Deep Fully Connected Model For Collective Activity Recognition

Jichao Liu¹, Chuanxu Wang¹, Yuting Gong¹ and Xue Hao¹
¹School of information science and technology, Qingdao University of Science & Technology, 266100 QingDao, China
Corresponding author: Chuanxu Wang (e-mail: wangchuanxu_qd@qust.edu.cn).

ABSTRACT Group activity recognition is a challenging task, because there are an exponentially large number of semantic and geometrical relationships among individuals. This makes it difficult to model these interactions and merge them as a whole for group activity classification. In this paper we propose a deep fully-connected model for group recognition, firstly we use the spatial-temporal model based on Convolution Neural Network (CNN) and Long Short-Term Memory Networks (LSTM) network to capture the dynamic features of each person. Then, we use the Fully-Connected Conditional Random Field (FCCRF) to learn the interactions between people. Finally, with FCCRF potential functions we re-fine the activity recognition predicted by the spatial-temporal model. Experimental results on Collective Activity Data-set and Collective Activity Extended Data-set show that our model improves recognition accuracy over baseline methods and gets competitive results in comparison to the state-of-the-art models.

INDEX TERMS Group behavior recognition, Fully-connected Conditional Random Field, Interaction modeling, potential function.

I. INTRODUCTION

In recent years, vision-based human activity recognition has become a hot research direction. For it is of great economic value in video surveillance, human-computer interaction, motion understanding and video retrieval. Especially, with the rising of deep learning, which has shown greater advantages in feature representation than hand-crafted methods, it promotes a rapid development of group activity recognition algorithms.

In [1], a hierarchical deep temporal model based on LSTM is proposed, the two-layer LSTM network is used to represent the behavior dynamics of each individual in the sequence, then their behavior information is aggregated, and finally the volleyball behavior is observed. Document [2] proposes a model based on CNN and LSTM networks, which comprehensively utilizes the characteristics of basketball players, inter-frame features and trajectory features to complete the classification of basketball behaviors. Document [3] proposes a Confidence-Energy Recurrent Network model based on two-layer LSTM network, using energy layer instead of ordinary softmax layer to predict their group activity under the condition of minimum energy. The above methods are all based on the models built by the deep learning network, generally speaking, their individual behavior classification methods are simple and with high accuracy, but they do not consider the interactions between individuals in case of group activity recognition.

In fact, interaction information among individuals plays an important role in terms of collective activity recognition. As shown in Figure1, if only considering the single person behavior, selected with red and orange bounding box respectively in two different scenarios, their behavior properties are the same as "standing"; after considering their interaction information in each contextual background, we can easily distinguish that the individual with the red bounding box in the left group in Figure1 is in "queuing" and the individual with the orange bounding box is in "discussing". It can be seen that the key cue to identify group activity is to describe the interactions among individuals in the group.

To solve the above problems, our paper proposes a CRF-based graphical model to describe the interactions between individuals pair-wisely in the group, and combines the deep learning network to identify group activity. Firstly, using CNN and LSTM network to obtain the dynamic information of the single-person behavior as well as the preliminary prediction of the group activity; then using
Conditional Random Field graphical model to describe the interactions between people in the group; finally integrating the knowledge from both CRF and deep learning network to refine the individual and group activity recognition. This means that our method not only gives an overall spatial-temporal representation for individuals, but also models long-range spatial-temporal interactions among them. Especially, we use Fully-Connected Conditional Random Field (FCCRF), which consider all the interactions among individuals in a video clip, and adjust the interaction strength accordingly based on the degree of their similarity.

The main contributions of this paper are as follows:

1) Proposing a graphical framework based on deep learning network to simulate the interactions between people in group activity;

2) Using fully-connected Conditional Random Field model to correct the prediction errors generated by the deep learning network;

The remaining contents are organized as below, in section 2 we review the related work of group activity recognition; in section 3 we describe the conditional random field model based on deep learning network; behavior classifications are analyzed in section 4; model training is briefly introduced in part 5; finally in section 6 we present the experimental results analysis and compare them to other models.

Figure 1. The same individual appearance in different group activity

II. RELATED WORK

Research methods on group activity recognition are diverse. It is proverbial that tracking algorithm is regarded as the basic preprocessing for group detection. Paolo Rota [4] et al. proposed the motion characteristics of marker-based interactive particles by learning the MLP neural network to create a consistent group, which can clearly represent the tracked object over time. Mykhaylo Andriluka [5] et al. proposed a video-based benchmark for human pose estimation and tracking to facilitate video analysis. Mohib Ullah [6] et al. proposed a directional sparse graphical model for multi-target tracking, by cascading spatial constraints with given weighting factors, they achieved a good tracking of multiple targets. Habib Ullah [7] et al. proposed a novel Gaussian Kernel based Integration Model (GKIM) for anomalous entities detection and localization in pedestrian flows, and the trained R-CRF model was used to detect and localize the anomalous entities. The above four methods have been very effective in detection, positioning and tracking of the human body, which are helpful to further group structure analysis and collective behavior recognition.

Traditional hand-crafted feature based methods were thoroughly explored and achieved fairly good group behavior recognition accuracy. Lan [8] et al. proposed a latent variable framework for capturing individual behaviors in groups and their interactions, an Action Context (AC) descriptor is used to eliminate ambiguity in group activity recognition. Choi [9] et al. proposed a group activity recognition framework, which is to track multiple people at the same time, and combine the detected trajectory information with video context information to analyze group activity at different semantic granular levels in the video. Amer [10] et al. constructed the HiRF model to locate the time period in which group activity occurred, and capture the time-space information of long-range and higher-order in video features to enhance recognition. But all of the above methods use artificially hand-crafted features, often relying on specific data sets. When the data source changes, the features need to be redesigned, so their adaption is poor.

In recent years, with the development of deep learning networks, more and more methods use neural networks to study group activity. Document [11] by adding a context-aware inference mechanism focused on capturing spatio-temporal interactions to build a hierarchical description for group activity understanding. Document [12-15] used CNN to model spatial feature in RGB images with various fusion techniques, and adopted multi-layer LSTM to analyze the motion information of human behavior in video sequences, a number of behavior recognition models from low-level classification to high-level combination are developed, and achieved good recognition results. These methods demonstrate the important contribution of deep networks to behavior recognition both in theory and application. Moreover, Deng et al. used a graphical model in both [12] and [13], and trained the entire model through a deep neural network. Specifically document [12] used a graphical model in conjunction with a convolutional network to propose a multi-step Message Passing Neural Network (MPNN), which considered the dependencies between single-person behavior and body posture to refine the group behavior category label. Document [13] used multi-step information transmission neural network for image analysis, and established a higher level of recurrent network. The recurrent network can perform reasoning on lower-level network output while simultaneously learning higher-level tasks effectively, and finally integrating the whole information to recognize group activity. In addition, Yan [14] et al. extended the Graphic Convolutional Networks (GCN) to Spatial-Temporal Graph Convolutional Networks (ST-GCN) and modeled the sequence of bone maps; then recognized the group behavior through the standard...
softmax classifier. In short, these deep learning networks combined with graphical models have yielded good results for behavioral recognition. In essence, our model follows a similar idea and proposes a probabilistic graphical model based on deep learning network. Specifically, the Fully-Connected Conditional Random Fields is used to analyze the interactions between people; it also corrects the prediction errors generated by the deep learning network. The whole framework is illustrated as follows.

### III. CONDITIONAL RANDOM FIELD MODEL BASED ON DEEP LEARNING NETWORK

#### A. OVERVIEW OF THE WHOLE ALGORITHM

The overall structure of our model is shown in Figure 2. First of all, the input RGB video images are fed to a spatial-temporal model based on CNN and LSTM, aiming to obtain posing and moving features for each person in the group, all of the information for person $i$ is denoted as $x_i$, and also the preliminarily predicted behavior category for each person is denoted as $y_i$. The Fully-Connected Conditional Random Field (FCCRF) is then used to model the rich interactions between people based on their above information $x_i$ and $y_i$, $i \in [1,q]$ ( $q$ stands for the number of persons in the scene). Next, potential functions of FCCRF are adopted to correct the behavior category predictions made in the previous stage and also give the classification of the group scene activity. These details are given in Sections 3.b and 3.c.

#### B. SPATIAL-TEMPORAL MODEL BASED ON CNN AND LSTM NETWORK

We fine-tuned the Alexnet model which was trained by the ImageNet dataset, video image is used as input, and LSTM network is used to describe the temporal sequence of each individual's actions. This temporal information of LSTM is a beneficial supplement to the observation information obtained by CNN, and it plays an important role in improving the performance of behavior recognition.

LSTMs, originally proposed by Hochreiter and Schmidhuber [16], was used in machine translation at first, then its application was expanded to computer vision. Yi Bin et al [17] proposed a method combining a bidirectional long-term memory network (BiLSTM) with a soft attention mechanism to achieve video subtitle generation, its BiLSTM architecture guaranteed continuous readability of video subtitles by preserving global time and visual information. More and more researchers are now applying the LSTM network to human behavior recognition [18][19].

The LSTM network formula used in this paper is as follows:

$$
\begin{align*}
    i_t &= \sigma(W_{ci} * x_{CNN,t} + W_{hi} * h_{t-1} + b_i) \\
    f_t &= \sigma(W_{cf} * x_{CNN,t} + W_{hf} * h_{t-1} + b_f) \\
    o_t &= \sigma(W_{co} * x_{CNN,t} + W_{ho} * h_{t-1} + b_o) \\
    g_t &= \phi(W_{cg} * x_{CNN,t} + W_{hg} * h_{t-1} + b_g) \\
    c_t &= f_t \otimes c_{t-1} + i_t \otimes g_t \\
    h_t &= o_t \otimes \phi(c_t)
\end{align*}
$$

Where the $\sigma$ represents Sigmoid function, and $\phi$ represents Tanh function. $x_{CNN,t}$ is the input from CNN, $h_t \in \mathbb{R}^N$ is the hidden state with $N$ hidden units, $c_t \in \mathbb{R}^N$ is the memory cell, $i_t, f_t \in \mathbb{R}^N$, $o_t, g_t \in \mathbb{R}^N$ are input gate, forget gate, output gate and input modulation gate at time $t$ respectively. $\otimes$ represents the multiplication of array elements. When modeling the actions of a single person, the hidden state $h_t$ is thought to be a simulation of an individual’s actions at time $t$. The output of memory cells changes over time according to the contents of past memories. Due to the deployment of control gates in the information flow, the hidden state will remember a person’s past behavior in a short period of time.

After the video image is processed through the spatial-temporal model based on CNN and LSTM network, the output obtained contains the preliminary observation information and behavior category of each person in the image. We define the spatial-temporal information for $q$ detected persons as $X = [x_1, x_2, \ldots, x_q]$, where $x_t$ represents the observation information of the $i^{th}$ individual, including pose appearance feature information, movement temporal information and position coordinate information. The corresponding behavior classification is defined as $Y = [y_1, y_2, \ldots, y_q]$, the set of behavior labels corresponding to each variable $y_i$ is $L = \{L_1, L_2, \ldots, L_M\}$, and $M$ is the number of behavior categories. The above spacial-temporal information for each individual is then further processed by our FCCRF for pair-wise interaction modeling.

#### C. FULLY CONNECTED CONDITIONAL RANDOM FIELD MODEL

The FCCRF model is shown in the later part of Figure 2, which is a graphical model $g = (V, E)$. Each person in the image is represented as a node $v$ and their relationship between two people acts as an edge $e$. All these probabilistic connections constitute a fully connected conditional random field $(X; Y)$ based on the variable $Y$.
under the condition variable $X$ (for clarity, only close edges are connected in Figure 2). Therefore, we refine the category label $y_t$ corresponding to the $i^{th}$ person by observing variable $x_t$ and its interaction relationship among all people (edges) nearby; and accomplish group activity refining with Gibbs energy functions.

1) FORMULAS OF FCCRF MODEL

In general, the conditional random field $(X; Y)$ is described by the Gibbs distribution:

$$P(Y | X) = \frac{1}{Z(X)} \exp(-\sum_{c \in C} \Phi_c(Y_c | X))$$  \hspace{1cm} (1)

Here, $C$ is the set of all unary and binary cliques in graph $g = (v, e); Z(X) = \sum_x \exp(-\Phi(Y | X))$ is a partition function for normalizing the Gibbs distribution and $Y$ is the derivative of $Y$; the exponent part in formula (1) is Gibbs Energy, which is related to the variable $Y$ based on the condition $X$, namely:

$$E(Y | X) = \sum_{c \in C} \Phi_c(Y_c | X)$$  \hspace{1cm} (2)

To make it easy to denote $\Phi_c(Y_c | X)$ under the CRF $(X; Y)$, the condition variable $X$ is omitted, that is, by default, probability $\Phi(Y_c)$ is used to represent $\Phi_c(Y_c | X)$ in the following formulas.

In order to simulate and analyze the interactions among all the people involved in the group activity, the Gibbs energy of the above formula (2) is defined as:

$$E(Y | X) = \sum_{i} \Phi_u(y_i) + \sum_{i \neq j} \Phi_p(y_i, y_j)$$  \hspace{1cm} (3)

Here, $\Phi_u(y_i)$ is a unary potential, designed to describe individual behavioral category information, $y_i$ indicates the behavioral category of the $i^{th}$ person. $\Phi_p(y_i, y_j)$ is a pairwise potential that describes the interaction between the $i^{th}$ and $j^{th}$ person. The strength of the pair-wise interaction relationship is dependent on observation information and relative position of the two specific persons. If their observation information is similar and the relative position is close, then there is more possibility to have strong interactive relationship and share the same behavioral category; otherwise not. The following definition of the potential function $\Phi$ will illustrate this for details.

2) DEFINITION OF POTENTIAL FUNCTION

The above unary potential function $\Phi_u(y_i)$ is based on the distribution of behavioral label $y_i$, which is calculated in spatial-temporal model. That is, each person has a unary potential function as below:

$$\Phi_u(y_i) = -\log(P(y_i))$$  \hspace{1cm} (4)

Here $P(y_i)$ is calculated by the softmax function in the spatial-temporal model, which represents the probability that the $i^{th}$ person is labeled $y_i$. Obviously, unary potential function only considers the category information of the current individual and ignores information from others.

On the other hand, the pairwise potential function $\Phi_p(y_i, y_j)$ is defined as:

$$\Phi_p(y_i, y_j) = \mu(y_i, y_j) k(f_{i}, f_{j})$$  \hspace{1cm} (5)

Here $\mu(y_i, y_j)$ is the label compatibility function, given by the Potts model [20], introducing a penalty mechanism for a pair of people, namely, for two persons with similar spacial-temporal information, if they are assigned two different behavior labels, then $\mu(y_i, y_j) = 0$; otherwise $\mu(y_i, y_j) = 1$. $k(f_{i}, f_{j})$ represents the Gaussian kernel function and is defined based on the position information $p_i$ and $p_j$, in addition to the spacial-temporal feature vector $f_i$ and $f_j$. Its formula is written as below:

$$k(f_{i}, f_{j}) = w_1 \exp \left( \frac{-|p_i - p_j|^2}{2\sigma_1^2} \right) + w_2 \exp \left( \frac{-|f_i - f_j|^2}{2\sigma_2^2} \right)$$  \hspace{1cm} (6)

Here $f_i$ and $f_j$ represent the feature vectors from CNN-LSTM model for the $i^{th}$ and $j^{th}$ individuals respectively; $p_i$ and $p_j$ are their positional information, which are normalized according to the corresponding detection frame and the intermediate position; $w_1$ and $w_2$ are CRF training weights. Equation (6) means that its Gaussian kernel function is partially affected by the observed spacial-temporal information $f_i$ and $f_j$. That is, when two people are of similar feature information in the same group, they may have a bigger potential function value indicating a strong interaction between them; simultaneously, $p_i$ and $p_j$ positional information are also taken into consideration to modulate Gaussian kernel function, $|p_i - p_j|$ stands for their Euclidean distance, that is, two spatially close persons may have a larger pairwise potential function value indicating a strong interaction relationship.

For example, the 3 persons in Figure 2 with red bounding boxes are observed. Their posture and the temporal information indicate they are all “talking”. Besides, their relative positions are close, therefore, they have stronger interactive relationship pair-wisely compared to the person in the yellow bounding box because his spacial-temporal information portrays “walking” which is different from “talking”. Additionally he is spatially not close to the other three persons.

D. STRENGTH VISUALIZATION OF THE FULLY CONNECTED RELATIONSHIP

Figure 3 shows the interaction structures between people in the graph model $g = (v, e)$ of each frame, each small red or orange circle represents a person node. Sub-picture (a) represents unary model indicating no connection with any other nodes; (b) represents interactions for the $i^{th}$ person.
(node) pair-wisely with other people (node) in a single-frame image; (c) indicates the full connections of the current $i^{th}$ person (node) in the video with all other people (nodes) spatially and temporally. In the graph structures shown in (b) and (c), their pairwise potential function is defined by the Gaussian edge potential in the equation (6), and therefore, the strength of the interaction between people is based on their similarities of observation information.

In the example (b), the current $i^{th}$ person has strong interactive relationship with the two red nodes in the vicinity, so their edges are denoted by solid lines; while the dash line is used to represent their weak interaction between the $i^{th}$ person and the person with yellow bounding box. In case of interaction structure (c), their connections are both spatially and temporally linked, which is called fully connected.

![Figure. 3. Three types of spatial and temporal interactions](image)

**IV. BEHAVIOR PREDICTIONS**

Given $M$ behavioral tags, the FCCRF behavior prediction for the $i^{th}$ person is as follows:

$$P(y_i) = \max(P(y_i | d_k))$$

(7)

Where $d_k$ represents the $k^{th}$ label in $M$ behavioral tags; $P(y_i | d_k)$ represents the probability of classifying the $k^{th}$ label to the $i^{th}$ person. However, the $P(y_i | d_k)$ is difficult to calculate, we use approximation algorithm [24-25] to achieve a more cost effective calculation that is, Cross Bilateral Filtering techniques and the Mean Field approximation and are adopted to simplify its quadratic nonlinear calculation to linear computation, thereby figuring out $P(y_i)$ as the maximum value.

Group behavior recognition is decided by the following formula:

$$g = \arg\max (p(y))$$

(8)

That is, if $N$ individuals are detected in an image, their largest proportion of behavioral categories is regarded as the final group behavioral category label $G$. For example, in Figure 3, of the four participants, three of them are labeled as "Talking" and one as "Walking", however the final group behavior is classified as "Talking".

**V. ALGORITHM TRAINING**

We use piece-wise training to learn model parameters. The model here consists of two parts: the spatial-temporal model based on CNN and LSTM; and the FCCRF model. They are trained independently. Theoretically speaking, piece-wise training can minimize an upper bound on the log partition function of the CRF model [21]. Moreover, the experimental results presented in Shotton et al. [22] and Mccallum et al. [23] show that piece-wise training often outperforms global training even with global inferences. Therefore, in our experiment, the unary potentials are trained first, and then learn a series of parameters of the binary potential functions.

In addition, since it is found in the experiment that $w_2$ and $\theta_3$ in formula (6) have little effect on the overall recognition, their appropriate default values are finally used for these two parameters. For the other parameters $\theta_1, \theta_2, w_1$ of the binary potential function (6), the grid search algorithm is used for learning on the training, and the parameters of all the kernel functions are cross-validated to find the optimal solutions.

**VI. EXPERIMENTAL CLASSIFICATION RESULTS AND ANALYSIS**

Collective Activity Dataset1 and its extension are used to test our algorithm. The experiment contents are organized as follows. First of all, we briefly introduce the videos contained in these two sets and give our sample split method in our learning and testing phases. Then we test the performances of three different CRF graph structures in both sets for verification of our FCCRF model. After that, we make recognition comparisons with other methods respectively on CAD1 and CAD2. A final analysis is made concerning all the observations and judgments made.

**A. DATASETS INFORMATION AND ITS SPLIT METHOD**

The first data-set used in our experiment is Collective Activity Dataset1 containing 44 video clips acquired using a low-resolution hand-held camera. It has five scene action labels: Crossing, Queuing, Walking, Talking, Waiting, which are performed by N people in each frame; also eight posture labels (not used in the experiment). In addition, each person has a behavioral label, and each frame image has a scene activity label.

The second data-set is Collective Activity Dataset2, which is an extension to Collective Activity Dataset1 that removes the Walking action from CAD1 and adds two additional actions, dancing and jogging. So there are 6 behaviors in CAD2, namely Crossing, Queuing, Dancing, Talking, Waiting, Jogging.

We use all video images in the above datasets. The split method is the same as [24], which is a third for training and the rest for verification and testing. Each video image is first input into the spatial-temporal model based on CNN and LSTM, and its output dimension is N*3008, N representing the number of people detected in each image. Among them there are 3002 dimensional observation information (indicated by $x_i$) and 6-dimensional behavioral labels (indicated by $y_i$). To elaborate, the 3002 dimensional observation information includes a 3000-dimensional observation information.
dimensional spatial temporal feature (indicated by $f_i$ in Equation 5) and a 2-dimensional position coordinates (indicated by $p_i$ in Equation 5). Then, the obtained output data is fed into the Fully-Connected Conditional Random Field, and the interaction relationship between persons is analyzed and the activity label is re-fined.

**B. RECOGNITION ACCURACY COMPARISONS FOR DIFFERENT GRAPH STRUCTURE MODELS**

In order to evaluate the rationality of our model, Table 1 illustrates the recognition accuracy comparisons for graph structure model (a), (b), and (c) in Figure 3 on CAD1 and CAD2. Recognition results of the of the (c) model, considering the interaction between people in the video image, is observed to be distinctly better than the (a) model without considering the interaction relationship. It is also worth noting that model (b) is not better than (a) due to its solely considering spatial interaction between individuals within one image. Model (b) produces low recognition accuracy results for group behavior of Walking and Crossing (see Table 2 for details) due to the two behaviors being spatially similar. In short, the FCCRF model is used to simulate and analyze the spatial and temporal interaction between people, which greatly improves group behavior recognition thereby proving model structure(c) to be the most effective.

**Table 1. Average recognition accuracy of 3 graph structure models on CAD1 and CAD2**

<table>
<thead>
<tr>
<th>Model</th>
<th>CAD1(%)</th>
<th>CAD2(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unary Only [a]</td>
<td>83.84</td>
<td>85.67</td>
</tr>
<tr>
<td>Connected Per Frame [b]</td>
<td>83.15</td>
<td>89.09</td>
</tr>
<tr>
<td>Fully Connected [c]</td>
<td>86.7</td>
<td>90.83</td>
</tr>
</tbody>
</table>

**C. RECOGNITION ACCURACY COMPARISONS WITH OTHER METHODS ON CAD1**

1) EXPERIMENT DATA ALIGNMENT ON CAD1

Table 2 lists recognition result comparisons of different methods on CAD1, the abbreviation MPCA stands for Mean Per Class Accuracy. As mentioned above, CAD1 includes 5 types of group activities, that is Crossing, Queuing, Walking, Talking, Waiting. According to document [29], the definition of Waiting is ambiguous because it is more like a single event than a group event; besides the difference between Walking and Crossing is slight in different people and different streets only. Therefore, in Table 2, this paper combines Walking and Crossing into Moving, and its recognition accuracy is the average of both.

<table>
<thead>
<tr>
<th>Class/Model</th>
<th>[8]</th>
<th>[27]</th>
<th>[28]</th>
<th>CNN+LSTM [b]</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moving</td>
<td>92</td>
<td>87</td>
<td>84</td>
<td>86.94</td>
<td>87.31</td>
</tr>
<tr>
<td>Waiting</td>
<td>69</td>
<td>75</td>
<td>84</td>
<td>64.09</td>
<td>52.71</td>
</tr>
<tr>
<td>Queuing</td>
<td>76</td>
<td>92</td>
<td>86</td>
<td>86.52</td>
<td>95.12</td>
</tr>
<tr>
<td>Talking</td>
<td>99</td>
<td>99</td>
<td>75</td>
<td>97.82</td>
<td>97.45</td>
</tr>
<tr>
<td>MPCA</td>
<td>84</td>
<td>85.8</td>
<td>82.8</td>
<td>83.84</td>
<td>83.15</td>
</tr>
</tbody>
</table>

**Table 2. Average recognition accuracy of different algorithms on CAD1**

2) ANALYSIS OF EXPERIMENT RESULTS ON CAD1

In Table 2, spatial-temporal model based on CNN and LSTM is used as a baseline method. Table 2 portrays that behavior “Waiting” accuracy of the baseline method and the model (Ours) is significantly lower than the other three methods, the reason being that pose label information in CAD1 is not used; in addition, the “Waiting” category always appears in both “Crossing” and “Walking”, which may be a factor in confusing predictions.

As can be seen from the comparison table, our average recognition accuracy is significantly higher than other models without using the full-connection Condition Random Field (FCCRF) model. At the same time, the recognition results of CNN and LSTM baseline model using simple deep learning networks are also significantly higher than those using traditional hand-crafted features [28], indicating that CNN and LSTM are more efficient than traditional methods in image feature extraction. Although the recognition result of the CNN and LSTM baseline model is slightly lower than the results in document [8] and [27], the two methods are based on handcrafted features, which is dependent on specific datasets, when the data source is changed, their model structures need to be redesigned, so they are not good at generality. In all, our model is more accurate in behavioral recognition and also has good adaptability due to deep learning capability.

**D. RECOGNITION ACCURACY COMPARISONS WITH OTHER METHODS ON CAD2**

1) RECOGNITION ACCURACY IN CAD2 AND COMPARISONS TO THAT IN CAD1

Table 3 is average recognition accuracy comparisons of our model with others on CAD2. Compared to the results in CAD1, all recognition accuracy is better. There are 2 reasons for this. The first reason is that the “Crossing” action in CAD1 is partially misjudged as Walking, due to the fact that “Walking” and “Crossing” are two very similar movements; actions which were captured from different view angles in a variety of environments resulting in low accuracy in CAD1. The second reason is that “Walking” is removed in CAD2, which avoids miscalculation between “Crossing” and “Walking”. Therefore, the recognition accuracy on CAD2 is higher than that in CAD1.

**Table 3. Average recognition accuracy comparisons to that in CAD1**

<table>
<thead>
<tr>
<th>Class/Model</th>
<th>[8]</th>
<th>[27]</th>
<th>[28]</th>
<th>CNN+LSTM [b]</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moving</td>
<td>92</td>
<td>87</td>
<td>84</td>
<td>86.94</td>
<td>87.31</td>
</tr>
<tr>
<td>Waiting</td>
<td>69</td>
<td>75</td>
<td>84</td>
<td>64.09</td>
<td>52.71</td>
</tr>
<tr>
<td>Queuing</td>
<td>76</td>
<td>92</td>
<td>86</td>
<td>86.52</td>
<td>95.12</td>
</tr>
<tr>
<td>Talking</td>
<td>99</td>
<td>99</td>
<td>75</td>
<td>97.82</td>
<td>97.45</td>
</tr>
<tr>
<td>MPCA</td>
<td>84</td>
<td>85.8</td>
<td>82.8</td>
<td>83.84</td>
<td>83.15</td>
</tr>
</tbody>
</table>
Table 3.
Recognition accuracy comparisons of different models on CAD2

<table>
<thead>
<tr>
<th>Model</th>
<th>CAD2(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combining Frame &amp; Track Cues [30]</td>
<td>90.23</td>
</tr>
<tr>
<td>Structure Inference Machines [31]</td>
<td>85.8</td>
</tr>
<tr>
<td>LRCN [32]</td>
<td>86.5</td>
</tr>
<tr>
<td>CNN + LSTM + FCCRF (Ours)</td>
<td>90.83</td>
</tr>
</tbody>
</table>

2) ANALYSIS OF COMPARISONS BETWEEN OURS AND OTHERS

Method [30] uses the trajectory information and contextual harmony as spatial temporal feature, and treats action recognition as a constrained multi-criteria objective function. Although its recognition rate is very effective, our model is still slightly higher than this method. This is because the handcrafted trajectory information is not easy to obtain in an actual scene, and its tracking error will greatly degrade recognition performance. Algorithm [31] combines graphical model and RNN network to simulate group interactions, its interaction information is modeled with RNN activation functions. Its drawback being the use of a CNN as the preliminary classifier to extract visual classifications for each individual, while our model uses CNN-LSTM model for spatial temporal feature extraction and primary action recognition. Therefore our model is more comprehensive in feature modeling. Method [32] uses Long-term Recurrent Convolutional Networks to learn temporal dynamics and convolutional perceptual representations, but it does not tackle interactions among individuals. To sum up, our method not only gives an overall spatial-temporal representation for individuals in the first stage, but also models spatial-temporal interactions among them, which is helpful to improve behavioral recognition accuracy.

VII. CONCLUSION

This paper constructs a comprehensive architecture combining deep learning network and probability graph model. It uses the spatial-temporal model based on CNN and LSTM network to learn the characteristic information of each person in the group activity and predict its behavior. It then proposes the fully-connected Conditional Random Field model to analyze the interactions between people in the group activities, including spatial interactions (single-frame images) and temporal interactions (continuous multi-frame images). Next, it utilizes potential function to get correct group behavior classification with different individual behavior attributes. In experiments, we use spatial-temporal model based on CNN and LSTM as a baseline, and make comparisons with other methods. The above result analysis proves the importance of interactions between people in group activity recognition. In the future, we plan to study the time complexity of the CRF graphical model and aim to semantically segment the complicated group into some subgroups, and optimize its structure, in the hope of predicting their behaviors respectively for complicated scenarios.

REFERENCES