HIERARCHICAL ATTENTION AND CONTEXT MODELING FOR GROUP ACTIVITY RECOGNITION

Longteng Kong¹, Jie Qin²∗, Di Huang¹, Yunhong Wang¹ and Luc Van Gool²

¹Beijing Advanced Innovation Center for Big Data and Brain Computing, Beijing 100191, China
²Computer Vision Laboratory, ETH Zürich, CH-8092 Zürich, Switzerland

ABSTRACT

Group activity recognition in videos is a challenging task, with two major issues, i.e., attending to those persons and their body parts that contribute significantly to the activity, and modeling contextual person structures in the group. Most previous approaches fail to provide a practical solution to jointly address both issues, however. In this paper, we propose to simultaneously deal with both issues via a hierarchical attention and context modeling framework based on Long Short-Term Memory (LSTM) networks. For the former, we propose ‘Hierarchical Attention Networks’ applied at the part/person level, capable of attending distinctively to different persons and their body parts. For the latter, we build ‘Hierarchical Context Networks’ that take the attentionally pooled person-level features as input and recurrently model intra/inter-group contextual structures. The attentive and contextual representations are concatenated and fed into another LSTM to generate high-level discriminative temporal representations for group activity recognition. Extensive experiments on two widely-used group activity datasets demonstrate the effectiveness and superiority of the proposed framework.

Index Terms— Group Activity Recognition, Visual Attention, Context Modeling, LSTM Networks

1. INTRODUCTION

Human activity recognition in videos [1, 2, 3, 4, 5] has attracted extensive research interests in the past few years, where recognition of high-level group activities is a very challenging task. Group activity recognition facilitates many real-world applications, e.g., intelligent video surveillance, anomalous event detection and tactics analysis in sports video. Previous approaches attempt to address this problem by modeling the contextual information using local contextual descriptors [6, 7] and graphical models [8, 9, 10]. Choi et al. [6] extract contextual descriptors from a person and the surrounding area to recognize group activities, which are further enhanced in [7] via structure learning. Lan et al. [9] propose a graphical model by considering interactions on the social role level. Moreover, Lan et al. [10] model an adaptive structure adjustable to the most discriminative interactions. However, all the above models are based on traditional learning strategies (e.g., linear models) using hand-crafted features, and thus suffer from representational limitations.

Recently, several deep learning approaches [11, 12, 13] have been proposed to model group contextual structures and these achieved better performances than traditional ones. Typically, they model individual actions and group activities sequentially using Recurrent Neural Networks (e.g., the LSTM networks [14]), where max/average pooling is adopted to aggregate person-level features. In other words, different people are paid comparable attention to. Yet, different people contribute to different degrees to the overall group activity. For instance, for a ‘serving’ action in volleyball, more attention ought to be paid to the server than the other players.

Based on the above considerations and inspired by the recent advance in document analysis [15, 16, 17], we propose a hierarchical soft-attention mechanism using LSTM to attach variable levels of importance to different persons and their distinct body parts, i.e., attention-aware pooling of part/person-level features. Unlike [13] who only consider person-level attention, we simultaneously also capture part-level attention, in the same way as attention based document analysis. There both sentence-level and word-level attentions are applied and performance was significantly improved by the hierarchical attention strategy [17].

In addition, as stated in [18], modeling intra-group contextual information (e.g., interaction within the same team in a volleyball game) is far from enough, and inter-group context (e.g., interaction between the two teams) needs to be explored as well. In [18], a recurrent encoding scheme is introduced to deal with intra/inter-group interactions. But it is not end-to-end trainable due to the additional context encoding step. In this paper, we perform the grouping/partition operation similar to [18], but in such a way that intra/inter-group context is modeled. More specifically, we propose ‘Hierarchical Context Networks’ (HCNs) that model intra/inter-group contextual information in a fully recurrent fashion without any additional operations, thus making it end-to-end trainable.

As shown in Fig. 1, we further integrate the above two networks in a sequential manner. We start with the Hi-
Hierarchical Attention Networks (HANs), which extend the original LSTM with a soft attention mechanism [19] and learn different levels of attention for different persons and their body parts as the group activity proceeds. Person-level features are pooled attentively with respect to the learned part/person-level attention weights. Subsequently, the HCNs take the person-level features as input and recurrently output intra/inter-group contextual features based on two-stage LSTM networks. Note that with the help of the attention mechanism, the persons/parts attended to explicitly enhance the contextual features. For example, in a 'set' activity, the setter and his/her arms contribute more than the surroundings and other body parts in building the contextual features. Finally, the person-level attentive features and group-level contextual features are concatenated into the final frame-level descriptions, which are then fed into another LSTM to generate a high-level temporal representation for recognizing group activities. Extensive experiments on two benchmarks (i.e. the Collective Activity Dataset and Volleyball Dataset) demonstrate the effectiveness and superiority of the proposed framework over the state-of-the-art.

2. APPROACH

Given video sequences of group activities, we first utilize the tracker by Danelljan et al. [20] to obtain human tracklets (e.g. a sequence of tracked human bounding boxes). Based on these tracklets, we propose a hierarchical attention and context modeling framework to extract attention/context-aware descriptions for group activity recognition.

Basically, we derive our HANs and HCNs from a variant of Recurrent Neural Networks, i.e. LSTM. Each LSTM cell is composed of three gates (i.e. input gate \(i\), output gate \(o\) and forget gate \(f\)) and a memory cell \(c_t\). In each time step \(t\), given the input \(x_t\) and the previous hidden state \(h_{t-1}\), the LSTM cell outputs an updated hidden state \(h_t\). Owing to the gates and memory cell, LSTM is capable of learning long-term dynamics. Please refer to [14] for more technical details.

2.1. Hierarchical Attention Networks

The HANs can automatically explore different degrees of importance for persons and their body parts. In particular, the networks can attend to salient parts of persons (part-level) as well as to relevant persons in the group (person-level). We use the attention LSTM cell [19] (AttLSTM in Fig. 1), a variant of the soft attention model from [21] because of its better performance, and extend it to accept video sequences as input.

When applying part-level attention, in each time step we equally divide each person into \(K\) parts and represent him/her as \(V_t = (v_{t,1}, \ldots, v_{t,K})\), where \(v_{t,i} \in \mathbb{R}^d\) indicates the feature of the \(i\)th part of a person. Given \(V_t\) and the hidden state \(h_t\) of the part-level AttLSTM, the scores \(\alpha_{t,k} = (\alpha_{t,1}, \ldots, \alpha_{t,K})\) indicating the importance of \(K\) parts are jointly obtained as follows:

\[
s_t = w_h^T (\tanh(W_v V_t + W_h h_t)), \quad \alpha_{t,k} = \frac{\exp(s_{t,k})}{\sum_{i=1}^K \exp(s_{t,i})},
\]

where \(W_v\), \(W_h\) and \(w_h\) are the learnable network parameters. Based on the above scores, the feature \(c_t\) of a person attended to can be computed as \(c_t = \sum_{i=1}^K \alpha_{t,i} v_{t,i}\). We further combine the current hidden state \(h_t\) and \(c_t\) in each time step to obtain the representation \(u_t\) of each person, i.e. \(u_t = c_t \oplus h_t\). The averaged outputs of all time steps are fed into a softmax layer (i.e. a fully connected layer with softmax activation function) to calculate the probability of an action:

\[
y_a = \text{softmax}(W_p (\frac{1}{T} \sum_{t=1}^T u_t)),
\]
where $T$ is the total number of time steps, and $W_p$ is the learnable weight parameter.

When learning person-level attention, we treat the activity of a group as a series of person-level actions. In each time step, a group can be represented by $U_t = (u_{t,1}, \ldots, u_{t,N})$, where $u_{t,j}$ is the feature of the $j^{th}$ person in the group of totally $N$ people. Similarly, the attention weights $\beta_t = (\beta_{t,1}, \ldots, \beta_{t,N})$ of different persons are calculated as

$$\hat{s}_t = \hat{w}_u^T \text{tanh}(W_u U_t + \hat{W}_h \hat{h}_t)), \beta_{t,n} = \frac{\exp(\hat{s}_{t,n})}{\sum_{j=1}^N \exp(\hat{s}_{t,j})},$$

(3)

where $W_u, \hat{W}_h$ and $\hat{h}_t$ are the network parameters to learn, and $\hat{h}_t$ is the current hidden state of the person-level AttLSTM. In each frame, the final representation of the $j^{th}$ person in a group after HANs is $p_{t,j} = \beta_{t,j} u_{t,j}$. Similar to Eq. (2), the probability of a group activity can be computed using a softmax layer:

$$y_g = \text{softmax}(\hat{W}_y (\frac{1}{T} \sum_{t=1}^T \sum_{j=1}^N p_{t,j} + \hat{h}_t))).$$

(4)

Note that the two-stage (i.e. part-level and person-level) attention networks can be trained jointly. We formulate the final objective function of HANs as a joint cross-entropy loss:

$$L = -\lambda_1 \sum_{n=1}^N \sum_{l_1=1}^{C_1} y_{a,n,l_1} \log \hat{y}_{a,n,l_1} - \lambda_2 \sum_{l_2=1}^{C_2} y_{g,l_2} \log \hat{y}_{g,l_2},$$

(5)

where $C_1$ and $C_2$ are the class numbers of individual actions and group actions respectively, $\hat{y}_{a,n,l_1}$ and $\hat{y}_{g,l_2}$ are the one-hot-encoded ground truth of actions and activities respectively, and $\lambda_1$ and $\lambda_2$ are trade-off parameters.

### 2.2. Hierarchical Context Networks

As [18] mentioned, besides modeling the intra-group contextual information (e.g. the evolution of person-level action dynamics within the same volleyball team), it is also critical to capture group to group context (e.g. interaction between two teams). To this end, we build the HCNs to model intra/inter-group contextual structures simultaneously. We first partition the original group into subgroups in a principled way (which will be elaborated in our experiments). Then, to model the context dependency within a group, we order the persons in a subgroup into a sequence and feed it into LSTM. We conduct a simple yet effective ordering operation, i.e. aligning person-level features by the $x$ or $y$ coordinates of the respective tracklets (we adopt the $x$ coordinate due to its better performance in our experiments). In the $t^{th}$ time step, the persons within the $m^{th}$ group can be depicted as $P_t^m = (p_{t,1}^m, \ldots, p_{t,N_m}^m)$, where $N_m$ denotes the total number of people in this group. $P_t^m$ is then fed into the intra-group LSTM networks to obtain the contextual representation of the $m^{th}$ subgroup.

The inter-group structure is modeled in a similar way. Specifically, the subgroup-level representations are first ordered by the $x$ or $y$ coordinates of the geometric centers of each subgroup and fed to inter-group LSTM networks, whose output serves as the group-level contextual feature $G_t$. In each time step, the global description of group activities consists of two parts, i.e. person-level attentive features of sub-groups ($Z_t$) and group-level contextual features ($G_t$). To obtain $Z_t$, features of all people within a subgroup are first attentively pooled and then concatenated across all subgroups to form $Z_t$. Finally, the global description passes through another LSTM layer. The hidden state $h_g$ of this LSTM layer carries high-level temporal information with visual attention and contextual structure. $h_g$ is fed into a softmax classification layer with cross-entropy loss to predict group activities.

### 3. EXPERIMENTAL RESULTS

We evaluate our framework on two widely-adopted benchmarks, i.e. the Collective Activity Dataset [6] and Volleyball Dataset [11]. We first depict our implementation details and then compare our method with the state-of-the-art ones.

**Implementation Details:** We adopt the GoogLeNet [22] pre-trained on the ImageNet [23] and extract the 1024×7×7 feature map for a person from the last convolutional layer. We train our hierarchical networks in two steps. For HANs, we pre-train the attention modules w.r.t. persons and their body parts, resp., to ensure the convergence. The whole training process of HANs includes three steps: training the part-level networks, fixing the parameters of part-level networks to train the person-level networks, and training the hierarchical networks jointly. As for HCNs, in each time step, we organize the attentively pooled person features into subgroups, and then feed them into HCNs to generate contextual features recurrently. The intra/inter-group features are concatenated and fed into the final LSTM networks followed by a softmax classification layer, which makes the whole context networks end-to-end trainable without any additional encoding steps as in [18]. In all the experiments, we set $\lambda_1 = 1$ and $\lambda_2 = 2$, and use stochastic gradient descent with ADAM [24], with the initial learning rate set to $10^{-5}$.

**Baselines:** In addition to the state-of-the-art methods, we compare our method with the following baselines:

1) **B1 (w/o HANs):** We replace HANs in our framework with 2-layer LSTM similar to [11], followed by our HCNs.
2) **B2 (w/o HCNs):** We remove HCNs in our framework. In other words, only the attention features are used for classification without any context modeling.

**Table 1:** Results on the Collective Activity Dataset.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structure Inference Machines</td>
<td>81.2%</td>
</tr>
<tr>
<td>Cardinality Kernel</td>
<td>83.4%</td>
</tr>
<tr>
<td>CERN-2 [12]</td>
<td>87.2%</td>
</tr>
<tr>
<td>Two-stage Hierarchical Model</td>
<td>81.5%</td>
</tr>
<tr>
<td>B1 (w/o HANs)</td>
<td>83.1%</td>
</tr>
<tr>
<td>B2 (w/o HCNs)</td>
<td>82.3%</td>
</tr>
<tr>
<td>Ours (HANs+HCNs)</td>
<td>84.3%</td>
</tr>
</tbody>
</table>
3.1. Results on the Collective Activity Dataset

The Collective Activity Dataset [6] contains 44 short video sequences of 5 different collective activities, and provides 8 pairwise interaction labels (not used in our work) and 6 person-level action labels. We adopt the same experimental setting as [26]. For tracklets grouping/partition, we employ the graph partition algorithm in [27]. There are 1024 hidden units in the LSTM layers of the two networks (HANs and HCNs) and 512 units in the last LSTM layer.

Table 1 shows the comparison results. Clearly, our hierarchical model outperforms the two baselines, which shows that incorporating either visual attention or contextual structure can improve the performance, and the combination of them further boosts the accuracy. The performance gain is more obvious w.r.t. B2, indicating the key role of our hierarchical context modeling. Meanwhile, our results are superior to conventional structure learning models and most deep learning models. Meanwhile, our results are superior to conventional structure learning models and most deep learning ones. Note that although [12] achieves better results, they use additional manually annotated context (i.e. interaction labels), which contributes significantly to recognizing collective activities. We show a qualitative example in Fig. 2 to illustrate our attention mechanism. As can be seen, important persons and body parts are paid more attention to. We also show the confusion matrix in Fig. 4 (a), where queue and talk are nearly 100% recognized. On the other hand, a small fraction of cross and wait is mistaken as walk, because they share similar visual features.

Table 2: Results on the Volleyball Dataset.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>CERN-2 [12]</td>
<td>83.3%</td>
</tr>
<tr>
<td>Two-stage Hierarchical Model [11]</td>
<td>81.9%</td>
</tr>
<tr>
<td>B1 (w/o HANs)</td>
<td>84.1%</td>
</tr>
<tr>
<td>B2 (w/o HCNs)</td>
<td>82.5%</td>
</tr>
<tr>
<td>Ours (HANs+HCNs)</td>
<td>85.1%</td>
</tr>
</tbody>
</table>

The thickness of each bounding box reflects the attention weight.

3.2. Results on the Volleyball Dataset

The Volleyball Dataset [11] consists of 4830 frames from 55 videos with 9 player actions and 8 group activities. We follow the train/test split and subgroup partition suggested by [11]. For each action/activity, we use a temporal window of length $T = 10$, which corresponds to 5 frames before the annotated frame, and 4 frames thereafter. We use 2048 hidden units for the LSTM layers of two networks (HANs and HCNs) and 1024 units for the last LSTM layer.

The accuracies of different methods are summarized in Table 2. It can be observed that the proposed approach clearly outperforms the state-of-the-art ones, indicating the effectiveness of the combination of visual attention and contextual structure. More interestingly, without attention networks (i.e. B1) we already achieve better performance than the state-of-the-art ones. In terms of baseline methods, the performance of B2 is improved prominently by modeling intra/inter-group context. This is mainly because activities in volleyball games involve more discriminative structural patterns. Fig. 3 illustrates an example of left-spike, where we can see more structured activities and the success of our attention mechanism. We also provide the confusion matrix in Fig. 4 (b) to illustrate our ability to recognize different individual actions.

4. CONCLUSION

In this paper, we propose a hierarchical attention and context modeling framework for group activity recognition. The HANs pay different levels of attention to different persons and their distinct body parts, and HCNs model both intra-group and inter-group contextual information. By integrating visual attention and contextual structure, the proposed framework can generate more discriminative descriptions for group activities. Extensive experiments on two datasets clearly demonstrate the superiority of the proposed framework.

**Acknowledgement**

This work was supported by the National Natural Science Foundation of China (61573045). Luc Van Gool acknowledges support by CHIST-ERA project MUSTER.

$^2$The thickness of each bounding box reflects the attention weight.
5. REFERENCES


