Deep Structured Models For Group Activity Recognition

by

Zhiwei Deng

B.Sc., Beijing University of Posts and Telecoms, 2013

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
School of Computing Science
Faculty of Applied Sciences

Approval

Name: **Zhiwei Deng**

Degree: **Master of Science (Computing Science)**

Title: ***Deep Structured Models For Group Activity Recognition***

Examining Committee: **Chair:** Richard Vaughan
Professor

Greg Mori
Senior Supervisor
Professor

Arash Vahdat
Supervisor
University Research Associate

Anoop Sarkar
Internal Examiner
Professor

Date Defended: December 18, 2015

Abstract

This thesis presents a deep neural-network-based hierarchical graphical model for individual and group activity recognition in surveillance scenes. As the first step, deep networks are used to recognize activities of individual people in a scene. Then, a neural network-based hierarchical graphical model refines the predicted labels for each activity by considering dependencies between different classes. Similar to the inference mechanism in a probabilistic graphical model, the refinement step mimics a message-passing encoded into a deep neural network architecture. We show that this approach can be effective in group activity recognition. The deep graphical model improves recognition rates over baseline methods.

Keywords: Deep Structured Model; Group Activity; Message Passing

Acknowledgements

First and foremost, I would like to thank my supervisor Dr. Greg Mori, for his great support and patient guidance. I'm lucky to have such an advisor who has always been positive, passionate and always cares for his students. Thanks to Greg for sharp comments and deep insights. He always tried to help me when I got lost, and gave me freedom to help me to become more independent on research.

I'm also deeply grateful to Dr. Arash Vahdat for his encouragement, advisement and support. Thanks for always encouraging me whenever I encountered problems, for being available despite your busy schedules and for being more than a supervisor but also a friend who often gives me good advice on research and life.

I would like to thanks to Guangtong Zhou, Wang Yan, Mengyao Zhai, Lei Chen, Hossein Hajimirsadeghi, Hexiang Hu and all other lab members for their generous help during the two years. Also great thanks to Tian Lan who mentored and helped me in my first project.

I would also like to express my attitude to committee members. Thanks for your helpful suggestions and insightful comments.

Last but not least, deepest thanks to my families for all their unconditional supports and encouragement.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Event understanding in videos is a key element of computer vision systems in the context of visual surveillance, human-computer interaction, sports interpretation, and video search and retrieval. Varied components with different characteristics are involved in corresponding events or scenes. Therefore events, activities, and interactions must be represented in such a way that retains all of the important visual information in a compact and rich structure. Accurate detection and recognition of atomic actions of each individual person in a video is the primary component of such a system, and also the most important, as it affects the performance of the whole system significantly. Although there are many methods to determine human actions in uncontrolled environments, this task remains a challenging computer vision problem, and robust solutions would open up many useful applications.

1.1 Features for Activity Recognition

The standard and yet state-of-the-art pipeline for activity recognition and interaction description in surveillance videos consists of extracting hand-crafted local feature descriptors either densely or at a sparse set of interest points (e.g., HOG, MBH, ...) in the context of a Bag of Words model [28]. These are then used as the input either to a discriminative or a generative model and achieved good results in many applications. In recent years, along with the emerging of deep learning field combined with new computational resources, many traditional computer vision tasks have witnessed a surprising jump in performance by using machine learning based methods. It also has been shown that deep learning techniques can achieve state-of-the-art results for action recognition in surveillance videos [24, 14]. However, deep learning has not yet proven successful for surveillance video analysis. Activities in videos are often complex, with high-level semantic meaning derived from often subtle distinctions and varied spatio-temporal interactions between people or people and objects.

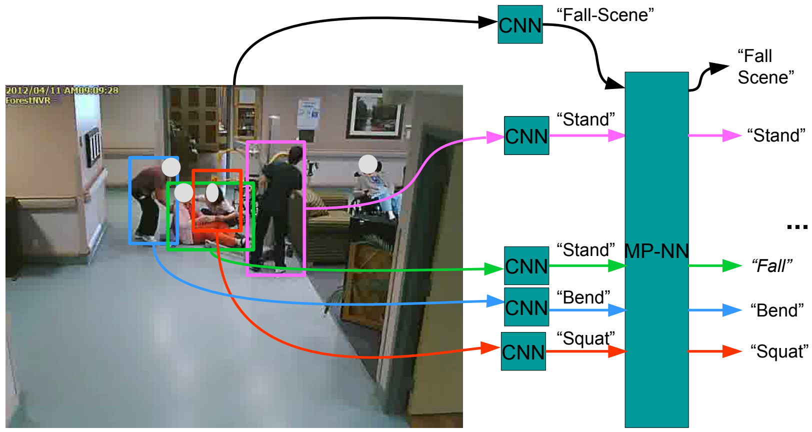

Figure 1.1: Recognizing individual and group activities in a deep network. Individual action labels are predicted via CNNs. Next, these are refined through a message passing neural network which considers the dependencies between the predicted labels.

1.2 Graphical Model In Group Activity Recognition

Understanding of complex visual events in a scene requires exploitation of richer information rather than individual atomic activities, such as recognizing local pairwise and global relationships in a social context and interaction between individuals and/or objects [18, 22, 30, 23, 5]. This complex scene description remains an open and challenging task. It shares all of difficulties of action recognition, interaction modeling[1], and social event description. Formulating this problem within the probabilistic graphical models framework provides a natural and powerful means to incorporate the hierarchical structure of group activities and interactions [20, 18]. Given the fact that deep neural networks can achieve very competitive results on the single person activity recognition tasks, they can, produce better results when they are combined with other methods, e.g. graphical models, in order to capture the dependencies between the variables of interest [26]. Following a similar idea of incorporating spatial dependency between variables into the deep neural network in a joint-training process presented [26], here we focus on learning interactions and group activities in a surveillance scene by employing a graphical model in a deep neural network paradigm.

[1]The term "interaction" refers to any kind of interaction between humans, and humans and objects that are present in the scene, rather than activities which are performed by a single subject.

1.3 Contribution

In this thesis, the main goal is to address the problem of *group activity understanding* and *scene classification* in complex surveillance videos using a deep learning framework. More specifically, we are focused on learning individual activities and describing the scene simultaneously while considering the pair-wise interactions between individuals and their global relationship in the scene. This is achieved by combining a Convolutional Network (ConvNet) with a probabilistic graphical model as additional layers in a deep neural network architecture into a unified learning framework. The probabilistic graphical models can be seen as a refining process for predicting class labels by considering dependencies between individual actions, body poses, and group activities. The probabilistic graphical model is modeled by a multi-step message passing neural network and the predicted label refinement is carried out through belief propagation layers in the neural network. Figure 1.1 depicts an overview of our approach for label refinement. Experimental results show the effectiveness of our algorithm in both activity recognition and scene classification. This work was published as [9]:

- Zhiwei Deng, Mengyao Zhai, Lei Chen, Yuhao Liu, Srikanth Muralidharan, Mehrsan Javan Roshtkhari, Greg Mori, Deep structured models for group activity recognition, BMVC, 2015

1.4 Overview of the Thesis

In this thesis, we proposed a model to combine deep neural networks with graphical models by mimicking the message passing process in the neural network. The rest of the thesis is organized as follows:

Chapter 2 illustrates the previous works related to the topic of group activity recognition and deep neural network model with structured outputs. Chapter 3 introduces and explains the proposed model for integrating graphical model and empowering deep learning with structured output. Chapter 4 presents the group activity analysis results based on our proposed pipeline on two challenging datasets. Chapter 5 concludes our method and proposes potential future work.

Chapter 2

Previous Work

Human activity analysis, as a line of research considering compositionality of basic elements towards more complex semantic higher-level understanding, has been actively studied for many years. A substantial amount of research on this topic has produced a diverse set of approaches and a rich collection of activity recognition algorithms. Readers can refer to recent surveys such as Poppe [21] and Weinland et al. [29] for a review. Many approaches concentrate on an activity performed by a single person, including state of the art deep learning approaches [24, 14].

2.1 Group Activity Analysis

In the context of scene classification and group activity understanding, the problem is typically formulated as a structured prediction problem that considers both individual actions, group activities and interactions among them. Graphical model as a standard method for modeling structure has been adopted by many previous works in varied ways. Hierarchical graphical models that represent activities and interactions for collective activity recognition have proven to be successful and effective [18, 3, 6, 19]. Choi et al. [6] has been focused on capturing spatio-temporal relationships between visual cues either by imposing a richer feature descriptor which accounts for context [27, 7] or a context-aware inference mechanism [6, 3]. Amer et al. [3] proposed a hierarchical graphical model with adaptive structure to cature the most discriminative interactions in a scene. Ryoo et al. [23] reasons over varied granularities of group activity by adopting powerful context-free grammar as representation. Lan et al. [18] adopted a graphical model reasons over actions, social roles and activities on sports videos. AND-OR graphs as a powerful method that is able to adaptively reason about structures of activities has been adopted to analyze group activity by a line of papers [2, 11]. And dynamic Bayesian networks are used to model spatial temporal human activities by considering context [30].

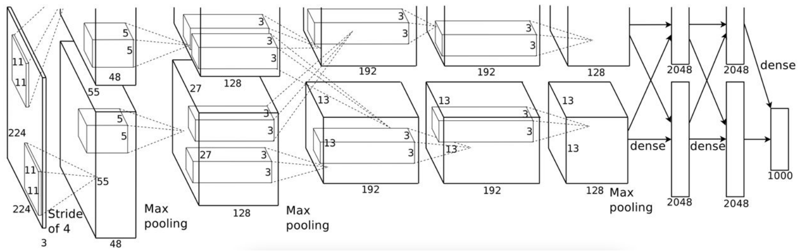

Figure 2.1: Convolutional neural network architecture proposed by Krizhevsky et al. [15].

2.2 Convolutional Neural Network

In recent years, deep learning models have been widely explored and adopted in a variety of computer vision applications and achieved impressing results. In image classification, Krizhenvsky et al. [15] proposed a deep convolutional network structure that extract effective features from raw pixels and achieved record-breaking results, the architecture is shown in fig Figure 2.1. Simonyan et al. [13] and Szegedy et al.[25] proposed deeper network structures with higher complexity to extract more discriminative features. Along with the success of deep learning method in image classification, varied applications have adopted convolutional neural network to explore the ability of deep neural network in more directions. In another classic task, Girshick et al. [10] expand the convolutional neural network image classification model to object detection and achieved state-of-the-art results. For deep learning model in action recognition, recent works [14, 24] have also shown effectiveness of deep learning model by exploiting information in temporal domain. Karpathy et al. [14] proposed to recognize atomic actions by varied fusion techniques. Simonyan et al. [24] proposed a two-stream framework which combine deep models for both optical flow and image inputs. However, no prior art in the CNN-based video description used activities and scene information jointly in a unified graphical representation for scene classification. Therefore, the main objective of this research is to develop a system for activity recognition and scene classification which simultaneously uses the action and scene labels in a neural network-based graphical model to refine the predicted labels via a multiple-step message passing procedure.

2.3 Deep Learning with Structured Output

For the structured output model on deep neural network, a few methods have been developed to adapt to applications respectively [26, 8, 16]. Deng et al. [8] propose an interesting solution to improve label prediction in large scale classification by considering relations between the predicted class labels. They employ a probabilistic graphical model with hard constraints on the labels on top of a neural network in a joint training process. Bottou et al. [16] proposed Graph Transformer Networks to jointy optimize subtasks. In this work, it was assumed that exact inference can be performed during a forward-backward pass. The more closely related work to our approach is combining graphical models with convolutional neural networks [26]. In the work of Tompson et al. [26], a one step message passing is implemented as a convolution operation in order to incorporate spatial relationship between local detection responses for human body pose estimation. In essence, our proposed algorithm follows a similar idea of considering dependencies between predicted labels for the actions, group activities, and the scene label to solve the group activity recognition problem. Here we focus on incorporating those dependencies by implementing the label refinement process via an inter-activity neural network, as shown in Figure 3.1. The network learns the weights during the message passing procedure and performs inference and learning in unified framework using back-propagation.

Chapter 3

Deep Structured Model

Considering the architecture of our proposed structured label refinement algorithm for group activity understanding (see Figure 3.1), the key part of the algorithm is a multi-step message passing neural network. In this chapter, we describe how to combine neural networks and graphical models by mimicking a message passing algorithm and how to carry out the training procedure.

3.1 Graphical Models in a Neural Network

Graphical models provide a natural way to hierarchically model group activities and capture the semantic dependencies between group and individual activities [20]. A graphical model defines a joint distribution over states of a set of nodes. For instance, one can use a factor graph, in which each ϕ_i corresponds to a factor over a set of related variable nodes x_i and y_i, and models interactions between these nodes in a log-linear fashion:

$$P(X,Y) \propto \prod_i \phi_i(x_i, y_i) \propto exp(\sum_k w_k f_k(x,y)) \tag{3.1}$$

where X are the inputs and Y the predicted labels, with weighted (w_k) feature functions f_k.

In order to do the inference in a graphical model, belief propagation is often adopted as a way to infer states or probabilities of the variables. In the belief propagation algorithm, each step of message passing involves two parts. At first the relevant information from the connected nodes to a factor node are collected. Those messages are then passed to the variable nodes by marginalizing over states of irrelevant variables.

Following this idea, we mimic the message passing process by representing every possible combination of states as a neuron in neural network, denoted as a “factor neuron”. While normal message passing calculates dependencies rigidly, a factor neuron can be used to learn and predict dependencies between states and pass messages to variable nodes. In the setting

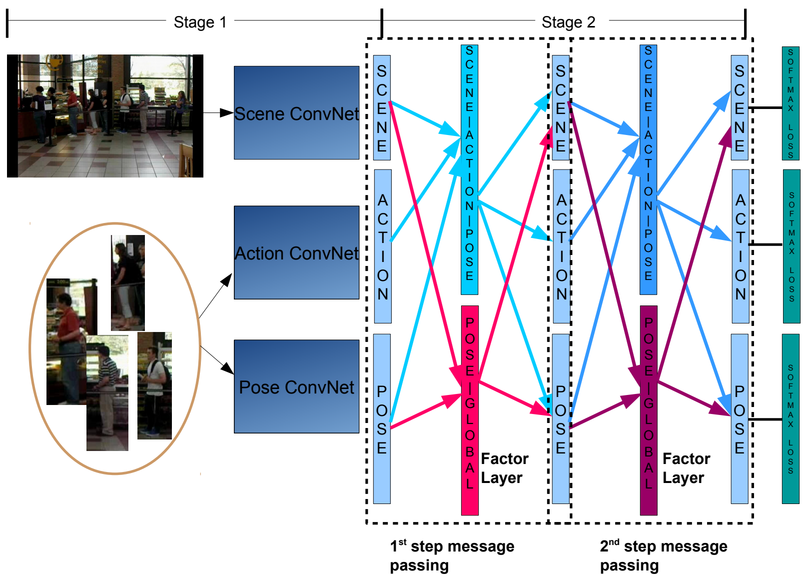

Figure 3.1: A schematic overview of our message passing neural network framework. Given an image frame and the detected bounding boxes around each person, our model predicts scores for individual actions and the group activities. The predicted labels are refined by applying a belief propagation-like neural network. This network considers the dependencies between individual actions and body poses, and the group activity. The model learns the message passing parameters and performs inference and learning in unified framework using back-propagation.

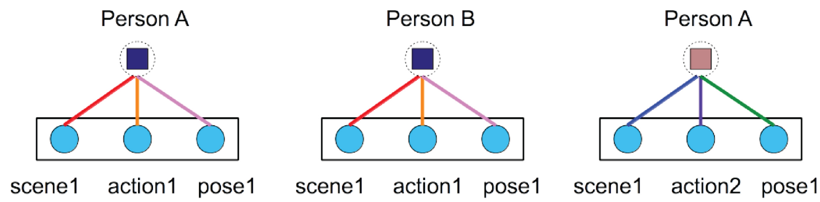

Figure 3.2: Weight sharing scheme in neural network. We use a sparsely connected layer to represent message passing between variable nodes and factor nodes. Each factor node only connects to its relevant nodes. And factor nodes of same type share a template of parameters. For example, factor node for person A and person B in this picture captures the same states combination, scene1, action1 and pose1, and share the same set of weights. And the third factor node shown in the picture captures the dependency of a different set of states of person A and the group activity of current frame, and adopts another set of weights.

of neural networks, this dependency representation becomes more flexible and can adopt varied types of neurons (linear, ReLU, Sigmoid, etc.). Moreover, by integrating graphical models into a neural network, the formulation of a graphical model allows for parameter sharing in the neural network, which not only reduces the number of free parameters to learn but also accounts for semantic similarities between factor neurons. Fig. 3.2 shows the parameter sharing scheme for different factor neurons.

3.2 Message Passing NN Architecture for Group Activity

Representing group activities and individual activities as a hierarchical graphical model has proven to be a successful strategy [20, 6, 2]. We adopt a similar structured model that considers group activity, individual activity, and group-individual interactions together. We introduce a new message passing Convolutional Neural Network framework as shown in Fig. 3.1. Our model has two main parts: (1) a set of fine-tuned CNNs that produce a *scene score* for an image, and *action scores* and *pose scores* for each individual person in that image; and (2) a message passing neural network which captures the dependencies between activities, poses, and scene labels.

Given an image I and a set of bounding boxes for detected persons $\{I_1, I_2, ..., I_M\}$[1], the first part of our model generates raw scores of scene. In addition, it produces the raw scores for the actions and poses of each of the M individuals in the image $\{I_i\}_{i=1}^{M}$. This is done by applying fine-tuned CNNs on the image and the detected bounding boxes. A soft-max normalization is then applied for each scene, activity, and pose score in order to produce the raw scores.

[1]It is assumed that the bounding box around each person is known. Those bounding boxes are obtained by applying a person detector on each image as a pre-processing step

The second part of our algorithm which does the label refinement takes those raw scores as the input. In our graphical model, outputs from CNNs correspond to unary potentials. The scene-level, and per-person action and pose-level unary potentials for the image I are represented by $s^{(k)}(I)$, $\mathbf{a}^{(k)}(I_m)$ and $\mathbf{r}^{(k)}(I_m)$ respectively. The super script (k) is the index of message passing step. We use G to denote all group activity labels, H to represent all the action labels and Z to denote all the pose labels. Then the group activity in one scene can be represented as g_I, $h_{I_1}, h_{I_2}, ..., h_{I_M}$, $g_{I_1}, g_{I_2}, ..., g_{I_M}$ where $g_I \in G$ is the group activity label for image I, h_{I_m} and z_{I_m} are action labels and pose labels for a person I_m.

Note that for training, the scene, action, and pose CNN models in the first part of our algorithm are fine-tuned from an AlexNet architecture pre-trained using ImageNet data. The architecture is similar to the one proposed by [1] for object classification with some minor differences, *e.g.* pooling is done before normalization. The network consists of five convolutional layers followed by two fully connected layers, and a softmax layer that outputs individual class scores. This network can be obtained from Caffe library [12]. We use the softmax loss, stochastic gradient descent and dropout regularization to train these three ConvNets.

In the second part of our algorithm, we use the method described in Sec. 3.1 to mimic the message passing in a hierarchical graphical model for group activity recognition in a scene. This stage can contain several steps of message passing. In each step, there are two types of passes: from outputs of step $k-1$ to factor layer and from factor layer to k step outputs. In the kth message passing step, the first pass computes dependencies between states. The inputs to the kth step message passing are

$$s_1^{(k-1)}(I), ..., s_{|G|}^{(k-1)}(I), a_1^{(k-1)}(I_1), ..., a_{|H|}^{(k-1)}(I_M), r_1^{(k-1)}(I_1), ..., r_{|Z|}^{(k-1)}(I_M) \tag{3.2}$$

where $s_g^{(k-1)}(I)$ is the scene score of image I for label g, $a_h^{(k-1)}(I_m)$ is the action score of person I_m for label h and $r_z^{(k-1)}(I_m)$ is the pose score of person I_m for label z. In the factor layer, the action, pose and scene interaction are calculated as:

$$\phi_j(s_g^{(k-1)}(I), a_h^{(k-1)}(I_m), r_z^{(k-1)}(I_m)) = f(\alpha_{\mathbf{g},\mathbf{h},\mathbf{z}}[s_g^{(k-1)}(I), a_h^{(k-1)}(I_m), r_z^{(k-1)}(I_m)]^T) \tag{3.3}$$

where $\alpha_{\mathbf{g},\mathbf{h},\mathbf{z}}$ is the scene score of image I for label g, action h and pose z, f is the activation function and j is the corresponding index of factor neuron. Similarly, pose interactions for all people in the scene are calculated as:

$$\psi_j(s_g^{(k-1)}(I), \mathbf{r}) = f(\beta_{\mathbf{d},\mathbf{g}}[s_g^{(k-1)}(I), \mathbf{r}]^T) \tag{3.4}$$

where $\mathbf{r}$ is all output nodes for all people, d is the factor neuron index for scene g, f is the activation function and j is the corresponding index of factor neuron. D latent factor neurons are used for a scene g. Note that parameters α and β are shared within factors that have the same semantic meaning. For the output of kth step message passing, the score for the scene label to be g can be defined as:

$$s_g^{(k)}(I) = s_g^{(k-1)} + \sum_{j \in \varepsilon_1^s} w_{ij} \phi(s_g^{(k-1)}, \mathbf{a}, \mathbf{r}; \alpha) + \sum_{j \in \varepsilon_2^s} w_{ij} \psi(s_g^{(k-1)}(I), \mathbf{r}; \beta) \tag{3.5}$$

where ε_1^s and ε_2^s are the set of factor nodes that are connected with scene g in first factor component (scene-action-pose factor) and second factor component (pose-global factor) respectively. Similarly, we also define action and pose scores after the kth message passing step as:

$$a_h^{(k)} = a_h^{(k-1)}(I_m) + \sum_{j \in \varepsilon_1^a} w_{ij} \phi_j(a_h^{(k-1)}(I_m), \mathbf{s}, \mathbf{r}; \alpha) \tag{3.6}$$

$$r_z^{(k)}(I_m) = r_z^{(k-1)}(I_m) + \sum_{j \in \varepsilon_1^r} w_{ij} \phi(r_z^{(k-1)}(I_m), \mathbf{a}, \mathbf{s}; \alpha) + \sum_{j \in \varepsilon_2^r} w_{ij} \psi(\mathbf{r}_z^{(k-1)}(I_m), \mathbf{r}; \beta) \tag{3.7}$$

where $\varepsilon = \{\varepsilon_1^s, \varepsilon_2^s, \varepsilon_1^a, \varepsilon_1^r, \varepsilon_2^r\}$ are connection configurations in the pass from factor neurons to output neurons, i is the index of output neurons. These connections are simply the reverse of the configurations in the first pass, from input to factors. The model parameters $\{\mathbf{W}, \alpha, \beta\}$ are weights on the edges of the neural network. Parameter $\mathbf{W}$ represents the concatenation of weights connected from factor layers to output layer (second pass), while α, β represent weights from the input layer of the k^{th} message passing to factor layers (first pass).

3.2.1 Components in Factor Layers

This section summarizes and explains all different components of our model, which are as follows:

Unary component: In our message passing model, the unary component corresponds to group activity scores for an image I, action and pose scores for each person I_m in frame I, represented as $s_g^{(k-1)}(I)$, $a_h^{(k-1)}(I_m)$ and $r_z^{(k-1)}(I_m)$ respectively. These scores are acquired from the previous step of message passing and are directly added to the output of the next message passing step.

Group activity-action-pose factor layer ϕ: A group's activity is strongly correlated to the participating individuals' actions. This component for the model is used to measure the compatibility between individuals and groups. An individual's activity can be described by both pose and action, and we use this ternary scene-pose-action factor layer to capture dependencies between a person's fine-grained action (e.g. talking facing front-left) and the scene label for a group of people. Note that in this factor layer we used the weight sharing scheme mentioned in Sec. 3.1 to mimic the belief propagation.

Poses-all factor layer ψ: Pose information is very important in understanding a group activity. For example, when all people are looking in the same direction, there is a high probability that it's a queueing scene. This component captures this global pose information for a scene. Instead of naively enumerate all combination of poses for all people, we exploit the sparsity of truly useful and frequent patterns, and simply use D factor nodes for one scene label. In our experiments, we simply set D to be 10.

3.3 Multi Step Message Passing NN Training

The steps of message passing depends on the structure of graphical model. In general, graphical models with loops or large number of levels will lead to more steps belief propagation for sharing local information globally. In our model, we adopt two message passing steps, as shown in Fig. 3.1.

Multi-loss training: Since the goal of our model is to recognize group activities through global features and individual actions in that group, we adopt an alternative strategy for training the model. For the k^{th} message passing step, we first remove the loss layers for actions and poses to learn parameters for group activity classification alone. In this phase, there is no back-propagation on action and pose classification. Since group activity heavily depends on an individual's activity, we then fix the softmax loss layer for scene classification and learn the model for actions and poses. The trained model is used for the next message passing step. Note that in each message passing step, we exploit the benefit of the neural network structure and jointly trained the whole network (including convolutional neural network and message passing neural network).

Learning semantic features for group activity: Traditional convolutional neural networks mainly focus on learning features for basic classification or localization tasks. However, in our proposed message passing NN deep model, we not only learn features, but also learn semantic high-level features for better representing group activities and interactions within the group. We explore different layers' features for this deep model, and results show that these semantic features can be used for better scene understanding and classification.

Implementation details: Firstly, in practice, it is not guaranteed that every frame has the same number of detections. However, the structure of neural network should be fixed. To solve this problem, denoting M_{max} as the maximum number of people contained in one frame, we do a dummy-image padding when the number of people is less than M_{max}. Then we filter out these dummy data by ignoring the activations or gradients generated by these dummy data. Secondly, after the first message passing step, instead of directly feeding the raw scores into the next message passing step, we first normalize the pose and action scores for each person and scene scores for one frame by a softmax layer, converting to probabilities similar to belief propagation.

Chapter 4

Experiments

Our models are implemented using the Caffe library [12] by defining two types of sparsely connected and weight shared inner product layers. One is from variable nodes to factor nodes, another is the reverse direction. We used TanH neurons as the non-linearity of these two layers. To examine the performance of our model, we test our model for scene classification on two datasets: (1) Collective Activity [7], (2) a nursing home dataset consisting of surveillance videos collected from a nursing home.

We trained an RBF kernel SVM on features extracted from the graphical model layer after each step of message passing model. These SVMs are used to predict scene labels for each frame, the standard task in these datasets.

4.1 Collective Activity Dataset

The Collective Activity Dataset contains 44 video clips acquired using low resolution hand-held cameras. Every person is assigned one of the following five action labels: crossing, waiting, queuing, walking and talking and one of the eight pose labels: right, front-right, front, front-left, left, back-left, back, back-right. Each frame is assigned one of the following five activities: crossing, waiting, queueing, walking, and talking. The activity category is attained by taking the majority of actions happening in one frame while ignoring the poses. We adopt the standard training test split used in [20].

In the Collective Activity dataset experiment, we further concatenate the global features for a scene with AC descriptors by HOG features [20]. We simply averaged AC descriptors features for all people and use this feature to serve as additional global information, namely this feature does not truly participated in the message passing process. The scene classification accuracy on the Collective Activity dataset by using a baseline AlexNet model is 63%. This additional global information assists in classification with the limited amount of training data available for this dataset.

We summarize the comparisons of activity classification accuracies of different methods in Tab. 4.1. The current best result using spatial information in graphical model is 79.1%, from Lan et al. [20], which adopted a latent max-margin method to learn graphical model with optimized structure. Our classification accuracies (the best is 80.6%) are competitive compared with the state-of-the-art methods. However, the benefits of the message passing are clear. Through each step of the message passing, the factor layer effectively captured dependencies between different variables and passing messages using factor neurons results in a gain in classification accuracy. Some visualization results are shown in Fig 4.1. The accuracy saturated after two steps of message passing.

	1 Step MP	2 Steps MP
Pure DL	73.6%	78.4%
SVM+DL Feature	75.1%	80.6%

Latent Constituent [4]	75.1%
Contextual model [20]	79.1%
Our Best Result	**80.6**%

Table 4.1: Scene classification accuracy on the Collective Activity Dataset.

4.2 Nursing Home Dataset

This dataset consists 80 videos and is captured in a nursing home, including a variety of rooms such as dining rooms, corridors, etc. The 80 surveillance videos are recorded at 640 by 480 pixels at 24 frames per second, and contain a diverse set of actions and frequent cluttered scenes. This dataset contains typical actions include walking, standing, sitting, bending, squatting, and falling. For this dataset, the goal is to detect falling people, thus we assign each frame one of two activity categories: fall and non-fall. A frame is assigned "fall" if any person falls and "non-fall" otherwise. Note that many frames are challenging, and the falling person may be occluded by others in the scene. We adopted a standard 2/3 and 1/3 training test split. In order to remove redundancy, we sampled 1 out of every 10 frames for training and evaluation. Since this dataset has a large intra-class diversity within actions, we used the action primitive based detectors proposed in [17] for more robust detection results.

Note that since this dataset has no pose attribute, we used the interaction between scene and actions instead to perform the two step message passing. For the SVM classifier, only deep learning features are used. We summarize the comparisons of activity classification accuracies of different methods in Table 4.2.

Ground Truth	Pure DL	SVM+DL Fea.
1 Step MP	82.5%	82.3%
2 Steps MP	84.1%	84.7%

Detection	Pure DL	SVM+DL Fea.
1 Step MP	74.4%	76.5%
2 Steps MP	75.6%	77.3%

Table 4.2: Classification accuracy on the nursing home dataset

.

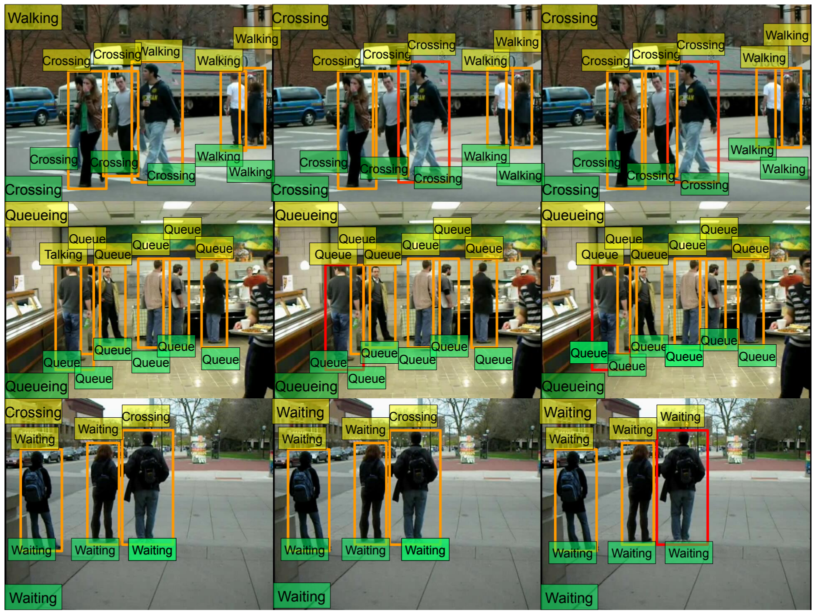

Figure 4.1: Results visualization for our model. Green tags are ground truth, yellow tags are predicted labels. From left to right is without message passing, first step message passing and second step message passing

The scene classification accuracy on the Nursing Home dataset by using a baseline AlexNet model is 69%. The results on scene classification for each step also shows gains. Note that in this dataset, accuracy on the second message passing gains an increase of around 1.5% for both pure deep learning or SVM prediction. We believe that this is due to the fact that the dataset only contains two scene labels, fall or non-fall, so scene variables are not as informative as scenes in the Collective Activity Dataset. The accuracy saturated after two steps of message passing.

Chapter 5

Conclusion and Future Works

In conclusion, we have presented a deep learning model for group activity recognition which jointly captures the group activity, the individual person actions, and the interactions between them. We propose a way to combine graphical models with a deep network by mimicking the message passing process to do inference mechanism. The model was successfully applied to real surveillance videos and the experiments showed the effectiveness of our approach in recognizing activities of a group of people.

The experiments show that the model can effectively conduct message passing in a neural network. And after each step of message passing which refines the label predictions by information sharing, there is an increase in the classification accuracies.

One limitation of this method is that temporal information is not considered in the whole pipeline. Temporal information is also an important element of video analysis. Both group activities and person actions have strong correlations along time axis and by incorporating temporal connections into graphical model, more interesting and powerful inference could be conducted. Hence, a possible future direction of this method is to build the temporal version of the message passing model by some typical temporally deep neural network model such as recurrent neural network or LSTM. For example, a recurrent neural network could be built on the last step of message passing output to serve as a temporal model that captures the changes along the time axis. Moreover, another possibility is to build recurrent neural networks on both output neurons and factor neurons to make it more powerful

Another possibility is to extend the current model into a more complex and bigger dataset with larger label space. More combinations of labels could be explored. And relevant techniques to reduce the enormous combinatorial states space should also be proposed in this situation. For example in a dataset with very large label space, to reduce the number of factor neurons need to be learned, L1 regularization or dropout techniques could be used to learn more compact and discriminative combinations of labels.

Bibliography

[1] Ilya Sutskever Alex Krizhevsky and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems (NIPS)*, 2012.

[2] Mohamed R. Amer, Dan Xie, Mingtian Zhao, Sinisa Todorovic, and Song-Chun Zhu. Cost-sensitive top-down / bottom-up inference for multiscale activity recognition. In *European Conference on Computer Vision (ECCV)*, 2012.

[3] Mohamed Rabie Amer, Peng Lei, and Sinisa Todorovic. Hirf: Hierarchical random field for collective activity recognition in videos. In *European Conference on Computer Vision (ECCV)*, pages 572–585, 2014.

[4] Borislav Antic and Bjorn Ommer. Learning latent constituents for recognition of group activities in video. In *European Conference on Computer Vision (ECCV)*, 2014.

[5] W. Brendel and S. Todorovic. Learning spatiotemporal graphs of human activities. In *International Conference on Computer Vision (ICCV)*, pages 778–785, 2011.

[6] W. Choi and S. Savarese. A unified framework for multi-target tracking and collective activity recognition. In *European Conference on Computer Vision (ECCV)*, 2012.

[7] Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *International Conference on Computer Vision Workshops on Visual Surveillance*, pages 1282–1289. IEEE, 2009.

[8] Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. Large-scale object classification using label relation graphs. In *European Conference on Computer Vision (ECCV)*, pages 48–64. Springer, 2014.

[9] Zhiwei Deng, Mengyao Zhai, Lei Chen, Yuhao Liu, Srikanth Muralidharan, Mehrsan Javan Roshtkhari, and Greg Mori. Deep structured models for group activity recognition. In *British Machine Vision Conference (BMVC)*, 2015.

[10] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, 2014.

[11] Abhinav Gupta, Praveen Srinivasan, Jianbo Shi, and Larry S. Davis. Understanding videos, constructing plots: Learning a visually grounded storyline model from annotated videos. In *Computer Vision and Pattern Recognition (CVPR)*, 2009.

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

[13] Simonyan Karen and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representation (ICLR)*, 2015.

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

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[16] Y. Bengio L. Bottou and Y. Le Cun. Global training of document processing systems using graph transformer networks. In *Computer Vision and Pattern Recognition (CVPR)*, 1997.

[17] Tian Lan, Chen Lei, Deng Zhiwei, Guang-Tong Zhou, and Greg Mori. Learning action primitives for multi-level video event understanding. In *International Workshop on Visual Surveillance and Re-Identification (at ECCV)*, 2014.

[18] Tian Lan, Leonid Sigal, and Greg Mori. Social roles in hierarchical models for human activity recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2012.

[19] Tian Lan, Yang Wang, Weilong Yang, Stephen Robinovitch, and Greg Mori. Discriminative latent models for recognizing contextual group activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 34(8):1549–1562, 2012.

[20] Tian Lan, Wang Yang, Yang Weilong, and Greg Mori. Beyond actions: Discriminative models for contextual group activities. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[21] R. Poppe. A survey on vision-based human action recognition. *IVC*, 28:976–990, 2010.

[22] Vignesh Ramanathan, Bangpeng Yao, and Li Fei-Fei. Social role discovery in human events. In *Computer Vision and Pattern Recognition (CVPR)*, 2013.

[23] M. S. Ryoo and J. K. Aggarwal. Stochastic representation and recognition of high-level group activities. *International Journal of Computer Vision (IJCV)*, 93(2):183–200, 2011.

[24] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 568–576. Curran Associates, Inc., 2014.

[25] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[26] J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[27] K.N. Tran, A. Gala, I.A. Kakadiaris, and S.K. Shah. Activity analysis in crowded environments using social cues for group discovery and human interaction modeling. *Pattern Recognition Letters*, 44:49–57, 2014.

[28] Heng Wang and C. Schmid. Action recognition with improved trajectories. In *International Conference on Computer Vision (ICCV)*, pages 3551–3558, 2013.

[29] D. Weinland, R. Ronfard, and E. Boyer. A survey of vision-based methods for action representation, segmentation and recognition. In *Computer Vision and Image Understanding (CVIU)*, 2010.

[30] Y. Zhu, N. Nayak, and A. Roy-Chowdhury. Context-aware modeling and recognition of activities in video. In *Computer Vision and Pattern Recognition (CVPR)*, 2013.