Human Activity Recognition using Deep Neural Network with Contextual Information

Li Wei and Shishir K. Shah
Computer Science Department, University of Houston, 3551 Cullen Blvd., Houston, TX 77204, U.S.A.

Keywords: Activity Recognition, Deep Learning, Context Information.

Abstract: Human activity recognition is an important yet challenging research topic in the computer vision community. In this paper, we propose context features along with a deep model to recognize the individual subject activity in the videos of real-world scenes. Besides the motion features of the subject, we also utilize context information from multiple sources to improve the recognition performance. We introduce the scene context features that describe the environment of the subject at global and local levels. We design a deep neural network structure to obtain the high-level representation of human activity combining both motion features and context features. We demonstrate that the proposed context feature and deep model improve the activity recognition performance by comparing with baseline approaches. We also show that our approach outperforms state-of-the-art methods on 5-activities and 6-activities versions of the Collective Activities Dataset.

1 INTRODUCTION

Human activity analysis is one of the most important problems that has received considerable attention from the computer vision community in recent years. It has various applications, spanning from activity understanding for intelligent surveillance systems to improving human-computer interactions. Recent approaches have demonstrated great performance in recognizing individual actions (Weinland et al., 2011; Tran et al., 2012). However, in reality, human activity can involve multiple people and to recognize such group activities and their interactions would require information more than the motion of individuals. This remains a challenging research topic largely due to the tremendous intra-class variation of human activities attributed to the visual appearance differences, subject motion variabilities, and viewpoint changes.

To solve these challenges, previous approaches in human activity recognition have focused on information about context. Context can be defined as information that is not directly related to the human activity itself, but it can be utilized to improve the traditional target-centered activity recognition (Wang and Ji, 2015). Lan *et. al.* (Lan et al., 2012) proposes action context to encode the human interactions among multiple people. Choi *et. al.* (Choi et al., 2011) uses spatio-temporal volume descriptor to capture nearby person actions.

Most existing approaches for human activity recognition mainly use people as context without richer context information, such as the scene information where the activity is performed, the location of person within the scene, etc. Further, previous approaches have either utilized the context directly as feature inputs to classifiers such as random forest (Choi et al., 2011) and support vector machine (Tran et al., 2013), or incorporated context through probabilistic models like conditional random fields (Tran et al., 2015). There is little work utilizing deep models and networks to capture the contexts for human activity recognition. Deep models have the potential to systematically incorporate multiple sources of contexts due to their multi-level deep structure, the capability of probabilistic reasoning, and the integration of hidden units to synthesize higher level representations of the raw input features (Wang and Ji, 2015). Therefore, in this work, we propose a deep neural network (DNN) based model to recognize the human activity by taking advantage of its probabilistic reasoning power and incorporate multiple sources of context information. We combine motion and context information. The motion information is encoded by using the low-level motion features and high-level mobility features. The context information is incorporated to represent the scene and the human interactions. The scene features encodes the attribute of the scene at global and local level, while the group fea-

Wei L. and K. Shah S.
Human Activity Recognition using Deep Neural Network with Contextual Information.
DOI: 10.5220/0006099500340043
In *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2017)*, pages 34-43
ISBN: 978-989-758-226-4

ture captures the human interaction similar to Tran *et al.* (Tran et al., 2013). For each feature, we carefully design the network structure to get the higher level representation of input features, and the combination of different representations. We demonstrate that the integration of our context features and deep model is able to achieve better performance than state-of-the-art approaches on the collective activity dataset (Choi et al., 2009).

In summary, the main contributions of this paper are:

- We introduce a two-level scene context descriptor. Beside the group context feature similar to many other works, we introduce a two-level scene context feature that describe the environment information of centered-target at the global and local levels.
- A deep model for human activity recognition. We present a deep neural network model that jointly captures multiple sources of context information, and achieves state-of-the-art performance over the collective activity dataset (Choi et al., 2009).

2 RELATED WORK

In human activity recognition systems, various low-level features are introduced to describe the activity observation. Schuldt *et. al.* (Schuldt et al., 2004) proposed a local space-time feature to represent the human movement observed in a video, and integrated such representations with SVM classification schemes for recognition. Laptev *et. al.* (Laptev et al., 2008a) proposed space-time feature point (STIP) and spatio-temporal bag-of-features as the descriptor for human motion. Tran *et. al.* (Tran et al., 2012) presented a framework for human action recognition based on modeling the motion of human body parts. They utilized a descriptor that combines both local and global representations of human motion, encoding the motion information as well as being robust to local appearance changes. The mentioned activity recognition methods mainly focus on recognizing the individual action. Their frameworks are difficult to scale to address real-world scenarios where multiple people activity and interaction are involved. Our approach represents the motion information using STIP feature similar to (Laptev et al., 2008a), but combines the rich context information that we extract from the video. By using the deep model, our method is able to: capture the extensive information about people motion and interactions; scale to recognize activity of each individual in the scene; and improve the accuracy of the overall activity recognition task.

Context based Activity Recognition. Context information is widely utilized in many video analysis applications (Wei and Shah, 2015; Wei and Shah, 2016). In the topic of human activity recognition, many approaches integrate contextual information by proposing new feature descriptors extracted from an individual and its surrounding area. Lan *et. al.* (Lan et al., 2012) proposed Action Context (AC) descriptor capturing the activity of the focal person and the behavior of other persons nearby. The AC descriptor is concatenating the focal person action probability vector with context action vectors that captures the nearby people action. Choi *et. al.* (Choi et al., 2009) propose Spatio-Temporal Volume (STV) descriptor, which captures spatial distribution of pose and motion of individuals in the scene to analyze group activity. STV descriptor centered on a person of interest is used to classify centered person's group activity. SVM with pyramid kernel is used for classification. The same descriptor is leveraged in (Choi et al., 2011), however, the random forest classification is used for group activity analysis. In (Lan et al., 2012; Choi et al., 2009; Choi et al., 2011), the nearby person that serves as context are selected according to the distance to the centered target. This does not necessarily ensure the existence of interactions among the selected persons. To address this issue, Tran *et. al.* (Tran et al., 2015) proposed group context activity descriptor similar to (Lan et al., 2012), but the people are first clustered into groups by modeling the social interaction among the persons. However, due to the noisy observation in videos, the group detection might not be robust or stable. Therefore, our approach utilizes the social interaction region to select the contextual people without a clustering process. Besides focusing on people as context, our approach also introduces scene information as context for the first time. The scene context describes the environment around the center target at the local and global levels. We utilize the existing place recognition method (Zhou et al., 2014) to provide scene context features that have semantic meanings.

Deep Model for Activity Recognition. In recent years, deep models including deep neural networks, convolution neural networks, and auto-encoders have been used in many applications. For human activity recognition (Ji et al., 2013; Karpathy et al., 2014), convolution neural networks and auto-encoder approaches (Hasan and Roy-Chowdhury, 2014) have been developed. However, these action/activity deep models are generally target-centered and do not consider any context information, which is important for human activity that involves multiple people. Comparatively, Wang *et. al.* (Wang and Ji, 2015) proposed

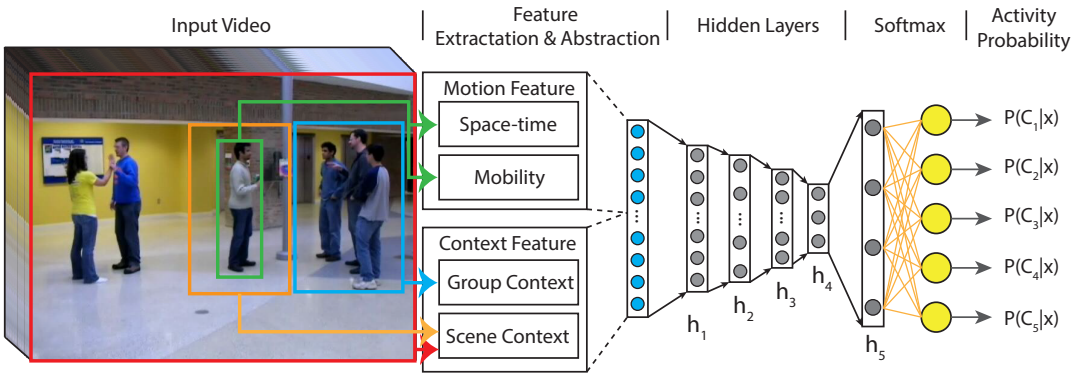

Figure 1: The overview of proposed neural network model.

an event recognition framework, which is a hierarchical context model that captures the context information in multiple levels. Similarly, our approach uses a deep-structure model that be trained using the contextual information extracted from human groups and video scene. However, our approach focuses on a different problem, which is recognizing the activity of each individual appearing in the scene, other than an overall event of the entire scene (Wang and Ji, 2015).

3 METHODOLOGY

To recognize the human activity, we a utilize deep neural network based approach. The structure of our network is as shown in Figure 1.

Human Activity Recognition. Given the input video with tracking information of each subject, our system recognizes activity of each individual person at every frame. Two distinct features are considered in this recognition network, the first based on human motion (Sec. 3.1) and the second based on the context (Sec. 3.2). The features are extracted and abstracted using dense fully connected hidden layers. The output hidden units of the two parts are combined and fed into another fully connected network, which has a soft-max layer to compute the probability of recognized activity from input observations.

The input datasets are videos. In addition, we assume that the people tracking information, an estimation of their 3D space location, and facing direction in 3D space is available. We denote persons appearing in video as $\{p_i\}$, the tracking 2D bounding box of i-th person at frame t as b_i^t, the estimated 3D location as $l_i^t \in \mathbb{R}^3$, estimated the facing direction quantized into 8 viewpoints as $d_i^t \in \{front, frontleft, left, backleft, back, backright, right, frontright\}$.

In the following, we discuss in detail about our proposed features, along with our training and inferencing approach.

3.1 Motion Features

Motion features we consider are the low-level observation of the movements within the video. In our approach, two motion features are used: Space-time features that capture the low-level motion observed in the video; and the mobility features that capture the movement of human as a whole part.

For an input video, we compute features for frames with interval β. That is, we extract features for the sample located at time t by computing the feature descriptors using a video segments comprising of frames in the interval $[t-\beta, t+\beta]$.

Space-time Features. There are various space-time features to describe human motions in the video. We choose space-time interest points (STIP) (Laptev et al., 2008b), because it can extract feature points in space-time dimension robustly, and it also has been applied in event recognition (Wang and Ji, 2015). STIP method detects interest points using a space-time extension of the Harris operator. For each interest point it computes descriptors of the associated space-time patch. In this paper, histograms of oriented gradient (HOG) and histograms of optical flow (HOF) feature are computed as the descriptors of the space-time patch. We obtain the feature words of both features by first detecting all the interest points over the entire videos data set, and then applying K-Means clustering to obtain K_i feature words for HOG features and HOF features.

To describe the motion of p_i at time t, we first collect all the interest points located within $\{b_i^k | k \in [t-\beta, t+\beta]\}$, as shown in Figure 2(a). Then we compute the histogram of gradient and optical flow given the collected interest points as shown Figure 2(b). Finally, it results in two K_m dimensional histogram vectors.

After normalization to ensure each vector can sum

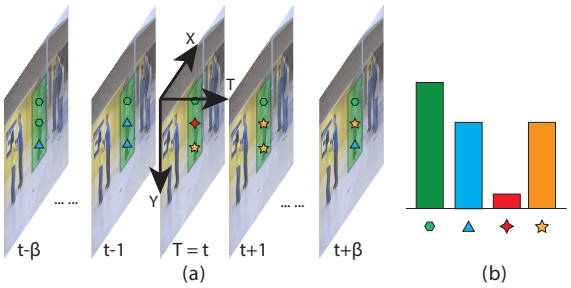

Figure 2: STIP feature histogram. (a) shows the video segment centered at time t, with length $2\beta+1$. The green boxes denote the bounding box areas of the subject. (b) shows the STIP histogram generated using the video segment (a).

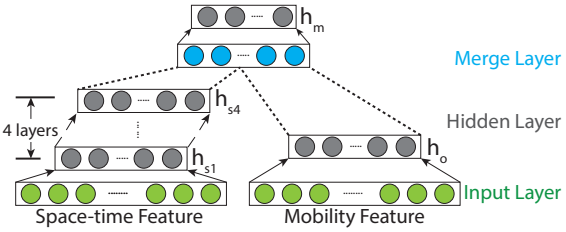

Figure 3: Motion feature layers. The green layers are network inputs; the gray layers are fully connected dense layers with hidden units; the blue layer is merge layer, which concatenates its inputs layers.

up to one, the concatenation of two vectors serves as the motion descriptor of the person. We denote this as S_i^t. If there are no interest points located in bounding boxes of the subject, the descriptor is a zero vector of dimension $2K_m$. The extracted feature forms the input into our network as shown in the left part of Figure 3, and is followed by four fully-connected layers with $(h_{s1}, h_{s2}, h_{s3}, h_{s4})$ hidden units at each layer. Finally, at the top we have a layer with h_{s4} hidden units to realize a response to be combined with mobility information described below.

Mobility Feature. As the estimation of people 3D location can be obtain using (Ess et al., 2008), we take the distance of movement in 3D space through the video segment as a description of human mobility. We compute subject movement at time t as $v_i^t = l_i^t - l_i^{t-1}$, where l_i^t denotes the location of p_i at time t. We denote p_i's mobility feature at time t is $V_i^t = [v_i^{t-\beta}, v_i^{t-\beta+1}, ..., v_i^t, ..., v_i^{t+\beta-1}, v_i^{t+\beta}]$, which is a vector of length $2\beta+1$. We input the extracted mobility feature into our network as shown in the right part of Figure 3. The input layer is fully connected with hidden layer that contains h_o units.

The hidden units of STIP features and mobility feature are concatenated to form a merge layer, which is fed into another fully connected layer of size h_m. These h_m hidden units abstract the overall motion information of the subject observed in the video at a sample frame.

3.2 Context Features

In our approach, context information plays an important role to improve the activity recognition accuracy. The context information includes two parts: the scene based context and group based context. Scene based context captures the environment information surrounding the subject, allowing the network to find the association between environment information and activities. Scene based context has two levels: **scene prior** describes the global scene attributes of the video frame; and **scene context** describes the scene around the person locally. The group of people that are physically near the subject also provide strong context information about human activity, as many approaches build various features to describe the people actions of near by humans to improve activity recognition accuracy (Choi et al., 2011; Tran et al., 2013). Similarly, the group based context contains two parts: **group action** describes the interaction observation of nearby people; and **group structure** captures the shape (positions, direction) of nearby humans relative to the target person.

3.2.1 Scene based Context

Some activities have strong association with the environment, so the environment information as a context can reduce the ambiguity in its recognition. For example, jogging or crossing activities are more likely to happen in outdoor scenes, and queuing will be more likely to happen if the scene appears to be a shop. In this paper, we extract the scene context information by looking into the image patch that extends around the bounding box of the tracked subject and use the descriptors of theses image patches as context features.

Rather than using low-level features such as appearance features to describe the image patches, we use a descriptor with semantic meaning. We utilize the existing place recognition methods to extract the semantic attribute of subject contexts. As deep convolution network gives the state-of-the-art performance in place recognition tasks, we use the Place-CNN (Zhou et al., 2014) to generate the image patch descriptor. Given an image patch to Place-CNN, it outputs the probability of given image belonging to 205 categories. An example of place recognition on context image patches is shown in Figure 4. We simply denote the recognition process of Place-CNN as function $Place(I_t)$, where I_t is the image frame at time t of a given video, $Place(.)$ returns the probability vector of given image being recognized as belonging

Scene Prior: crosswalk:0.54, gas_station:0.30
Scene Context #1: crosswalk:0.70, parking_lot:0.07
Scene Context #2: phone_booth:0.20, lobby:0.16

Figure 4: The scene prior and scene context. The green box is the bounding box of tracked people, with people id inside it. The yellow boxes are the scene context areas of persons. The red box which bounds the whole image is the scene prior area. We input images into Place-CNN to recognize place probability. The top two likely places of the above scene and scene context of person 1 and 2 are shown below the figure.

to the place categories.

Scene Prior. The scene prior gives the environment context information at a global level for each video frame. To extract the scene prior feature, L^t, for all the subjects that appear at time t, we compute

$$L^t = \frac{1}{2\beta+1}\sum_{k=t-\beta}^{t+\beta} Place(I^k), \qquad (1)$$

where scene prior feature $L^t \in \mathbb{R}^{205}$ and $\sum_{s=1}^{205} L_s^t = 1$.

Scene Context. Besides the scene prior as global information for all the subjects appearing in the video frame, for each individual subject, we also build local scene features that capture the local environment information.

We denote the scene context image patch of p_i at time t as T_i^t, which is the region surrounding the bounding box b_i^t. Both T_i^t and b_i^t have the same center location, while width and height of T_i^t is 3 and 1.5 times the width and height of b_i^t, respectively. The scene context feature of p_i at time t is denoted as Q_i^t, which is computed as follow:

$$Q_i^t = \frac{1}{2\beta+1}\sum_{k=t-\beta}^{t+\beta} Place(T_i^k), \qquad (2)$$

Where $Q_i^t \in \mathbb{R}^{205}$ and $\sum_{s=1}^{205} Q_s^t = 1$.

After we compute scene prior and scene context features, we input the two features into the network as shown in Figure 5. We first concatenate the two features prior to feeding them to two fully connected layers h_{t1} and h_{t2}. The intent is to capture the interaction between global scene prior and local scene context. The hidden units in layer h_{t2} serve to provide the scene context information.

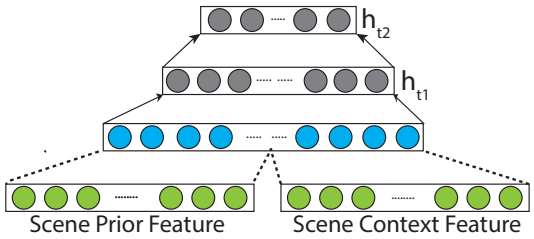

Figure 5: The network of combining scene prior and context information.

3.2.2 Group based Context

As people tend to form groups in various social behaviors, many approaches use the information from persons that physically are near the subject of interest to infer the activity. In our approach, we simply define the group as people within the social interaction area.

There are two group based context information that are extracted: *group interaction context* captures the activity interaction of subject with group members; *group structure context* describe the spatial distribution of positions and directions of group members.

Group Interaction Context. The group interaction context captures the activity interactions between the centered subject and group members. We use concepts from proxemics (Was et al., 2006; Tran et al., 2013), and define interaction region as an area where the people are able to make social interaction with the centered subject. Interaction region is an ellipse $E(c_i, a, b)$, where the center of ellipse is c_i and (a, b) is the major and minor axis of ellipse, respectively. In our implementation, we use $c_i = l_i + 0.3d_i$, $a = 3.35$, $b = 2.0$ as suggested in (Was et al., 2006). We are able to detect group members by finding the person within social interaction region, as shown in Figure 6. We denote the group members of subject p_i at time t as $N(p_i, t)$.

To generate the group interaction context feature for p_i at time t, we first compute the space-time features S_i^t, which is a bag-of-feature histogram of motion features as discussed in Sec. 3.1. Then we compute the average space-time feature U_i^t for all persons within the interaction region $N(p_i, t)$ as follow:

$$U_i^t = \frac{1}{|N(p_i,t)|}\sum_{p_j \in N(p_i,t)} S_j^t \qquad (3)$$

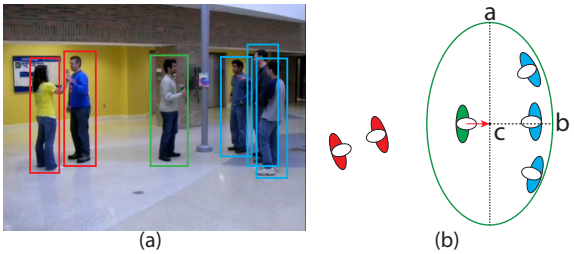

Figure 6: Interaction region. (a) The centered subject is in the green box, where the group members of target subject are in blue boxes, non-group members are in red boxes. (b) top view of persons 3D locations estimation of (a), the interaction region of the centered subject is displayed as the green ellipse, with center c and major a, minor b marked at ellipse.

We generate a 2D histogram as $B_i^t = S_i^{t\,\mathsf{T}} * U_i^t$ that captures the co-occurrence frequencies of S_i^t and U_i^t. We normalize the 2D histogram B_i^t to ensure that all elements in the matrix sum to 1 and build a group interaction context feature by flattening the matrix into a K_i^2 dimension vector. If K_i is large, then we can recreate word bags for STIP features by clustering all the motion features of the data set. In our implementation we use $K_a = \sqrt{K_i}$ as number of bags for group interaction context feature extraction.

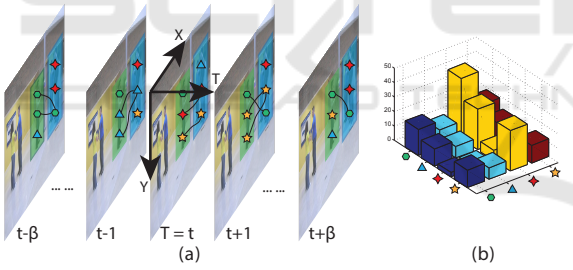

Figure 7: Group interaction context feature. (a) shows a video segment, where the green part covers the bounding box of target subject, the blue part covers the interacting group members. (b) shows the 2D co-occurrence histogram of target subject in the video segment (a).

Group Structure Context. The group structure context describes the relative positions and directions of people within interaction regions. For different activities, the shape of the group and the interactions between group members can be different. For example, group talking activity would have more than two people positioned in front of each other, face to face, while queuing activity most likely has more than two people standing in a line and facing the same direction. Therefore, we design group structure context feature to capture the positions and facing direction of the group.

To describe the position information, we construct a local coordinate centered at the target subject, as shown in Figure 8(a), and form a histogram of angles to represent the position distribution of group members. We denote the function $Ang(\vec{V_1},\vec{V_2})$ that re-

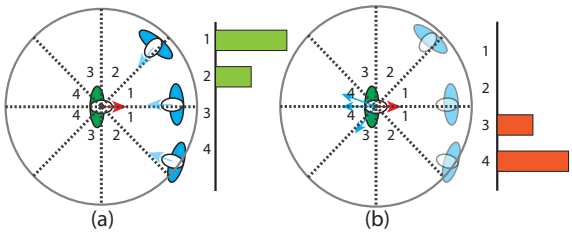

Figure 8: Group position histogram and direction histogram. (a) shows the position histogram; (b) shows the direction histogram. In this figure, the angle space is split into 4 sub-range in order to compute histogram.

turns the angles between vector $\vec{V_1}$ and $\vec{V_2}$. The group member position distribution of p_i at time t is obtain by computing the normalized histogram of angle set $\{Ang(d_i^k, l_j^k - l_i^k) | p_j \in N(p_i,k), k \in [t-\beta, t+\beta]\}$. To capture the direction information, we calculate the angles between the direction of centered subjects and other group members, then form a histogram of directions that represent the direction distribution of interacting neighbors, as shown in Figure 8(b). The group member direction distribution of p_i at time t is obtain by computing the normalized histogram of angle set $\{Ang(d_i^k, d_j^k) | p_j \in N(p_i,k), k \in [t-\beta, t+\beta]\}$. Both position and direction histogram have K_s bins.

After the position histogram and direction histograms are concatenated, we have the group structure context feature. We denote it as G_i^t, which is a $2K_s$ dimension feature, where K_s is bin size of angle histogram.

Finally, position histogram and direction histogram are input into two fully connected hidden layers (the hidden units number are h_{i1} and h_{i2} for group interaction feature; h_{r1} and h_{r2} for group structure feature), followed by a merge layer. The hidden units at the top represent the group context information.

After the final hidden layer for scene context information and group context information, we use a merge layer to concatenate hidden units from the two layers, as shown in Figure 9. The merge layer is fed into a fully connected layer for further abstraction. The top h_c hidden units form the representation for the overall context information of a given observation.

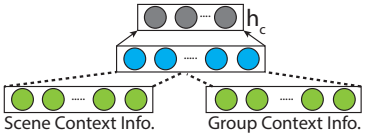

Figure 9: The network of group context informations.

The hidden units of motion information in Figure 3 and context information in Figure 9 are further concatenated, and input into the network shown in Figure 1, which includes four fully connected layers (with hidden units number h_1, h_2, h_3, h_4), and a soft-max layer at the end to calculate the probabilities of given observation for a set of activities. So far we have presented our deep neural network model, and in the following we are going to present the method for training and inference using our model.

3.3 Learning and Inference

Model Learning. The proposed model is a neural network with parameter W, which includes the weights matrix and bias parameters of all dense layers in the network. We denote $X = \{x_i^t, A_i^t | i = 1,...,N, t = 1+\beta,...\}$ as the training data, where $x_i^t = (S_i^t, V_i^t, L^t, Q_i^t, B_i^t, G_i^t)$ includes all the individual features, and A_i is the ground truth human activity label. The output of the network is the probability of given observation belonging to each class of activity label. We denote the forward propagation as $F(W, x_i^t) = \{P(C_k|x_i^t), k = 1...M\}$, where M is the number of activity categories. In the training phase, we compute and minimize the categorical cross-entropy between predictions and ground truth:

$$E(W,X) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} 1(A_i = C_j)\log(P(C_j|x_i)) \quad (4)$$

We optimize the loss function using Stochastic Gradient Descent (SGD) updates with Nesterov momentum (Nesterov et al., 2007). In each iteration, the model parameters W are updated as follow:

$$\triangle W_t = \mu * \triangle W_{t-1} - lr * \nabla_W E(W_t + \mu * \delta W_{t-1}, Z) \quad (5)$$

$$W_{t+1} = W_t + \triangle W_t \quad (6)$$

Where μ is the momentum, lr is learning rate, ∇_W is the gradient of the model parameter W, and Z is a random subset of training data for computing gradient in each iteration. We initialize the parameters of the network using Glorot weight initialization (Glorot and Bengio, 2010).

Model Inference. Given query human activity observation x, our model recognizes the activity category $C^\star$ by finding the maximal posterior probability given the observations from both motion feature space and context feature space through Equation 7.

$$C^\star = \arg\max_k P(C_k|x) \quad (7)$$

We implement our network using Lasagne (Dieleman et al., 2015) with GPU acceleration.

4 RESULTS

In this section we describe the experiments that evaluate the performance of the proposed model for human activity recognition.

4.1 Datasets

Our human activity recognition model is tested using Collective Activity dataset introduced by Choi *et. al.* (Choi et al., 2009). Other datasets (e.g, CAVIAR, VIRAT, or UT-Interaction) either focus on single person activity or the semantic labels provided are agnostic to scene context.

Collective Activity dataset comprises of two versions. The first version of data set contains five activities (*Cross*, *Wait*, *Queue*, *Walk* and *Talk*) and we denote this as Data-Act-5. The second version of dataset includes two additional activities (*Dance* and *Jog*) and removes the *Walk* activity, since the *Walk* activity is an individual activity rather than a collective activity. We denoted the second version as Data-Act-6. HOG based human detection and head pose estimation along with a probabilistic model is used to estimate camera parameters (Choi et al., 2009). Extended Kalman filtering is employed to extract 3D trajectories and head pose estimates are provided as part of the dataset. In general, this dataset represents real-world, noisy observation with occlusions and automatic person detection and trajectory generation. We use the 4-fold cross-validation scheme similar to (Choi and Savarese, 2012) to test the performance of our approach. To minimize the overfitting in training phrase, we split data of non-training fold randomly into validation data set (30%) and testing data set (70%). In each interaction of parameter updates, the accuracy of validation data set is computed. When the accuracy over the training data set increases, but the accuracy over the validation data set stays the same or decreases, the neural network is over-fitting and we stop training.

4.2 Experiments and Comparison

In this section, we demonstrate the effectiveness of the proposed human activity recognition model that integrates both motion features and multiple sources of context information. The neural network to be evaluated has configuration as shown in Table 1.

The experiments are performed on both versions of Collective Activity dataset. The performance of proposed model on both versions of the dataset is shown in Figure 10. The low value of the non-diagonal elements implies that our model is highly

Table 1: Experiments Network Configuration

h_{s1}	150	h_m	25	h_{r1}	10	h_3	25
h_{s2}	100	h_{t1}	100	h_{r2}	10	h_4	25
h_{s3}	100	h_{t2}	20	h_c	10	h_5	25
h_{s4}	100	h_{i1}	10	h_1	50		
h_o	10	h_{i2}	10	h_2	50		

discriminative with low decision ambiguity between activities.

The confusion matrix of Data-Act-5 in Figure 10 (left) also shows that the confusion between *Walk* and *Cross* is reasonably low, despite the fact that both activities are *Walk* activity but with different scene semantics. Our model captures the scene context information and recognizes *Walk* activity better than baseline approaches as shown in Table 2 and other state-of-the-art approaches as shown in Table 3.

Compare with Baseline Approaches. To investigate the contribution of each individual information that builds up the feature, we separate the features into three parts: *Motion* part denotes the space-time feature and mobility feature; *Scene* part denotes the scene prior and scene context feature; *Group* part denotes the group interaction feature and group structure context. We use the following combinations of above three parts **(Motion, Motion-Scene, Motion-Group, Motion-Scene-Group)** to train the deep neural network (*DNN*) model and compare their performance to validate the contribution of each individual part. When one part of feature is not involved in the training, we remove the nodes and layers related to that part within the network. To evaluate the discriminative power of proposed deep model, we take the same feature combinations and train the *Support Vector Machine (SVM)* classifier (Chang and Lin, 2011) and compare its performance with DNN model. We conduct the experiments on both Data-Act-5 and Data-Act-6 and the results are summarized in Table 2.

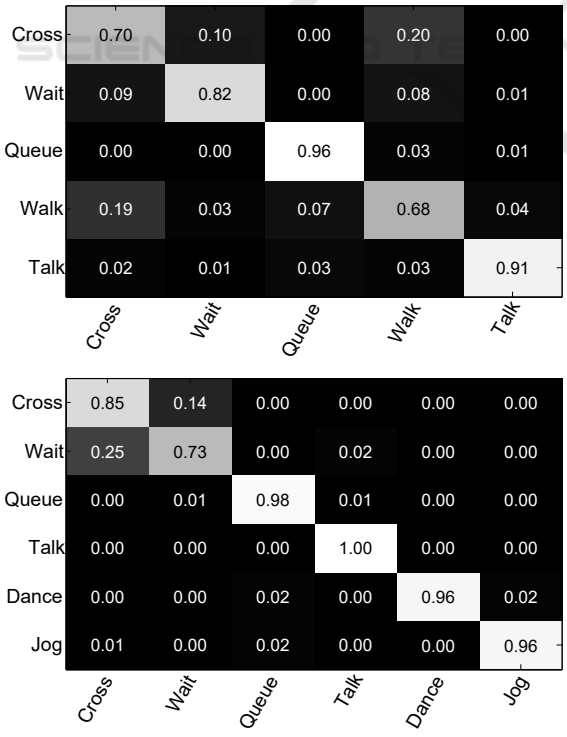

Figure 10: Confusion matrix of Collective Activity Dataset. 5 activities version (top) and 6 activities version (bottom).

By looking into the average accuracy, **DNN-Motion-Scene** outperforms the **DNN-Motion** by 13.4% in Data-Act-5 and 19.2% in Data-Act-6, the activities that lead to significant accuracy improvements are *Talk* (31.0%) and *Queue* (16.2%) in Data-Act-5, *Queue* (42.3%) and *Wait* (34.4%) in Data-Act-6, respectively. **DNN-Motion-Group** outperforms the **DNN-Motion** by 25.7% in Data-Act-5 and 17.5% in Data-Act-5 in average. Interestingly, *Queue* and *Talk* lead to most significant accuracy improvements in both datasets: *Queue* improves 48.4% in Data-Act-5 and 42.8% in Data-Act-5, *Talk* improves 42.4% in Data-Act-5 and 42.8% in Data-Act-6. The observed improvements are reasonable because queuing and talking activities have relatively stable group structures and interaction patterns, and these improvements indicate that our proposed group context feature captures the meaningful information for group structure and interaction. **DNN-Motion-Scene-Group** outperforms **DNN-Motion-Scene** by 17.9% in Data-Act-5 and 1.1% in Data-Act-6. It also outperforms **DNN-Motion-Group** by 4.3% in Data-Act-5 and 2.9% in Data-Act-6. This indicates that both scene context information and group context information contribute to the final performance improvements of the combined feature. However, the contribution rate of scene context information and group context information may vary among different datasets.

By comparing the accuracy of SVM classifier and deep neural network model that is trained using the same features, we are able to evaluate the discriminative power of the proposed deep model. Overall, the accuracy of DNN based model outperforms the SVM model by 13.3% in Data-Act-5 and by 11.1% in Data-Act-6. This clearly indicates that our proposed DNN model also contributes to higher performance of activity recognition task.

Compare with State-of-the-art. We also compare our results with other approaches that have state-of-the-art performance on Collective Activity dataset. For Data-Act-5, we compare our results with Spatio-Temporal Volume descriptor of Choi *et. al.* (Choi

Table 2: Comparison with state-of-the-art approaches.

Accuracy(%)

5-Activites	Walk	Cross	Queue	Wait	Talk	Jog	Dance	Avg.(5 Act.)
SVM-Motion	36.2	64.8	52.0	28.3	20.5	-	-	40.4
DNN-Motion	46.4	63.9	49.1	44.0	53.6	-	-	51.4
SVM-Motion-Scene	42.5	65.2	65.7	43.3	51.0	-	-	53.5
DNN-Motion-Scene	49.5	64.7	65.3	54.8	84.6	-	-	64.8
SVM-Motion-Group	33.4	68.3	78.9	36.1	93.7	-	-	62.1
DNN-Motion-Group	39.3	74.0	97.4	78.8	96.0	-	-	77.1
SVM-Motion-Scene-Group	37.5	67.2	82.3	40.3	93.7	-	-	64.2
DNN-Motion-Scene-Group	67.6	70.2	96.2	81.6	91.5	-	-	81.4
6-Activites	Walk	Cross	Queue	Wait	Talk	Jog	Dance	Avg.(6 Act.)
SVM-Motion	-	63.9	51.8	29.3	21.1	98.4	95.3	60.0
DNN-Motion	-	78.6	57.4	45.8	56.1	93.1	95.3	71.0
SVM-Motion-Scene	-	67.2	65.9	42.2	51.5	97.7	96.5	70.1
DNN-Motion-Scene	-	86.5	99.6	80.2	79.0	98.4	97.8	90.2
SVM-Motion-Group	-	75.2	79.1	46.9	88.9	99.9	99.7	81.6
DNN-Motion-Group	-	75.6	99.2	70.9	98.9	90.1	96.2	88.5
SVM-Motion-Scene-Group	-	83.1	81.9	50.6	93.4	99.9	99.7	84.8
DNN-Motion-Scene-Group	-	85.4	97.9	72.6	99.6	96.4	96.1	91.3

Table 3: Comparison with state-of-the-art approaches.

Accuracy(%)

Approaches	Year	Walk	Cross	Queue	Wait	Talk	Jog	Dance	Avg.(5 Act.)
Choi et. al. (Choi et al., 2009)	2009	57.9	55.4	63.3	64.6	83.6	-	-	65.9
Lan et. al. (Lan et al., 2012)	2012	**68.0**	65.0	96.0	68.0	**99.0**	-	-	79.1
Our Method (5 Act.)		67.6	**70.2**	**96.2**	**81.6**	91.5	-	-	**81.4**
Approaches	Year	Walk	Cross	Queue	Wait	Talk	Jog	Dance	Avg.(6 Act.)
Choi et. al. (Choi et al., 2011)	2011	-	76.5	78.5	78.5	84.1	94.1	80.5	82.0
Amer et. al. (Amer and Todorovic, 2011)	2011	-	69.9	96.8	74.1	**99.8**	87.6	70.2	83.1
Amer et. al. (Amer et al., 2012)	2012	-	77.2	95.4	78.3	98.4	89.4	72.3	85.1
Khai et. al. (Tran et al., 2015)	2015	-	60.6	89.1	**80.9**	93.1	93.4	95.4	85.4
Our Method (6 Act.)		-	**85.4**	**97.9**	72.6	99.6	**96.4**	**96.1**	**91.3**

et al., 2009) and Action Context descriptor of Lan *et. al.* (Lan et al., 2012). For Data-Act-6, the following methods are compared: the approach by Tran *et. al.* (Tran et al., 2015) that uses group context descriptor, the approach by Amer *et. al.* (Amer and Todorovic, 2011) that uses a chain model for group activities recognition and (Amer et al., 2012) that utilize top-down/bottom-up inference for activity recognition; and the approach by Choi *et. al.* (Choi et al., 2011) that uses random forest for activities recognition.

The results are shown in Table 3. We can see that our approach performs best in 3 out of 5 activities in Data-Act-5, and 4 out of 6 activities in Data-Act-6. Our approach also gives the best average accuracy for both datasets. Finally, our approach outperforms other approaches by 2% in Data-Act-5 and at least by 5% in Data-Act-6.

5 CONCLUSION

In conclusion, we propose a deep neural network model for human activity recognition from video. The input features of the deep network include motion feature and context feature. We design the scene prior feature and scene context feature to capture the environment around the subject of interest global and local levels. We demonstrate that our model is able to outperform state-of-the-art human activity recognition methods in the collective activities dataset.

REFERENCES

Amer, M. and Todorovic, S. (2011). A chains model for localizing participants of group activities in videos.

In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 786–793.

Amer, M. R., Xie, D., Zhao, M., Todorovic, S., and Zhu, S.-C. (2012). Cost-sensitive top-down/bottom-up inference for multiscale activity recognition. In *Computer Vision–ECCV 2012*, pages 187–200. Springer.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

Choi, W. and Savarese, S. (2012). A unified framework for multi-target tracking and collective activity recognition. In *Computer Vision–ECCV 2012*, pages 215–230. Springer.

Choi, W., Shahid, K., and Savarese, S. (2009). What are they doing? : Collective activity classification using spatio-temporal relationship among people. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 1282–1289.

Choi, W., Shahid, K., and Savarese, S. (2011). Learning context for collective activity recognition. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*.

Dieleman, S., Schluter, J., Raffel, C., Olson, E., Snderby, S. K., Nouri, D., Maturana, D., Thoma, M., Battenberg, E., Kelly, J., Fauw, J. D., Heilman, M., diogo149, McFee, B., Weideman, H., takacsg84, peterderivaz, Jon, instagibbs, Rasul, D. K., CongLiu, Britefury, and Degrave, J. (2015). Lasagne: First release.

Ess, A., Leibe, B., Schindler, K., and Van Gool, L. (2008). A mobile vision system for robust multi-person tracking. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *International conference on artificial intelligence and statistics*, pages 249–256.

Hasan, M. and Roy-Chowdhury, A. K. (2014). Continuous learning of human activity models using deep nets. In *Computer Vision–ECCV 2014*, pages 705–720. Springer.

Ji, S., Xu, W., Yang, M., and Yu, K. (2013). 3d convolutional neural networks for human action recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):221–231.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1725–1732. IEEE.

Lan, T., Wang, Y., Yang, W., Robinovitch, S. N., and Mori, G. (2012). Discriminative latent models for recognizing contextual group activities. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(8):1549–1562.

Laptev, I., Marszałek, M., Schmid, C., and Rozenfeld, B. (2008a). Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE.

Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B. (2008b). Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8.

Nesterov, Y. et al. (2007). Gradient methods for minimizing composite objective function. Technical report, UCL.

Schuldt, C., Laptev, I., and Caputo, B. (2004). Recognizing human actions: a local svm approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 32–36 Vol.3.

Tran, K. N., Bedagkar-Gala, A., Kakadiaris, I. A., and Shah, S. K. (2013). Social cues in group formation and local interactions for collective activity analysis. In *VISAPP*, pages 539–548.

Tran, K. N., Kakadiaris, I. A., and Shah, S. K. (2012). Part-based motion descriptor image for human action recognition. *Pattern Recognition*, 45(7):2562–2572.

Tran, K. N., Yan, X., Kakadiaris, I. A., and Shah, S. K. (2015). A group contextual model for activity recognition in crowded scenes. In *VISAPP*.

Wang, X. and Ji, Q. (2015). Video event recognition with deep hierarchical context model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4418–4427.

Was, J., Gudowski, B., and Matuszyk, P. J. (2006). Social distances model of pedestrian dynamics. In *Cellular Automata*, pages 492–501. Springer.

Wei, L. and Shah, S. K. (2015). Subject centric group feature for person re-identification. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 28–35.

Wei, L. and Shah, S. K. (2016). Person re-identification with spatial appearance group feature. In *2016 IEEE Symposium on Technologies for Homeland Security (HST)*, pages 1–6.

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

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A. (2014). Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*, pages 487–495.