A Hierarchical Deep Temporal Model for Group Activity Recognition

Mostafa S. Ibrahim*, Srikanth Muralidharan*, Zhiwei Deng, Arash Vahdat, Greg Mori
School of Computing Science, Simon Fraser University, Burnaby, Canada
{msibrahi, smuralid, zhiweid, avahdat}@sfu.ca, mori@cs.sfu.ca

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 person-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 compared to baseline methods.

1. Introduction

What are the people in Figure 1 doing? This question can be answered at numerous levels of detail – in this paper 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.

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, 21, 27, 1]. 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. Deep representations have overcome this limitation and yielded state of the art results in several computer vision benchmarks [18, 33, 16].

*Equal Contribution

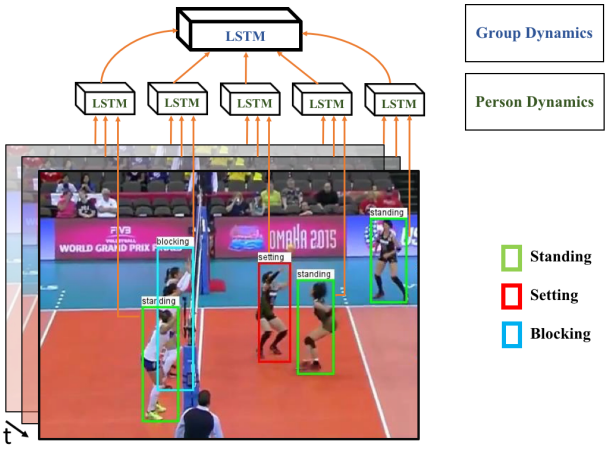

Figure 1: Group activity recognition via a hierarchical model. Each person in a scene is modeled using a temporal model that captures his/her dynamics, these models are integrated into a higher-level model that captures scene-level activity.

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, frames 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. [36]) 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 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 paper is the proposal of a novel deep architecture that models group activities in a principled structured temporal framework. 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, and 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.

This paper is organized as follows. In Section 2, we provide a brief overview of the literature related to activity recognition. In Section 3, we elaborate details of the proposed group activity recognition model. In Section 4, we tabulate the performance of approach, and end in Section 5 with a conclusion of this work.

2. Related Work

Human activity recognition is an active area of research, with many existing algorithms. Surveys by Weinland et al. [40] and Poppe [26] explore the vast literature in activity recognition. Here, we will focus on the group activity recognition problem and recent related advances in deep learning.

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. [23] 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. [27] explore the idea of social roles, the expected behaviour of an individual person in the context of group, in fully supervised and weakly supervised frameworks respectively. Choi and Savarese [3] have unified tracking multiple people, recognizing individual actions, interactions and collective activities in a joint framework. In other work [5], a random forest structure is used to sample discriminative spatio-temporal regions from input video fed to 3D Markov random field to localize collective activities in a scene. Shu et al. [30] detect group activities from aerial video using an AND-OR graph formalism. The above-mentioned methods use shallow hand crafted features, and typically adopt a linear model that suffers from representational limitations.

Sport Video Analysis: Previous work has extended group activity recognition to team activity recognition in sport footage. Seminal work in this vein includes Intille and Bobick [13], who examined stochastic representations of American football plays. Siddiquie et al. [31] proposed sparse multiple kernel learning to select features incorporated in a spatio-temporal pyramid. Morariu et al. [24] track players, infer part locations, and reason about temporal structure in 1-on-1 basketball games. Swears et al. [35] used the 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. [20] optimize based on a rule-based depiction of interactions between people.

Deep Learning: Deep Convolutional Neural Networks (CNNs) have shown impressive performance by 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 [18, 33] and action recognition [32, 16]. 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 to a variety of tasks [9, 10, 25, 38]. For instance, in Donahue et al. [9], the so-called Long term Recurrent Convolutional network, formed by stacking an LSTM on top of pre-trained CNNs, is proposed for handling sequential tasks such as activity recognition, image description, and video description. In Karpathy et al. [15], structured objectives are used to align CNNs over image regions and bi-directional RNNs over sentences. A deep multimodal RNN architecture is used for generating image descriptions using the deduced alignments.

In this work, we aim at building a hierarchical struc-

tured model that incorporates a deep LSTM framework to recognize individual actions and group activities. Previous work in the area of deep structured learning includes Tompson et al. [37] for pose estimation, and Zheng et al. [42] and Schwing et al. [29] 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 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. In [41], a new dataset is introduced 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.

Datasets: Popular datasets for activity recognition include the Sports-1M dataset [15], UCF 101 database [34], and the HMDB movie database [19]. 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 activity. In this paper, 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.

3. Proposed Approach

Our goal in this paper 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. In this paper we consider three cues that can aid in determining 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 [39, 28, 23, 22, 27]. Inspired by the success of deep learning based solutions, in this paper, a novel hierarchical deep learning based model is proposed that is potentially capable of learning low-level image features, person-level actions, their 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.

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 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 :

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \quad (1)$$
$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \quad (2)$$
$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \quad (3)$$
$$g_t = \phi(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (4)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (5)$$
$$h_t = o_t \odot \phi(c_t) \quad (6)$$

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 temporal features for a person. Various pooling strategies can be used to aggregate these features over all people in the scene at each time step.

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

The output of the pooling layer forms our representation for the group activity. The second LSTM network, working on top of the 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:

$$P_{tk} = x_{tk} \oplus h_{tk} \quad (7)$$
$$Z_t = P_{t1} \diamond P_{t2} \ldots \diamond P_{tk} \quad (8)$$

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 kth person at time t. We concatenate these two features (represented by $\oplus$) to obtain the temporal feature representation P_{tk} for kth 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 subsection, 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 [14]. Similar to other approaches [9, 7, 38], 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 sub-

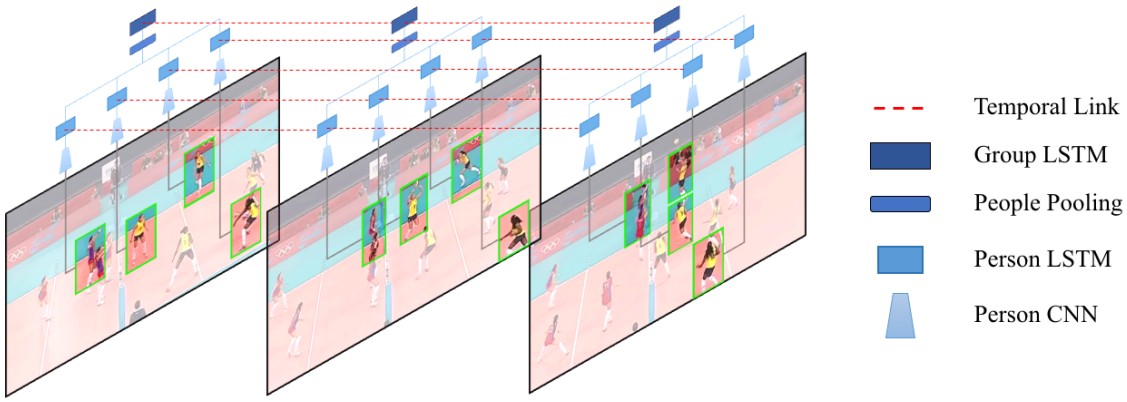

Figure 2: Our two-stage model for a volleyball match. Given tracklets of K-players, we feed each tracklet in a CNN, followed by a person LSTM layer to represent each player's action. We then pool over all people's temporal features in the scene. The output of the pooling layer is feed to the second LSTM network to identify the whole teams activity.

jects in a scene, we used the tracker by Danelljan et al. [6], implemented in the Dlib library [17].

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 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 features are pooled over all people, and are 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 is a temporal extension of the first baseline. It examines the idea of feeding image level features directly to a LSTM model to recognize group activities. In this baseline, the AlexNet model is deployed on the whole image and resulting fc7 features are fed to a LSTM model. This baseline can be considered as a reimplementation of Donahue et al. [9].

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

6. **Two-stage Model without LSTM 1:** This baseline is a variant of our model, omitting the person-level temporal model (LSTM 1). Instead, the person-level classification is done only with the fine-tuned person CNN.

7. **Two-stage Model without LSTM 2:** This baseline is a variant of our model, omitting the group-level temporal model (LSTM 2). In other words, we do the final classification based on the outputs of the temporal models for individual person action labels, but without an additional group-level LSTM.

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 [1, 7, 2]. This dataset consists of 44 videos, eight person-level pose labels (not used in our work), five person level action labels, and five group-level activities. A scene is assigned a group activity label based on the majority of what 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
B6-Two-stage Model without LSTM 1	70.1
B7-Two-stage Model without LSTM 2	76.8
Two-stage Hierarchical Model	**81.5**

Table 1: Comparison of our method with baseline methods on the Collective Activity Dataset.

Method	Accuracy
Contextual Model [23]	79.1
Deep Structured Model [7]	80.6
Our Two-stage Hierarchical Model	81.5
Cardinality kernel [11]	**83.4**

Table 2: Comparison of our method with previously published works on the Collective Activity Dataset.

In Table 1, the classification results of our proposed architecture 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-stage model. It is interesting to observe that adding temporal information using LSTMs improves the performance of these baselines.

Table 2 compares our method with state of the art methods for group activity recognition. The performance of our two-stage model is comparable to the state of the art methods. Note that only Deng et al. [7] is a previously published deep learning model. We postulate that there would be a significant improvement in the relative performance of our model if we had a larger dataset for recognizing group activities. In contrast, the cardinality kernel approach [11] outperformed our model. It should be noted that this approach works on hand crafted features fed to a model highly optimized for a cardinality problem (i.e. counting the number of actions in the scene) which is exactly the way group activities are defined in this dataset.

4.2.1 Discussion

The confusion matrix obtained for the Collective Activity Dataset using our two-stage model is shown in Figure 3. We observe that the model performs almost perfectly for the talking and queuing classes, and gets confused between crossing, waiting, and walking. Such behaviour is perhaps due to a 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 the 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 improved performance with compared to baselines. The temporal information improves performance. Further, finding and describing the elements of a video (i.e. persons) provides benefits over utilizing frame level features.

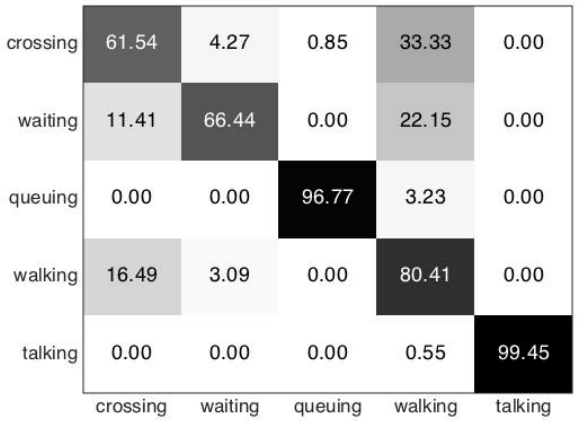

Figure 3: Confusion matrix for the Collective Activity Dataset obtained using our two-stage model.

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 that were handpicked from 15 videos with seven player action labels and six team activity labels. We used frames from 2/3rd of the videos for training, and the remaining 1/3rd for testing. The list of action and activity labels and related statistics are tabulated in Tables 3 and 4.

Figure 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).

From the tables, we observe that the group activity labels are relatively more balanced compared to the player action labels. This follows from the fact that we often have people present in static actions like standing compared to dynamic actions (setting, spiking, etc.). Therefore, our dataset presents a challenging team activity recognition task, where we have interesting actions that can directly determine the group activity occur rarely in our dataset. The dataset will be made publicly available to facilitate future comparisons [2].

In Table 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.

[2] https://github.com/mostafa-saad/deep-activity-rec

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 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: Statistics of the action labels in the volleyball dataset.

However, compared to the baselines, the performance gain using our model is more modest. This is likely because we can infer group activity in volleyball by using just a few frames. Therefore, in the 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 a 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, showing 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
B6-Our Two-stage Model without LSTM 1	48.8
B7-Our Two-stage Model without LSTM 2	49.7
Our Two-stage Hierarchical Model	**51.1**

Table 5: Comparison of the team activity recognition performance of baselines against our model evaluated on the volleyball dataset.

In both datasets, an observation from the tables is that while both LSTMs contribute to overall classification performance, having the first layer LSTM (B7 baseline) is relatively more critical to the performance of the system, compared to the second layer LSTM (B6 baseline).

All the reported experiments use max-pooling as mentioned above. However, we also tried both sum and average pooling, but their performance was consistently lower compared to their max-pooling counterpart.

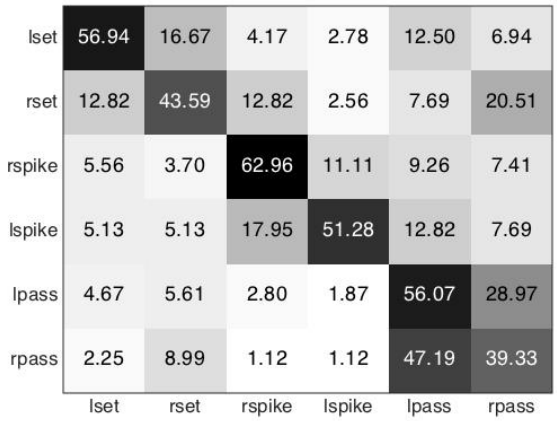

Figure 5: Confusion matrix for the volleyball dataset obtained using our two-stage hierarchical model.

4.3.1 Discussion

Figures 4 and 5 show visualizations of our detected activities and the confusion matrix obtained for the volleyball dataset using our two-stage model. From the confusion matrix, we observe that our model generates consistently accurate high level action labels. Nevertheless, our model has some confusion between set and pass activities, as these activities often may look similar.

5. Conclusion

In this paper, we presented a novel deep structured architecture to deal with the group activity recognition problem. Through a two-stage process, we learn a temporal representation of person-level actions and combine the representation of individual people to recognize the group 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.

Acknowledgements

This work was supported by grants from NSERC and Disney Research.

References

[1] M. R. Amer, P. Lei, and S. Todorovic. Hirf: Hierarchical random field for collective activity recognition in videos. In *Computer Vision–ECCV 2014*, pages 572–585. Springer, 2014.

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

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

[4] W. Choi, K. Shahid, and S. 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] W. Choi, K. Shahid, and S. Savarese. Learning context for collective activity recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3273–3280. IEEE, 2011.

[6] M. Danelljan, G. Häger, F. Shahbaz Khan, and M. Felsberg. Accurate scale estimation for robust visual tracking, 2014. BMVC.

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

[9] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. *arXiv preprint arXiv:1411.4389*, 2014.

[10] A. Graves and N. 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] H. Hajimirsadeghi, W. Yan, A. Vahdat, and G. Mori. Visual recognition by counting instances: A multi-instance cardinality potential kernel. *CVPR*, 2015.

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

[13] S. S. Intille and A. Bobick. Recognizing planned, multi-person action. *Computer Vision and Image Understanding (CVIU)*, 81:414–445, 2001.

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

[15] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *CVPR*, 2015.

[16] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. 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.

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

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[19] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: a large video database for human motion recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2556–2563. IEEE, 2011.

[20] S. Kwak, B. Han, and J. H. Han. Multi-agent event detection: Localization and role assignment. In *CVPR*, 2013.

[21] T. Lan, L. Sigal, and G. 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.

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

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

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

[25] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. *CVPR*, 2015.

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

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

[28] C. Schüldt, I. Laptev, and B. 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.

[29] A. G. Schwing and R. Urtasun. Fully connected deep structured networks. *arXiv preprint arXiv:1503.02351*, 2015.

[30] T. Shu, D. Xie, B. Rothrock, S. Todorovic, and S.-C. Zhu. Joint inference of groups, events and human roles in aerial videos. In *CVPR*, 2015.

[31] B. Siddiquie, Y. Yacoob, and L. Davis. Recognizing plays in american football videos. Technical report, Technical report, University of Maryland, 2009.

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

[33] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[34] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[35] E. Swears, A. Hoogs, Q. Ji, and K. Boyer. Complex activity recognition using granger constrained dbn (gcdbn) in sports and surveillance video. In *Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[36] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

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

[38] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*, 2014.

[39] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3169–3176. IEEE, 2011.

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

[41] S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, and L. Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *arXiv preprint arXiv:1507.05738*, 2015.

[42] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr. Conditional random fields as recurrent neural networks. In *International Conference on Computer Vision (ICCV)*, 2015.