Spatiotemporal Deformable Part Models for Action Detection

Yicong Tian[1] Rahul Sukthankar[2,1] Mubarak Shah[1]
ytian@crcv.ucf.edu rahuls@cs.cmu.edu shah@crcv.ucf.edu
[1]Center for Research in Computer Vision, University of Central Florida [2]Google Research

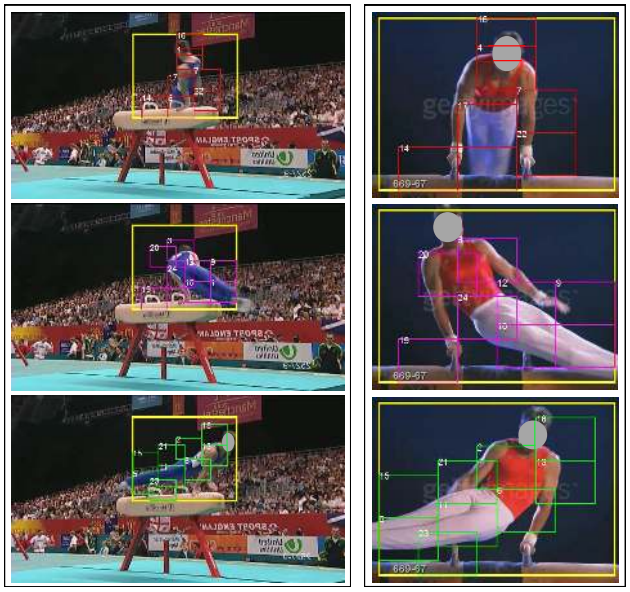
Abstract

Deformable part models have achieved impressive performance for object detection, even on difficult image datasets. This paper explores the generalization of deformable part models from 2D images to 3D spatiotemporal volumes to better study their effectiveness for action detection in video. Actions are treated as spatiotemporal patterns and a deformable part model is generated for each action from a collection of examples. For each action model, the most discriminative 3D subvolumes are automatically selected as parts and the spatiotemporal relations between their locations are learned. By focusing on the most distinctive parts of each action, our models adapt to intra-class variation and show robustness to clutter. Extensive experiments on several video datasets demonstrate the strength of spatiotemporal DPMs for classifying and localizing actions.

1. Introduction

Action recognition in video continues to attract significant attention from the computer vision community, with the bulk of the research focusing primarily on whole-clip video classification, where approaches derived from bag-of-words dominate [13, 14, 20, 24]. This paper focuses on the related problem of action detection [7, 21], sometimes termed action localization [12] or event detection [8, 9], where the goal is to detect every occurrence of a given action within a long video, and to localize each detection both in space and time. As observed by others [1, 8, 28], the action detection problem can be viewed as a spatiotemporal generalization of 2D object detection in images; thus, it is fruitful to study how successful approaches pertaining to the latter could be extended to the former. Analogous to the manner in which Ke *et al*. [8] investigate spatiotemporal extensions of Viola-Jones [23], we study how the current state-of-the-art method for object detection in images, the deformable part model (DPM) [6] should best be generalized to spatiotemporal representations (see Fig. 1).

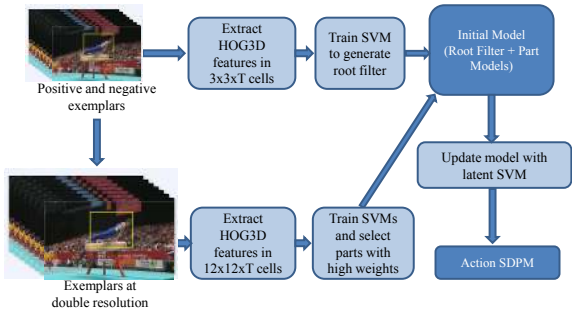
Deformable part models for object detection in images



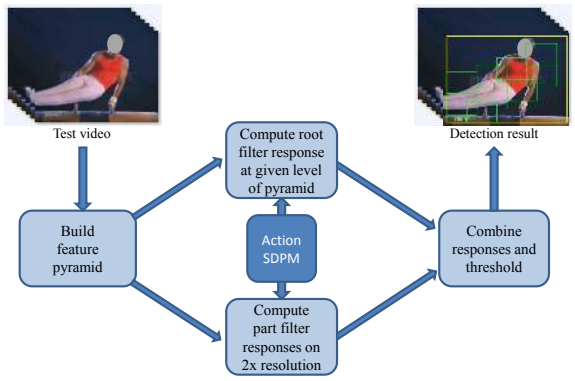
(a) training (b) testing

Figure 1. An example of "Swing Bench" SDPM (left) and its localization result in a test video from UCF Sports (right). This model consists of several parts across three temporal stages (middle frame of each stage shown in each row). The large yellow rectangle indicates the area under the root filter and the small red, magenta, and green ones denote parts. Although trained in videos with cluttered background at a different scale, the SDPM successfully localizes the target action in both space and time.

were proposed by Felzenszwalb *et al*. [6] and are not detailed here due to space limitations. Niebles *et al*. explored a temporal (but not spatiotemporal) extension for video [15]. Two straightforward spatiotemporal generalizations of the DPM approach to action detection in video would be to: 1) treat action detection as a set of image-level detection problems addressed using DPMs, and 2) detect actions as spatiotemporal volumetric patterns that can be captured by a global template and set of 2D parts, each represented using the standard histograms of oriented gradients (HOG) features [4]. Unfortunately, the first is not sufficiently expressive to distinguish between similar actions and the second is



(a) Training: like DPM, we automatically select discriminative parts.



(b) Testing: the action is detected with root denoted by yellow rectangle and parts indicated by green rectangles.

Figure 2. The SDPM framework retains the overall structure of DPM but the volumetric parts are organized in temporal stages.

unable to capture the intra-class spatiotemporal variation of many actions [12]. Clearly, a more sophisticated approach is warranted and in this paper, we propose a spatiotemporal deformable part model (SDPM) that stays true to the structure of the original DPM (see Fig. 2) while generalizing the parts to capture spatiotemporal structure. In SDPM, both the global (yellow rectangle) and the part (smaller green rectangles) templates employ the volumetric HOG3D descriptor [10]. Our automatically selected parts are organized into multiple temporal stages (seen in Fig. 1) that enable SDPM to capture how the appearance of parts changes through time during an action. A key difference between SDPM and earlier approaches is that our proposed model employs volumetric parts that displace in both time and space; this has important implications for actions that exhibit significant intra-class variation in terms of execution and also improves performance in clutter.

The primary aim of this paper is to comprehensively evaluate spatiotemporal extensions of the deformable part model to understand how well the DPM approach for object detection generalizes to action detection in video. For this reason, we restrict ourselves to HOG-like features and resist the temptation of augmenting our method with features such as person detection, dense flow, or trajectories [2, 24] or enhancements like the mixture model. Although SDPM achieves state-of-the-art performance on both controlled and real-world datasets, we stress that it was not engineered for that goal. We believe that a hybrid action detection system that incorporates our ideas could achieve further gains.

2. Related Work

Bag-of-words representations [13, 14, 20, 24] have demonstrated excellent results in action recognition. However, such approaches typically ignore the spatiotemporal distribution of visual words, preventing localization of actions within a video. With bag-of-words representations, Neibles *et al*. [16] and Wong *et al*. [27] apply pLSA to capture the spatiotemporal relationship of visual words. Although some examples of action localization are shown, the localization is performed in simple or controlled settings and no quantitative results on action detection are presented.

Earlier work proposes several strategies for template matching approaches to action localization. Rodriguez *et al*. [18] generalize the traditional MACH filter to video and vector-valued data, and detect actions by analyzing the response of such filters. Kläser *et al*. [11] localize human actions by a track-aligned HOG3D action representation, which (unlike our method) requires human detection and tracking. Ke *et al*. [9] introduce the notion of parts and efficiently match the volumetric representation of an event against oversegmented spatiotemporal video volumes; however, these parts are manually specified using prior knowledge and exhibit limited robustness to intra-class variation.

There has been recent interest in learning parts directly from data. Lan *et al*. [12] detect 2D parts frame-by-frame followed by a CRF with tracking constraints. Brendel and Todorovic [2] construct spatiotemporal graphs over tubes to represent the structure of primitive actions. Raptis *et al*. [17] embed parts obtained by grouping trajectories into graphical model. However, SDPM differs from these in the following four respects. First, SDPM includes an explicit model to capture intra-class variation as a deformable configuration of parts. By contrast, the model in [2] is not flexible enough to handle speed variation within an action. Second, both the global template and set of part templates in SDPM are spatiotemporal volumes, and we search for the best fit across scale, space and time. As a 3D subvolume, each part jointly considers appearance and motion information spanning several frames, which is better suited for actions than 2D parts in a single frame [12] that primarily capture pose. Third, we employ a dense scanning approach that matches parts to a large state space, avoiding the potential errors caused by hard decisions on video segmentation, which are then used for matching parts [17]. Finally, we focus explicitly on demonstrating the effectiveness of action detection within a DPM framework, without

resorting to global bag-of-words information [12, 17], trajectories [17] or expensive video segmentation [2, 9].

3. Generalizing DPM from 2D to 3D

Generalizing deformable part models from 2D images to 3D spatiotemporal volumes involves some subtleties that stem from the inherent asymmetry between space and time that is often ignored by volumetric approaches. Briefly: 1) Perspective effects, which cause large variation in observed object/action size do not affect the temporal dimension; similarly, viewpoint changes affect the spatial configuration of parts while leaving their temporal orderings unchanged. 2) The units of space (pixels) and time (frames) in a video are different and should not be treated interchangeably. Additionally, we make several observations that are specific to deformable part models.

First, consider the difference between a bounding box circumscribing an object in a 2D image and the corresponding cuboid enclosing an action in a video. In the former, unless the object is unusually shaped or wiry, the majority of pixels contained in the bounding box correspond to the object. By contrast, for actions — particularly those that involve whole-body translation, such as walking, or large limb articulations such as kicking or waving — the bounding volume is primarily composed of background pixels. This is because enclosing the set of pixels swept during even a single cycle of the action requires a large spatiotemporal box (see Fig. 3). The immediate consequence of this phenomenon, as confirmed in our experiments, is that a detector without parts (solely using the root filter on the enclosing volume) is no longer competitive. Finding discriminative parts is thus more important for action detection than learning the analogous parts for DPMs for 2D objects.

To quantify the severity of this effect, we analyze the masks in the Weizmann dataset and see that for nine out of ten actions, the percentage of pixels occupied by the actor in a box bounding a *single cyle* of the action is between 18% to 30%; the highest is 'pjump' with 35.7%. These are all dramatically smaller than 80%, which is the fraction of the bounding box image occupied by object parts in DPM [6]. This observation drives our decision during training to select a set of parts such that in total they occupy 50% of the action cycle volume.[1] Naively using the same settings as DPM would force SDPM to form parts from background or unreliable regions, impairing its overall accuracy.

Second, in the construction of spatiotemporal feature pyramids that enable efficient search across scale, we treat space and time differently. This is because, unlike its size, the duration of an action does not change with its distance from the camera. The variation in action duration is principally caused by differences between actors, is relatively small and better handled by shifting parts. Thus, our feature pyramids employ multiple levels in space but not in time.

Finally, the 2D HOG features in the original DPM must be replaced with their volumetric counterparts. To maximize reproducibility, rather than proposing our own generalization of HOG, we employ Kläser *et al*.'s HOG3D [10].

[1] Since SDPM parts are themselves rigid cuboids that contain background pixels, the total volume they occupy in the bounding volume should be higher than the fraction of pixels that correspond solely to the actor.

4. Deformable Part Models

Inspired by the 2D models in [6], we propose a spatiotemporal model with deformable parts for action detection. The model we employ consists of a root filter F_0 and several part models. Each part model is defined by a part filter F_i, an anchor position (x_i,y_i,t_i) and coefficients of deformation cost $d_i = [d_{i1}, d_{i2}, d_{i3}, d_{i4}, d_{i5}, d_{i6}]$. Here $i \in (1, N)$, where N is the number of parts.

4.1. HOG3D feature descriptor

Kläser *et al*. propose the HOG3D [10] descriptor based on a histogram of oriented spatiotemporal gradients as a volumetric generalization of the popular HOG [4] descriptor. The effectiveness of HOG3D as a feature is evidenced in [25]. We briefly summarize the HOG3D descriptor that we use to build fixed-length representations of each volume, along with our minor modifications.

We divide each video volume into a fixed number of non-overlapping cuboid cells. First, gradients are computed along x, y and t directions at every pixel. For each pixel, gradient orientation is quantized to a 20-dimensional vector by projecting the (dx, dy, dt) vector on to a regular icosahedron with the gradient magnitude as its weight. Then for each cell, a 3D Gaussian filter (σ is determined by the size of cell) placed at the centre of the cell is used to smooth the weighted gradients. These gradients are then accumulated into histograms with 20 bins (corresponding to the 3D gradient directions defined by the icosahedron) and normalized using L2 norm within each cell. The final descriptor is obtained by concatenating the histograms of all cells, which is different with the interest point based HOG3D descriptor in [10]. Thus, the dimension of the computed descriptor is determined by the number of cells, but is independent of the dimensions of the input volume.

This spatiotemporal feature jointly encodes both appearance and motion information, but is invariant to changes in illumination and robust to small deformations. During training, we extract HOG3D features over an action cycle volume to train root filter and part filters. During detection, HOG3D features of the whole test video volume are used to form feature maps and construct a feature pyramid to enable efficient search through scale and spatiotemporal location.

4.2. Root filter

We follow the overall DPM training paradigm, as influenced by the discussion in Section 3: During training, for positive instances, from each video we select a single box enclosing one cycle of the given action. Volumes of other actions are treated as negative examples. These negatives are supplemented with random volumes drawn at different scales from videos that do not contain the given action to help better discriminate the given action from background.

The root filter captures the overall information of the action cycle and is obtained by applying an SVM on the HOG3D features of the action cycle volume. How to divide the action volume is important for good performance. Too few cells will decrease the distinctiveness of the feature in each cell. On the other hand, dividing the volume into too many cells, means that each cell cannot capture enough appearance or motion information since it contains too few pixels or frames. In our experiments, to train the root filter, we have experimentally determined that dividing the spatial extent of an action cycle volume into 3×3 works well. However, the temporal division is critical since cycles for different actions may vary from only 6 frames (short actions) to more than 30 frames (long actions). This is an instance of the asymmetry between space and time discussed in Section 3 since the observed spatial extent of an action varies greatly with camera pose but is similar across actions, while temporal durations are invariant to camera pose but very dependent on the type of action.[2] Dividing all of them into the same number of temporal stages would, of course, be too brittle. Thus, the number of stages T is determined automatically for each action type according to its distribution of durations computed over its positive examples, such that each stage of the model contains 5–10 frames. In summary, we adopt a $3\times3\times T$ scheme and the resulting root filter F_0 is a vector with $3\times3\times T\times20$ weights. Fig. 3 shows an example root filter with $3\times3\times3$ cells.

4.3. Deformable parts

As discussed in Section 3 and seen in Fig. 3(a), only a small fraction of the pixels in a bounding action volume correspond to the actor. The majority of pixels correspond to background and can detrimentally impact detection accuracy, particularly in dynamic environments with cluttered backgrounds. As confirmed by our experiments, these issues are more serious in volumetric action detection than in images, so the role of automatically learned deformable parts in SDPM to address them is consequently crucial.

The same training examples, including random negatives, and the same number of temporal stages T is employed for training part models. Our experiments confirm that extracting HOG3D features for part models at twice the resolution and with more cells in space (but not time) enables the learned parts to capture important details; this is consistent with Felzenszwalb *et al.*'s observation [6] for DPMs in images. Analogous to the parts in DPM, we allow the parts selected by SDPM to overlap in space.

[2] As observed by [19], the correlation between action type and duration can cause researchers to overestimate the accuracy of action recognition when testing on temporally segmented video, since features inadvertently encode duration. This supports our decision to detect actions in raw video.

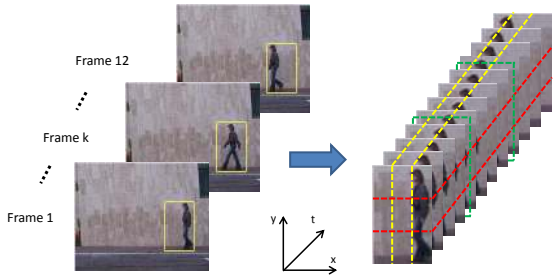


Figure 3. Example of computing HOG3D features for root filter. Left: 12 consecutive frames consisting one cycle of walking (annotations in yellow). Right: spatial area corresponding to the bounding volume, which (for this action type) is divided into 3 cells in x (yellow), 3 cells in y (red), 3 cells in t (green) to compute the HOG3D features for the root filter. The resulting feature descriptor is a $3\times3\times3\times20$ vector. (Part filters not shown here.)

After applying SVM to the extracted features, subvolumes with higher weights, which means they are more discriminative for the given action type, are selected as parts, while those with lower weights are ignored. In our setting, the action volume is divided into $12\times12\times T$ cells to extract HOG3D features and each part is a subvolume occupying $3\times3\times1$ cells. Then, we greedily select the N parts with the highest energy such that their union fills 50% of the action cycle volume. Here we define energy as the sum of positive weights in all cells of a subvolume. The weights in a subvolume are cleared after that subvolume has been selected as a part, and this process continues until all N parts are determined.

In our model, each part represents a spatiotemporal volume. It captures both appearance and motion information spanning several frames. Weights for each part filter are initialized by weights from corresponding cells forming this part. So each part filter is a vector with $3\times3\times1\times20$ weights. In addition, an anchor position (x_i,y_i,t_i) for the ith part is determined, where x_i, y_i and t_i are indices of the cell in the middle of the ith part. Anchor positions define spatiotemporal configuration of parts. For example, $x_i < x_j$ means that the ith part occurs to the left of the jth part, and $t_i < t_j$ means that the ith part occurs before the jth part in time.

Additionally, to address the high degree of intra-class variability in each action type, we allow each part of the model to shift within a certain spatiotemporal region. The cost for the ith part's deformation is a quadratic function of the distance between the placement (x'_i, y'_i, t'_i) and the an-

chor position (x_i, y_i, t_i): $\varepsilon(i, X_i) = d_i \cdot X_i^T$, where $X_i = [|x'_i - x_i|, |y'_i - y_i|, |t'_i - t_i|, |x'_i - x_i|^2, |y'_i - y_i|^2, |t'_i - t_i|^2]$ records the displacement of the ith part. d_i is the learned coefficient of deformation cost for the ith part, and is initialized to $[0, 0, 0, 0.1, 0.1, 0.1]$.

Fig. 4 illustrates an example model for "lifting" trained on UCF Sports (on clean background for clarity). An action cycle is divided into three temporal stages, with each stage containing several frames. In this case, HOG3D features for root filter are computed by dividing the action cycle volume into 3×3×3 cells. (a), (b) and (c) show middle frames of the first, second and third stage in time, respectively. The large yellow rectangle indicates the region covered by the root filter; the small red, magenta, and green ones are the selected parts in each temporal stage. Each part's index is shown at the top left corner of its corresponding rectangle. A low index denotes that the part was selected early and is therefore more discriminative. We observe that our learned parts cover the essential portions of the action, both in terms of appearance and motion, and that SDPM eliminates the majority of the background. Crucially, these results hold in complex scenes (*e.g.*, Fig. 1) because the background clutter is not consistently discriminative for the given action.

4.4. Model update using latent SVM

After obtaining our initial model, we train it using latent SVM with hard negative mining, as in a standard DPM. The exact position of the ith part (x'_i, y'_i, t'_i) is treated as latent information. Thus, filters and deformation cost coefficients d_i are updated to better capture action characteristics.

5. Action detection with SDPM

Given a test video volume, we build a spatiotemporal feature pyramid by computing HOG3D features at different scales, enabling SDPM to efficiently evaluate models in scale, space and time. As discussed in Section 3, the pyramid has multiple scales in space but only one in time. We denote the HOG3D features at level l of the pyramid as $\phi(l)$.

We employ a sliding window approach for template matching during detection (where the sliding window is actually a sliding subvolume). The aspect ratio of the template is determined by the mean of aspect ratios of positive training examples. Score maps for root and part filters are computed at every level of the feature pyramid using template matching. For level l, the score map $S(l)$ of each filter can be obtained by correlation of filter F with features of the test video volume $\phi(l)$,

$$S(l, i, j, k) = \sum_{m,n,p} F(i, j, k)\ \phi(i + m, j + n, k + p, l). \tag{1}$$

At level l in the feature pyramid, the score of a detection volume centered at (x, y, t) is the sum of the score of the root filter on this volume and the scores from each part filter on the best possible subvolume:

$$\text{score}(x, y, t, l) = F_0 \cdot \alpha(x, y, t, l) + \sum_{1 \le i \le n} \max_{(x', y', t') \in Z} \left[F_i \cdot \beta(x'_i, y'_i, t'_i, l) - \varepsilon(i, X_i) \right], \tag{2}$$

where F_0 is the root filter and F_i are part filters. $\alpha(x, y, t, l)$ and $\beta(x', y', t', l)$ are features of a $3{\times}3{\times}T$ volume centered at (x, y, t) and $3{\times}3{\times}1$ volume centered at part location (x'_i, y'_i, t'_i) respectively, at level l of the feature pyramid. Z is the set of all possible part locations and $\varepsilon(i, X_i)$ is the corresponding deformation cost. We choose the highest score from all possible placements in the detection volume as the score of each part model, and for each placement, the score is computed by the filter response minus deformation cost. If a detection volume scores above a threshold, then that action is detected at the given spatiotemporal location.

We perform a scanning search with a step stride equal to the cell size. This strikes an effective balance between exhaustive search and computational efficiency, covering the target video volume with sufficient spatiotemporal overlap.

As with DPM, our root filter expresses the overall structure of the action while part filters capture the finer details. The scores of part filters are computed with different cell size for HOG3D features and at twice the resolution compared with the root filter. This combination of root and part filters ensures good detection performance. In experiments, we observe that the peak of score map obtained by combining root score and part scores is more distinct, stable and accurate than that of only root score map. Since the parts can ignore the background pixels in the bounding volume and focus on the distinctive aspects of the given action, the part-based SDPM is significantly more effective.

6. Experimental Methodology and Results

Since most of previously published results on actions in video are on whole-clip recognition rather than localization, we choose to evaluate SDPM using both criteria, while stressing that the former is not the focus of our work. Where possible, we also present direct comparisons to published localization results on standard datasets. More importantly, since this paper's primary goal is to study the correct way to generalize DPM to spatiotemporal settings, we stress reproducibility by employing standard features, eschewing parameter tweaking and making our source code available.[3]

We present evaluations on three standard datasets, Weizmann, UCF Sports and MSR-II. The main advantage of the first is that the controlled conditions under which actions are performed and the availability of pixel-level actor masks enable us to directly assess the impact of design choices and

[3]SDPM: `http://www.cs.ucf.edu/~ytian/sdpm.html`

(a) (b) (c)

Figure 4. SDPM for "lifting" in UCF Sports, with parts learned in each of the temporal stages. There are in total 24 parts for this SDPM and the index of each part is indicated at the left top corner of corresponding small rectangle. See Fig. 1 for example in clutter.

better understand how SDPM root and part filters work in spatiotemporal volumes. SDPM achieves 100% recognition (without use of masks) and localizes every action occurrence correctly, which is an excellent sanity check.

The second dataset is much more challenging and is drawn from broadcast videos with realistic actions performed in dynamic, cluttered environments. Our results on UCF Sports demonstrate that SDPM achieves state-of-the-art localization in challenging video.

The third dataset contains videos recorded in complex environments and is particularly well suited for cross-dataset experiments. We evaluate action detection on MSR-II Dataset using SDPMs trained solely on the KTH Dataset. Our results on MSR-II confirm that parts are critical for action detection in crowded and complex scenes.

For action detection (spatiotemporal localization), we employ the usual "intersection-over-union" criterion, generate ROC curves when overlap criterion equals 0.2 and also summarize ROC curves with different overlap criterions by the area-under-curve (AUC) measure when necessary for space constraints. For MSR-II, we show precision-recall curves following [3].

For action recognition (whole clip, forced-choice classification), we apply an SDPM for each action class to each clip and assign the clip to that class with the highest number of detections. We provide action recognition results mainly to show that SDPM is also competitive on this task, even though detection is our primary goal.

6.1. Experiments on Weizmann Dataset

The Weizmann dataset [1] is a popular action dataset with nine people performing ten actions. This dataset is considered easy because the actor in each clip is filmed against a static background, with little variation in viewpoint, scale and illumination. We use it primarily to understand the relative contribution of SDPM root vs. part filters.

Weizmann does not come with occurrence-level annotations so we annotate a single action cycle from each video clip to provide positive training instances; as usual, negatives include such instances from other classes augmented with randomly-sampled subvolumes from other classes.

For recognition, we follow the experimental methodology from [1]. SDPM achieves 100% recognition accuracy. While perfect recognition has also recently been achieved by others (*e.g.*, [5, 22, 26]), these all perform recognition through silhouettes. To the best of our knowledge, we are the first to achieve 100% recognition on Weizmann in a detection-based framework that operates only on raw video. When SDPM is learned using root filter alone, recognition accuracy drops to 92.4%, confirming our hypothesis in Sec. 3 that parts are important, even under "easy" conditions. The feature pyramid does not contribute much on this dataset since actions are roughly at the same scale.

On detection, SDPM also achieves perfect results, correctly localizing every occurrence with no false positives. But, SDPM without parts performs poorly: only 66.7% of occurrences are correctly localized! Table 1 compares the detection rate for SDPM with and without parts.

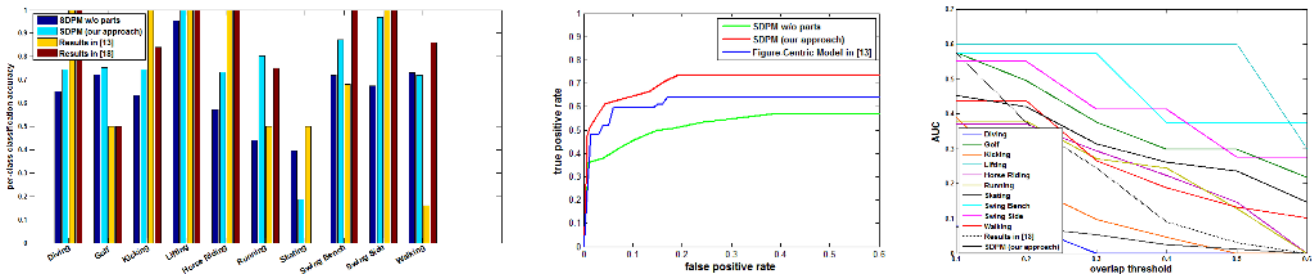
Table 1. Detection rate on Weizmann, showing impact of parts.

	bend	jack	jump	pjump	run	side	skip	walk	wav1	wav2
SDPM	100	100	100	100	100	100	100	100	100	100
w/o parts	100	75	43.8	78.6	80	95.7	27.3	67.5	85	52.9

6.2. Experiments on UCF Sports Dataset

The UCF Sports Dataset [18] consists of videos from sports broadcasts, with a total of 150 videos from 10 action classes, such as golf, lifting and running. Videos are captured in realistic scenarios with complex and cluttered background, and actions exhibit significant intra-class variation. From the provided frame-level annotations, we create a new large bounding volume that circumscribes all of the annotations for a given action cycle. We train the SDPM using these bounding boxes.

Following Lan *et al.*'s experimental methodology [12], we split the dataset into disjoint training and testing sets. For action recognition (not our primary goal), SDPM's forced-choice classification accuracy, averaged over action classes is 75.2%, which is between 73.1% from [12] and 79.4% in [17]. Our recognition results are competitive, considering that we restrict ourselves to HOG-like features and do not employ trajectories or bag-of-words [12, 17]. When SDPM is trained without parts, the recognition ac-



(a) (b) (c)

Figure 5. Direct comparisons on UCF Sports vs. [12, 17]. (a) classification; (b) detection, ROC at overlap threshold of $\theta = 0.2$; (c) detection, AUC for θ from 0.1 to 0.6. The black solid curve shows the average performance of SDPM and the black dotted curve shows the average performance of [12]. Other curves show SDPM results for each action. (Best viewed in color.)

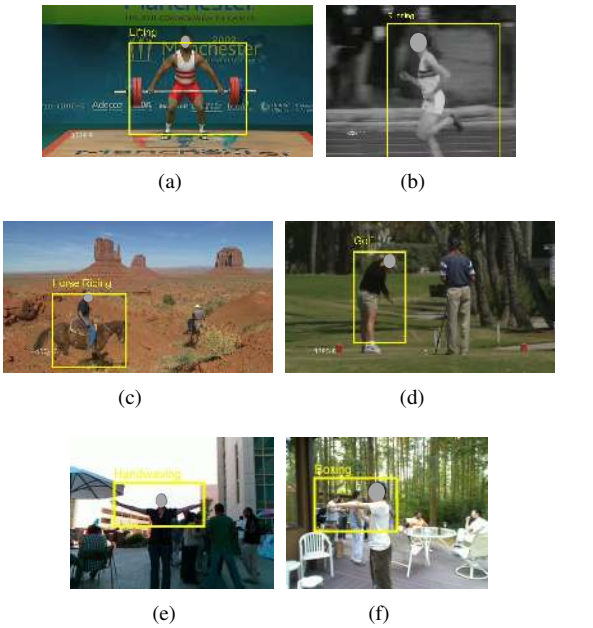
curacy drops to 64.9%; the drop of 10.3% is greater than the 7.6% observed on Weizmann, supporting our hypothesis that parts are more important in complex videos. The per-class classification accuracy comparison among all of these methods is summarized in Fig. 5(a).

We evaluate action localization using the standard "intersection-over-union" measure. Following [12], an action occurrence is counted as correct when the measure exceeds 0.2 and the predicted label matches. Fig. 5(b) shows the ROC curve for overlap score of 0.2; Fig. 5(c) summarizes results (as AUC) for overlap scores ranging from 0.1 to 0.6. In direct comparisons, SDPM clearly outperforms Lan *et al*. [12] on action detection; we are unable to directly compare detection accuracy against Raptis *et al*. [17] because they do not provide bounding-box level evaluations.

Fig. 6 shows several sample detections from UCF Sports and MSR-II datasets in a diverse set of complex scenes.

6.3. Experiments on MSR-II Dataset

MSR-II [3] includes 54 video sequences recorded in crowded and complex scenes, with each video containing several instances of boxing, handclapping and handwaving. Following the cross-dataset paradigm in [3], we train on actions from KTH and test on MSR-II. For each model, the training set consists of a single action cycle from each KTH clip (positives) and instances from the other two classes (negatives). Fig 7 shows a direct comparison[4] between SDPM and Cao *et al*. [3]. Surprisingly, SDPM outperforms [3] even though we perform no explicit domain adaptation. We attribute this robustness to SDPM's ability to capture the intrinsic spatiotemporal structure of actions.



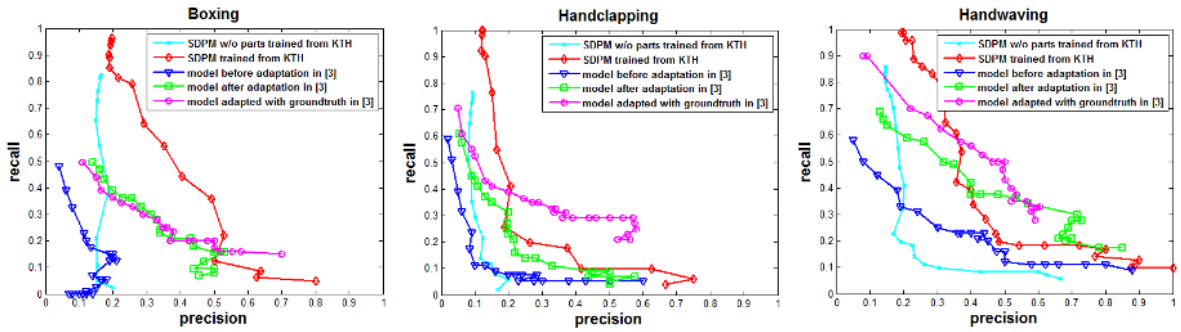
(a) (b) (c) (d) (e) (f)

Figure 6. Detection examples on UCF Sports and MSR-II. (a)–(d) are examples with lifting, running, horse riding and golf SDPMs, respectively. (e) and (f) are examples with handwaving and boxing SDPMs. Actions are detected correctly even in complex scenarios.

7. Conclusion

We present SDPM for action detection by extending deformable part models from 2D images to 3D spatiotemporal volumes. Naive approaches to generalizing DPMs fail because the fraction of discriminative pixels in action volumes is fewer than that in corresponding 2D bounding boxes. We show that SDPM parts are critical, both to focus on important regions in the volume as well as to handle the significant intra-class variation in real-world actions. We are the first to demonstrate perfect recognition and localization results on Weizmann in an unconstrained detection setting and achieve state-of-the-art recognition and localization results on both UCF Sports as well as MSR-II datasets. We conclusively demonstrate that DPMs (when extended correctly) can achieve state-of-the-art results in video, even with simple HOG-like features. A natural direction for

[4]We note that the description of precision and recall in [3] is reversed. In our evaluation, we employ the correct expression.



(a) (b) (c)

Figure 7. Action detection on MSR-II. SDPM outperforms model w/o parts as well as baselines in [3]. Comparison of average precision by SDPM and the best baseline in [3]: 0.3886 vs. 0.1748 (Boxing), 0.2391 vs. 0.1316 (handclapping), 0.4470 vs. 0.2671 (handwaving).

future work would be to integrate the SDPM framework with video-specific features; our open-source implementation will enable others to effectively explore such directions.

Acknowledgments

This research is supported in part by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20071. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

[1] M. Blank et al. Actions as space-time shapes. In *ICCV*, 2005.
[2] W. Brendel and S. Todorovic. Learning spatiotemporal graphs of human activities. In *ICCV*, 2011.
[3] L. Cao, Z. Liu, and T. S. Huang. Cross-dataset action detection. In *CVPR*, 2010.
[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.
[5] A. Fathi and G. Mori. Action recognition by learning mid-level motion features. In *CVPR*, 2008.
[6] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9), 2010.
[7] Y. Hu et al. Action detection in complex scenes with spatial and temporal ambiguities. In *ICCV*, 2009.
[8] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *ICCV*, 2005.
[9] Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. In *ICCV*, 2007.
[10] A. Kläser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3D-gradients. In *BMVC*, 2008.
[11] A. Kläser, M. Marszałek, C. Schmid, and A. Zisserman. Human Focused Action Localization in Video. In *ECCV Workshop on Sign, Gesture, and Activity*, 2010.
[12] T. Lan, Y. Wang, and G. Mori. Discriminative figure-centric models for joint action localization and recognition. In *ICCV*, 2011.
[13] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.
[14] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos "in the wild". In *CVPR*, 2009.
[15] J. C. Niebles, C.-W. Chen, and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *ECCV*, 2010.
[16] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *IJCV*, 79(3), 2008.
[17] M. Raptis, I. Kokkinos, and S. Soatto. Discovering discriminative action parts from mid-level video representations. In *CVPR*, 2012.
[18] M. Rodriguez, J. Ahmed, and M. Shah. Action MACH a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008.
[19] S. Satkin and M. Hebert. Modeling the temporal extent of actions. In *ECCV*, 2010.
[20] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In *ICPR*, 2004.
[21] H. Seo and P. Milanfar. Action recognition from one example. *PAMI*, 33(5), 2011.
[22] D. Tran and A. Sorokin. Human activity recognition with metric learning. In *ECCV*, 2008.
[23] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001.
[24] H. Wang, A. Kläser, C. Schmid, and L. Cheng-Lin. Action Recognition by Dense Trajectories. In *CVPR*, 2011.
[25] H. Wang, M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.
[26] D. Weinland and E. Boyer. Action recognition using exemplar-based embedding. In *CVPR*, 2008.
[27] S.-F. Wong et al. Learning motion categories using both semantic and structural information. In *CVPR*, 2007.
[28] A. Yilmaz and M. Shah. Actions sketch: A novel action representation. In *CVPR*, 2005.