MULTI-LEVEL ACTION DETECTION VIA LEARNING LATENT STRUCTURE

Behzad Bozorgtabar

1Vision & Sensing, HCC Lab, ESTeM
University of Canberra
Email: Behzad.Bozorgtabar@canberra.edu.au

Roland Goecke

2IHCC, RSCS, CECS
Australian National University
Email: roland.goecke@ieee.org

ABSTRACT

Detecting actions in videos is still a demanding task due to large intra-class variation caused by varying pose, motion and scales. Conventional approaches use a Bag-of-Words model in the form of space-time motion feature pooling followed by learning a classifier. However, since the informative body parts motion only appear in specific regions of the body, these methods have limited capability. In this paper, we seek to learn a model of the interaction among regions of interest via a graph structure. We first discover several space-time video segments representing persistent moving body parts observed sparsely in video. Then, via learning the hidden graph structure (a subset of the graph), we identify both spatial and temporal relations between the subsets of these segments. In order to seize the more discriminative motion patterns and handle different interactions between body parts from simple to composite action, we present a multi-level action model representation. Consequently, for action classification, the classifier learned through each action model labels the test video based on the action model that gives the highest probability score. Experiments on challenging datasets, such as MSR II and UCF-Sports including complex motions and dynamic backgrounds, demonstrate the effectiveness of the proposed approach that outperforms state-of-the-art methods in this context.

Index Terms— Action detection, pooling regions, multi-level video representation, latent structure

1. INTRODUCTION AND RELATED WORK

Action detection approaches have mostly concentrated on developing action representation models, where the objective is to specify when and where certain actions appear. In recent years, action detection has become important in comprehending human motion from video. Compared with action recognition, it not only determines the action, but also localises a video subvolume encompassing the action of interest. Notable methods in this context include: The Spatio-temporal Deformable Part Model (SDPM) \(^\text{[1]}\) performing action detection by extending deformable part models from 2D images to 3D space-time volumes. The Figure-Centric model \(^\text{[2]}\) recovers the latent 2D parts localising action of interest in each frame with smooth constraints. \(^\text{[3]}\) propose an action bank, conceptually simple yet powerful, as a high-level action representation. In \(^\text{[4]}\), hierarchical space-time segments form a new representation for both action recognition and localisation. The majority of these methods aims to discover the human body parts structure and their correlations in a space-time domain. Our motivation here is as follows:

1. Since the discriminative spatio-temporal pooling regions only reside in a small portion of the video volume (e.g. correlated moving body parts in the Swing-Bench action), irrelevant features from background movements must be suppressed efficiently so as not to degrade the overall performance. On the other hand, some static regions may contain useful information (e.g. configuration of the body parts) and should be preserved.

2. For distinct actions, the local spatio-temporal features enclosing the video subvolumes exhibit different levels of interactions from a simple to composite movements between the body parts as can be observed in the diving action. Such relationships between the local spatio-temporal pooling regions can be separated into a number of action primitives, each representing a different level of interaction.

Motivated by these observations, we propose an enhanced scheme in which instead of computing a global pooling of features that ignores useful information on the spatio-temporal location of moving body parts, the correlation among different video subvolumes is considered leading to a discriminative structure of local spatio-temporal regions robust to background effects. In the proposed method, we model the interaction between body parts via a graph structure that aims to find the subset of graph nodes and edges that distinguish the corresponding action class from other action classes. Specifically, we first extract space-time video segments containing the potential moving body parts and partition them into the canonical clusters serve as the graph nodes. In fact, each graph node attribute specifies the motion feature pooling computed on all segments of the corresponding cluster. Finally, to account for the large differences in scale, pose and viewpoints, we propose clusters with different scales, providing a set of more precise local video representations. The learned graph in each level of the hierarchy represents one specific kind of interaction for the body parts movement in a certain action. The overall framework is illustrated in Figure \(^\text{[1]}\).

2. PROPOSED MULTI-LEVEL ACTION DETECTION

Section \(^\text{[2.1]}\) briefly reviews how space-time video segments (pooling regions) are extracted. Section \(^\text{[2.2]}\) presents our approach to obtain a multi-level representation of the action video as a latent graph modelling interactions between the local pooling regions (Section \(^\text{[2.3]}\)).

2.1. Space-Time Video segments (Pooling Regions)

Motivated by the successful work in action localisation \(^\text{[4]}\), we first extract a set of hierarchical space-time video segments \(S\) for the given video. This representation includes two steps: for the first step, an unsupervised method is utilised for each video frame producing a set of segment trees \(T^i\), which preserve both static and nonstatic characteristics of the body parts. Each segment tree \(T^i\) is...
Finally, using our multi-level representation, localise the specific body parts (red bounding boxes) contributing to the diving space-time segments reveal consistent moving body parts (blue and yellow bounding boxes show root and parts of segment tree, respectively).

Fig. 1. The overall framework for building a multi-level action representation. First, extract segment trees in each frame. Then, video space-time segments reveal consistent moving body parts (blue and yellow bounding boxes show root and parts of segment tree, respectively). Finally, using our multi-level representation, localise the specific body parts (red bounding boxes) contributing to the diving action.
\[ \phi(X, z) = [p_1, \ldots, p_L; \psi_{1,2}, \ldots, \psi_{L-1,L-2}], \] where \( L \) is the number of candidate nodes. With the known action class label (classifier \( w_k \)) for the corresponding video \( X \), the latent subgraph \( z \) can be estimated by the scoring function of Eq. \( 3 \) including maximisation over latent variables representing the space-time region of the interest. We utilise an efficient algorithm, TRW-S \([7]\) to solve this NP-hard Quadratic Integer Programming problem. As a result, the discriminative pooling volumes are inferred yielding a distinctive representation of the action class.

\[ z^* = \max_{z \in Z} f_w(X, z) \quad \text{s.t.} \quad f_w(X, z) = w_k^T \phi(X, z) \] (3)

2.4. Learning Multi-level Classifiers

As discussed in the previous section, we utilise classifier \( w_k, k \in \mathcal{T} = \{1, \ldots, K\} \) to find the latent configuration graph modelling correlation of the interacting body parts. However, to take these facts into account, we propose a multi-level action representation by learning multiple latent graphs (each of them learned in a separate level), where all classifier models \( w_k \) learned from these discrete levels are stacked together to create the final set of action models. Our experiments reveal that less than three nodes suffices for the latent graphs to cope with different interactions. Given a set of training videos with binary labels \([((X_i, y_i))_{i=1}^m, (y_i \in \{+1, -1\})\], we use a maximum margin based optimisation to learn the classifiers over the defined latent graph structure \( z \) in each layer:

\[ \min_{w_k, \xi_k \geq 0, \xi_i \geq 0} \frac{1}{2} \sum_{k=1}^K ||w_k||^2 + C_1 \sum_{i=1}^M \xi_i + C_2 \sum_{y_i = +1} \xi_i \] (4)

\[ \text{s.t.} \forall i=1, 2, \ldots, M, \max_{\xi, \xi \in \mathcal{X}} \max_{x_k \in \mathcal{T}} w_k^T \phi(X_i, z) \geq 1 - \xi_i \] (5)

\[ \forall y_i = +1, k \in \mathcal{T}, \max_{\xi, \xi \in \mathcal{X}} \max_{x_k \in \mathcal{T}} w_k^T \phi(X_i, z) \geq 1 - \xi_i \] (6)

\[ \forall k, 1 \leq k \in \mathcal{T}, \sum_{y_i = +1} \frac{1}{2} \sum_{x_k \in \mathcal{X}} (w_k^T \phi(X_i, z))^2 - w_k^T \phi(X_i, z)) \leq 0 \] (8)

where \( J_i \) denotes the number of candidate graphs in the latent space \( Z_i \). The first constraint correlates to a multiple-instance based marginal for each bag \( X \). In this framework, each positive bag \( X_{P,\text{pos}} \) should carry at least one latent structure (instance of the bag) explained correctly by the discriminative models \( w_k \), while on the other side, the negative bags \( X_{N,\text{neg}} \) should not contain any latent configuration of the graph indicated by the classifiers. The second set of constraints enforces diversity among the latent structures where only the instances from the positive bags play a part in the margin between the classifiers. The last set of constraints enforces the class stability to avoid insignificant (trivial) solutions of the latent structures that are allocated to one classifier. Motivated by the observation of the learning algorithm in \([8]\) and since the constraints are the difference between two convex functions, we implement the Convex-Concave Cutting Plane (CCCP) algorithm \([9]\) to deal with the training procedure. This algorithm seeks (1) for the latent structures explained in the previous section and then (2) solves a structured SVM problem according to the cutting plane method.

2.5. Action Recognition

In the training process, we obtain an action model (a set of all classifiers learned for the multi-level graphs configurations) for each action class. Given a new test video, we follow the same process, extract the latent graphs for the different levels of the space-time video segments, then measure the confidence score of the estimated latent graphs in the current video with respect to the obtained classifier models. In fact, these scores indicate how likely the trained latent configuration model exists in the new test video. Precisely, for the \( l \)th layer of the test video, we obtain the confidence scores by maximising the discriminative function \( f_w(X, z) \) over the latent space \( Z_i \) as \( w_k^* = \max_{w_k \in \mathcal{T}} w_k^T \phi(X, z) \), where \( w_k \) and \( w_k^* \) are the \( k \)th trained classifier model and the acquired confidence score, respectively. Finally, we concatenate the confidence scores from all layers into a single vector as \( \Theta = [\Gamma_1, \Gamma_2, \Gamma_3] \), where \( \Gamma_i = [w_k^*]_{i=1}^G \) to achieve the final video representation. A test video label is chosen from the action model that gives the highest probability score.

3. EXPERIMENTS AND DISCUSSION

Experimental Settings. We evaluate our algorithm on two widely used action recognition and localisation benchmark datasets: UCF Sports and MSR-II. In addition, we assess the performance of the proposed approach with several state-of-the-art methods, such as action localisation via Hierarchical Space-Time Segments (HSTS) \([4]\) and Spatio-temporal Deformable Part Models for Action Detection (SDPM) \([11]\). For a fair comparison, all results are computed using code and annotations provided by the authors. The parameters of the proposed framework are fixed in all experiments. A few parameters are found empirically and most remaining unique parameters for all datasets do not have a significant impact on the final results.

We extract features computed along improved dense trajectories \([10]\) and compute low-level feature descriptors (i.e. HOG, HOF, MBH and trajectory) with the same parameters as \([11]\). We use localised soft assignment quantisation (LSAQ) \([12]\) to code all low-level features. Then, we apply K-means to obtain the individual dictionaries (codebook size = 3000) where 100k randomly sampled features are employed from each type of descriptor. Each graph node is represented with a weighted-pooling region of the coded features, with \( L^2 \) normalisation. The parameters \( (r_x, r_y, r_t) \) are set to 6.

3.1. Results on UCF Sports Dataset

The UCF Sports dataset \([2]\) includes 150 videos from sports broadcasts from 10 action classes including running, kicking, lifting and so on. Due to the large intra-class variations such as pose changes and background clutter, this dataset is more challenging than other standard datasets. For our experiments, we follow the experimental setting of \([2]\) for the disjoint splitting of the training and testing samples. Furthermore, we utilise Leave-One-Out splitting scheme. Regarding action recognition (not our primary focus), with the provided frame-level annotation, our approach, Multi-level action detection (MLAD) achieves significantly higher average per-class accuracy compared with the state-of-the-art result (81.7% in \([4]\)) as shown in Table \( 1 \). We also evaluate the action localisation using the ‘intersection-over-union’ measure where the correct detection is defined as the detection with the true predicted label and its intersection measure must be higher than 0.2. We first plot the ROC curve for each action class of UCF datasets according to the different overlap threshold at the left bottom of Figure \( 2 \). In addition, we compare the proposed method with two recently state-of-the-art methods: figure-centric model \([2]\) and spatio-temporal deformable part model \([11]\). Considering the ROC curve for the overlap

\begin{footnotesize}
\begin{itemize}
  \item We consider each video as a bag of instances (latent graph structures).
  \item We set three layers in our experiments to achieve a better computational cost without sacrificing the results accuracy.
\end{itemize}
\end{footnotesize}
measure with the varying thresholds in Figure 2, our approach outperforms the others and achieves the state-of-the-art performance on this dataset.

3.2. Results on MSR-II Dataset

The MSR-II dataset contains 54 video sequences from three action classes: hand-waving, boxing and hand-clapping. The dataset is recorded with a background of outdoor traffic, parties and walking people devised for action detection in complex scenes. Following the cross-dataset paradigm strategy in [13], we train action models using a subset of the KTH [16] for training and test the model on the MSR-II dataset. For each model, the training set includes positive examples from each KTH clip and negative samples from the other two classes. For evaluating the action detection performance, we use the same strategy as [13] and estimate the average precision (AP) for each action class. We demonstrate the precision-recall (PR) curves in Figure 2 and outline the average precision (AP) for each class in Table 1.

Table 1. Mean per-class classification accuracies on the UCF-Sports dataset based on two different splitting strategies.

<table>
<thead>
<tr>
<th>Method</th>
<th>Leave-One-Out</th>
<th>Strategy of [12]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raptis et al. [14]</td>
<td>NA</td>
<td>79.4%</td>
</tr>
<tr>
<td>Lan et al. [13]</td>
<td>NA</td>
<td>73.1%</td>
</tr>
<tr>
<td>Sadanand et al. [3]</td>
<td>95.0%</td>
<td>NA</td>
</tr>
<tr>
<td>Wang et al. [11]</td>
<td>89.1%</td>
<td>NA</td>
</tr>
<tr>
<td>Kovashka et al. [15]</td>
<td>87.3%</td>
<td>NA</td>
</tr>
<tr>
<td>Ma et al. [4]</td>
<td>NA</td>
<td>81.7%</td>
</tr>
<tr>
<td>Proposed MLAD</td>
<td>97.0%</td>
<td>82.5%</td>
</tr>
</tbody>
</table>

Table 2. Performance comparison on the MSR-II dataset, showing APs and mean AP (mAP) for all classes. Comparison with GMM methods with or without adaption [13] and SDPM [11].

<table>
<thead>
<tr>
<th>Video Clip</th>
<th>Boxing</th>
<th>Hand-waving</th>
<th>Hand-clapping</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>17.4%</td>
<td>26.7%</td>
<td>13.1%</td>
<td>19.1%</td>
</tr>
<tr>
<td>SDPM</td>
<td>38.8%</td>
<td>44.7%</td>
<td>23.9%</td>
<td>35.8%</td>
</tr>
<tr>
<td>MLAD (one level)</td>
<td>39.4%</td>
<td>75.9%</td>
<td>45.6%</td>
<td>53.6%</td>
</tr>
<tr>
<td>MLAD (two level)</td>
<td>41.5%</td>
<td>77.5%</td>
<td>48.4%</td>
<td>55.8%</td>
</tr>
</tbody>
</table>

For the benchmark, we evaluate results against the GMM adaption method [13] (baseline) and SDPM [11]. Our approach surpasses these state-of-the-arts methods in the different action classes, especially for hand-waving where we are capable to handle intra-class variations using our multi-level representation.

4. CONCLUSIONS

This paper has presented an enhanced approach for action detection (including both recognition and localisation) in video by using graphs with different layers to cope with varying interactions (from simple to complex ones) of human body parts in different classes. Instead of considering all body parts, our representation reveals only the specific body parts that contribute to the corresponding action class by modelling their space-time relationship. We have demonstrated the potential and effectiveness of the proposed approach on two challenging datasets for action classification. In addition, we show that our method can achieve state-of-the-art results in action localisation in videos (Figure 2).
5. REFERENCES


