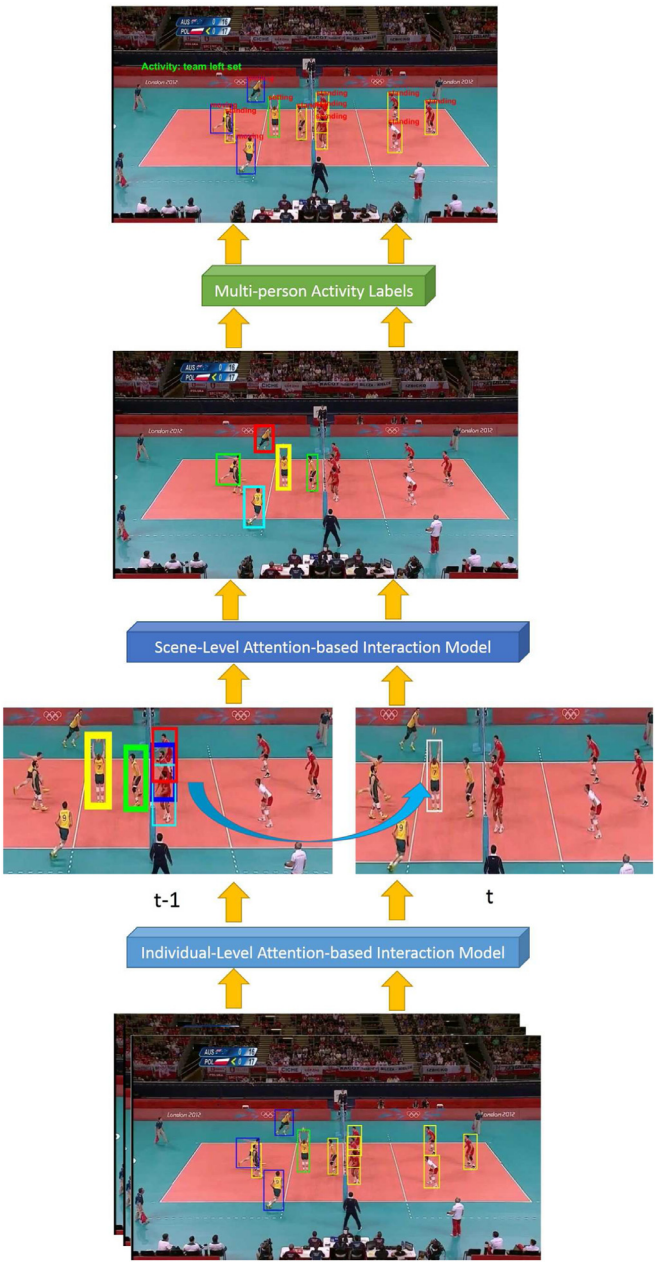


Fig. 3. The detailed flow of our proposed model. Our model takes as inputs a set of person detections for each video frame, and outputs labels of high-level activities. The individual-level attention-based interaction model encodes interactions among persons while updating their states at each time step. And the scene-level attention-based interaction model encodes interactions between individuals' actions and the actual activity while attentively pooling all individuals' states in a scene at each time step. Finally, our model predicts labels of all individuals' actions and the high-level activity (best viewed in color and the thickness of bounding boxes).

ded representation. These embedding individuals' features are then fed into our hierarchical attention-based interaction model. At each time step, the individual-level attention-based interaction model jointly encodes interactions among persons, associates persons in temporal domain without tracking, and infers action for each detection, making preparations for high-level multi-person activity recognition. Particularly, our individual-level attention mechanism is conditioned on human articulated pose features represented by a series of 2D joints, derived from [42]. Then on top of the individual-level attention-based interaction model, our scene-level attention-based interaction model jointly analyzes interaction relationships between low-level individuals' actions and high-level multi-person activities, and predicts labels of multi-person activities. We implement our model using a two-stage subtly modified Gated Recurrent Units (GRUs) network to model a large range of low-level and high-level dynamics defined on top of individuals and the entire group. In the following sections, we will describe each ingredient of our model in details.

3.2. Theindividual-level attention-based interaction model

The key to multi-person activity recognition is reasoning about interaction relationships. We model these interactions in multi-person activities from two perspectives: exploring interactions among individuals within a scene at individual level; and exploiting interactions between low-level individuals' actions and high-level activities at scene level.

3.2.1. Analyzeindividual-level interactions

In this section, we focus on how to infer interaction relationships among persons in a frame at individual level. We argue that in videos or image sequences interactions lie in the evolution of each individual at every time step. That is to say, each individual behaves relying on not only spatio-temporal features of itself but also interactions propagated from individuals within the scene. Furthermore, behaviors of others can facilitate the model to disambiguate the action of the anchor one. For example, in a volleyball match, persons standing there will jump to block the ball if an opponent attacker spikes the ball. Tactfully, we cast the inference of interactions as an indispensable part of the evolution of each individual.

We also observe that it is not wise to equally treat interactions among individuals. Take a volleyball match as an example, for the anchor one spiking the ball, persons standing nearby have stronger interaction effects than those who are far in the distance, persons blocking the ball have more contributions than those who just stand there, and so on. Therefore we propose a novel pose-based attention mechanism to infer individual-level interactions. For each anchor individual, our attention mechanism attaches different levels of importance to different individuals conditioned on computed attention weights to capture interactions propagated by other individuals in the scene. Based on these captured interactions, our model can temporally update each individual at every time step without tracking. To update each individual at time step t, our individual-level attention-based interaction model takes into consideration all individuals within the scene at time step $t-1$, and makes use of the pose-based attention mechanism to weight individuals' interactions to this update. It can be indicated that the pose-based attention mechanism is the key to capturing individual-level interactions. Next we will clarify our pose-based attention mechanism and our individual-level interaction model in detail.

3.2.2. Pose representations

We believe that human poses are essential for action recognition, and a large set of actions can be recognized solely from pose features. Pose features are robust to illumination and so on, and can explicitly model dynamics of actions. These characteristics enable our pose based attention network to think highly of the action itself in place of context information redundant. In our model, we represent each human pose by a set of 2D coordinates of joints derived from pose estimation [42]. In order to preserve the strong neighborhood relationships in the human body, we follow the work [23] to travel all body joints along the topological structure of human body (see Fig. 4). This travel strategy guarantees the spatial relationships of body joints by accessing all joints at least twice. Finally, all persons we obtain the pose representations

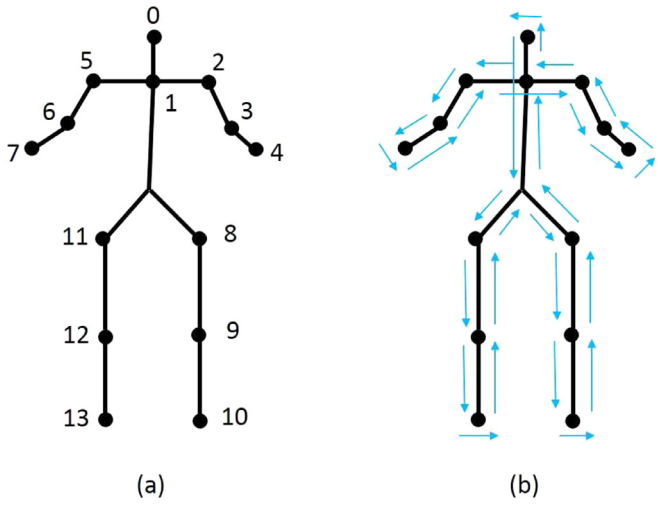


Fig. 4. (a) The physical structure of 14 body joints [23]. (b) Travel along the topological order of human body joints. The order of the sequence is the same as the visiting order of the arrow, starting from the point 1. The travel sequence is 1-5-6-7-6-5-1-2-3-4-3-2-1-0-1-11-12-13-12-11-8-9-10-9-8-1.

$x_t^i = \{x_1, \ldots, x_j\}, i = \{1, \ldots, N\}, j = \{1, .., J\}, t = \{1, \ldots, T\}$, where J is the number of joints and T is the number of frames in a video or an image sequence.

3.2.3. Our pose-based attention mechanism for individual-level interactions

Our individual-level attention-based interaction model takes as inputs the embedding representations $e_t^i \in \mathbb{R}^{D_e}$ and pose features x_t^i of each individual at every video frame t, and outputs a set of refined actions' classification scores at each time step. The key pose-based attention mechanism reasons about individual-level interactions and facilitates our model to temporally associate individuals without tracking. For each anchor individual, our attention mechanism attaches different levels of importance to different individuals conditioned on computed attention weights to capture interactions propagated by other individuals in the scene. Based on these captured interactions, our model can temporally update each individual at every time step without tracking.

For each individual, our pose-based attention mechanism takes advantage of all individuals and weights them by attention coefficients in pose feature space, depicted in Fig. 5. In some detail, on top of pose representations x_t^i, our pose-based attention mechanism computes pairwise attention weights conditioned on the distances between the anchor person at time step t and all persons at time step $t-1$ in pose feature space. Then a softmax layer is added to normalize these attention weights which finally denoted as $\omega_t^{i,j}$. For the anchor person at time step t, our individual-level attention-based interaction model adaptively captures individual-level interactions passed from other persons by attaching different levels of importance to different persons at time step $t-1$ on top of these normalized attention weights. In addition, an exciting side effect is that for each person at time step t, we can acquire its last state at time step $t-1$ by a weighted sum of all persons' states at time step $t-1$. That is to say for each person at time step t, we adopt the weighted persons' features at time step $t-1$ to update its state at time step t. It can be seen that thanks to the pose-based attention mechanism we can associate persons in temporal domain without tracking.

To handle sequential data in videos or image sequences, we are inclined to employ Recurrent Neural Networks (RNNs) to merge and propagate information in the long-range temporal domain. Inspired by Bagautdinov et al. [4], we ground our individual-level attention-based interaction model by a modified Gated Recurrent Units (GRUs) network. We note that the traditional GRUs network maps states for entities through tracking in time domain, which means that the hidden states $ph_{t-1}^i \in \mathbb{R}^{D_h}$ and $ph_t^i \in \mathbb{R}^{D_h}$, where D_h is the number of features in the hidden state, need to correspond to the same person. In order to adapt the traditional GRUs network to our individual-level attention-based interaction model, we modify its access to each person's last state at time step $t-1$ by substituting our pose-based attention mechanism for tracking. Generally, our model can take any efficient RNNs such as Long Short-Term Memory (LSTM), Gated Recurrent Units (GRUs) and so on. In order to compare the performance of LSTM and GRUs, we deploy check experiments in Baselines (described in Sections 4.2.1 and 4.2.2).

In what follows, we formulate our model in details:

$$\omega_t^{i,j} \propto \exp(-\parallel x_t^i - x_{t-1}^j \parallel_2^2), \sum_j \omega_t^{i,j} = 1 \tag{1}$$

$$ph_{t-1}^i = \sum_j \omega_t^{i,j} ph_{t-1}^j \tag{2}$$

$$ph_t^i = GRU(e_t^i, ph_{t-1}^i) \tag{3}$$

Finally, the refined updates are based on both embedding features of its own and iteratively weighted interaction information provided by all individuals. Then we simply feed the outputs of the individual-level model at each time step to a softmax classification layer to predict the single-person action for each person detection.

Our pose-based attention mechanism has some advantages: (1) It is time-varying and content-dependent. In every iteration of refinement, weights are recomputed based on pose features for every person detection. (2) It enables us to jointly handle interaction relationships among individuals within a scene and control the evolution of each individual in a single forward procedure. (3) It makes our model be free of tracking, which means that the hidden states $ph_{t-1}^i \in \mathbb{R}^{D_h}$ and $ph_t^i \in \mathbb{R}^{D_h}$, where D_h is the number of features in the hidden state, are not necessarily corresponding to the same person. Experiments in baselines show that our tracking-free model based on the attention mechanism is comparable to tracking-based algorithm (illustrated in Sections 4.2.1 and 4.2.2). 4) Thanks to pose features of their powerful characteristics such as coordinate positions, postures, and so on, our pose-based attention mechanism is more reasonable position-based attention mechanism (earnestly depicted in Section 4.3.1).

In this stage, we first train our model to predict individual-level actions, and then pass the hidden states of the GRUs layer to the second stage for multi-person activity recognition, as discussed in the next section. The first-trained person-level model is used as a pre-trained model and is finetuned in the second stage making the whole hierarchical model be trained in an end-to-end way.

3.3. Thescene-level attention-based interaction model

3.3.1. Analyze scene-level interactions

Interactions between persons' actions in a scene and the high-level activity are crucial to understand high level activity. Sad to say, it is troublesome to explicitly capture the interactions between people and scene as they vary with different situations. People active in a scene can do the same thing at the same time or even different persons or subgroups do different things confined to a global high-level goal. In addition, persons within a frame matter with the group activity, but different people contribute differently to the actual activity. Therefore, when it comes to high-level activity recognition, equally treating all persons may introduce significant noises. For instance, in an activity "team right spiking", persons spiking the ball and persons blocking the ball contribute more to the actual activity, while persons standing there are less relevant or even bring ambiguities. In conclusion, a favorable model needs

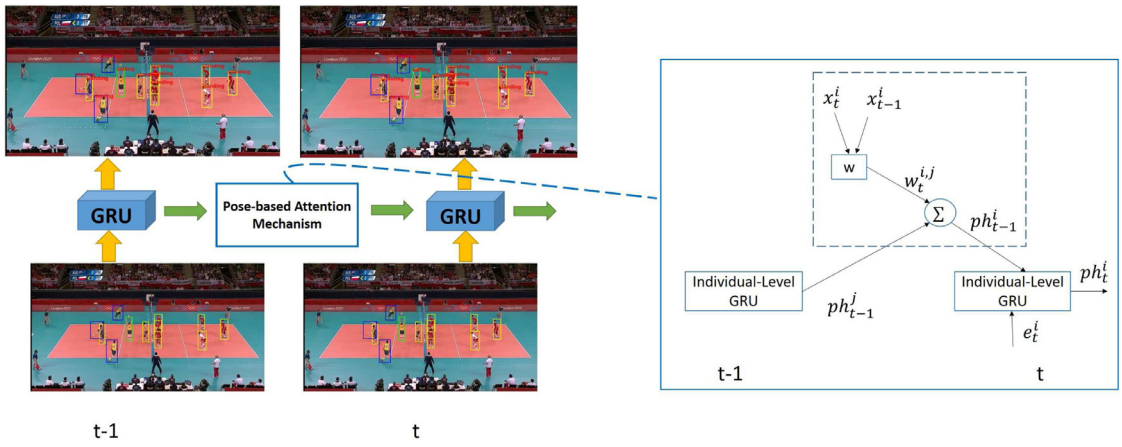


Fig. 5. Our proposed individual-level attention-based interaction model. At each iteration of the modified GRUs network, it updates individuals' states based on the interaction information provided by the pose-based attention network (described in dotted rectangle).

to bridge the gap between individuals' actions and the high-level activity, and deduce different levels of individuals' responsibilities for the actual activity.

3.3.2. Attention-based pooling strategy for scene-level interactions

We observe that interactions between individuals' actions and the high-level activity can be rationally modeled by a well-designed pooling algorithm. That is to say, we cast modeling scene-level interaction relationships into constructing a favorable pooling policy. In our model, the attention-based pooling strategy encodes scene-level interactions by giving different degrees of importance to different individuals while pooling their features into scene-level representations. Various pooling strategies can be used to aggregate features over all persons in the scene at each time step, such as max pooling [16,43], adaptive Scan Pooling [44], generalized rank pooling [45], average pooling [46]. Differ from these pooling strategies, we propose a scene-level pooling strategy based on an attention mechanism, which is also grounded by a GRUs network.

On top of the individual-level attention-based interaction model, we concatenate the embedding representations e_t^i and the hidden states of the first stage GRUs ph_t^i (represented by $\oplus$) to obtain the spatio-temporal features p_t^i for ith person. Our scene-level attention mechanism is simple, comprising two fully-connected layers and a softmax layer. It computes different weights γ_t^i on top of features p_t^i for detected persons in a frame and scene-level state h_{t-1}^e at last time step $t-1$, namely the hidden layer's features of the scene-level GRUs network. At every time step of scene-level GRUs network, our attention-based pooling network aggregates weighted individuals' features into the scene-level representation a_t (best viewed in Fig. 6). Remember that our scene-level attention mechanism is content-dependent and it evolves with the activity.

$$p_t^i = e_t^i \oplus ph_t^i \tag{4}$$

$$\gamma_t^i = softmax(fc(fc(p_t^i, h_{t-1}^e))) \tag{5}$$

$$a_t = \sum_{i=1}^{N} \gamma_t^i p_t^i \tag{6}$$

$$h_t^e = GRU(h_{t-1}^e, a_t) \tag{7}$$

Finally, based on these scene-level representations, a softmax classifier is added to get final predictions for multi-person activities. The single-person actions are predicted by a separate softmax classifier on top of each person detection. The loss is defined as follows:

$$L_{CI} = -\frac{1}{T \cdot N_C} \sum_{t,c} \hat{p_{C,c}^t} \log p_{C,c}^t - \omega_I \frac{1}{T \cdot N \cdot N_I} \sum_{t,n,a} \hat{p_{I,n,a}^t} \log p_{I,n,a}^t \tag{8}$$

In Eq. (8), T is the number of frames, N_C, N_I are the numbers of labels for group and individual actions, N is the number of detections, and $\hat{p_*}$ is the one-hot-encoded ground truth. The weight ω_I allows us to balance the two tasks differently. In our experiments, we set $\omega_I = 2$.

4. Experiments

In this section, we evaluate our model which focuses on multi-person activity recognition. We compare our two-level attention-based interaction model with several baselines and some typical works on the Volleyball dataset [16].

4.1. Implementation details

We train our model in an end-to-end way. At first, we pre-train our network only with the individual-level attention-based interaction model on single frames, to predict single-person actions. These single-person actions are predicted by a separate softmax classifier on top of person detections. Secondly, we extend the model by adding a scene-level attention-based interaction network. We jointly finetune the individual-level attention-based interaction model and train the scene-level attention-based interaction model based on the loss function illustrated in In equation (8). We use a temporal window of length $T = 10$, which corresponds to 5 frames before the annotated frame, and 4 frames after. Our two-stage GRUs network is the key of our implementation of the proposed model. In our experiments, the number of hidden units in the individual-level GRUs network is 1024, the same as the scene-level GRUs network. The number of parameters and FLOPS of our model are respectively about 1.2×10^7 (80.4 MB) and 7.8×10^9. In the training phase, the model can handle about 20 frames in one second.

Our implementation is based on *TensorFlow* and a single *NVIDIA TITAN X GPU*. All our models are trained using back propagation using the same optimization scheme: for all the experiments, we use stochastic gradient descent with ADAM [28], with the initial learning rate set to 10^{-5}, and fixed hyper parameters to $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$. We use a fixed learning rate of 0.00001 and a momentum of 0.9. For tracking persons in videos or image sequences,

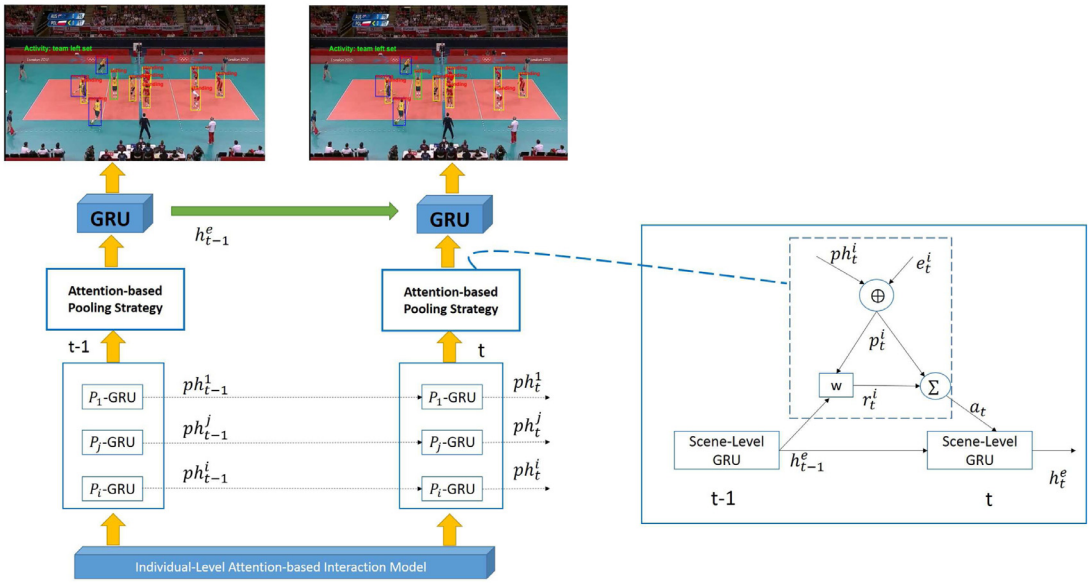


Fig. 6. Our scene-level attention-based interaction model. On top of the individual-level attention-based interaction model, it pools individuals' features into a scene-level spatio-temporal representation based on weights provided by the attention mechanism at every iteration of GRUs network.

we use the tracker by Danelljan et al. [12], implemented in the Dlib library [19].

4.2. Experiments on volleyball dataset

We evaluate our framework on Volleyball Dataset [16], which is publicly available dataset for multi-person activity recognition that is relatively large-scale and contains labels for people locations, as well as their collective and individual actions. The dataset consists of 55 volleyball games with 4830 labelled frames, where each player is annotated with the bounding box and one of the 9 individual actions, and the whole scene is assigned with one of the 8 collective activity labels, which define which part of the game is happening. For each annotated frame, there are multiple surrounding unannotated frames available. To get the ground truth locations of people for those, we resort to the same appearance-based tracker as proposed by the authors of the dataset [16]. We used frames from $\frac{2}{3}$rd of the videos for training, and the remaining $\frac{1}{3}$rd for testing.

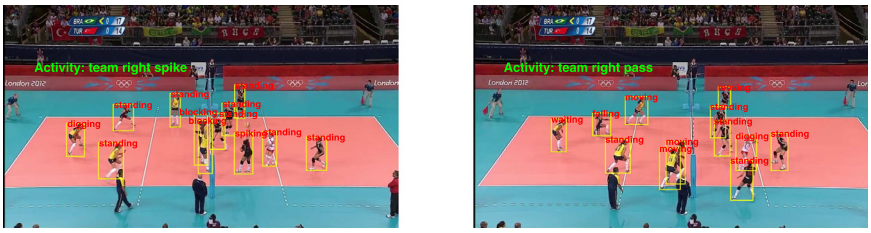
4.2.1. Baselines

The following baselines are considered in all our experiments:

1. B1 – image-based model – this model is based on global image-level features. It examines the idea of feeding image-level features directly to a LSTM model to recognize group activities. In this baseline, the Inception-v3 network [47] is pre-trained on ImageNet and fine-tuned to predict collective activities on whole images, without taking into account locations of individuals and person-level features.
2. B2 – asingle-level temporal model through tracking – this model is based on person-level features and implemented by a single-stage traditional GRUs network to encode temporal information at person level. Then it just adopts the max-pooling to aggregate individuals' features into frame-level representations, and makes use of these frame-level representations to predict the final activity labels.
3. B3 – the individual-level attention-based interaction model – this key to this model is the individual-level attention mechanism implemented in pose feature space. We deploy this baseline, the first part of our two-level model, to demonstrate the effectiveness of our individual-level attention-based model. Similar to B2, this model adopts max-pooling to compute frame-level representations.
4. B4 – a two-level temporal model through tracking – this model is based on person-level features and implemented by a two-stage traditional GRUs network to encode a large range of temporal information through tracking. In this baseline, the Inception-v3 network [47] is deployed for each person to get high-resolution fixed-sized features conditioned on the ground truth detections. Features are pooled over all people in a frame, and then fed into a softmax classifier to recognize activity for each single frame. This model is differ from our two-level attention-based interaction model with the key fact: this model associates persons in temporal domain through tracking instead of attention mechanism, therefore it cannot model powerful interactions to refine multi-person activity predictions.
5. B5 – a two-level attention-based interaction model on LSTM – this baseline is a implementation of our proposed model through a two-stage Long Short-Term Memory (LSTM) Networks. Since we want to use Recurrent Neural Networks (RNNs) to capture temporal dynamics in videos or image sequences, it is naturally to adopt the commonly used Long Short-Term Memory (LSTM) network or Gated Recurrent Units (GRUs) network. Here, we ground our proposed two-level attention-based interaction model by LSTM to compare the performance of LSTM and GRUs in our specific multi-person activity recognition task.
6. Our two-level attention-based interaction model on GRUs – a complete implementation of our proposed model on two novel attention mechanisms, which is grounded by a two-stage GRUs network.

4.2.2. Comparison with baselines

In Table 1, the performance of our proposed model is compared against the baselines. Obviously, our two-level attention-based interaction model outperforms these baselines. Here we discuss these comparisons from several aspects: (1) Comparing B1 with other baselines, it can be concluded that explicitly encoding individuals' features instead of the image-level features is necessary for obtaining better performance, since the background is



(a) (b)

Fig. 7. Visualization of the single-person action and multi-person activity labels predicted by our proposed model.

Table 1
Comparison of multi-person activity recognition performance of baselines against our model evaluated on the volleyball dataset. We report the mean average accuracy for multi-person activity.

Method	Accuracy
B1-image-based model	46.7
B2-a single-level temporal model through tracking	83.8
B3-the individual-level attention-based interaction model	86.3
B4-a two-level temporal model through tracking	88.8
B5-a two-level attention-based interaction model on LSTM	91.0
Our two-level attention-based interaction model on GRUs	91.2

Table 2
Memory and time consumption of LSTM and GRUs implementations of our model.

Unit	# of Units	# of Parameters	FLOPS
LSTM	1024	$\approx 2.0 \times 10^7$ (81.4MB)	$\approx 1.3 \times 10^{10}$
GRU	1024	$\approx 1.2 \times 10^7$ (80.4MB)	$\approx 7.8 \times 10^9$

cluttered and even corrupts the temporal dynamics of the foreground. (2) The comparison of B2 and B3 and comparison of B4 and our proposed model can respectively validate the effectiveness of our two attention mechanisms, and demonstrate the inferior performance of our individual-level attention-based interaction model and scene-level attention-based interaction model. Therefore, our idea that multi-person activity recognition can benefit from modeling interactions at both individual and scene level is forcefully confirmed. Particularly, comparing B2 with B3, it can be seen that thanks to the pose-based attention mechanism, our individual-level attention-based interaction model can achieve better performance than tracking-based models. In my opinion, the critical reason for this is that our pose-based attention mechanism can effectively encode interactions at individual level thanks to powerful expressive force of pose features. Visualization results can be seen in Section 4.3. 3) For completeness, we deploy B5 compared with our proposed model, it can be seen that both LSTM and GRUs work well for our end task – multi-person activity recognition. Considering the memory and time consumption (depicted in Table 2), we ground our model by a two-stage GRUs network in our experiments.

4.2.3. Comparison with the state-of-the-art methods

We compare our two-level attention-based interaction model with some state-of-the-art methods, including social scene understanding [4], a hierarchical deep temporal model [16], SBGAR [17]. Table 3 offers the quantitative results of our proposed model and some typical models on the Volleyball dataset [16]. Our model outperforms these existing typical models and improves performances to 91.2%. Visualization results are depicted in Fig. 7.

Fig. 8 shows the confusion matrix obtained for the Volleyball dataset [16] using our proposed model. From the confusion matrix, we observe that our model generates consistently accurate high-level activity labels. Nevertheless, it has confusions between some classes, such as set and pass activities, as these activities often look similar.

Table 3
Comparison of multi-person activity recognition performance of some typical models against our model evaluated on the volleyball dataset [16]. We report mean average accuracy for multi-person activity.

Method	Accuracy
A hierarchical deep temporal model [16]	51.1
SBGAR [17]	66.9
Social scene understanding [4]	90.6
Our two-level attention-based interaction model	91.2

Table 4
Comparison of different pooling strategies.

Method	Accuracy
Distance-based attention mechanism	90.6
Pose-based attention mechanism	91.2

4.3. Analyze attention mechanism

We have quantitatively verified in Section 4.2.1 above that our two-level attention mechanism which can encode interactions at both individual and scene levels is practical to multi-person activity recognition. For completeness, we deploy more check experiments and supply more intermediate results to analyze our attention mechanisms. Intuitively, we also visualize performances of our attention mechanisms.

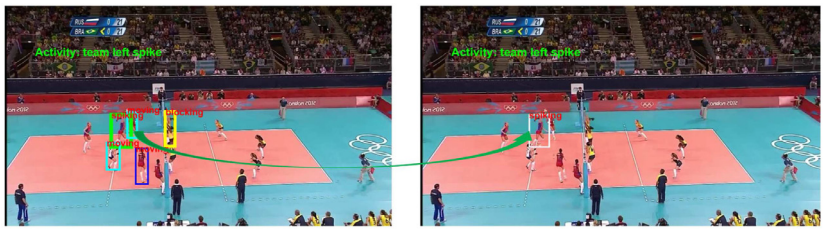
4.3.1. Analyze individual-level attention mechanism

At the individual level, each individual behaves relying on not only the spatio-temporal features of itself but also interaction effects given by others. In order to encode these interactions, we propose an individual-level attention-based interaction model on top of a pose-based attention mechanism. Our pose-based attention mechanism encode interactions among individuals by attaching different importance to different individuals at last time step while updating each individual at each time step. To confirm the efficiency of our pose-based attention mechanism, we design check experiments, including distance-based attention mechanism which computes weights based on distances between bounding boxes of individuals, and our pose-based attention mechanism. Results are shown in Table 4.

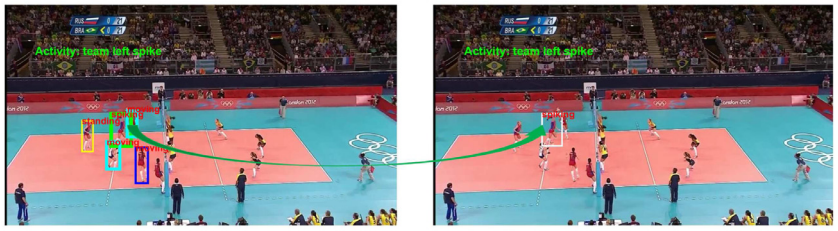
Fig. 9a and Fig. 9b respectively show the performance of our pose-based attention mechanism and the distance-based attention mechanism. It can be seen that, for the anchor person in the next frame, the distance-based model overlooks the person blocking the ball in the activity “team left spike”, since this person is far from the anchor one in coordinate feature space. However, according to

	lwinpoint	rwinpoint	lpass	rpass	lset	rset	lspike	rspike
lwinpoint	95.33	4.01	0.06	0.07	0.12	0.17	0.17	0.07
rwinpoint	3.75	94.01	0.53	1.07	0.06	0.06	0.34	0.17
lpass	0.77	0.07	88.50	5.70	2.63	1.90	0.03	0.40
rpass	1.33	1.36	4.57	**85.57**	2.90	3.13	0.80	0.33
lset	0.67	0.67	2.33	5.44	85.26	4.43	0.37	0.83
rset	0.33	0.33	1.47	3.00	3.43	**89.33**	0.83	1.27
lspike	0.40	0.30	1.13	0.74	0.10	0.30	94.63	2.40
rspike	0.43	0.23	0.23	0.04	0.37	0.93	1.10	96.67

Fig. 8. Confusion matrix for the volleyball dataset obtained using our two-level attention-based model.

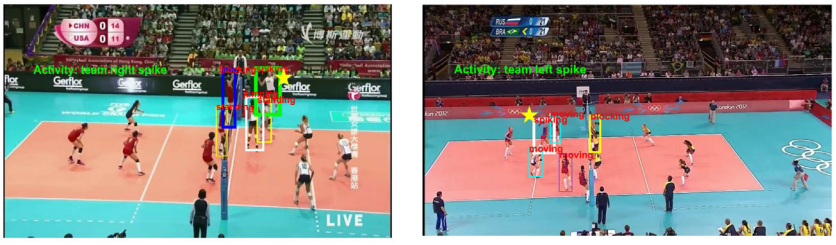


(a) The performance of our pose-based attention mechanism.



(b) The performance of the distance-based attention mechanism.

Fig. 9. Visualization of the performance of our individual-level attention mechanism. For simplicity, we just supply the top 5 results and different weights are better viewed in color and thickness of bounding boxes' borders. The anchor person in the next frame is denoted in white.



(a) (b)

Fig. 10. Visualization of the performance of our scene-level attention mechanism. For simplicity, we just supply the top 5 results and different attention weights are better viewed in color and thickness of bounding boxes' borders. The most important is labeled by the star.

our human senses, persons blocking the ball should take considerable effect. Fortunately, our pose-based attention mechanism can perceive interactions among individuals in accordance with human senses. In our opinion, the primary reason under this is that human poses are consist of a series of joint coordinates which are more detailed and more powerful than bounding boxes of persons. Therefore features can provide not only low-level characters such as coordinate positions of persons but also high-level semantic characters such as body orientations, postures and so on, and a large number of actions can be recognized on top of the evolutions of a series of poses.

4.3.2. Analyze scene-level attention mechanism

Multi-person activity generally consists of multiple persons but different persons contribute unequally to the actual high-level activity. That is to say, interaction effects between individuals' actions and the actual activity should be perceived. Therefore, we propose a scene-level attention-based interaction model to encode these interactions. The key to our scene-level model is the attention-based pooling strategy, which can encode interactions by giving different importance to different individuals while pooling individuals' features into a global scene-level representation. To confirm the performance of model, we deploy some check experiments

Table 5
Comparison of different pooling strategies.

Method	Accuracy
Max-pooling strategy	89.1
Average-pooling strategy	90.3
Our attention-based pooling strategy	91.2

including the max-pooling strategy, the average-pooling strategy and our attention-based pooling strategy. In these experiments, we hold our individual-level attention-based interaction model, and make use of different pooling strategy to compute frame-level representations for final representations. Results are shown in Table 5.

In addition, we show some visualization results of our scene-level attention mechanism. For instance, in Fig. 10a, individuals behave the action "spiking the ball" and "blocking the ball" take more important part in the activity "team right spike". This result conform to our human perceptions, and confirm the performance of our simple scene-level attention mechanism.

5. Conclusion

We have proposed a two-level attention based network for modeling interaction relationships in multi-person activity recognition at both person and scene levels. Based on pose features, our person level attention network can infer interactions among individuals while updating individuals' states at every time step. Our scene level attention network attaches various importance to individuals while pooling individuals' features to activity representations. Explicitly modeling interaction relationships at these two levels enable our model to gain state-of-the-art performance in multi-person activity recognition, involving many persons behaving differently but confined to a common purpose. Future work exists in extending our work to more general multi-person activity recognition.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (No. 61273273), by the National Key Research and Development Plan (No. 2017YFC0112001).

References

[1] N.-G. Cho, S.-H. Park, J.-S. Park, U. Park, S.-W. Lee, Compositional interaction descriptor for human interaction recognition, Neurocomputing 267 (2017) 169–181.
[2] M.R. Amer, P. Lei, S. Todorovic, HIRF: hierarchical random field for collective activity recognition in videos, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer, 2014, pp. 572–585.
[3] Z. Deng, A. Vahdat, H. Hu, G. Mori, Structure inference machines: recurrent neural networks for analyzing relations in group activity recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2016, pp. 4772–4781.
[4] T. Bagautdinov, A. Alahi, F. Fleuret, P. Fua, S. Savarese, Social scene understanding: end-to-end multi-person action localization and collective activity recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 3425–3434.
[5] W. Choi, Y.-W. Chao, C. Pantofaru, S. Savarese, Discovering groups of people in images, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer, 2014, pp. 417–433.
[6] H. Hajimirsadeghi, G. Mori, Learning ensembles of potential functions for structured prediction with latent variables, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), IEEE, 2015, pp. 4059–4067.
[7] T. Lan, L. Sigal, G. Mori, Social roles in hierarchical models for human activity recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 1354–1361.
[8] M. Wang, B. Ni, X. Yang, Recurrent modeling of interaction context for collective activity recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 7408–7416.
[9] F. Han, B. Reily, W. Hoff, H. Zhang, Space-time representation of people based on 3d skeletal data: a review, Comput. Vis. Image Underst. 158 (2017) 85–105.
[10] F.C. Heilbron, W. Barrios, V. Escorcia, B. Ghanem, SCC: semantic context cascade for efficient action detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 1454–1463.
[11] J.C. Caicedo, S. Lazebnik, Active object localization with deep reinforcement learning, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), IEEE, 2015, pp. 2488–2496.
[12] C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. Xu, et al., Look and think twice: capturing top-down visual attention with feedback convolutional neural networks, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), IEEE, 2015, pp. 2956–2964.
[13] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, A. Courville, Describing videos by exploiting temporal structure, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), IEEE, 2015, pp. 4507–4515.
[14] K. Gregor, I. Danihelka, A. Graves, D.J. Rezende, D. Wierstra, Draw: a recurrent neural network for image generation, arXiv:1502.04623 (2015).
[15] V. Ramanathan, J. Huang, S. Abu-El-Haija, A. Gorban, K. Murphy, L. Fei-Fei, Detecting events and key actors in multi-person videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2016, pp. 3043–3053.
[16] M.S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, G. Mori, A hierarchical deep temporal model for group activity recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2016, pp. 1971–1980.
[17] X. Li, M.C. Chuah, Sbgar: semantics based group activity recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 2876–2885.
[18] H. Hajimirsadeghi, W. Yan, A. Vahdat, G. Mori, Visual recognition by counting instances: a multi-instance cardinality potential kernel, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2015, pp. 2596–2605.
[19] T. Lan, Y. Wang, W. Yang, S.N. Robinovitch, G. Mori, Discriminative latent models for recognizing contextual group activities, IEEE Trans. Pattern Anal. Mach. Intell. 34 (8) (2012) 1549–1562.
[20] M. Ryoo, J. Aggarwal, Stochastic representation and recognition of high-level group activities, Int. J. Comput. Vis. 93 (2) (2011) 183–200.
[21] T. Shu, D. Xie, B. Rothrock, S. Todorovic, S. Chun Zhu, Joint inference of groups, events and human roles in aerial videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2015, pp. 4576–4584.
[22] Y. Zhu, N.M. Nayak, A.K. Roy-Chowdhury, Context-aware modeling and recognition of activities in video, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2013, pp. 2491–2498.
[23] F. Baradel, C. Wolf, J. Mille, Pose-conditioned spatio-temporal attention for human action recognition, arXiv:1703.10106 (2017).
[24] S. Yeung, O. Russakovsky, G. Mori, L. Fei-Fei, End-to-end learning of action detection from frame glimpses in videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2016, pp. 2678–2687.
[25] S. Sharma, R. Kiros, R. Salakhutdinov, Action recognition using visual attention, arXiv:1511.04119 (2015).
[26] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv:1409.0473 (2014).
[27] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: neural image caption generation with visual attention, in: Proceedings of the International Conference on Machine Learning (ICML), ACM, 2015, pp. 2048–2057.
[28] S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, L. Fei-Fei, Every moment counts: Dense detailed labeling of actions in complex videos, Int. J. Comput. Vis. 126 (2–4) (2015) 375–389.
[29] H. Liu, M. Yuan, F. Sun, Rgb-d action recognition using linear coding, Neurocomputing 149 (2015) 79–85.
[30] Z. Xu, R. Hu, J. Chen, C. Chen, H. Chen, H. Li, Q. Sun, Action recognition by saliency-based dense sampling, Neurocomputing 236 (2017) 82–92.
[31] T. Qi, Y. Xu, Y. Quan, Y. Wang, H. Ling, Image-based action recognition using hint-enhanced deep neural networks, Neurocomputing 267 (2017) 475–488.
[32] C. Wang, Y. Wang, A.L. Yuille, An approach to pose-based action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2013, pp. 915–922.
[33] J. Weng, C. Weng, J. Yuan, Spatio-temporal Naive–Bayes nearest-neighbor (ST-NBNN) for skeleton-based action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 4171–4180.
[34] Q. Ke, M. Bennamoun, S. An, F. Sohel, F. Boussaid, A new representation of skeleton sequences for 3d action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 4570–4579.
[35] H. Wang, L. Wang, Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 3633–3642.
[36] Y. Ji, G. Ye, H. Cheng, Interactive body part contrast mining for human interaction recognition, in: Proceedings of the IEEE International Conference on Multimedia and Expo Workshops (ICMEW), IEEE, 2014, pp. 1–6.
[37] J. Wang, Z. Liu, Y. Wu, J. Yuan, Mining actionlet ensemble for action recognition with depth cameras, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 1290–1297.

[38] Y. Yacoob, M.J. Black, Parameterized modeling and recognition of activities, Comput. Vis. Image Underst. 73 (2) (1999) 232–247.
[39] Y. Du, W. Wang, L. Wang, Hierarchical recurrent neural network for skeleton based action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2015, pp. 1110–1118.
[40] V. Veeriah, N. Zhuang, G.-J. Qi, Differential recurrent neural networks for action recognition, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), IEEE, 2015, pp. 4041–4049.
[41] A. Shahroudy, J. Liu, T.-T. Ng, G. Wang, Ntu rgb+ d: a large scale dataset for 3d human activity analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2016, pp. 1010–1019.
[42] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 7291–7299.
[43] I.C. Duta, B. Ionescu, K. Aizawa, N. Sebe, et al., Spatio-temporal vector of locally max pooled features for action recognition in videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 3205–3214.
[44] A. Kar, N. Rai, K. Sikka, G. Sharma, Adascan: adaptive scan pooling in deep convolutional neural networks for human action recognition in videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 3376–3385.
[45] A. Cherian, B. Fernando, M. Harandi, S. Gould, Generalized rank pooling for activity recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 1581–1590.
[46] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, L. Van Gool, Temporal segment networks: towards good practices for deep action recognition, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer, 2016, pp. 20–36.
[47] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2016, pp. 2818–2826.

Lihua Lu received the B.S. degree from Qingdao Agricultural University, in 2013. She is currently pursuing the Ph.D. degree in Computer Science at Beijing Institute of Technology, Beijing, China. Her main research interests include collective activity recognition, action recognition and video segmentation.

Huijun Di joined School of Computer Science at Beijing Institute of Technology in fall of 2012. He received his B.E. degree and Ph.D. degree in computer science from Tsinghua University in 2002 and 2009, respectively. He was a visiting scholar at Siemens Cooperate Research, Munich, Germany from 2008 to 2009. His postdoctoral research was carried out at Department of Computer Science and Technology, Tsinghua University from 2009 to 2012. His research interests include computer vision, pattern recognition and machine learning.

Yao Lu received the B.S. degree in electronics from Northeast University, Shenyang, China, in 1982 and the Ph.D. degree in computer science from Gunma University, Gunma, Japan, in 2003. He was a Lecturer and an Associate Professor with Hebei University, China, from 1986 to 1998, and a foreign researcher with Gunma University in 1999. In 2003, he was an invited professor with the Engineering Faculty, Gunma University, a Visiting Fellow of University of Sydney, Australia. He is currently a Professor with the Department of Computer Science, Beijing Institute of Technology, Beijing, China. He has published more than 100 papers in international conferences and journals. His research interests include neural network, image processing and video analysis, and pattern recognition.

Lin Zhang received the B.S and the M.S degrees in Control Science and Engineering from North China University of Technology in 2012 and 2015 respectively. She is currently pursuing the Ph.D. degree in Computer Science at Beijing Institute of Technology, China. Her main research interests include video saliency and segmentation.

Shunzhou Wang received the B.S. degree in Electrical Engineering and the M.S. degree in Control Engineering from Shanghai Institute of Technology, China, in 2015 and in 2018 respectively. He is currently pursuing the Ph.D. degree in Computer Science at Beijing Institute of Technology, China. His main research interests include multiple object tracking and crowd counting.