stagNet: An Attentive Semantic RNN for Group Activity Recognition

Mengshi Qi[1], Jie Qin[2,3], Annan Li[1], Yunhong Wang[1⋆],
Jiebo Luo[4], and Luc Van Gool[2]

[1] Beijing Advanced Innovation Center for Big Data and Brain Computing,
School of Computer Science and Engineering, Beihang University, China
[2] Computer Vision Laboratory, ETH Zurich, Switzerland
[3] Inception Institute of Artificial Intelligence, UAE
[4] Department of Computer Science, University of Rochester, USA

Abstract. Group activity recognition plays a fundamental role in a variety of applications, *e.g.* sports video analysis and intelligent surveillance. How to model the spatio-temporal contextual information in a scene still remains a crucial yet challenging issue. We propose a novel attentive semantic recurrent neural network (RNN), dubbed as stagNet, for understanding group activities in videos, based on the spatio-temporal attention and semantic graph. A semantic graph is explicitly modeled to describe the spatial context of the whole scene, which is further integrated with the temporal factor via structural-RNN. Benefiting from the 'factor sharing' and 'message passing' mechanisms, our model is capable of extracting discriminative spatio-temporal features and capturing inter-group relationships. Moreover, we adopt a spatio-temporal attention model to attend to key persons/frames for improved performance. Two widely-used datasets are employed for performance evaluation, and the extensive results demonstrate the superiority of our method.

Keywords: Group Activity Recognition · Spatio-temporal Attention · Semantic Graph · Scene Understanding

1 Introduction

Understanding dynamic scenes in sports and surveillance videos has a wide range of applications, such as tactics analysis and abnormal behavior detection. How to recognize/understand group activities within the scene, such as 'team spiking' in a volleyball match [23] (see Figure 1), is an important yet challenging issue, due to cluttered backgrounds and confounded relationships, *etc.*

Extensive efforts [33, 28, 51, 5, 44, 4, 31, 39, 38] have been made to address the above issue in the computer vision community. Fundamentally, spatio-temporal relations between people [17, 23, 25] are important cues for group activity recognition. There are two major issues in representing such information. One is the representation of visual appearance, which plays an important role in identifying

[⋆] Corresponding author: yhwang@buaa.edu.cn.

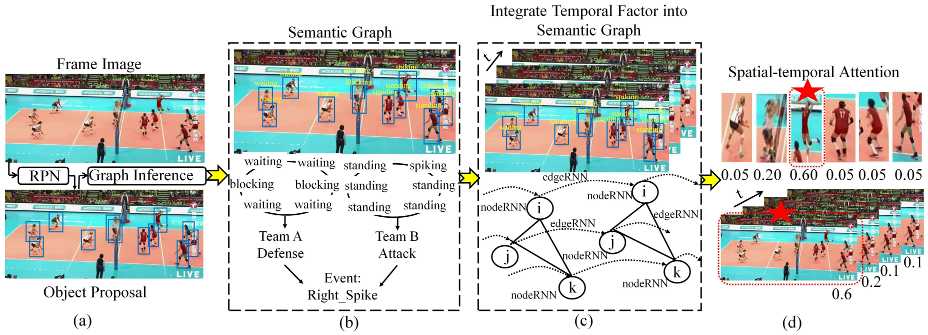


Fig. 1. Pipeline of the semantic graph-based group activity recognition. From left to right: (a) object proposals are extracted from raw frames by a region proposal network [14]; (b) the semantic graph is constructed from text labels and visual data; (c) temporal factor is integrated into the graph by using a structural-RNN, and the semantic graph is inferred via message passing and factor sharing mechanisms; (d) finally, a spatio-temporal attention mechanism is adopted for detecting key persons/frames (denoted with a red star) to further improve the performance.

people and describing their action dynamics. The other is the representation of spatial and temporal movement, which describes the interaction between people.

Traditional approaches for modeling the spatio-temporal information in group activity recognition can be summarized as a combination of hand-crafted features and probabilistic graph models. Hand-crafted features used in group activity recognition include motion boundary histograms (MBH) [16], histogram of gradients (HOG) [15], the cardinality kernel [19], *etc.* Markov Random Fields (MRFs) [8] and Conditional Random Fields (CRFs) [26] have been adopted to model the inter-object relationships.

An obvious limitation of the above approaches is that the low-level features they adopted fall short of representing complex group activities and dynamic scenes. With the success of convolutional neural networks (ConvNets) [27, 42, 20], deep feature representations have demonstrated their capabilities in representing complex visual appearance and achieved great success in many computer vision tasks. However, typical ConvNets regard a single frame of a video as input and output a holistic feature vector. With such architectures, spatial and temporal relations between consecutive frames cannot be explicitly discerned. The spatio-temporal relations [17, 23, 25] among the people are important cues for group activity recognition in the scene. They consist of the spatial appearance and temporal action of the individuals and their interaction. Recurrent Neural Networks (RNNs) [22, 11] are able to capture the temporal features from the video, and to represent dynamic temporal actions from the sequential data. Therefore, it is highly desirable to explore a RNN based network architecture that is capable of capturing the crucial spatio-temporal contextual information.

Moreover, automatically describing the semantic contents in the scene is helpful for better understanding the overall hierarchical structure of the scene (*e.g.* sports matches and surveillance videos). Yet, this task is very difficult, because the semantic description not only captures the personal action, but also expresses how these people relate to each other and how the whole group event occurs. If the above RNN based network can also describe the semantics in the scene, we can have a substantially clearer understanding of the dynamic scene.

In this paper, to address the above-mentioned issues, we propose a novel attentive semantic recurrent neural network named *stagNet* for group activity recognition, based on the spatial-temporal attention and semantic graph. In particular, individual activities and their spatial relations are inferred and represented by an explicit semantic graph, and their temporal interactions are integrated by a structural-RNN model. The network is further enhanced by a spatio-temporal attention mechanism to attach various levels of importance to different persons/frames in video sequences. More importantly, the semantic graph and spatio-temporal attention is collaboratively learned in an end-to-end fashion. The main contributions of this paper include:

- We construct a novel semantic graph to explicitly represent individuals' actions, their spatial relations, and group activity with a 'message passing' mechanism. To the best of our knowledge, we are the first to output a semantic graph for understanding group activities.
- We extend our semantic graph model to the temporal dimension via a structural-RNN, by adopting the 'factor sharing' mechanism in RNN.
- A spatio-temporal attention mechanism, which places emphasis on the key persons/frames in the video, is further integrated for better performance.
- Experiments on two benchmark datasets show that the performance of our framework is competitive with that of the state-of-the-art methods.

2 Related Work

Group Activity Recognition. Traditional approaches [28, 51, 5, 44, 41, 35, 3, 2] usually extract hand-crafted spatio-temporal features (*e.g.* MBH and HOG), followed by graph models for group activity recognition. Lan *et al.* [28] introduced an adaptive structure algorithm to model the latent structure. Amer *et al.* [5] formulated Hierarchical Random Field (HiRF) to model grouping nodes and the hidden variables in a scene. Shu *et al.* [44] conducted joint inference of groups, events and human roles with spatio-temporal AND-OR graph [4]. However, these approaches employed shallow features that could not encode higher-level information, and often lost temporal relationship information.

Recently, several deep models [23, 17, 43, 50, 6, 30] have been proposed for group activity recognition. Deng *et al.* [17] proposed a joint graphical model learned by gates between edges and nodes. Wang *et al.* [50] proposed a recurrent interaction context framework, which unified the features of individual person, intra-group and inter-group interactions. However, most of these works either extracted individual features regardless of the scene context or captured the context in an implicit manner without any semantic information. In this paper, we attempt to *explicitly* model the scene context via an intuitive spatio-temporal *semantic* graph [37] with RNNs. Moreover, we adopt a spatio-temporal attention model to attend to key persons/frames in the scene for better performance.

Deep Structure Model. Many researches have been conducted to make deep neural networks more powerful by integrating graph models. Chen *et al.* [10] combined Markov Random Fields (MRFs) with deep learning to estimate complex representations. Liu *et al.* [32] addressed semantic image segmentation by solving MRFs using Deep Parsing Network. In [55, 29, 49], structured-output learning was performed using deep neural networks for human pose estimation. Zheng *et al.* [57] integrated CRF-based probabilistic graphic model with RNN for semantic segmentation. Zhang *et al.* [56] improved object detection with deep ConvNets based on Bayesian optimization [46]. Most of these works were task-specific, however, they might fail to handle spatio-temporal modeling and extract interaction information from dynamic scenes. In [25], the Structural-RNN was proposed by combining high-level spatio-temporal graphs and Recurrent Neural Networks. Inspired by [25], we explicitly exploit a semantic spatio-temporal structure graph by injecting specific semantic information, such as inter-object and intra-person relationships, and space-time dynamics in the scene.

Attention Mechanism. Attention mechanisms [24, 34, 7, 9, 53, 54] have been successfully applied in the field of vision and language. An early work [24] introduced the saliency-based visual attention model for scene recognition. Mnih *et al.* [34] were the first to integrate RNNs with visual attention, and their model could extract selected regions by sequence. The mechanism proposed by [9] could capture visual attention with deep neural networks on special objects in images. Xu *et al.* [53] introduced two kinds of attention mechanisms for image caption. A temporal attention mechanism was proposed in [54] to select the most relevant frames based on text-generation RNNs. In this work, we integrate our spatio-temporal semantic graph and spatio-temporal attention into a joint framework, which is collaboratively trained in an end-to-end manner to attend to more relevant persons/frames in the video.

3 The Proposed Approach

The framework of the proposed approach for group activity recognition is illustrated in Figures 1 and 2. We utilize two-layer RNN and integrate two kinds of RNN units (*i.e.* nodeRNN and edgeRNN) into our framework, which is trained in an end-to-end fashion. In particular, the first part is to construct the semantic graph from input frames, and then we integrate the temporal factor by using a structural RNN. The inference is achieved via 'message-passing' and 'factor sharing' mechanisms. Finally, we adopt a spatio-temporal attention mechanism to detect key persons and frames to further improve the performance.

3.1 Semantic Graph

In this subsection, we introduce the semantic graph and the mapping from visual data to the graph. We inference the semantic graph to predict person's affiliations based on their positions and visual appearance. As shown in Figure 1(b), the semantic graph is built by parsing a scene with multiple people into a set of bounding boxes associated with the corresponding spatial positions. Each bounding box of a specific person is defined as a node of the graph. The graph edge that describes pairwise relations is determined by the spatial distance and temporal correlation, which will be introduced in Section 3.2.

To generate a set of person-level proposals (bounding boxes) from the t-th frame I^t in video I, we employ the region proposal network (RPN), which is part of the region-based fully convolutional networks [14]. The RPN outputs position-sensitive score maps as the relative position, and connects a position-sensitive region-of-interest (RoI) pooling layer on top of the fully convolutional layer. These proposals are regarded as input of the graph inference procedure. Throughout the graph modeling, three types of information are inferred: 1) the personal action label for each person, 2) the inter-group relationships in each frame, and 3) the group activity label of the whole scene.

In frame I^t, we denote a set of K bounding boxes as $B_{I^t} = (x_{t,1}, ..., x_{t,K})$, and the inter-person relationship set as R (*e.g.* whether two players belong to the same team on the Volleyball dataset). Given the group activity or scene labels set C_{scene}, and personal action labels set C_{action}, we denote $y^t \in C_{scene}$ as the scene class label, $x_i^{act} \in C_{action}$ as the action class label of the i-th person proposal, x_i^{pos} as its spatial coordinates, and $x_{i \to j} \in R$ as the predicted relationship between the i-th and j-th proposal boxes. Meanwhile, we denote the set of all variables to be $x = \{x_i^{act}, x_i^{pos}, x_{i \to j} \mid i = 1, ..., K; j = 1, ..., K; j \neq i\}$. Specifically, the semantic graph is built up by finding the optimal x^* and y^{t*} that maximize the following probability function:

$$\begin{aligned} &< x^*, y^{t*} >= \arg\max_{x, y^t} Pr(x, y^t \mid I^t, B_{I^t}), \\ &Pr(x, y^t \mid I^t, B_{I^t}) = \prod_{i,j \in K} \prod_{j \neq i} Pr(y^t, x_i^{act}, x_i^{pos}, x_{i \to j} \mid I^t, B_{I^t}). \end{aligned} \tag{1}$$

In the following, we will introduce how to infer the frame-wise semantic graph structure in detail.

3.2 Graph Inference

Inspired by [52], the graph inference is performed by using the mean field and computing the hidden states with Long Short-Term Memory (LSTM) network [22], which is an effective recurrent neural network. Let the semantic graph be $G = (S, V, E)$, where S is the scene node, and V and E are the object nodes and edges respectively. Specifically, S represents the global scene information in a video frame, an object node $v_i \in V$ $(i = 1, ..., K)$ indicates the person-level proposal, and the edge E corresponds to the spatial configuration of object nodes V in the frame. In the mean field inference, we approximate $Pr(x, y^t \mid \cdot)$ by $Q(x, y^t \mid \cdot)$, which only depends on the current states of each node and edge. The hidden state of the LSTM unit is the current state of each node and edge in the semantic graph. We define h^t as the current hidden state of scene node,

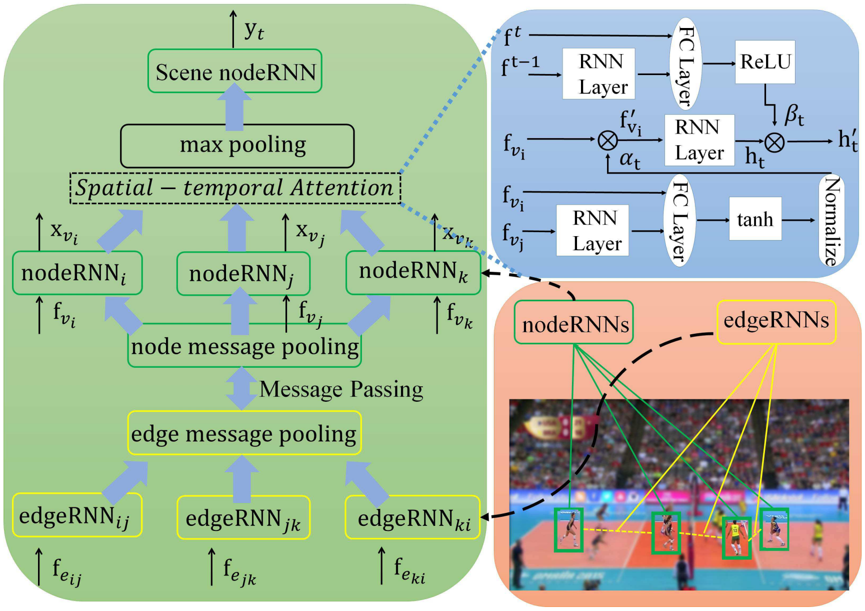


Fig. 2. Illustration of our nodeRNN and edgeRNN model. The model first extracts visual features of nodes and edges from a set of object proposals, and then takes the visual features as initial input to the nodeRNNs and edgeRNNs. We introduce the node/edge message pooling to update the hidden states of nodeRNNs and edgeRNNs. The input of nodeRNNs is the output of the edgeRNNs, and nodeRNNs also output the labels of personal actions. The max pooling is performed subsequently. Furthermore, a spatio-temporal attention mechanism is incorporated into our architecture. Finally, the top-most nodeRNN (*i.e.* Scene nodeRNN) outputs the label of group activity.

and h_{v_i} and $h_{e_{ij}}$ as the current hidden state of node i and edge $i \to j$, respectively. Notably, all the nodeRNNs share the same set of parameters and all the edgeRNNs share another set of parameters. The solution to $Q(x, y^t \mid I^t, B_{I^t})$ can be obtained by computing the mean field distribution as follows:

$$\begin{aligned} & Q(x, y^t \mid I^t, B_{I^t}) \\ = & \prod_{i=1}^{K} Q(x_i^{act}, x_i^{pos}, y^t \mid h_{v_i}, h^t) Q(h_i \mid f_{v_i}) Q(h^t \mid f^t) \\ & \prod_{j \neq i} Q(x_{i \to j} \mid h_{e_{ij}}) Q(h_{e_{ij}} \mid f_{e_{ij}}), \end{aligned} \tag{2}$$

where f^t is the convolutional feature of the scene in the t-th frame, f_{v_i} is the feature of the i-th node, and $f_{e_{ij}}$ is the feature of the edge connecting the i-th node and j-th node, which is the unified bounding box over two nodes. The feature $f_{e_{ij}}$ has six elements by computing the basic distances and direction vectors, which include $< |dx|, |dy|, |dx + dy|, \sqrt{(dx)^2 + (dy)^2}, \arctan(dy, dx), \arctan2(dy, dx) >$.

All of these features are extracted by the RoI pooling layer. Then the messages aggregated from other previous LSTM units are fed into the next step.

As shown in Figure 2, the edgeRNNs provide contextual information for the nodeRNNs, and the max pooling is performed over the nodeRNNs. The nodeRNN concatenates the node feature and the outputs of edge-RNN accordingly. The edgeRNN passes the summation of all edge features that are connected to the same node as the message. The edgeRNNs and nodeRNNs take the visual features as initial input and produce a set of hidden states. The model iteratively updates the hidden states of the RNN. Finally, the hidden states of the RNN are used to predict the frame-wise scene label, personal action label, person position information and inter-group relationships.

Message passing [52] can iteratively improve the efficacy of inference in the semantic graph. In the graph topology, the neighbors of the egdeRNNs are nodeRNNs. Passing messages through the whole graph involves two subgraphs: *i.e.* node-centric sub-graph and edge-centric sub-graph respectively. For node-centric sub-graph, the nodeRNN receives messages from its neighboring edgeRNNs. Similarly, for edge-centric sub-graph, the edgeRNN gets messages from its adjacent nodeRNNs. We adopt an aggregation function called message pooling to learn adaptive weights for modeling the importance of passed messages. We compute the weight factors for each incoming message and aggregate the messages via a total weight for representation. It is demonstrated that this method is more effective than average pooling or max pooling [52].

Specifically, we denote the update message input to the i-th node v_i as m_{v_i}, and the message to the edge between the i-th and j-th node e_{ij} as $m_{e_{ij}}$, respectively. Then, we compute the message passed into the node considering its own hidden state h_{v_i} and the hidden states of its connected edges $h_{e_{ij}}$ and $h_{e_{ji}}$, and obtain the message passed into the edge with respect to the hidden state of its adjacent nodes h_{v_i} and h_{v_j}. Formally, m_{v_i} and $m_{e_{ij}}$ are computed as

$$\begin{aligned} m_{v_i} &= \sum_{j:i\rightarrow j} \sigma(U_1^T[h_{v_i}, h_{e_{ij}}])h_{e_{ij}} + \sum_{j:j\rightarrow i} \sigma(U_2^T[h_{v_i}, h_{e_{ji}}])h_{e_{ji}}, \\ m_{e_{ij}} &= \sigma(W_1^T[h_{v_i}, h_{e_{ji}}])h_{v_i} + \sigma(W_2^T[h_{v_j}, h_{e_{ij}}])h_{v_j}, \end{aligned} \tag{3}$$

where W_1, W_2, U_1 and U_2 are parameters to be learned, σ is a sigmoid function, and $[\cdot, \cdot]$ means the concatenation of two hidden vectors. Finally, we utilize these messages to update the hidden states of nodeRNN and edgeRNN iteratively. Once finishing updating, the hidden states are then employed to predict personal action categories, bounding box offsets and relationship types.

3.3 Integrating Temporal Factors

With the semantic graph of a frame, temporal factors are further integrated to form the spatio-temporal semantic graph (see Figure 1(c)). Particularly, we adopt the structural-RNN [25] to model the spatio-temporal semantic graph. Based on the graph definition in Sections 3.1 and 3.2, we add a temporal edge E_T, such that $G = (S, V, E_S, E_T)$, where E_S refers to the spatial edge. The node

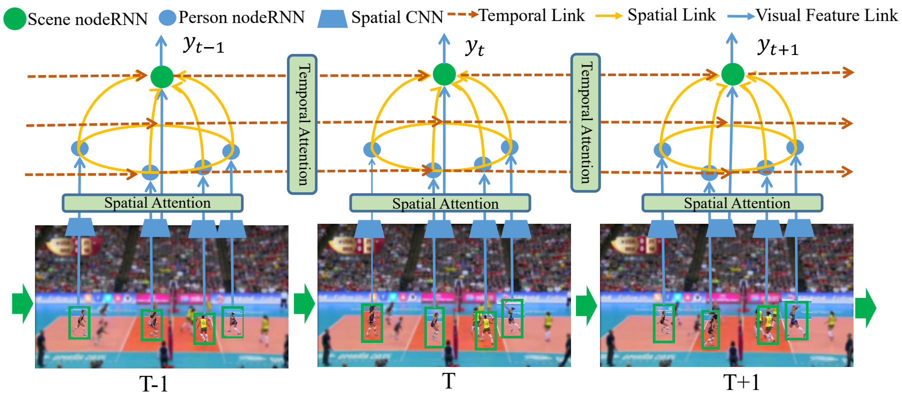


Fig. 3. Hierarchical semantic RNN structure for a volleyball match. Given object proposals and tracklets of all players, we feed them into spatial CNN, followed by a RNN to represent each player's action and appearance of the whole scene. Then we adopt structural-RNN to establish temporal links for a sequence of frames. Furthermore, we integrate the LSTM based spatio-temporal attention mechanism into the model. The output layer classifies the whole team's group activity.

$v_i \in V$ and edge $e \in E_S \cup E_T$ in the spatio-temporal semantic graph enrolls over time. Specifically, the nodes at adjacent time steps, *e.g.* the node v_i at time t and time $t+1$ are connected with the temporal edge $e_{ii} \in E_T$. Denote the node label as y_v^t and the corresponding feature vectors for node and edge are denoted as f_v^t, f_e^t at time t, respectively. We introduce a 'factor sharing' mechanism, which indicates that the nodes denoting the same person and the edges representing the same relationship tend to share factors (*e.g.* parameters, original hidden states of RNNs) across different video frames. Figure 3 shows an example of structural-RNN across three time steps in a volleyball game video. Please refer to [25] for more technical details about structural-RNN.

We define two kinds of edges (edgeRNN) in the spatio-temporal graph. One is spatial-edgeRNN representing the spatial relationship. It is formed by the spatial message pooling in each frame and computed from the neighbor player's nodeRNN using the Euclidean distance. The other is temporal-edgeRNN that connects neighbor frames of the same player to represent the temporal information. It is formed by sharing factors between players' nodeRNNs in a video sequence. We incorporate the features of the spatial edgeRNNs between two consecutive frames into the temporal edgeRNN, resulting in 12 additional features.

During the training phase, the errors of predicting the labels of scene nodes and object nodes are back-propagated through the sceneRNNs, nodeRNNs and edgeRNNs. The passed messages represent the interactions between nodeRNNs and edgeRNNs. The nodeRNN is connected to the edgeRNN, and outputs the personal action labels. Every edgeRNN simultaneously models the semantic interaction between adjacent nodes and the evolution of interaction over time.

3.4 Spatio-Temporal Attention Mechanism

The group activity involves multiple persons, but only few of them play decisive roles in determining the activity. For example, the 'winning point' in a volleyball match often occurs with a specific player spiking the ball and another player failing to catch the ball. For a better understanding of the group activity, it is necessary to attend higher levels of importance to key persons. Inspired by [40, 47], we attend to a set of features of different regions at each time step, which contain key persons or objects, with a spatio-temporal soft attention mechanism. With the attention model, we can focus on specific persons in specific frames to improve the recognition accuracy of the group activity.

Since person-level attention is often affected by the evolution and state of the group activity, the context information needs to be taken into consideration. Particularly, we combine the proposals of the same person with KLT trackers [36]. The whole representation of a player can be extracted by incorporating the context information from a sequence of frames.

Person-Level Spatial Attention We apply a spatial attention model to assign weights to different persons via LSTM networks. Specifically, given one frame that involves K players $x_t = (x_{t,1}, ..., x_{t,K})$, we define the scores $s_t = (s_{t,1}, ..., s_{t,K})^T$ as the importance of all person-level actions in each frame:

$$s_t = W_s \tanh(W_{xs} x_t + U_{hs} h^s_{t-1} + b_s), \tag{4}$$

where W_s, W_{xs}, U_{hs} are the learnable parameter matrices, and b_s is the bias vector. h^s_{t-1} is the hidden variable from an LSTM unit. For the k-th person, the spatial attention weight is computed as a normalization of the scores:

$$\alpha_{t,k} = \frac{\exp(s_{t,k})}{\sum_{i=1}^{K} \exp(s_{t,i})}. \tag{5}$$

Subsequently, the input to the LSTM unit is updated as $x'_t = (x'_{t,1}, ..., x'_{t,K})^T$, where $x'_{t,k} = \alpha_{t,k} x_{t,k}$. Then the representation of the attended player can be used as the input to the RNN nodes in the spatio-temporal semantic graph described in Section 3.1.

Frame-Level Temporal Attention We adopt a temporal attention model to discover the key frames. For T frames in a video, the temporal attention model is composed of an LSTM layer, a fully connected layer and a nonlinear ReLU unit. The temporal attention weight of the t-th frame can be computed as

$$\beta_t = \mathrm{ReLU}(W_{x\beta} x_t + U_{h\beta} h^{\beta}_{t-1} + b_\beta), \tag{6}$$

where x_t is the current input and h^{β}_{t-1} is the hidden variables at time step t-1. The temporal attention weight controls how much information of every frame can be used for the final recognition. Receiving the output z_t of the main LSTM network and the temporal attention weight β_t at each time step t, the important scores for C_{scene} classes are the weighted summation w.r.t. all time steps:

$$o = \sum_{t=1}^{T} \beta_t \cdot z_t, \tag{7}$$

where $o = (o_1, o_2, \cdots, o_{C_{scene}})^T$. The probability that a video I belongs to the i-th class is

$$p(C^i_{scene}|I) = \frac{e^{o_i}}{\sum_{j=1}^{C_{scene}} e^{o_j}}. \tag{8}$$

3.5 Joint Objective Function

Finally, we formulate the overall objective function with a regularized cross-entropy loss, and combine the semantic graph modeling and the spatio-temporal attention network learning as

$$\begin{aligned} L = &- \sum_{i=1}^{C_{scene}} y^i \log \hat{y^i} - \frac{1}{K} \sum_{i=1}^{K} x^*_i \log \hat{x^*_i} + \\ &\lambda_1 \sum_{k=1}^{K} (1 - \frac{\sum_{t=1}^{T} \alpha_{t,k}}{T})^2 + \frac{\lambda_2}{T} \sum_{t=1}^{T} \|\beta_t\|_2 + \lambda_3 \|W\|_1, \end{aligned} \tag{9}$$

where y^i and x^*_i denote the ground-truth label of group activity and personal action, respectively. If a video sequence is classified as the i-th category, $y^i = 1$ and $y^j = 0$ for $j \neq i$. $\hat{y^i} = p(C^i_{scene}|I)$ is the probability that a sequence is classified as the i-th category. $\hat{x}^*_i = p(C^i_{action}|B_{I^t})$ is the probability that a personal action belongs to the i-th category. For classification, we perform max pooling over the hidden representations followed by a softmax classifier. λ_1, λ_2 and λ_3 denote regularization terms. The third regularization term ensures to attend to more persons in the spatial space, and the fourth term regularizes the learned temporal attention via ℓ_2 normalization. The last term regularizes all the parameters of the spatio-temporal attention mechanism [47].

4 Experiments

We evaluate our framework on two widely-adopted benchmarks, *i.e.* the Collective Activity dataset for group activity recognition, and the Volleyball dataset for group activity recognition and personal action recognition.

Collective Activity [13] contains 44 video clips (about 2,500 frames captured by low-resolution cameras), in which there are five group activities: *crossing, waiting, queueing, walking* and *talking*, and six individual actions: *N/A, crossing, waiting, queueing, walking* and *talking*. The group activity label is predicted based on the majority of people's actions. Following the same experimental setting in [28], we use the tracklet data provided in [12]. The scene is modeled as a bag of individual action context feature descriptors, and we select 1/3 of the video clips for testing and the rest for training.

Volleyball [23] contains 55 volleyball videos with 4,830 annotated frames. Each player is labeled with a bounding box and one of the nine personal action labels: *waiting, setting, digging, falling, spiking, blocking, jumping, moving* and *standing*. The whole frame is annotated with one of the eight group activity labels: *right set, right spike, right pass, right winpoint, left winpoint, left pass, left spike* and *left set*. Following [23], we choose 2/3 of the videos for training and the remaining 1/3 for testing. Particularly, we split all the players in each frame into two groups using the strategy in [23], and define four additional team-level activities: *attack, defense, win* and *lose*. The labeled data are beneficial for training our semantic RNN model.

4.1 Implementation Details

Our model is implemented using the TensorFlow [1] library. We adopt the VGG-16 model [45] pre-trained on ImageNet, which is then fine-tuned on the Collective Activity and Volleyball datasets, respectively. Based on [14], we only employ the convolution layers of VGG-16 and concatenate a 1024-d 1 × 1 convolutional layer. As such, each frame is represented by a 1024-d feature vector. Specifically, a person bounding box is represented as a 2805-d feature vector, which includes 1365-d appearance information and 1440-d spatial information. Based on the RPN detector [14], the appearance features can be extracted by feeding the cropped and resized bounding box through the backbone network, and utilizing spatially pooling to obtain the response map from a lower layer. To represent the bounding box at multiple scales, we follow [14] and employ spatial pyramid pooling [14], with respect to a 32 × 32 spatial histogram.

The LSTM layers used as nodes and edges contain 1024-d hidden units, and they are trained by adding a softmax loss on top of the output at each time step. We use a softmax layer to produce the score maps for the group activity class and action class. The batch size for training the bottom layer of LSTM and fully connected layer of RPN is 8, and the training is performed within 20,000 iterations. The top layer of LSTM is trained in 10,000 iterations with a batch size of 32. For optimization, we adopt RMSprop [21] with a learning rate ranging from 0.00001 to 0.001 for mini-batch gradient descent. In practice, we set $\{\lambda_1,\lambda_2,\lambda_3\}$ as $\{0.001,0.0001,0.0001\}$ for Collective Activity, and $\{0.01,0.001,0.00001\}$ for Volleyball. Besides, the training and output semantic graph in our paper is recorded as a JavaScript Object Notation (JSON) file, which is a popular tool for extracting structure data.

4.2 Compared Methods

We compare our approach with VGG-16 Network [45], LRCN [18], HDTM [23], Contextual Model [28], Deep Structure Model [17], Cardinality Kernel [19],

Table 1. Performance comparison of our method and the state-of-the-art approaches.

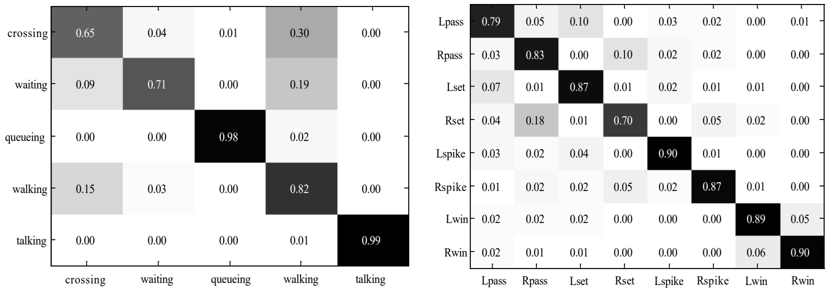
Methods	Semantic?	Accuracy		
		Collective Activity	Volleyball (Group)	Volleyball (Personal)
VGG-16-Image [45]	×	68.3	71.7	-
VGG-16-Person [45]	×	71.2	73.5	-
LRCN-Image [18]	×	64.2	63.1	-
LRCN-Person [18]	×	64.0	67.6	-
HDTM (1 group) [23]	×	81.5	70.3	75.9
HDTM (2 groups) [23]	×	-	81.9	-
Contextual Model [28]	×	79.1	-	-
Deep Structure Model [17]	×	80.6	-	-
Cardinality kernel [19]	×	83.4	-	-
CERN-1 (1 group) [43]	×	84.8	34.4	69.0
CERN-2 (1 group) [43]	×	87.2	73.5	-
CERN-2 (2 groups) [43]	×	-	83.3	-
SSU-temporal (MRF) [6]	×	-	87.1	-
SSU-temporal (GT) [6]	×	-	89.9	82.4
Ours w/o attention (PRO)	√	85.6	85.7	79.6
Ours w/ attention (PRO)	√	87.9	87.6	-
Ours w/o attention (GT)	√	87.7	87.9	81.9
Ours w/ attention (GT)	√	89.1	89.3	-

'PRO' and 'GT' indicate that we use proposal-based and ground-truth bounding boxes [23], respectively. The best performance is highlighted in red and the second best in blue.

CERN [43] and SSU [6]. Particularly, in Table 1, 'VGG-16-Image' and 'LRCN-Image' utilize the holistic image features in a single frame for recognition. 'VGG-16-Person' and 'LRCN-Person' predict group activities with features pooled over all fixed-size individual person-level features. 'HDTM' and 'CERN' conduct experiments on the Volleyball Dataset using the grouping strategy, which divides all persons into one or two groups. 'SSU-temporal' models adopted two kinds of detection methods on the Volleyball Dataset, with one using the ground truth (GT) bounding boxes, and the other using Markov Random Fields (MRF) based detection. Note that 'LRCN', 'HDTM' and 'Deep Structure Model' adopt the AlexNet [27] as the backbone, and 'SSU' employs the Inception-V3 [48] framework, while 'CERN' and our model utilize the VGG-16 architecture.

4.3 Results and Analysis

Results on the Collective Activity Dataset. The experimental results of group activity recognition are shown in Table 1. As can be seen, our model with the attention model achieves the best performance among the compared state-of-the-art methods, regardless of using the proposal-based or ground-truth bounding boxes. For instance, our model achieves ≈15% higher in accuracy than image-level and person-level classification methods, mostly because of our RNN-based semantic graph with the iteratively message passing scheme. Meanwhile,



(a) Collective Activity (b) Volleyball

Fig. 4. Confusion matrices for the two group activity datasets.

our method is the only one that incorporates semantics into the model. The improved performance also indicates that the spatio-temporal semantic graph is beneficial for improving the recognition performance. Note that the cardinality kernel approach [19] achieves the best performance among non-deep learning methods. This approach predicts the group activity label by directly counting the numbers of individual actions based on hand-crafted features. In addition, we draw the confusion matrix based on our model with the spatio-temporal attention in Figure 4(a). We can observe that nearly 100% recognition accuracies can be obtained in terms of ‘queueing’ and ‘talking’, proving the effectiveness of our framework. However, there are also some failure cases, which is probably due to that some action classes share high similarities, such as ‘walking’ and ‘crossing’. More training data are needed for distinguishing these action categories.

Results on the Volleyball Dataset. The recognition results of our method and the state-of-the-art ones are shown in Table 1. As we can see, the group activity and personal action recognition accuracies of our model are superior to most state-of-the-art methods, and also highly competitive to the best ‘SSU’ method. It should be noted that ‘SSU’ obtains the bounding boxes by a much more sophisticated multi-scale method and adopts the more advanced Inception-V3 as the backbone. In contrast, we just employ the basic VGG-16 model, and the ‘ground-truth’ bounding boxes provided by [23] are obtained with a relatively simple strategy. Hence, it can be expected that our performance could be further improved by adopting more advanced backbone networks. Besides, our model outperforms other RNNs based methods by about $5 \sim 8\%$ w.r.t. group activity recognition, since our semantic graph with structural-RNN can capture spatio-temporal relationships. Integrating the attention model can further improve the recognition performance, indicating that key persons’ visual features are crucial for recognizing the whole scene label. It is also worth noting that all the other methods, including ‘SSU’, could not extract the semantic structural information to describe the scene context. On the contrary, our method can output the semantic description of the scene owing to our semantic graph model. We visually depict the recognition results in Figure 5, including semantic

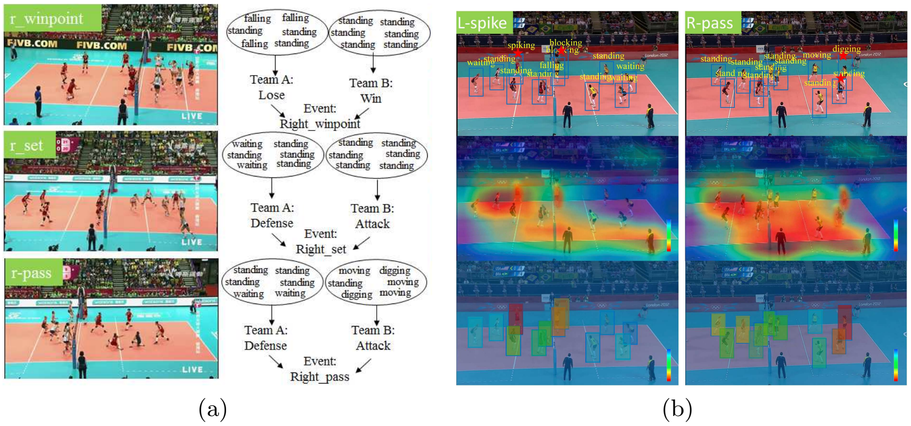

r_winpoint
r_set
r-pass
falling falling standing standing falling standing
standing standing standing standing standing standing
Team A: Lose
Team B: Win
Event: Right_winpoint
waiting standing standing standing waiting standing
standing standing standing standing standing standing
Team A: Defense
Team B: Attack
Event: Right_set
standing standing standing standing waiting waiting
moving digging standing moving digging moving
Team A: Defense
Team B: Attack
Event: Right_pass
L-spike
R-pass
(a)
(b)

Fig. 5. Visualization of results on the Volleyball dataset. (a) Semantic graphs obtained by our method. (b) From top to bottom: group activity and personal action recognition results; attention heat maps using proposal-based bounding boxes; attention heat maps using ground-truth bounding boxes. The important persons are denoted with red stars. The attention weights decrease along with the colors changing from red to blue.

graphs and attention heat maps. In addition, the confusion matrix using our method is shown in Figure 4(b). As we can see from the figure, our method can achieve promising recognition accuracies (≥87%) in terms of the majority of group activities.

5 Conclusion

In this paper, we presented a novel RNN framework (*i.e.* stagNet) with semantic graph and spatio-temporal attention for group activity recognition. The stagNet could explicitly extract spatio-temporal inter-object relationships in a dynamic scene with a semantic graph. Through the inference procedure of nodeRNNs and edgeRNNs, our model could simultaneously predict the label of the scene and inter-person relationships. By further integrating the spatio-temporal attention mechanism, our framework attended to important persons or frames in the video, leading to enhanced recognition performance. Extensive results on two widely-adopted benchmarks showed that our framework achieved competitive results to the state-of-the-art methods, whilst uniquely outputting the semantic description of the scene.

Acknowledgements This work was partly supported by the National Natural Science Foundation of China (No. 61573045) and the Foundation for Innovative Research Groups through the National Natural Science Foundation of China (No. 61421003). Jiebo Luo would like to thank the support of New York State through the Goergen Institute for Data Science and NSF Award (No. 1722847). Mengshi Qi acknowledges the financial support from the China Scholarship Council.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M.: Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv (2016)
2. Amer, M.R., Todorovic, S.: Sum product networks for activity recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(4), 800–813 (2016)
3. Amer, M.R., Todorovic, S., Fern, A., Zhu, S.C.: Monte carlo tree search for scheduling activity recognition. In: ICCV. IEEE (2013)
4. Amer, M.R., Xie, D., Zhao, M., Todorovic, S., Zhu, S.C.: Cost-sensitive top-down/bottom-up inference for multiscale activity recognition. In: ECCV. Springer International Publishing (2012)
5. Amer, M.R., Lei, P., Todorovic, S.: Hirf: Hierarchical random field for collective activity recognition in videos. In: ECCV. Springer International Publishing (2014)
6. Bagautdinov, T., Alahi, A., Fleuret, F., Fua, P., Savarese, S.: Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In: CVPR. IEEE (2017)
7. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: ICLR (2015)
8. Bengio, Y., LeCun, Y., Henderson, D.: Globally trained handwritten word recognizer using spatial representation, convolutional neural networks, and hidden markov models. In: NIPS. MIT Press (1994)
9. Cao, C., Liu, X., Yang, Y., Yu, Y.: Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In: ICCV. IEEE (2015)
10. Chen, L.C., Schwing, A.G., Yuille, A.L., Urtasun, R.: Learning deep structured models. ICLR (2014)
11. Cho, K., Van Merriënboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: Encoder-decoder approaches. arXiv (2014)
12. Choi, W., Savarese, S.: A unified framework for multi-target tracking and collective activity recognition. In: ECCV. Springer International Publishing (2012)
13. Choi, W., Shahid, K., Savarese, S.: What are they doing? : Collective activity classification using spatio-temporal relationship among people. In: ICCV Workshops. IEEE (2009)
14. Dai, J., Li, Y., He, K., Sun, J.: R-fcn: Object detection via region-based fully convolutional networks. In: NIPS. MIT Press (2016)
15. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. IEEE (2005)
16. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: ECCV. Springer International Publishing (2006)
17. Deng, Z., Vahdat, A., Hu, H., Mori, G.: Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In: CVPR. IEEE (2016)
18. Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M.: Long-term recurrent convolutional networks for visual recognition and description. In: CVPR. IEEE (2015)
19. Hajimirsadeghi, H., Yan, W., Vahdat, A., Mori, G.: Visual recognition by counting instances: A multi-instance cardinality potential kernel. In: CVPR. IEEE (2015)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. IEEE (2016)

21. Hinton, G., Srivastava, N., Swersky, K.: Neural networks for machine learning-lecture 6a-overview of mini-batch gradient descent
22. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997)
23. Ibrahim, M.S., Muralidharan, S., Deng, Z., Vahdat, A., Mori, G.: A hierarchical deep temporal model for group activity recognition. In: CVPR. IEEE (2016)
24. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **20**(11), 1254–1259 (1998)
25. Jain, A., Zamir, A.R., Savarese, S., Saxena, A.: Structural-rnn: Deep learning on spatio-temporal graphs. In: CVPR. IEEE (2016)
26. Krahenbuhl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: NIPS. MIT Press (2011)
27. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. MIT Press (2012)
28. Lan, T., Wang, Y., Yang, W., Robinovitch, S.N., Mori, G.: Discriminative latent models for recognizing contextual group activities. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(8), 1549–62 (2012)
29. Li, S., Zhang, W., Chan, A.B.: Maximum-margin structured learning with deep networks for 3d human pose estimation. In: ICCV. IEEE (2015)
30. Li, X., Chuah, M.C.: Sbgar: Semantics based group activity recognition. In: CVPR. IEEE (2017)
31. Liu, J., Carr, P., Collins, R.T., Liu, Y.: Tracking sports players with context-conditioned motion models. In: CVPR. IEEE (2013)
32. Liu, Z., Li, X., Luo, P., Loy, C.C., Tang, X.: Semantic image segmentation via deep parsing network. In: ICCV. IEEE (2015)
33. Lu, W.L., Ting, J.A., Little, J.J., Murphy, K.P.: Learning to track and identify players from broadcast sports videos. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(7), 1704–1716 (2013)
34. Mnih, V., Heess, N., Graves, A., Kavukcuoglu, K.: Recurrent models of visual attention. In: NIPS. MIT Press (2014)
35. Mori, G.: Social roles in hierarchical models for human activity recognition. In: CVPR. IEEE (2012)
36. Munkres, J.: Algorithms for the assignment and transportation problems. Journal of the society for industrial and applied mathematics **5**(1), 32–38 (1957)
37. Qi, M., Wang, Y., Li, A.: Online cross-modal scene retrieval by binary representation and semantic graph. In: MM. ACM (2017)
38. Qin, J., Liu, L., Shao, L., Ni, B., Chen, C., Shen, F., Wang, Y.: Binary coding for partial action analysis with limited observation ratios. In: CVPR (2017)
39. Qin, J., Liu, L., Shao, L., Shen, F., Ni, B., Chen, J., Wang, Y.: Zero-shot action recognition with error-correcting output codes. In: CVPR (2017)
40. Ramanathan, V., Huang, J., Abuelhaija, S., Gorban, A., Murphy, K., Li, F.F.: Detecting events and key actors in multi-person videos. In: CVPR. IEEE (2016)
41. Ryoo, M.S., Aggarwal, J.K.: Stochastic representation and recognition of high-level group activities: Describing structural uncertainties in human activities. International Journal of Computer Vision **93**(2), 183–200 (2011)
42. S, R., K, H., R, G., J, S.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(6), 1137 (2017)
43. Shu, T., Todorovic, S., Zhu, S.C.: Cern: Confidence-energy recurrent network for group activity recognition. In: CVPR. IEEE (2017)

44. Shu, T., Xie, D., Rothrock, B., Todorovic, S.: Joint inference of groups, events and human roles in aerial videos. In: CVPR. IEEE (2015)
45. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
46. Snoek, J., Larochelle, H., Adams, R.P.: Practical bayesian optimization of machine learning algorithms. NIPS **4**, 2951–2959 (2012)
47. Song, S., Lan, C., Xing, J., Zeng, W., Liu, J.: An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In: AAAI. AAAI (2017)
48. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR (2016)
49. Tompson, J., Jain, A., Lecun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. In: NIPS. MIT Press (2014)
50. Wang, M., Ni, B., Yang, X.: Recurrent modeling of interaction context for collective activity recognition. In: CVPR. IEEE (2017)
51. Wang, Z., Shi, Q., Shen, C., Anton, V.D.H.: Bilinear programming for human activity recognition with unknown mrf graphs. In: CVPR. IEEE (2013)
52. Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L.: Scene graph generation by iterative message passing. In: CVPR. IEEE (2017)
53. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: ICML. ACM (2015)
54. Yao, L., Torabi, A., Cho, K., Ballas, N.: Describing videos by exploiting temporal structure. In: ICCV. IEEE (2015)
55. Zhang, N., Paluri, M., Ranzato, M., Darrell, T., Bourdev, L.: Panda: Pose aligned networks for deep attribute modeling. In: CVPR. IEEE (2014)
56. Zhang, Y., Sohn, K., Villegas, R., Pan, G., Lee, H.: Improving object detection with deep convolutional networks via bayesian optimization and structured prediction. In: CVPR. IEEE (2015)
57. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.S.: Conditional random fields as recurrent neural networks. In: ICCV. IEEE (2015)