A Hierarchical Deep Temporal Model for Group Activity Recognition

by

Srikanth Muralidharan

B. Tech., Indian Institute of Technology Jodhpur, 2014

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

SIMON FRASER UNIVERSITY
Spring 2016

Approval

Name:	**Srikanth Muralidharan**
Degree:	**Master of Science (Computing Science)**
Title:	***A Hierarchical Deep Temporal Model for Group Activity Recognition***
Examining Committee:	**Chair:** Dr. Richard Vaughan Associate Professor

Dr. Greg Mori
Senior Supervisor
Professor

Dr. Mehrsan Javan Roshtkhari
Supervisor
Sportlogiq Inc.

Dr. Jiannan Wang
Internal Examiner
Assistant Professor

Date Defended: 12 April 2016

Abstract

In group activity recognition, the temporal dynamics of the whole activity can be inferred based on the dynamics of the individual people representing the activity. We build a deep model to capture these dynamics based on LSTM (long-short term memory) models. To make use of these observations, we present a 2-stage deep temporal model for the group activity recognition problem. In our model, a LSTM model is designed to represent action dynamics of individual people in a sequence and another LSTM model is designed to aggregate human-level information for whole activity understanding. We evaluate our model over two datasets: the collective activity dataset and a new volleyball dataset. Experimental results demonstrate that our proposed model improves group activity recognition performance as compared to baseline methods.

Keywords: Group Activity Recognition; Recurrent Neural Network; Hierarchical Models

Acknowledgements

I would like to thank my supervisors , Dr. Greg Mori and Dr. Mehrsan Javan for helping me through the course of my Masters. I am grateful for the opportunity I got to be part of this work. I thank all my team members Mostafa S. Ibrahim, Arash Vahdat, and Zhiwei Deng for their great co-ordination with me. I also thank my parents and friends for standing by me and supporting my pursuit for Computer Vision research.

Table of Contents

Approval ii

Abstract iii

Acknowledgements iv

Table of Contents v

List of Tables vii

List of Figures viii

1 Introduction 1

2 Related Work 4

2.1 Group Activity Recognition 4

2.2 Sport Video Analysis 7

2.3 Deep Learning 8

2.4 Datasets 11

3 Proposed Approach 12

3.1 Temporal Model of Individual Action 13

3.2 Hierarchical Model for Group Activity Recognition 14

3.3 Implementation Details 16

4 Experiments 17

4.1 Baselines 17

4.2 Experiments on the Collective Activity Dataset 18

4.2.1 Discussion 19

4.3 Experiments on the Volleyball Dataset 19

5 Conclusion 24

5.1 Limitations and Future Work 24

Bibliography **25**

List of Tables

Table 4.1 Comparison of our method with baseline methods on the Collective Activity Dataset. 18

Table 4.2 Comparison of our method with previously published works on the collective activity dataset. 18

Table 4.3 Statistics of the group activity labels in the volleyball dataset. 20

Table 4.4 Statistics of the action labels in the volleyball dataset. 20

Table 4.5 Comparison of the team activity recognition performance of baselines against our model evaluated on the volleyball dataset. 21

List of Figures

Figure 1.1 Group activity recognition via a hierarchical model. Each person in a scene is modeled using a temporal model that captures his/her dynamics, these models are integrated into a higher-level model that captures scene-level activity. 2

Figure 2.1 ©2012 IEEE. Overview of Lan et al. [22] social roles model 5

Figure 2.2 ©2013 IEEE. Overview of Ramanathan et al. [29] social roles model 5

Figure 2.3 Overview of Choi et al. [3] model 6

Figure 2.4 Overview of Deng et al. [7] deep structured model 7

Figure 2.5 ©2011 IEEE. Overview of [26] 8

Figure 2.6 ©2015 IEEE. Overview of Video captioning network [9]. 9

Figure 2.7 ©2014 IEEE. Overview of visual-linguistic alignment stage in karpathy et al. [14] 10

Figure 3.1 ©2015 IEEE. Schematic representation of an LSTM (Figure borrowed from Donahue et al. [9] 13

Figure 3.2 Our two-stage model for a volleyball match. Given tracklets of K-players, we feed each tracklet in a CNN, followed by a person LSTM layer to represent each player's action. We then pool over all people's spatio-temporal features in the scene. The output of the pooling layer is feed to the second LSTM network to identify the whole teams activity. 15

Figure 4.1 **Confusion matrix for the Collective Activity Dataset obtained using our two-stage model.** 19

Figure 4.2 Confusion matrix for the volleyball dataset obtained using our two-stage hierarchical model. 21

Figure 4.3 Confusion matrix for the volleyball dataset obtained using our two-stage hierarchical model without team considerations. 21

Figure 4.4 Visualizations of the generated scene labels using our model. Green denotes correct classifications, red denotes incorrect. The incorrect ones correspond to the confusion between different actions in ambiguous cases (h and j examples), or in the left and right distinction (i example). 23

Chapter 1

Introduction

What are the people in Figure 1.1 doing? This question can be answered at numerous levels of detail – in this work, we focus on the group activity, a high-level answer such as “team spiking acivity”. We develop a novel hierarchical deep model for group activity recognition.

Human activity recognition is a challenging computer vision problem, and, it has received a lot of attention in the past. It is a challenging problem due to considerable variations arising from the variability within action class, background clutter, high overlap between different action classes, to name a few. Group activity recognition is relatively even more challenging than person level activity recognition, and finds a lot of applications in the context of video surveillance, sport analytics, video search and retrieval.

Popular approaches that address activity recognition use hand crafted feature representation including HOG, MBH both in a dense and sparse fashion [42] [30]. In the context of group activity recognition, it is shown that representing a group activity using a hierarchical model yields better performance than flat representations [22]. Concretely, a semantic representation that captures different levels of interaction (e.g. person-person interactions) distinctly is likely to be more successful than a model that does not possess such semantic structuring. However, since they rely on shallow hand crafted feature representation, they are limited by their representational abilities to model a complex learning task. On the other hand, deep representations overcame this limitation and yielded state of the art results in several computer vision benchmarks [35] [15] [19].

A key cue for group activity recognition is the spatio-temporal relations among the people in the scene. Determining where individual people are in a scene, analyzing their image appearance, and aggregating these features and their relations can discern which group activity is present. A volume of research has explored models for this type of reasoning [4, 23, 29, 2]. However, these approaches have focused on probabilistic or discriminative models built upon hand-crafted features. Since they rely on shallow hand crafted feature representations, they are limited by their representational abilities to model a complex learning task.

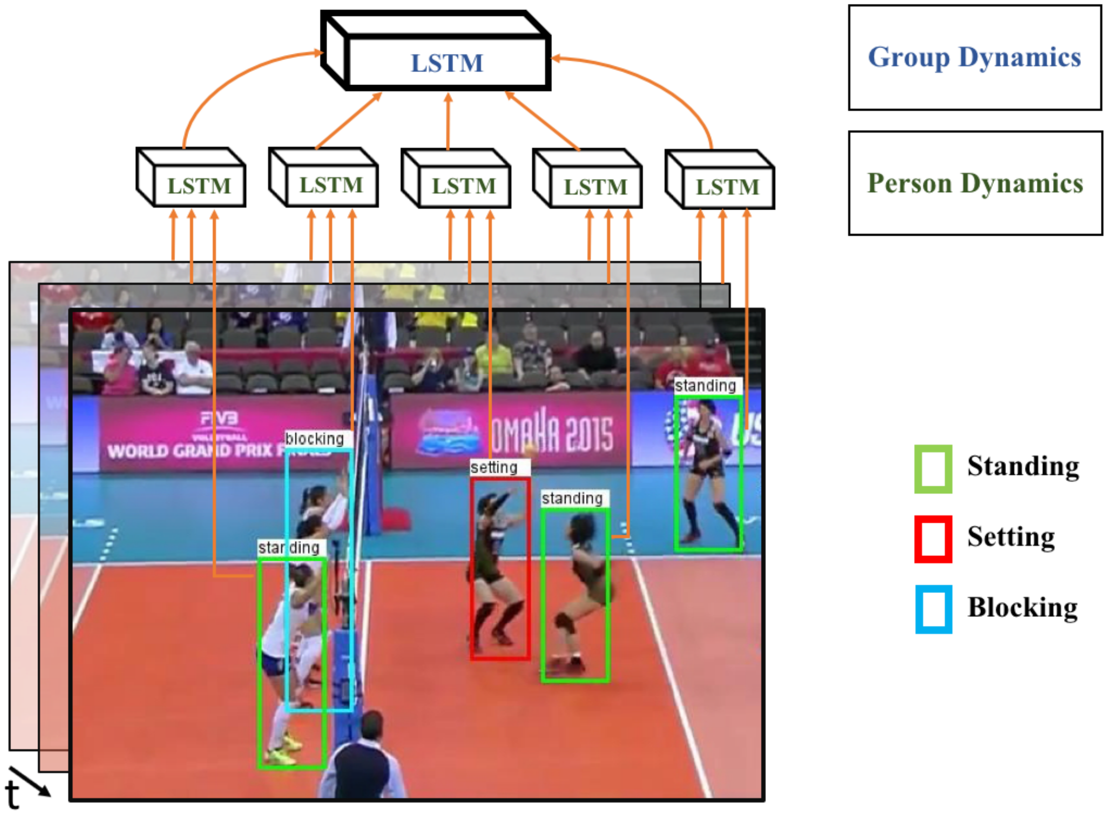


Figure 1.1: Group activity recognition via a hierarchical model. Each person in a scene is modeled using a temporal model that captures his/her dynamics, these models are integrated into a higher-level model that captures scene-level activity.

A naive approach to group activity recognition with a deep model would be to simply treat an image as an holistic input. One could train a model to classify this image according to the group activity taking place. However, it isn't clear if this will work given the redundancy in the training data: with a dataset of volleyball videos, frame will be dominated by features of volleyball courts. The differences between the different classes of group activities are about spatio-temporal relations between people, beyond just global appearance. Forcing a deep model to learn invariance to translation, to focus on the relations between people presents a significant challenge to the learning algorithm. Similar challenges exist in the object recognition literature, and research often focuses on designing pooling operators for deep networks (e.g. [39]) that enable the network to learn effective classifiers.

Group activity recognition presents a similar challenge – appropriate networks need to be designed that allow the learning algorithm to focus on differentiating higher-level classes of activities. Hence, we develop a novel hierarchical deep spatio-temporal model that reasons over individual people. Given a set of detected and tracked people, we run temporal deep networks (LSTMs) to analyze each individual person. These LSTMs are aggregated over the people in a scene into a higher level deep temporal model. This allows the deep model to learn the relations between the people (and their appearances) that contribute to recognizing a particular group activity.

The main contribution of this work is the proposal of a novel deep architecture that models group activities in a principled structured spatio-temporal framework. In this work, we assume that we have a single group activity happening in a given video. Our 2-stage approach models individual person activities in its first stage, and then combines person level information to represent group activities. The model's temporal representation is based on the long short-term memory (LSTM): recurrent neural networks such as these have recently demonstrated successful results in sequential tasks such as image captioning [9] and speech recognition [10]. Through the model structure, we aim at constructing a representation that leverages the discriminative information in the hierarchical structure between individual person actions and group activities. The model can be used in general group activity applications such as video surveillance, sport analytics, video search and retrieval.

To cater the needs of our problem, we also propose a new volleyball dataset that offers person detections, and both the person action label, as well as the group activity label. The camera view of the selected sports videos allows us to track the players in the scene. Experimentally, the model is effective in recognizing the overall team activity based on recognizing and integrating player actions.

Chapter 2

Related Work

Human activity recognition is an active area of research, with many existing algorithms. One can refer to the surveys [43], and [28] for exploring the vast literature in activity recognition. Here, we will focus on the group activity recognition problem and recent related advances in deep learning.

2.1 Group Activity Recognition

Group activity recognition has attracted a large body of work recently. Most previous work has used hand-crafted features fed to structured models that represent information between individuals in space and/or time domains.

Lan et al. [24] proposed an adaptive latent structure learning that represents hierarchical relationships ranging from lower person-level information to higher group-level interactions.

Lan et al. [22] and Ramanathan et al. [29] introduced the idea of social roles, where the expected behaviour of an individual person in the context of group, is exploited in fully supervised and weakly supervised frameworks respectively.

In Lan et al. [22], the model maps the individual features to group activity by constructing a hierarchical map consisting of individual action, role based unary components, through pairwise roles, through scene level group activities. The interactions and unary roles/activities are represented using an undirected graphical model. The parameters of this model are learnt using structured SVM formulation that operates on max margin framework, and operates under completely supervised settings. This model is shown in Figure 2.1.

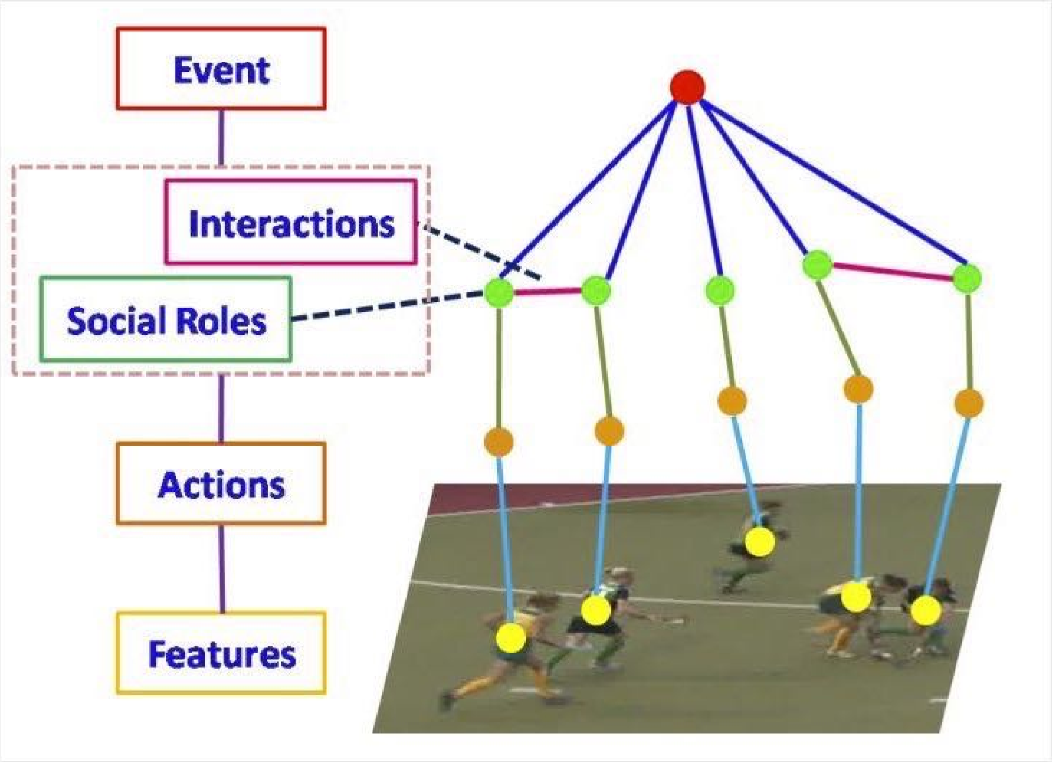


Figure 2.1: ©2012 IEEE. Overview of Lan et al. [22] social roles model

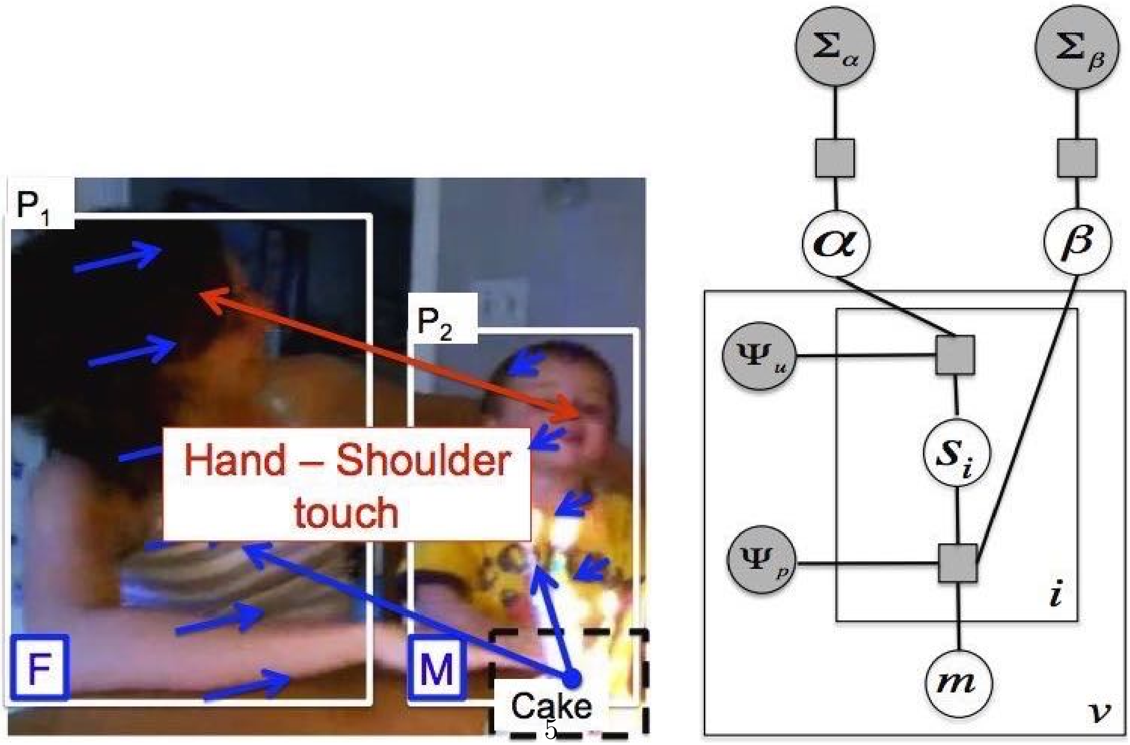


Figure 2.2: ©2013 IEEE. Overview of Ramanathan et al. [29] social roles model

In Ramanathan et al. [29], CRF based social role model is deployed under weakly supervised setting. To learn model parameters and role labels, a joint variational inference procedure is adapted. HOG3D [17], spatio-temporal feature [42], object interaction feature [25], and social role feature [46] are used as unary component representations. A subsequent layer consisting of pairwise spatio-temporal interaction features is used to refine the noisy unary component features. Finally, variational inferencing is used to learn the unknown role labels and model parameters. This model is shown in Figure 2.2.

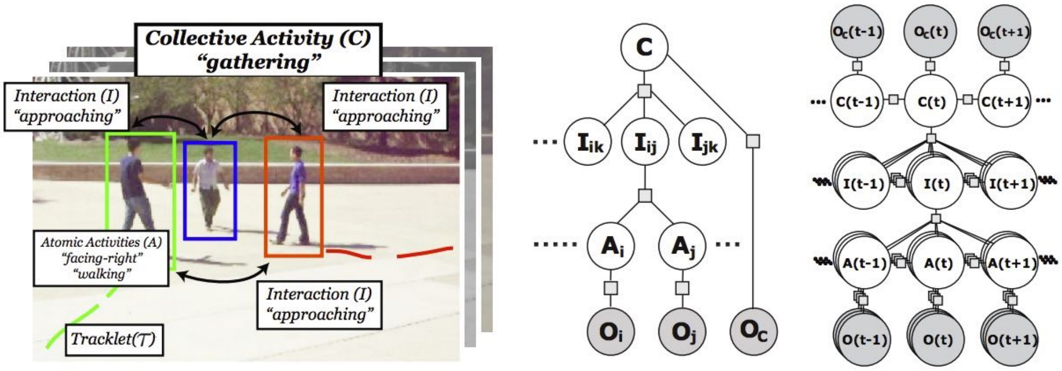


Figure 2.3: Overview of Choi et al. [3] model

Choi and Savarese [3] have unified tracking multiple people, recognizing individual actions, interactions and collective activities in a joint framework. The model is based on the premise that strong correlation exists between an individual's activity, and the activities of the other nearby people. Following this intuition, they come up with a hierarchical structure of activity types that maps the individual activity to overall group activity. In this process, they simultaneously track atomic activities, interactions and overall group activities. The parameters of this model (and the inferencing) are learnt by combining belief propogation with branch/bound algorithm. Figure 2.3 summarizes this model.

Choi et al. [5] adopted a random forest structure to sample discriminative spatio-temporal regions from input video that were fed to 3D Markov random field to localize collective activities in a scene. Shu et al. [33] detect group activities from aerial video using an AND-OR graph formalism. The above-mentioned methods use shallow hand crafted features, and typically adapt a linear model that suffers from representational limitations.

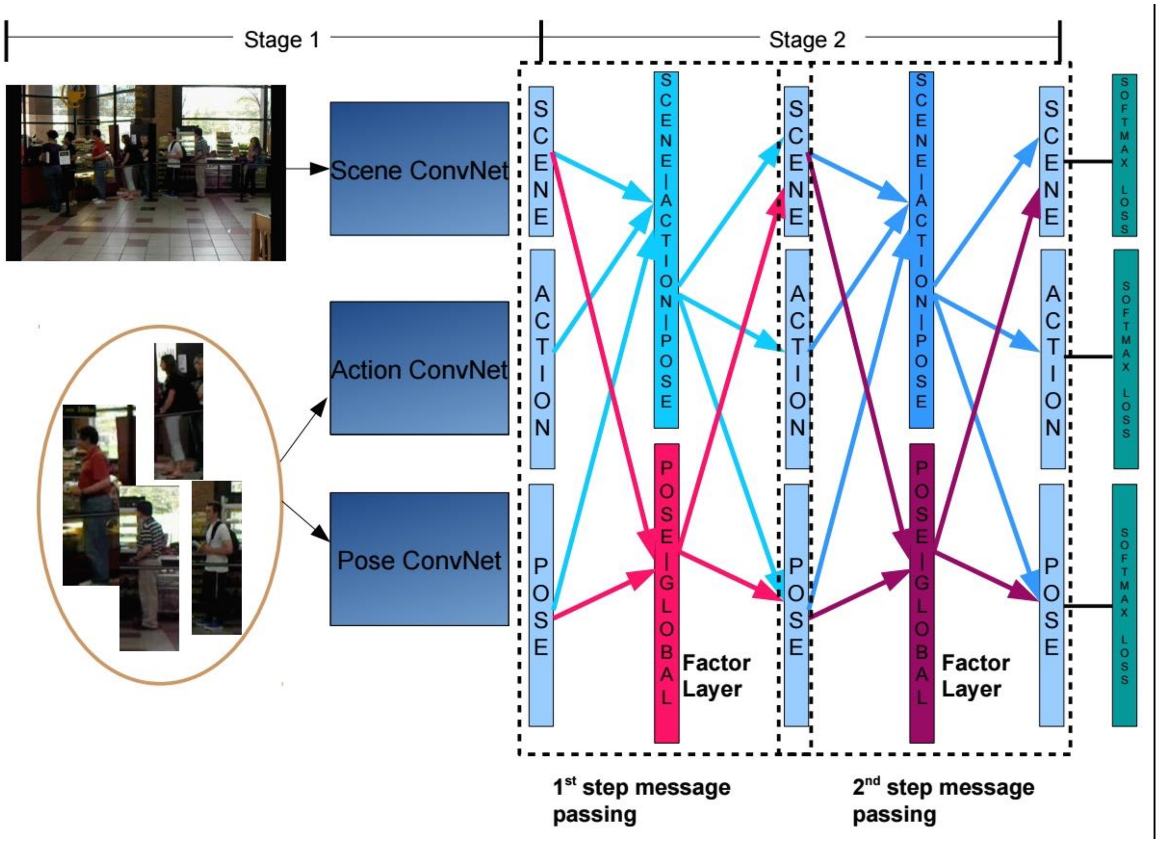


Figure 2.4: Overview of Deng et al. [7] deep structured model

In Deng et al. [7] a similar framework is used for group activity recognition, where a neural network-based hierarchical graphical model refines the person action labels, and learns to predict the group activity simultaneously. In this model, factor neurons are employed, that perform the message passing operation to learn the relationship between different aspects of data such as pose, action and scene. The model consists of individual pose/action unary components, in addition to an overall scene based unary component. The unary components are represented as class probabilities obtained by passing the input (individual person / overall scene appropriately) to alexnet. This is further refined by a couple of bifactor layers that help in refining information obtained from unary component. The final model comprises of blocks of unary component layers and bifactor layers. A schematic representation of this model is shown in Figure 2.4.

2.2 Sport Video Analysis

Several previous work have extended group activity recognition to team activity recognition in sport footage.

Siddiquie et al. [34] proposed sparse multiple kernel learning to select features incorporated in spatio-temporal pyramid. In this work, quantized hand crafted features are used for pyramid kernel based classification.

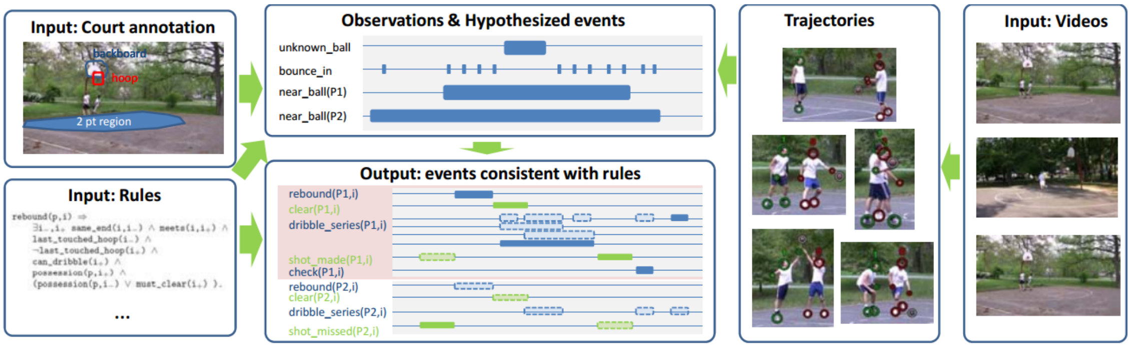


Figure 2.5: ©2011 IEEE. Overview of [26]

Morariu et al. [26] track players, infer part locations, and reason about temporal structure in 1-on-1 basketball games. In this method, trajectories are first extracted to generate video observations, which are then used to construct event hypothesis. Finally, an event label is assigned to a test video if its observations correlate strongest to the hypothesis corresponding to the label. The model is shown in 2.5.

Swears et al. [38] used Granger Causality statistic to automatically constrain the temporal links of a Dynamic Bayesian Network (DBN) for handball videos. Direkoglu and O'Connor [8] solved a particular Poisson equation to generate a holistic player location representation. Kwak et al. [21] optimize based on a rule-based depiction of interactions between people.

2.3 Deep Learning

Deep Convolutional Neural Networks (CNNs) have shown successful performance mainly due to unifying feature and classifier learning and the availability of large labeled datasets. Successes have been demonstrated on a variety of computer vision tasks including image classification [19, 36] and action recognition [35, 15].

More flexible recurrent neural network (RNN) based models are used for handling variable length space-time inputs. Specifically, LSTM [12] models are popular among RNN models due to the tractable learning framework that they offer when it comes to deep representations. These LSTM models have been applied in a variety of tasks [9, 10, 27, 41].

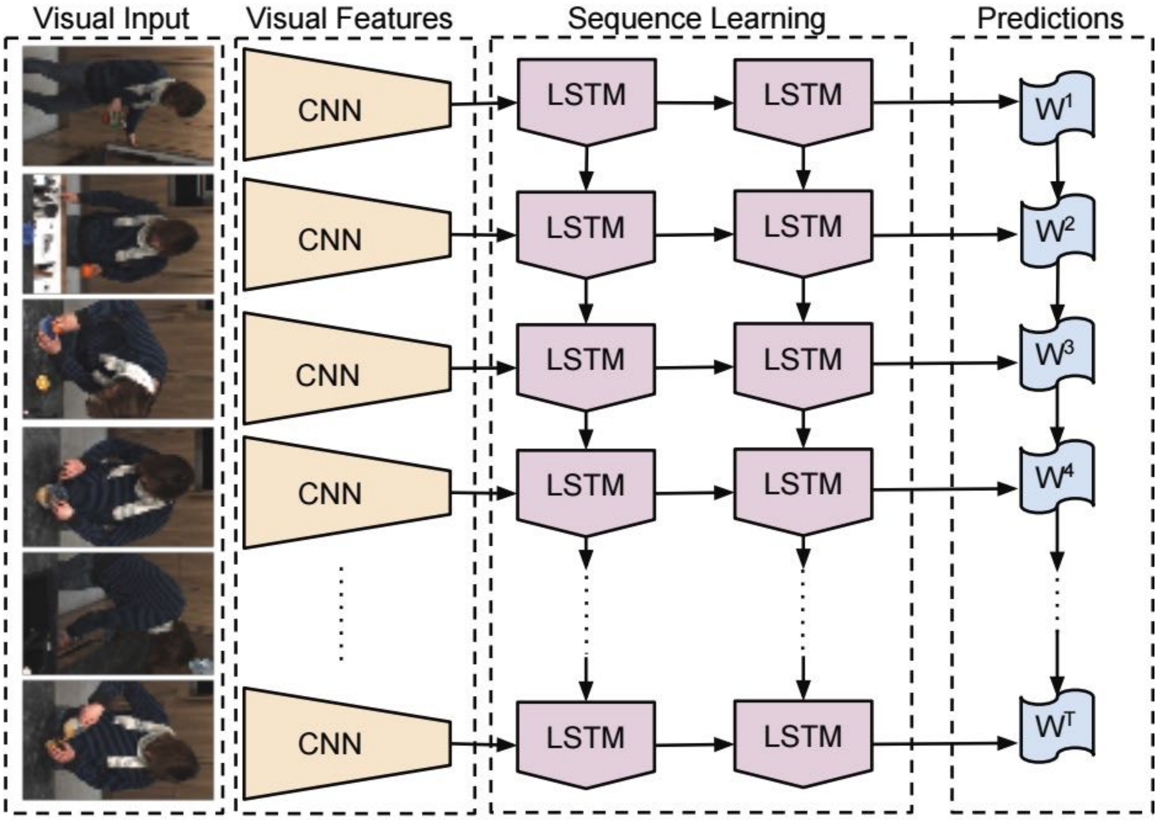


Figure 2.6: ©2015 IEEE. Overview of Video captioning network [9].

For instance, in Donahue et al. [9], the so-called Long term Recurrent Convolutional network formed by stacking LSTM on top of pre-trained CNNs, is proposed for handling sequential tasks such as activity recognition, image description, and video description. In this work, they showed that it is possible to jointly train LSTMs along with convnets and acheive comparable results to the state of the art time varying tasks. For video captioning, they first construct a semantic representation of the video using maximum aposteriori estimate of a CRF which is then used to construct a natural sentence using LSTMs, as against the statistical machine translation used in Koehn et al. [18] for constructing natural sentences. A schematic view of this model is shown in Figure 2.6.

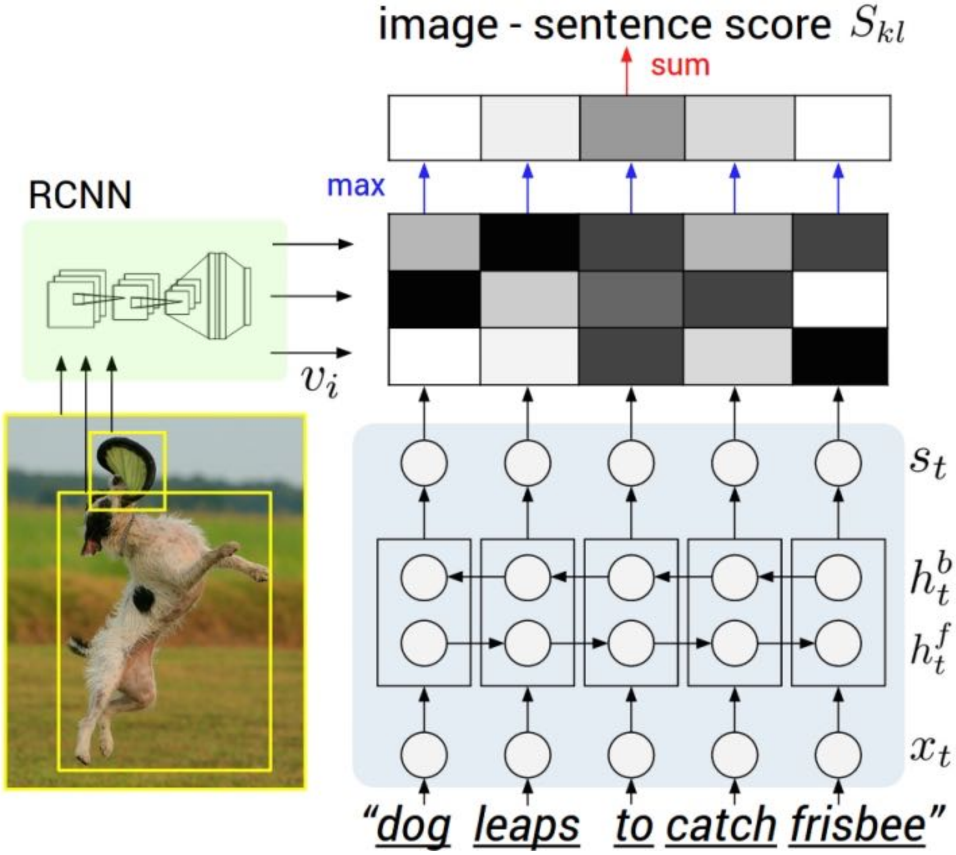


Figure 2.7: ©2014 IEEE. Overview of visual-linguistic alignment stage in karpathy et al. [14]

In Karpathy et al. [14], structured objectives are used to align CNNs over image regions and bi-directional RNNs over sentences and a deep multi-modal RNN architecture is used for generating image description using the deduced alignments. In the first stage, words and image regions are embedded onto an alignment space. Image regions are represented by RCNN embeddings, and words are represented using bi-directional recurrent neural network [31] embeddings. Figure 2.7 illustrates the first stage of this model. In the second stage, using the image regions and textual snippets, or full image and sentence descriptions, a generative model based on RNN is constructed, that outputs probability map of the next word.

In this work, we aim at building a hierarchical structured model that incorporates a deep LSTM framework to recognize individual actions and group activities. Previous work in the area of deep structured learning include Tompson et al. [40] for pose estimation, Zheng et al. [45] and Schwing et al. [32] for semantic image segmentation. In Deng et al. [7] a similar framework is used for group activity recognition, where a neural network-based hierarchical graphical model refines the person action labels, and learns to predict the group activity simultaneously. While these methods use neural network-based graphical representations, in our current approach, we leverage LSTM-based temporal modelling to learn discriminative information from time varying sports activity data. Yueng et al. [44] introduced a new dataset that contains dense multiple labels per frame for underlying action, and a novel Multi-LSTM is used to model the temporal relations between labels present in the dataset.

2.4 Datasets

Popular datasets for activity recognition include the Sports-1M dataset [14], UCF 101 database [37], and the HMDB movie database [20]. These datasets started to shift the focus to unconstrained Internet videos that contain more intra-class variation, compared to a constrained dataset. While these datasets continue to focus on individual human actions, in our work, we focus on recognizing more complex group activities in sport videos. Choi et al. [4] introduced the Collective Activity Dataset consisting of real world pedestrian sequences where the task is to find the high level group activities. In this work, we experiment with this dataset, but also introduce a new dataset for group activity recognition in sport footage which is annotated with player pose, location, and group activities to encourage similar research in the sport domain.

Chapter 3

Proposed Approach

Our goal in this work is to recognize activities performed by a group of people in a video sequence. The input to our method is a set of tracklets of the people in a scene. The group of people in the scene could range from players in a sports video to pedestrians in a surveillance video. Generally speaking, there are three elements that can be considered to determine what a group of people is doing:

- **Person-level actions** collectively define a group activity. Person action recognition is a first step toward recognizing group activities.

- **Temporal dynamics of a person's action** is higher-order information that can serve as a strong signal for group activity. Knowing how each person's action is changing over time can be used to infer the group's activity.

- **Temporal evolution of group activity** represents how a group's activity is evolving over time. For example, in a volleyball game a team may move from defence phase to pass and then attack.

Many classic approaches to the group activity recognition problem have modeled these elements in a form of structured prediction based on hand crafted features [42, 30, 24, 22, 29]. Inspired by the success of deep learning based solutions, in this work, a novel hierarchical deep learning based model is proposed that is potentially capable of learning low-level image features, person-level actions, their spatio-temporal relations, and temporal group dynamics in a unified end-to-end framework.

Given the sequential nature of group activity analysis, our proposed model is based on a Recurrent Neural Network (RNN) architecture. RNNs consist of non-linear units with internal states that can learn dynamic temporal behavior from a sequential input with arbitrary length. Therefore, they overcome the limitation of CNNs that expect constant length input. This makes them widely applicable to video analysis tasks such as activity recognition.

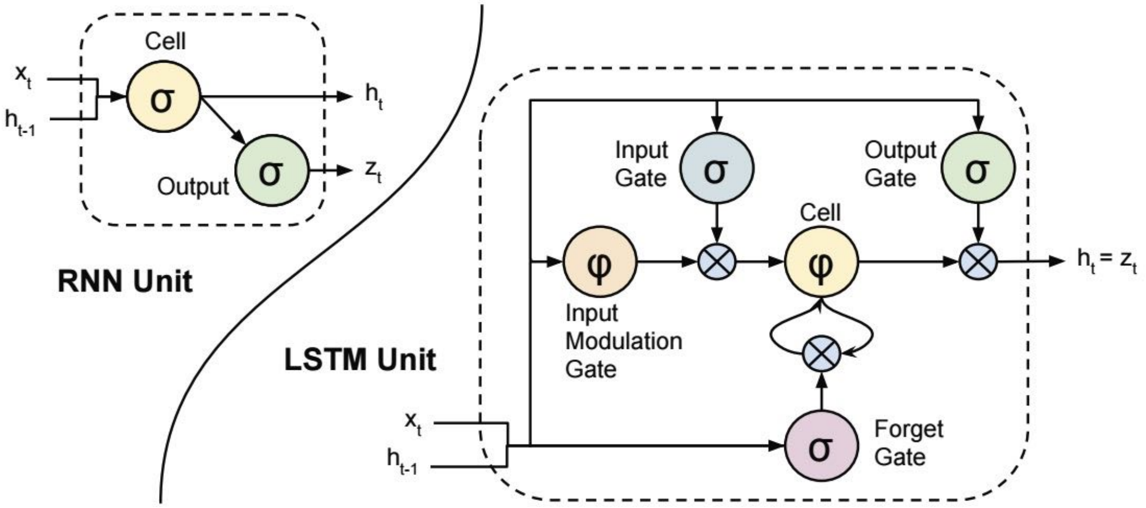


Figure 3.1: ©2015 IEEE. Schematic representation of an LSTM (Figure borrowed from Donahue et al. [9]

Our model is inspired by the success of hierarchical models. Here, we aim to mimic a similar intuition using recurrent networks. We propose a deep model by stacking several layers of RNN-type structures illustrated in Figure 3.1 to model a large range of low-level to high-level dynamics defined on top of people and entire groups. We describe the use of these RNN structures for individual and group activity recognition next.

3.1 Temporal Model of Individual Action

Given tracklets of each person in a scene, we use long short-term memory (LSTM) models to represent temporally the action of each individual person. Such temporal information is complementary to spatial features and is critical for performance. LSTMs, originally proposed by Hochreiter and Schmidhuber [12], have been used successfully for many sequential problems in computer vision. Each LSTM unit consists of several cells with memory that stores information for a short temporal interval. The memory content of a LSTM makes it suitable for modeling complex temporal relationships that may span a long range.

The content of the memory cell is regulated by several gating units that control the flow of information in and out of the cells. The control they offer also helps in avoiding spurious gradient updates that can typically happen in training RNNs when the length of a temporal input is large. This property enables us to stack a large number of such layers in order to learn complex dynamics present in the input in different ranges.

We use a deep Convolutional Neural Network (CNN) to extract features from the bounding box around the person in each time step on a person trajectory. The output of the CNN, represented by x_t, can be considered as a complex image-based feature describing the spatial region around a person. Assuming x_t as the input of an LSTM cell at time t, the cell activition can be formulated as :

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \tag{3.1}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \tag{3.2}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \tag{3.3}$$

$$g_t = \phi(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{3.4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{3.5}$$

$$h_t = o_t \odot \phi(c_t) \tag{3.6}$$

Here, σ stands for a sigmoid function, and ϕ stands for the tanh function. x_t is the input, $h_t \in R^N$ is the hidden state with N hidden units, $c_t \in R^N$ is the memory cell, $i_t \in R^N$, $f_t \in R^N$, $o_t \in R^N$, and, $g_t \in R^N$ are input gate, forget gate, output gate, and input modulation gate at time t respectively. $\odot$ represents element-wise multiplication.

When modeling individual actions, the hidden state h_t could be used to model the action a person is performing at time t. Note that the cell output is evolving over time based on the past memory content. Due to the deployment of gates on the information flow, the hidden state will be formed based on a short-range memory of the person's past behaviour. Therefore, we can simply pass the output of the LSTM cell at each time to a softmax classification layer[1] to predict individual person-level action for each tracklet.

The LSTM layer on top of person trajectories forms the first stage of our hierarchical model. This stage is designed to model **person-level actions and their temporal evolution**. Our training proceeds in a stage-wise fashion, first training to predict person level actions, and then pasing the hidden states of the LSTM layer to the second stage for group activity recognition, as discussed in the next section.

3.2 Hierarchical Model for Group Activity Recognition

At each time step, the memory content of the first LSTM layer contains discriminative information describing the subject's action as well as past changes in his action. If the memory content is correctly collected over all people in the scene, it can be used to describe the group activity in the whole scene.

Moreover, it can also be observed that direct image-based features extracted from the spatial domain around a person carries a discriminative signal for the ongoing activity. Therefore, a deep CNN model is used to extract complex features for each person in addition to the temporal features captured by the first LSTM layer.

At this moment, the concatenation of the CNN features and the LSTM layer represent spatio-temporal features for a person. Various pooling strategies can be used to aggregate

[1] More precisely, a fully connected layer fed to softmax loss layer.

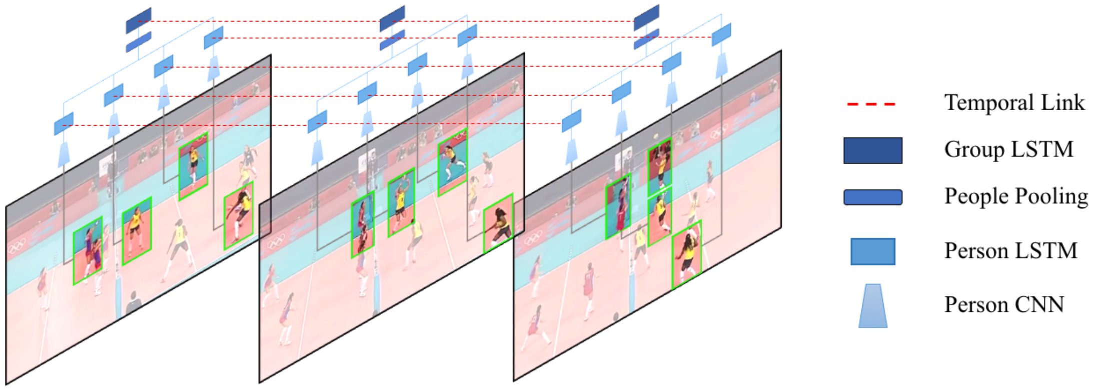


Figure 3.2: Our two-stage model for a volleyball match. Given tracklets of K-players, we feed each tracklet in a CNN, followed by a person LSTM layer to represent each player's action. We then pool over all people's spatio-temporal features in the scene. The output of the pooling layer is feed to the second LSTM network to identify the whole teams activity.

these features over all people in the scene at each time step. The output of the pooling layer forms our representation for the group activity. The second LSTM network, working on top of the spatio-temporal representation, is used to directly model the **temporal dynamics of group activity**. The LSTM layer of the second network is directly connected to a classification layer in order to detect group activity classes in a video sequence.

Mathematically, the pooling layer can be expressed as the following:

$$P_{tk} = x_{tk} \oplus h_{tk} \tag{3.7}$$

$$Z_t = P_{t1} \diamond P_{t2} \ldots \diamond P_{tk} \tag{3.8}$$

In this equation, h_{tk} corresponds to the first stage LSTM output, and x_{tk} corresponds to the AlexNet fc7 feature, both obtained for the k$^{\text{th}}$ person at time t. We concatenate these two features (represented by $\oplus$) to obtain the spatio-temporal feature representation P_{tk} for k$^{\text{th}}$ person. We then construct the frame level feature representation Z_t at time t by applying a max pooling operation (represented by $\diamond$) over the features of all the people. Finally, we feed the frame level representation to our second LSTM stage that operates similar to the person level LSTMs that we described in the previous section, and learn the group level dynamics. Z_t, passed through a fully connected layer, is given to the input of the second-stage LSTM layer. The hidden state of the LSTM layer represented by h_t^{group} carries temporal information for the whole group dynamics. h_t^{group} is fed to a softmax classification layer to predict group activities.

3.3 Implementation Details

We trained our model in two steps. In the first step, the person-level CNN and the first LSTM layer are trained in an end-to-end fashion using a set of training data consisting of person tracklets annotated with action labels. We implement our model using Caffe [13]. Similar to other approaches [9, 7, 41], we initialize our CNN model with the pre-trained AlexNet network and we fine-tune the whole network for the first LSTM layer. 9 timesteps and 3000 hidden nodes are used for the first LSTM layer and a softmax layer is deployed for the classification layer in this stage.

After training the first LSTM layer, we concatenate the fc7 layer of AlexNet and the LSTM layer for every person and pool over all people in a scene. The pooled features, which correspond to frame level features, are fed to the second LSTM network. This network consists of a 3000-node fully connected layer followed by a 9-timestep 500-node LSTM layer which is passed to a softmax layer trained to recognize group activity labels.

For training all our models (that include both the baseline models and both the stages of the two-stage model), we follow the same training protocol. We use a fixed learning rate of 0.00001 and a momentum of 0.9. For tracking subjects in a scene, we used the tracker by Danelljan et al. [6], implemented in the Dlib library [16].

Chapter 4

Experiments

In this section, we evaluate our model by comparing our results with several baselines and previously published works on the Collective Activity Dataset [4] and our new volleyball dataset. First, we describe our baseline models. Then, we present our results on the Collective Activity Dataset followed by experiments on the volleyball dataset.

4.1 Baselines

The following five baselines are considered in all our experiments:

1. **Image Classification:** This baseline is the basic AlexNet model fine-tuned for group activity recognition in a single frame.

2. **Person Classification:** In this baseline, the AlexNet CNN model is deployed on each person, fc7 is pooled over all people, and is fed to a softmax classifier to recognize group activities in each single frame.

3. **Fine-tuned Person Classification:** This baseline is similar to the previous baseline with one distinction. The AlexNet model on each player is fine-tuned to recognize person-level actions. Then, fc7 is pooled over all players to recognize group activities in a scene without any fine-tuning of the AlexNet model. The rational behind this baseline is to examine a scenario where person-level action annotations as well as group activity annotations are used in a deep learning model that does not model the temporal aspect of group activities. This is very similar to our two-stage model without the temporal modeling.

4. **Temporal Model with Image Features:** This baseline which is the temporal extension of the first baseline, examines the idea of feeding image level features directly to a LSTM model to recognize group activities. In this baseline, AlexNet model is deployed on the whole image, and, fc7 is fed to a LSTM model. This baseline can be considered as reimplementation of Donahue et al.'s work [9] for activity recognition.

5. **Temporal Model with Person Features:** This baseline which is the temporal extension of the second baseline, feeds fc7 pooled over all people to a LSTM model to recognize group activities.

4.2 Experiments on the Collective Activity Dataset

The Collective Activity Dataset [4] has been widely used for evaluating group activity recognition approaches in the computer vision literature [2, 7, 1]. This dataset consists of 44 videos, eight person-level pose labels (not in use), five person level action labels, and five group-level activities. A scene is simply assigned with what the majority of people are doing. We follow the train/test split provided by [11]. In this section, we present our results on this dataset.

Method	Accuracy
B1-Image Classification	63.0
B2-Person Classification	61.8
B3-Fine-tuned Person Classification	66.3
B4-Temporal Model with Image Features	64.2
B5-Temporal Model with Person Features	62.2
Our Two-stage Hierarchical Model	**81.5**

Table 4.1: Comparison of our method with baseline methods on the Collective Activity Dataset.

Method	Accuracy
Contextual Model [24]	79.1
Deep Structured Model [7]	80.6
Our Two-stage Hierarchical Model	81.5
Cardinality kernel [11]	**83.4**

Table 4.2: Comparison of our method with previously published works on the collective activity dataset.

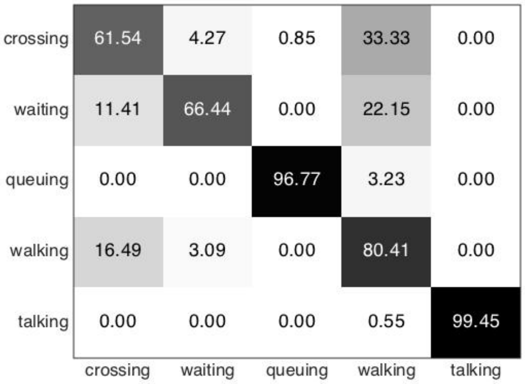
In Table 4.1, the classification results of our proposed architectures is compared with the baselines. As shown in the table, our two-stage LSTM model significantly outperforms the baseline models. An interesting comparison can be made between temporal and frame-based counterparts including B1 vs. B4, B2 vs. B5 and B3 vs our two-sage model. It is interesting to observe that adding temporal information using LSTM model improves the performance of these baselines.

Table 4.2 contains the comparison between our model and other top performing group activity recognition models. As shown in the table, our model fares better than most of the previous methods, and is less than 2% behind the state of the art cardinality kernel [11] based approach, that assumes that group activity representation specific to the dataset (i.e., group activity is defined as what majority are performing in the scene).

4.2.1 Discussion

The confusion matrix obtained for collective activity dataset using our two-stage model in shown in Figure 4.1. From the figure, we could observe that the model performs almost perfectly in case of talking and queuing class, and gets confused between crossing, waiting, and walking. We observe such a behaviour perhaps due to lack of consideration of spatial relations between people in the group, which is shown to boost the performance of previous group activity recognition methods. e.g. crossing involves walking action, but is confined in a path, which people perform in orderly fashion. Therefore, our model that is designed only to learn the dynamic properties of group activities often gets confused with the walking action.

It is clear that our two-stage model has robust performance with compare to baselines. From one side, the temporal information helps much as refereed before. On the other side, finding and describing the elements of a video (e.g. persons) seems reveal clearer information than just utilizing frame level.



	crossing	waiting	queuing	walking	talking
crossing	61.54	4.27	0.85	33.33	0.00
waiting	11.41	66.44	0.00	22.15	0.00
queuing	0.00	0.00	96.77	3.23	0.00
walking	16.49	3.09	0.00	80.41	0.00
talking	0.00	0.00	0.00	0.55	99.45

Figure 4.1: **Confusion matrix for the Collective Activity Dataset obtained using our two-stage model.**

4.3 Experiments on the Volleyball Dataset

In order to evaluate the performance of our model for team activity recognition on sport footage, we collected a new dataset based on publicly available YouTube volleyball videos. We annotated 1525 frames collected from 15 videos with seven player action labels, and six team activity labels. The list of action and activity labels and related statistics are tabulated in Tables 4.3 and 4.4.

Group Activity Class	No. of Instances
Right set	229
Right spike	187
Right pass	267
Left pass	304
Left spike	246
Left set	223

Table 4.3: Statistics of the group activity labels in the volleyball dataset.

Action Classes	Average No. of Instance per Frame
Waiting	0.30
Setting	0.33
Digging	0.57
Falling	0.21
Spiking	0.28
Blocking	0.58
Others	9.22

Table 4.4: Statistics of the action labels in the volleyball dataset.

From the tables, we observe that the group activity labels are relatively more balanced, compared to the player action labels. This feature follows from the fact that we have more people often present in static actions like standing pose compared to dynamic actions (setting, spiking, etc.). Therefore, our dataset presents a challenging team activity recognition task, where we have the interesting actions that solely determine the group activity occur rarely in our dataset. The dataset will be made publicly available to facilitate direct comparisons.

In Table 4.5, the classification performance of our proposed model is compared against the baselines. Similar to the performance in the collective activity dataset, our two-stage LSTM model outperforms the baseline models. However, compared to the baselines, performance gain using our model is quite modest unlike in the case of collective activity dataset. This follows directly from the fact that in general, we can infer group activity in volleyball by using just a few frames. Therefore, in case of volleyball dataset, our baseline B1 is closer to the actual model's performance, compared to the collective activity dataset. Moreover, explicitly modeling people is necessary for obtaining better performance in this dataset, since the background is rapidly changing due to fast moving camera, and therefore it corrupts the temporal dynamics of the foreground. This could be verified from the performance of our baseline model B4, which is a temporal model, that does not consider people explicitly, shows inferior performance compared to the baseline B1, which is a non-temporal image classification style model. On the other hand, baseline model B5, which is a temporal model that explicitly considers people performs comparably to the image classification baseline, in spite of the problems that arise due to tracking and motion artifacts.

Method	Accuracy
B1-Image Classification	46.7
B2-Person Classification	33.1
B3-Fine-tuned Person Classification	35.2
B4-Temporal Model with Image Features	37.4
B5-Temporal Model with Person Features	45.9
Our Two-stage Hierarchical Model	**51.1**

Table 4.5: Comparison of the team activity recognition performance of baselines against our model evaluated on the volleyball dataset.

Analysis

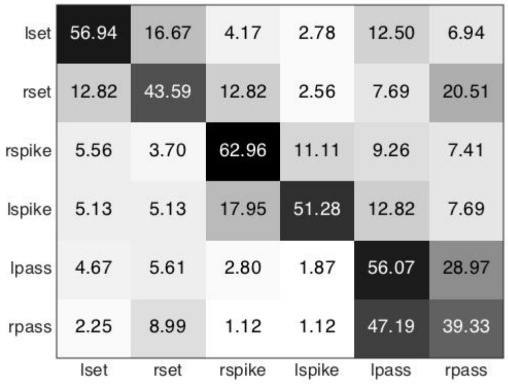


Figure 4.2: Confusion matrix for the volleyball dataset obtained using our two-stage hierarchical model.

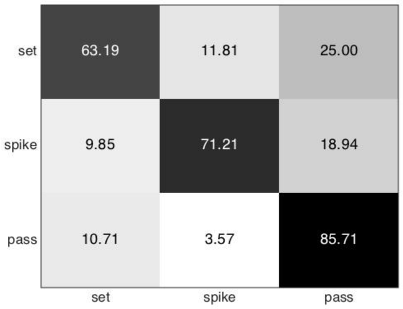
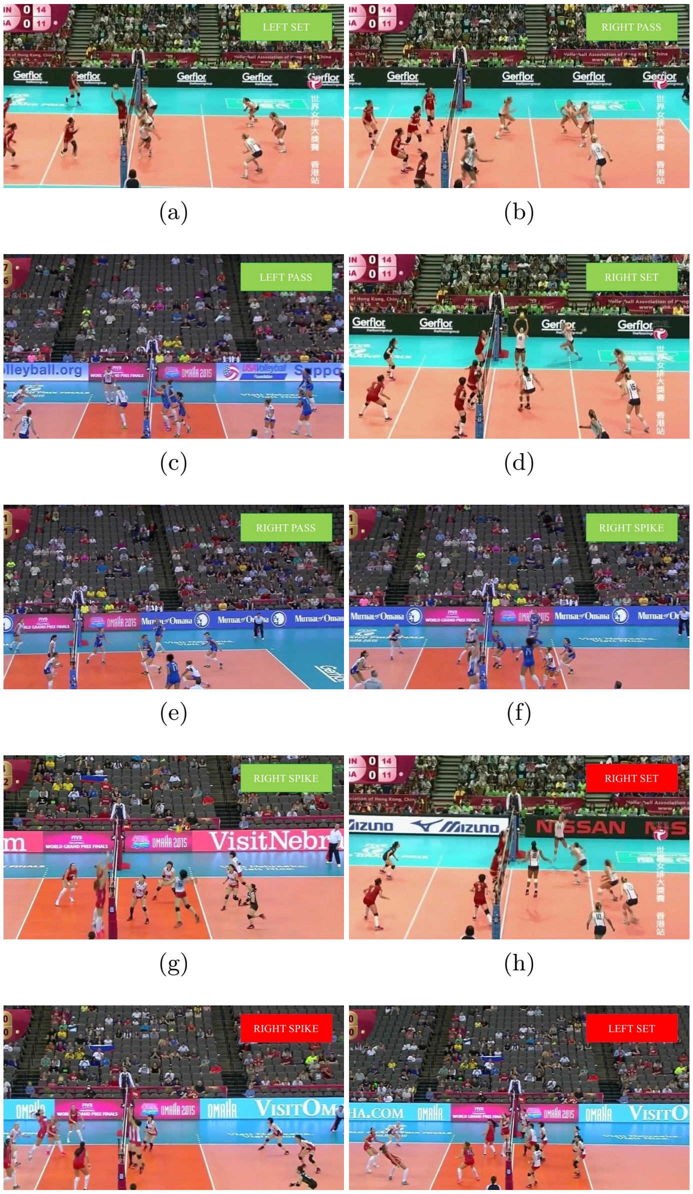


Figure 4.3: Confusion matrix for the volleyball dataset obtained using our two-stage hierarchical model without team considerations.

Figure 4.2 shows the confusion matrix obtained for the volleyball dataset using our two-stage model. From the confusion matrix, we observe that our model generates consistent high level action labels. Nevertheless, our model has some confusion between set and pass activities.as these activities typically may look similar. Further, it is evident from better numbers obtained with no team considerations as shown in confusion matrix in Figure 4.3. It means that our model gets confused on the side that is performing the action, and makes some mistakes as a result of it. Some examples of classification using our model is shown in Figure 4.4.



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Figure 4.4: Visualizations of the generated scene labels using our model. Green denotes correct classifications, red denotes incorrect. The incorrect ones correspond to the confusion between different actions in ambiguous cases (h and j examples), or in the left and right distinction (i example).

Chapter 5

Conclusion

In this work, we presented a novel deep structured architecture to deal with the group activity recognition problem. Through a two-stage process, we learn spatio-temporal representation of people actions and combine the representation of individual people to recognize the whole activity. We also created a new volleyball dataset to train and test our model, and also evaluated our model on the Collective Activity Dataset. Results show that our architecture can improve upon baseline methods lacking hierarchical consideration of individual and group activities using deep learning.

5.1 Limitations and Future Work

In this model, we considered a very rudimentary pooling techniques as we focused on the learning architecture. Specifically, these pooling operators limit the performance of the overall model. Therefore, we aim to incorporate other pooling operation that have limited loss of information and thereby improving the overall performance.

Yet another interesting direction is to apply a joint training approach where we train the entire model together under weakly supervised setting. This is a significant direction of research as it reduces the cost incurred due to human annotations substantially. We would also conduct a detailed analysis of the current model as to examine the role of each LSTM layer in order to learn more interesting properties.

A more sophisticated approach might involve incorporating ideas like [29] in deep learning framework so as to learn interesting roles and higher level activity simultaneously under deep weakly supervised settings, and apply it to tasks that include group activity recognition, event recognition etc.

Bibliography

[1] 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 *Computer Vision–ECCV 2012*, pages 187–200. Springer, 2012.

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

[3] Wongun Choi and Silvio Savarese. A unified framework for multi-target tracking and collective activity recognition. In *Computer Vision–ECCV 2012*, pages 215–230. Springer, 2012.

[4] Wongun Choi, Khuram Shahid, and Silvio Savarese. 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. IEEE, 2009.

[5] Wongun Choi, Khuram Shahid, and Silvio Savarese. Learning context for collective activity recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3273–3280. IEEE, 2011.

[6] Martin Danelljan, Gustav Häger, Fahad Khan, and Michael Felsberg. Accurate scale estimation for robust visual tracking. In *British Machine Vision Conference, Nottingham, September 1-5, 2014*. BMVA Press, 2014.

[7] Z. Deng, M. Zhai, L. Chen, Y. Liu, S. Muralidharan, M. Roshtkhari, , and G. Mori. Deep structured models for group activity recognition. In *British Machine Vision Conference (BMVC)*, 2015.

[8] Cem Direkoglu and Noel E O'Connor. Team activity recognition in sports. In *Computer Vision–ECCV 2012*, pages 69–83. Springer, 2012.

[9] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. *arXiv preprint arXiv:1411.4389*, 2014.

[10] Alex Graves and Navdeep Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1764–1772, 2014.

[11] Hossein Hajimirsadeghi, Wang Yan, Arash Vahdat, and Greg Mori. Visual recognition by counting instances: A multi-instance cardinality potential kernel. *arXiv preprint arXiv:1502.02063*, 2015.

[12] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[13] Y. Jia. Caffe: An open source convolutional architecture or fast feature embedding, 2013. http://caffe.berkeleyvision.org/.

[14] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *arXiv preprint arXiv:1412.2306*, 2014.

[15] Andrej Karpathy, George Toderici, Sachin Shetty, Tommy Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1725–1732. IEEE, 2014.

[16] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[17] Alexander Klaser, Marcin Marszałek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC 2008-19th British Machine Vision Conference*, pages 275–1. British Machine Vision Association, 2008.

[18] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics, 2007.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[20] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2556–2563. IEEE, 2011.

[21] Suha Kwak, Bohyung Han, and Joon Hee Han. Multi-agent event detection: Localization and role assignment. In *CVPR*, 2013.

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

[23] Tian Lan, Leonid Sigal, and Greg Mori. Social roles in hierarchical models for human activity recognition. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1354–1361. IEEE, 2012.

[24] 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*, 34(8):1549–1562, 2012.

[25] Li-Jia Li, Hao Su, Li Fei-Fei, and Eric P Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Advances in neural information processing systems*, pages 1378–1386, 2010.

[26] Vlad I. Morariu and Larry S. Davis. Multi-agent event recognition in structured scenarios. In *Computer Vision and Pattern Recognition (CVPR)*, 2011.

[27] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. *arXiv preprint arXiv:1503.08909*, 2015.

[28] Ronald Poppe. A survey on vision-based human action recognition. *Image and vision computing*, 28(6):976–990, 2010.

[29] Vignesh Ramanathan, Bangpeng Yao, and Li Fei-Fei. Social role discovery in human events. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2475–2482. IEEE, 2013.

[30] Christian Schüldt, Ivan Laptev, and Barbara Caputo. 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. IEEE, 2004.

[31] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *Signal Processing, IEEE Transactions on*, 45(11):2673–2681, 1997.

[32] Alexander G Schwing and Raquel Urtasun. Fully connected deep structured networks. *arXiv preprint arXiv:1503.02351*, 2015.

[33] Tianmin Shu, Dan Xie, Brandon Rothrock, Sinisa Todorovic, and Song-Chun Zhu. Joint inference of groups, events and human roles in aerial videos. In *CVPR*, 2015.

[34] Behjat Siddiquie, Yaser Yacoob, and Larry Davis. Recognizing plays in american football videos. Technical report, Technical report, University of Maryland, 2009.

[35] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576, 2014.

[36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[37] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[38] Eran Swears, Anthony Hoogs, Qiang Ji, and Kim Boyer. Complex activity recognition using granger constrained dbn (gcdbn) in sports and surveillance video. In *Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[39] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[40] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1799–1807. Curran Associates, Inc., 2014.

[41] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*, 2014.

[42] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3169–3176. IEEE, 2011.

[43] Daniel Weinland, Remi Ronfard, and Edmond Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding*, 115(2):224–241, 2011.

[44] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *arXiv preprint arXiv:1507.05738*, 2015.

[45] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip H. S. Torr. Conditional random fields as recurrent neural networks. In *International Conference on Computer Vision (ICCV)*, 2015.

[46] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE, 2012.