Action Detection by Implicit Intentional Motion Clustering

Wei Chen
CSE, SUNY at Buffalo
wchen23@buffalo.edu

Jason J. Corso
EECS, University of Michigan
jjcorso@eecs.umich.edu

Abstract

Explicitly using human detection and pose estimation has found limited success in action recognition problems. This may be due to the complexity in the articulated motion human exhibit. Yet, we know that action requires an actor and intention. This paper hence seeks to understand the spatiotemporal properties of intentional movement and how to capture such intentional movement without relying on challenging human detection and tracking. We conduct a quantitative analysis of intentional movement, and our findings motivate a new approach for implicit intentional movement extraction that is based on spatiotemporal trajectory clustering by leveraging the properties of intentional movement. The intentional movement clusters are then used as action proposals for detection. Our results on three action detection benchmarks indicate the relevance of focusing on intentional movement for action detection; our method significantly outperforms the state of the art on the challenging MSR-II multi-action video benchmark.

1. Introduction

Action requires an actor; action requires intention; action requires movement [9, 6]. In short, action requires the *intentional movement*, or movement to achieve some active purpose, of an actor, such as a human or animal. Good actor detection and pose estimation can clearly lead to state of the art computer vision systems [29]. Jhuang et al. [15], for example, demonstrate that action-recognition representations built from accurate actor-pose (from ground-truth) outperform low- and middle-level feature-based representations. And, various video understanding problems, such as surveillance [13, 22], video-to-text [16, 8], and group-based activity understanding [18, 20], depend explicitly on detecting the actors or humans in the video.

Yet, in works on individual action understanding like action recognition and action detection, the explicit use of human detection and subsequent processing seems not necessary. The highest performing methods, e.g., Peng et al. [23], do not use any explicit human detection and instead rely on low-level features like dense trajectories [33] or banks of templates [26]. The use of human pose estimation and human detection as an explicit measure for understanding action in video has only minimally been used, e.g., [36, 31, 34]. Why?

Consider action recognition based on human-pose. Jhuang et al's [15] strong results rely on ground-truth pose. When using automatic actor-pose the performance drops or is comparative to non-pose methods: Xu et al. [36] use a bag of pose [37] and achieve weak performance unless fusing the pose-based detector with low-level features, Brendel and Todorovic [3] learn a sparse activity-pose codebook for yielding then-competitive performance and Wang et al. [31] optimize the pose estimation and integrate local-body parts and a holistic pose representation to achieve comparative performance. Neither of these works are evaluated on the larger action recognition datasets like HMDB51 [17].

Human-pose estimation is hard; is performance too weak still? Unfortunately, the picture is similar to the comparatively simpler human detection as with pose estimation for action understanding. Aside from Wang et al. [34] who develop dynamic-poselets for action detection successfully, most works completely ignore human detection or find it underperforms. For example, Chen et al. [6] achieve significantly better performance for ranking action-regions using an ordinal random field model on top of low-level features rather than a DPM-based human detector method [11].

Perhaps the most successful use of human detection in action understanding to date is the improved dense trajectory work [33] in which human detection is used to filter *out* trajectories on human regions when estimating inter-frame homographies. Ironically, in that work, human detection is not directly used to drive the recognition performance.

This thorough evidence suggest that direct use of human detectors and pose-estimators should be avoided for action recognition, at least until pose estimation methods improve. A similar argument could be made for action detection: e.g., both early action detection methods like ST-DPM [30] and recent Tubelets [14], do not use any explicit human detection or tracking. But the evidence is weaker as this is a newer problem.

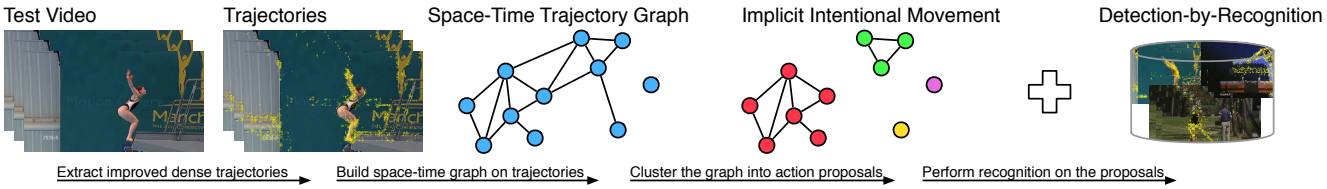


Figure 1. Illustration of our method. Given trajectories in a testing video, the spatio-temporal trajectory graph is used to select action proposals based on our notion of implicit intentional movement. Each cluster on the graph gives rise to an action proposal. The action classifier trained by videos for the action recognition task can be used to achieve action detection on these proposals, in our action detection-by-recognition framework.

Our goals are twofold. First, we seek to understand the role that human detection (whether explicit or implicit) can play in action detection. Second, to improve action detection performance, we seek to leverage the fact that action requires intentional motion [9, 6], which is distinct from the human detection or human mask. For example, various actions, like running, are detectable when only viewing partial information such as running legs or waving hands as in Fig. 4 bottom-center.

We achieve these goals in a systematic fashion. First, we thoroughly quantitatively analyze the properties that dense trajectories [33] exhibit in space-time video regions of explicit intentional motion, i.e., regions where a human is performing an action. We find that trajectories from intentional motion are significantly densely localized in space and time. Second, we propose a method that leverages this finding to compute implicit intentional motion, which is a group of trajectories that obey the properties observed for cases of explicit intentional motion but for which we have not explicitly detected or extracted humans; our method cluster a space-time trajectory graph and then performs action detection-by-recognition on the clusters of this graph (Fig. 1 illustrates this method). Raptis et al. [24] proposed a similar space-time trajectory clustering, but they compute a hierarchical clustering on trajectories to yield action parts and then build detection models based on those parts. In contrast, we leverage our findings of intentional motion to directly cluster on the space-time trajectory graph to yield action proposal clusters. Furthermore, our detection results significantly outperform theirs.

Third, we thoroughly analyze our proposed method as well as a human-detection-and-tracking method on the three recognized action detection benchmarks: UCF Sports [25], sub-J-HMDB [15], MSRII[17]. Our findings suggest that, although explicit human-detection-based action detection has weak performance, our proposed implicit intentional movement-based representation performs comparably or superiorly to the state of the art on all three benchmarks.

We discuss the quantitative study of trajectories for intentional movement in Section 2, our proposed implicit intentional movement-based action detection-by-recognition in Section 3, our experimental findings in Section 4, and discuss related methods in Section 5.

Table 1. Types of Trajectories with respect to the intentional movement bounding box.

Types	Descriptions
AbsPos	All trajectory-points lie in the box.
CenPos	Center trajectory-point lies in the box.
FstPos	First trajectory-point lies in the box.
LstPos	At least one trajectory-point lies in the box.

2. Trajectories and Intentional Movement

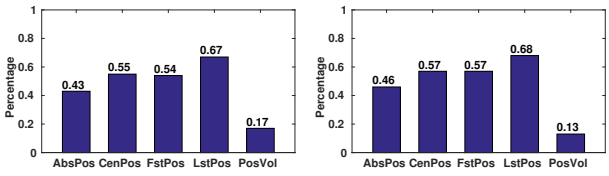
Trajectories are frequently used in action understanding [33] and motion segmentation [4]. However, the relationship between trajectories and intentional motion is unknown. In this section, we systematically quantify this relationship.

Since we are not aware of a dataset that explicitly labels intentional versus non-intentional motion, we use the UCF Sports[1] [25] and sub-J-HMDB[2] [15] datasets, which both have detailed annotations on the locations of the humans. They are action detection datasets and hence we make the assumption that the annotated humans are the acting ones; so, this define our proxy for intentional motion. In order to maintain a consistent labeling, the human masks in the sub-J-HMDB dataset are extended to human bounding boxes as in the UCF Sports dataset. We extract improved dense trajectories [33] without human detection for trajectory extraction on these datasets (using default parameters).

We analyze the spatiotemporal relationship between the trajectories and the intentional motion region. Consider four types of such relationships explained in Table 1. The center point of the trajectory is a virtual point, located at the arithmetic mean of all trajectory points in space-time. Our goal in defining FstPos and CenPos types is to study which point can well represent the spatial and temporal information of a trajectory. FstPos, AbsPos and LstPos types will elucidate how well the articulating human motion is captured by the trajectories. These four types trajectories include intentional motion in different degrees.

For each of the trajectory types, we compute the percentage that each type occupies with respect to the total trajectories in that video. Then, we average those percentages.

[1] http://crcv.ucf.edu/data/UCF_Sports_Action.php
[2] http://jhmdb.is.tue.mpg.de/



(a) UCF Sports (b) Sub-J-HMDB

Figure 2. The percentages of different types of trajectories in UCF Sports and sub-J-HMDB datasets. PosVol indicates the ratio of the positive motion volume to the whole video volume. FstPos, CenPos, AbsPos and LstPos are four types of defined trajectories (Table 1).

Fig. 2 shows these averages for each type of trajectory.

From these statistics, we can summarize several points. First, the trajectories are tracked well in regions of intentional motion. In the UCF Sports dataset, there are 54% FstPos trajectories and 43% AbsPos trajectories. Around 80% of the FstPos trajectories *stick* on the human actor, and roughly 20% of FstPos trajectories drift from human agent to background. There are 67% LstPos trajectories in UCF Sports dataset, which indicates that 28% non-FstPos trajectories drift from the background to the human. This drift is one source of error in a method relying on trajectories to isolate the human prior to an action understanding process. A similar situation is observed on the sub-J-HMDB dataset. Although not intended as such, this quantitative scoring could serve as an evaluation metric for trajectory extraction.

Second, the statistical results of FstPos and CenPos trajectories are similar (less than 1% on both datasets), which implies that when a trajectory is participating in the intentional motion, it is not articulating to a high degree. We note that the CenPos statistics are dependent on the trajectory length as well as the articulation of the motion since the arithmetic mean of the trajectory points need not lie on the trajectory itself.

Third and most importantly, the trajectories extracted by the IDT method [33] include intentional motion information. LstPos trajectories occupy more than 67% of all the trajectories in UCF Sports and sub-J-HMDB datasets. Since these two datasets have only one action per video, the LstPos statistics imply that action and hence intentional motion is the main source of the trajectories.

Finally, we compute PosVol, the ratio of the volume of the actor-acting to the volume of the whole video. Fig. 2 plots this value with respect to the other values indicating that although more than 67% of the trajectories merely touch the action-box, less than 17% of the whole video volume is in the action-box.

Hence, intentional movement is characterized by a high-density of trajectories that, in the majority, will remain a part of the action through their duration. Although this result is intuitive, we have quantified it and in the next section, drive an action proposal approach inspired by it.

3. Method

Our intentional movement analysis provides the rationale for a new action detection approach that seeks to find dense clusters of trajectories in novel videos and then inspect these for actions. Inspired by earlier work in clustering on trajectories for object segmentation [4] and discovering action parts [24], we propose a method that constructs a space-time trajectory graph to capture the interrelations among trajectories, partitions the graph into clusters and then uses action recognition methods on each cluster for detection.

3.1. Space-time Trajectory Graph

For a video $\mathbf{V}$, denote the set of m trajectories extracted from it as $\mathbf{T} = \{T^1, \dots T^m\}$. Each trajectory T^i, lasts n continuous frames and includes n points $\{\mathbf{p}_1^i, \dots \mathbf{p}_n^i\}$ where n is common across all $\mathbf{T}$. The kth point in trajectory i, denoted $\mathbf{p}_k^i$, is a vector $[\mathbf{x}_k^i, f_k^i]^\mathsf{T} \in \mathbb{Z}^3$ indicating the 2D spatial point location $\mathbf{x}$ and the temporal frame f in the video.

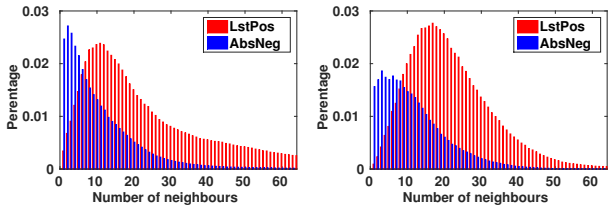
We generate the trajectory graph $\mathbf{G}$ for a video, $\mathbf{G} = \{\mathbf{T}, \mathbf{E}\}$, where trajectory in $\mathbf{T}$ becomes a vertex in this trajectory graph. Edge set $\mathbf{E}$ captures the relationship among pairs of trajectories; we define a novel distance between trajectories to emphasize the spatiotemporal overlap as motivated by our intentional movement analysis. The distance d^{ij} between from trajectory T^i and trajectory T^j is defined as

$$d^{ij} = \begin{cases} \sum_k \|\mathbf{x}_k^i - \mathbf{x}_{k-o^{ij}}^j\|_2 & 0 \le o^{ij} < n \\ \infty & \text{otherwise} \end{cases} \quad (1)$$

where k is the index of the point in a trajectory and the offset o^{ij} is computed once for the pair by

$$o^{ij} = \min_{\hat{o}} \left(n - \sum_k \mathbb{1}\left[f_k^i = f_{k-o}^j \right] \right) \quad (2)$$

where $\mathbb{1}$ is the indicator function. In other words, we compute the distance only for those frames that are overlapping in time. Distance among other frames for these two trajectories is irrelevant. The overlap o captures the temporal relationship among the two trajectories and is retained. For example, if o is 0, then the trajectories exactly align in time; if o is n it indicates zero overlap in time. $n - o$ is the magnitude of their overlap. This distance may seem awkward as the distance will be greater for trajectories with more overlapping frames; however, before we finally cluster on this



(a) UCF Sports (b) Sub-J-HMDB

Figure 3. The empirical cardinality distributions of LstPos trajectories, which touch the intentional motion region, and AbsNeg trajectories, which are the complement of LstPos and do not touch the intentional motion region. Since the cardinality distribution is long-tailed, we count the number of neighbors up to 1024 in UCF Sports and 512 in sub-J-HMDB datasets. When computing the distributions, the bin sizes for UCF Sports and sub-J-HMDB datasets are 16 and 8. The spatial threshold τ_s is 16 pixels and the temporal threshold τ_t is 1 frame.

graph, we will explicitly account for this (Eq. 4). Furthermore, our distance function is unique among similar methods [24, 4] as we temporally align the trajectories and then compute spatial distance only for aligned frames whereas other methods directly compute full trajectory distance.

To enforce the spatiotemporal density expected and observed in intentional motion, we add edges among trajectory pairs based on their spatiotemporal locality. First, candidate pairs are filtered based on their temporal overlap by enforcing a threshold τ_t of the minimum number of overlapping frames. Second, candidate pairs are filtered based on the mean spatial distance in their overlapping frames against threshold τ_s. Finally, the edge set $\mathbf{E}$ is defined as

$$\mathbf{E} = \left\{ E^{ij} : o^{ij} > \tau_t \ \wedge \ \frac{d^{ij}}{n - o^{ij}} < \tau_s \right\} . \qquad (3)$$

We now verify the spatiotemporal locality and density of our graph continuing the analysis from the previous Section 2 by comparing the cardinality of trajectories from the LstPos set and its complement, which we denote AbsNeg. Recall, LstPos trajectories are those that touch the intentional motion region in at least one frame. Figure 3 shows the results on the UCF Sports and sub-J-HMDB datasets. Clearly the modes for these two distributions are different verifying our implicit intentional motion approach. In both cases, the cardinality mode for LstPos trajectories more than doubles that for AbsNeg trajectories. However, we note the high degree of overlap among these two distributions; the Bayes error rate would be high if we simply tried to distinguish the trajectories based on their cardinality.

3.2. Implicit Intentional Motion Clustering

The space-time trajectory graph has been constructed to implicitly capture intentional motion. Now, we execute spectral clustering on the graph to partition it into intentional motion clusters which then serve as action proposals.

We convert the graph $\mathbf{G}$ into a similarity matrix $\mathbf{S}$ using the edge set $\mathbf{E}$ as follows. First, initialize $\mathbf{S}$ to be zeros. Then, for each edge $E^{ij} \in \mathbf{E}$, similarity S^{ij} is

$$S^{ij} = \exp\left(-\frac{d^{ij} + (o\gamma)^2}{n^2 \sigma} \right) , \qquad (4)$$

where σ is a scale parameter, which is automatically-tuned by the spectral clustering method, and γ is the fixed distance for the trajectory-points that are not aligned in time. The $o\gamma$ term accounts for the unaligned points remaining after the temporal alignment in Eq. 2. Recall that o indicates the number of frames two trajectories are out of alignment.

Our similarity explicitly incorporates the locations of points in trajectory and their distance. The length of trajectory and the velocity of each point in trajectory is implicitly considered since their computation is directly related to the position of trajectories. In this way, we avoid the noise amplification from tracking trajectories. For example, if the velocities for two trajectories are similar, then the distance between points will be nearly constant.

We use the efficient spectral clustering method from Chen et al. [7] to compute the eigen-decomposition of the normalized Laplacian on $\mathbf{S}$ and then subsequently cluster the projected trajectories with k-means as is standard. Despite tens of thousands of trajectories per video, we find the clustering to be fast and run in less than ten minutes per video with eight cores.

3.3. Action Detection-by-Recognition

Given the construction of the space-time trajectory graph, we expect each cluster to contain a candidate action. The clusters, by construction, are connected sets of trajectories. Our observations on the characteristics of intentional motion suggest that the spatiotemporal density of trajectories involved in the action and the way we have computed distance (Eq. 1) will lead to compact clusters of (implicit) intentional movement.

To evaluate the action detection, we run a standard action recognition method on each proposal. We use a nonlinear SVM action classifiers with RBF-χ^2 kernel [32] on trajectory displacement features (a weak feature). Our training process for the action classifiers is different than convention. Since we rely on the implicit intentional motion clustering in our work, we discard the bounding boxes in the training videos. Although we could directly use the bounding boxes, we make a weaker assumption: each training video has a single action within it. We hence use trajectories from the dominant action proposal (clustering is unsupervised) for positive or negative samples depending on whether the video is positive or negative, respectively. The whole training process is hence weakly supervised and bounding box free; our assumption is weaker than comparable state of the art methods, which rely on bounding boxes.

4. Experiments

We extensively evaluate our method to achieve out stated goals of assessing the role that human detection and intentional motion detection can play in action detection. We use three standard action detection benchmarks: UCF Sports [25], sub-J-HMDB [15] and MSR-II [5] datasets.

UCF Sports [25] comprises 150 realistic videos captured in dynamic and cluttered environments from sports broadcasts and 10 action categories with one action per video. The UCF Sports dataset provides the bounding boxes of the human actors.

Sub-J-HMDB [15] is a subset of HMDB51 dataset [17] containing 316 videos from 12 categories with one action per video. This dataset provides the human actor masks.

MSR-II [5] consists of 54 videos recorded in a crowded environment, with many unrelated objects (people, cars) moving in the background. There are three types of actions in the dataset, boxing, hand-clapping and hand-waving, but each video may contain many instances of the actions. Bounding sub-volumes of action are provided in the ground truth. Another challenge for this dataset is that all the positive training videos come from KTH dataset [27].

Comparisons We compare our proposed implicit intentional motion-based action detection against the state of the art methods for action detection [30, 19, 34, 14, 5, 33], against the methodologically similar Raptis et al. [24], and against a human detection-based baseline. The human detection-based baseline uses DPM [11] to detect humans in each video-frame. Then, it links together these detections in time based on common trajectories they share as a means of tracking; detection with no trajectories and detections less than five frames are discarded. Each DPM-based space-time set of detections forms an action proposal against which we evaluate our action classifier (for a fair comparison varying only the action proposal mechanism). This baseline assesses the question of whether direct human detection should be used in action detection.

Visual Results Fig. 4 shows visual action detection results from all three datasets using our method.

4.1. One Action Per Video

In the UCF Sports and sub-J-HMDB datasets, there is one action per video. So, we expect the largest cluster from the implicit intentional clustering to be this action and extract it as the sole action proposal. Here, the spatial threshold τ_s is set to 16 pixels, and the temporal threshold τ_t is set to 8 frames.

Category-Independent Action Detection Fig. 5 shows the quantitative comparisons of our method against baselines. In addition to the DPM baseline, in this case, we also use the moving background subtraction (MBS) method from Shiekh et al. [28]. The foreground is assumed as the action in the video. The IOU scores for the detections are generated and the overlap threshold varies from 0.1 to 0.6. On both datasets, our method is above 0.9 when the threshold is 0.1. Because of the limited duration and the complex

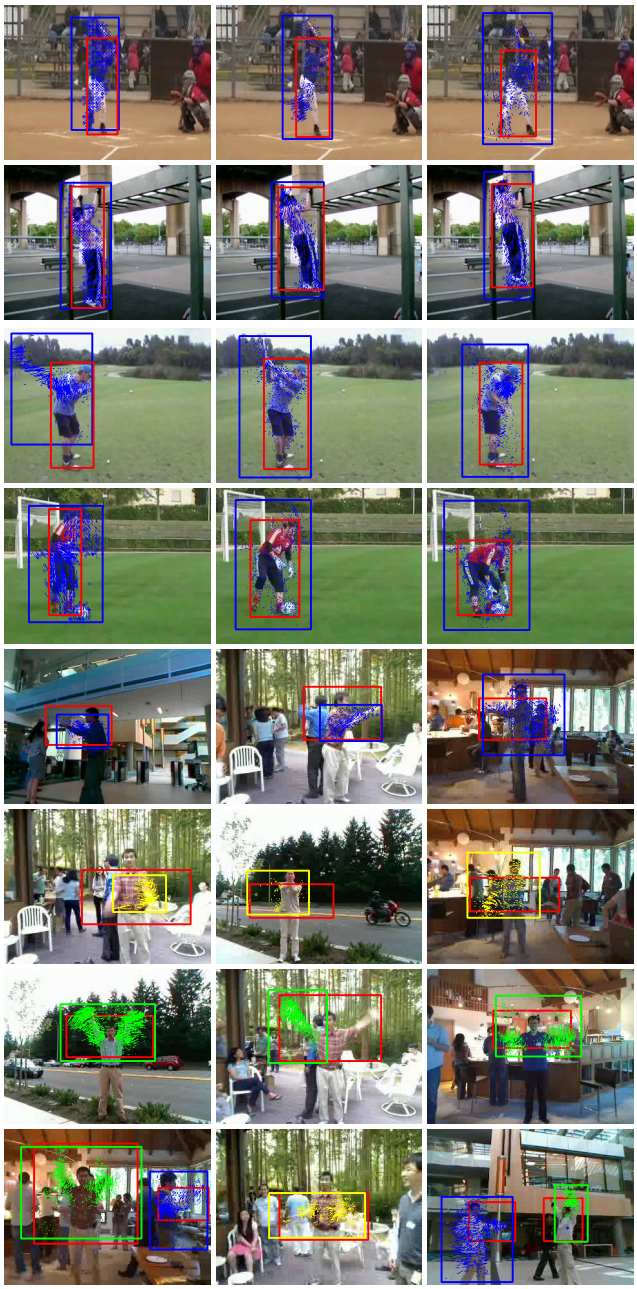
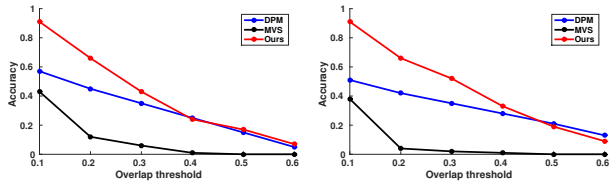
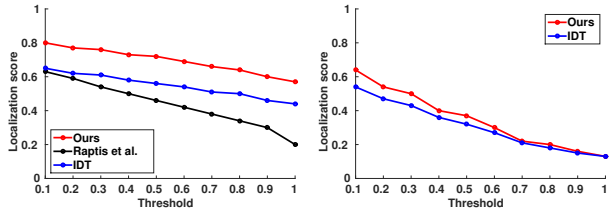


Figure 4. Examples of action detection on three datasets. Red color indicates groud-truth. The first 4 rows are for UCF Sports and sub-J-HMDB datasets. The detection results are labelled in blue bounding boxes. The last 4 rows are for MSRII dataset. Blue indicates boxing, yellow indicates hand-clapping and green indicates hand-waving.



(a) UCF Sports (b) Sub-J-HMDB

Figure 5. Comparative unsupervised action detection on UCF Sports and sub-J-HMDB datasets against baselines.



(a) UCF Sports (b) Sub-J-HMDB

Figure 6. The average localization scores for the trajectory clusters on UCF Sports and -sub-J-HMDB datasets.

motion in the video, MBS does not perform well, even when the threshold is 0.1. DPM performs better than MBS: actor detection is useful for action detection. But the DPM-based method is worse than our method likely due to the its limited ability to deal with pose variability.

Trajectory Localization We evaluate the relevance of the selected trajectories to the action. As in Raptis et al. [24], each point in a trajectory can be classified as a positive or negative point based on whether it is located in the action bounding box. Then each trajectory can be assigned a length-normalized score. We count the average number of trajectories that have length-normalized overlap with the bounding box higher than a threshold. The results are in Fig. 6. The trajectories extracted by IDT [33] have higher quality than the method in [24], but our method performs better than IDT.

Interest Action Detection Whereas Fig. 5 depicts category-independent action detection, here, we consider joint action detection and classification, often called *interest action detection*. Here, we use those action proposals from our implicit intentional motion clustering that contain more than 150 trajectories. We follow the experimental setting and the criterion for measurement from Lan et al. [19]. An instance is considered as detected if the action is correctly predicted by the classifier and also the detection score is larger than the threshold based on the IOU criterion. We compute the AUC with false positive rates varying from 0 to 0.6.

The performance of our method on UCF Sports is shown in Fig. 7 with the disjoint train-test split suggested in [19]. We compare our approach with several recently published methods, spatiotemporal deformable part model (SDPM) [30], figure-centric model (FC) [19], Relational Dynamic-Poselets model (RDP) [34] and Tubelets [14]. The performance of RDP, Tubelets and our method are similar with variation in ranking as a function of the overlap threshold.

The details of our method on each action class is shown in the middle. Our method achieves a high detection rate for many action classes, such as lifting, running, walking, swing-bench and skating. There is a significant difference between our method and other methods. For most other methods, the performance decreases with respect to the increasing of overlap threshold, while our method increases in several cases. This is because our classifier comes from action recognition method, a cluster with large overlap with the ground truth has more chance to be correctly classified. We show ROC curves of different methods in the right of Fig. 7. The classifiers in our method have space to improve, especially when the false positive rate is small, and we note that we strictly use the trajectory displacement feature for the classifier in this work.

The performance of our method on J-HMDB dataset is shown in Fig 7 with the default data split provided in the dataset. The average AUC and AUC-per-class are shown in the left and middle part of the figure. Our method performs better than IDT+FV and RDP methods in most cases. The ROC curves of these methods are shown in the right. Although our action classifier is simpler than that in IDT+FV, our method is significant better than IDT+FV method, which comes from the higher quality action proposal (IDT+FV simply uses a sliding window strategy).

4.2. Multiple Actions Per Video

The MSR-II dataset has several different kinds of action in each video and is clearly the more challenging benchmark. Furthermore, the training and testing are on different datasets. We follow the measurement criteria in [5] for evaluation. Considering that the resolution of KTH and MSR-II is low, when extracting the trajectories, the minimum distance between sampling points for trajectory extraction is set as 3 pixels.

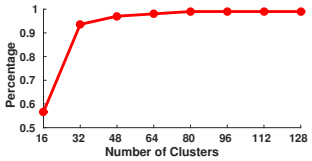


Figure 8. MSR-II ground-truth actions our action proposals cover.

Given the multiple actions per video, we need to consider more than one *best* action proposal from our clustering process. In total, there are 203 actions in all MSR-II videos, the relationship between the number of our action proposals (clusters) per video and the action coverage is shown in Fig. 8. As expected, with more proposals, more actions will be covered in our method. But, no single number of proposals achieves best performance across the whole dataset because the number

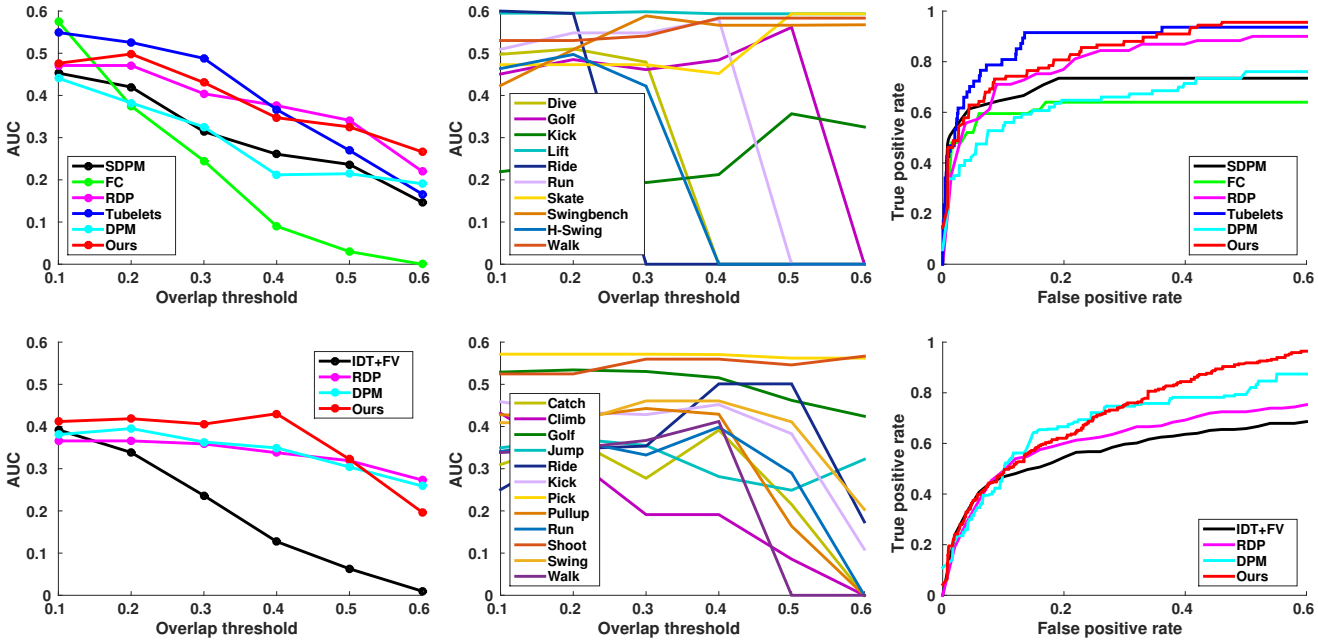


Figure 7. Results on the UCF Sports (top) and sub-J-HMDB (bottom) datasets. Left: We compare our method with other state of the art methods and the DPM baseline method. Center: We plot the AUC per class of our detection result with a varying overlap thresholds. Right: ROC curves of all the methods are shown here, when the overlap threshold is set as 0.2.

of proposals is related to spatiotemporal scale. Therefore, we combine proposals from multiple clustering processes together (ranging from 16 to 128 clusters) to ensure we can cover various spatiotemporal scales.

The threshold for the action classifier affects the performance of our method. Fig. 9 shows precision-recall curves when varying the threshold. The AUC under precision-recall is shown in Fig. 9(a). Increasing the threshold leads to better performance of our method.

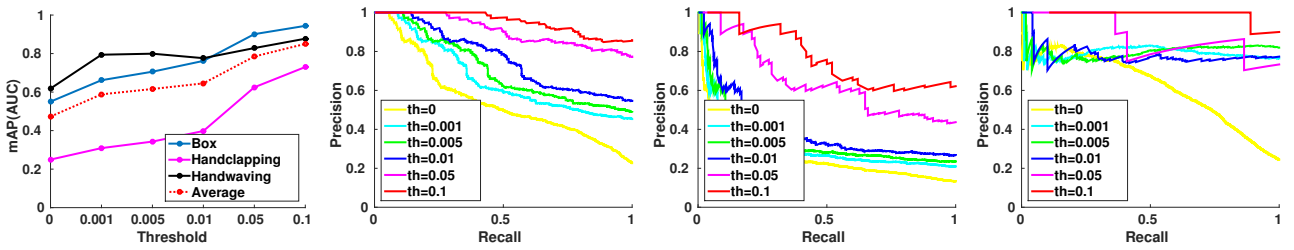
We compare the performance of our method to the state of the art in Table 2, using a threshold of 0.1. Our method achieves significantly better performance on boxing and hand-waving actions and comparable best performance on the hand-clapping action. According to the Fig. 9, when the threshold of the classifiers is larger than 0.01, the average performance of our method is better than all the other methods, which demonstrates the generalization ability of our method for the action detection task. The precision-recall curves is shown in Fig. 10.

Failure Modes We have identified two distinct failure modes of our method. The first failure mode is due to the strictly motion-based representation of our method; we two actions occur nearby in space and/or time, it is difficult for our method to distinguish them from one another. The second failure mode is due to our implicit intentional motion assumptions: when the IDT trajectories violate these assumptions, then our method fails to make the correct action proposals.

5. Related Work

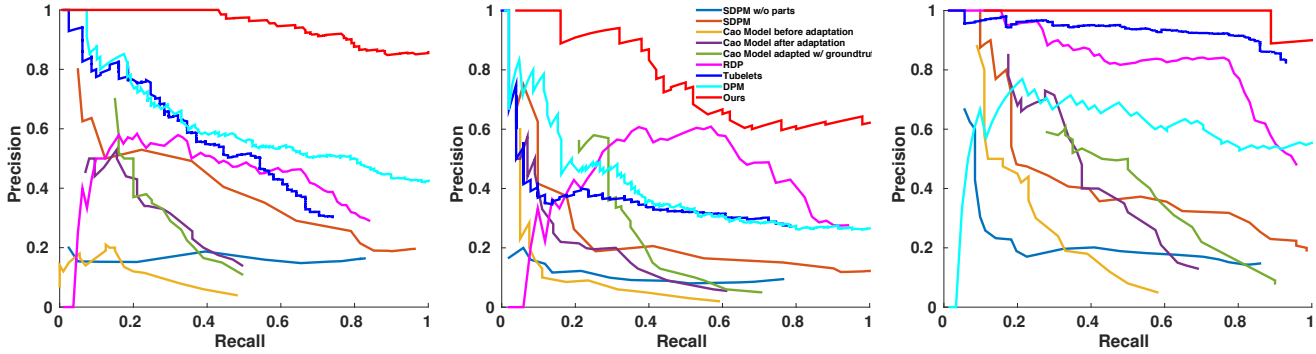
Action recognition has been extensively studied in recent years [1, 35, 13]. This section only covers the works related to the action detection task. Early work in action detection focuses on motion information from human labeling for action modeling. The template-based methods manually chose templates [2, 10, 12] and apply templates to exhaustively search for the action in the video. The different strategies [38, 21] for combining templates are designed to deal with the variations of the action. Relying on sliding window templates that are often manually chosen limits the potential of these methods in contrast to our approach that directly seeks action proposals by finding intentional movement. To overcome this single template approach, the space-time deformable parts model [30] automatically learns sets of parts that can move in space-time. These can clearly capture greater degrees of space-time articulation. But, the latent SVM used requires significant computation and data.

Recently, human pose information as a middle level representation [34] has been applied for action detection, and achieves good performance. Inspired by the unsupervised object proposal in still images, the action proposals are generated by an extension of the hierarchical clustering method based on video segmentation [14]. This direction is more closely related to our approach as it directly incorporates



(a) Average AUC (b) Boxing (c) Handclapping (d) Handwaving

Figure 9. Performance of our method on MSR-II dataset with respect to the variation of classifier threshold.



(a) Boxing (b) Handclapping (c) Handwaving

Figure 10. Results on the MSR-II dataset. We plot the PR curves for the three action classes: boxing, hand-clapping, and hand-waving. We compare our results with GMM methods with or without adaption [5], SDPM [30], RDP [34] and Tubelets [14]. Best viewed in color.

Table 2. Average precision for MSR-II

Method	Boxing	Handclapping	Handwaving
Cao et al.	17.5	13.2	26.7
SDPM	38.9	23.9	44.4
Tubelets	46.0	31.4	85.8
RDP	41.7	50.2	80.9
DPM	60.5	39.5	59.5
Ours	**94.4**	**73.0**	**87.7**

human pose motion, but the limited accuracy in human pose estimation is the primary concern for this method whereas our approach does not require explicit human pose estimation but rather implicitly focuses on intentional motion.

Our work is also related to motion segmentation [4, 24] for object detection. Just like the importance of the similarity definition for clustering methods, an effective similarity in spatial and temporal domains for trajectories is needed. But, these methods have not focused on actions explicitly; we push the direction further into action.

6. Conclusion

We have systematically assessed the role that human detection can play in action detection and found that explicitly incorporating human detection performs worse than implicitly incorporating information about the intentional movement of the acting human. This implicit intentional movement is the technical contribution of our work. We quantified the relationship between intentional movement and the spatiotemporal statistics of trajectories within the action region. Based on our findings, we developed a new trajectory distance and clustering method that, when coupled with a simple action classifier, achieves state of the art performance on challenging action detection benchmarks.

Ultimately, our finding is that implicitly incorporating information about the acting human by way of customizing trajectory clustering to seek intentional movement action proposals leads to state of the art performance. Furthermore, our proposed method is only weakly supervised and is bounding box free. In contrast, the other methods against which we compared all directly use the bounding boxes. In the future, we plan to focus on further properties of intentional movement leading to better extraction as well as improving the classifiers underlying our action detection-by-classification method.

Acknowledgement

This work was partially supported by the National Science Foundation CAREER grant (IIS-0845282) and the Army Research Office grant (W911NF-15-1-0354).

References

[1] J. Aggarwal and M. Ryoo. Human activity analysis: A review. *ACM Computing Surveys*, 43(16), 2011. 7

[2] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. In *TPAMI*, 2001. 7

[3] W. Brendel and S. Todorovic. Activities as time series of human postures. In *ECCV*, 2010. 1

[4] T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*, 2010. 2, 3, 4, 8

[5] L. Cao, Z. Liu, and T. S. Huang. Cross-dataset action detection. In *CVPR*, 2010. 5, 6, 8

[6] W. Chen, C. Xiong, R. Xu, and J. J. Corso. Actionness ranking with lattice conditional ordinal random fields. In *CVPR*, 2014. 1, 2

[7] W.-Y. Chen, Y. Song, H. Bai, C.-J. Lin, and E. Y. Chang. Parallel spectral clustering in distributed systems. *TPAMI*, 33(3):568–586, 2011. 4

[8] P. Das, C. Xu, R. F. Doell, and J. J. Corso. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *CVPR*, 2013. 1

[9] D. Davidson. Actions, reasons and causes (1963). In *Essays on Actions and Events*. Clarendon Press, Oxford, 2001. 1, 2

[10] K. G. Derpanis, M. Sizintsev, K. Cannons, and R. P. Wildes. Efficient action spotting based on a spacetime oriented structure representation. In *CVPR*, 2010. 7

[11] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *TPAMI*, 2010. 1, 5

[12] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *TPAMI*, 29(2247-2253), 2007. 7

[13] W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 2004. 1, 7

[14] M. Jain, J. van Gemert, H. Jegou, P. Bouthemy, and C. Snoek. Action localization by tubelets from motion. In *CVPR*, 2014. 1, 5, 6, 7, 8

[15] H. Jhuang, J. Gall, J. Zuff, C. Schmid, and M. J. Black. Towards understanding action recognition. In *ICCV*, 2013. 1, 2, 5

[16] N. Krishnamoorthy, G. Malkarnenkar, R. J. Mooney, K. Saenko, and S. Guadarrama. Generating natural-language video descriptions using text-mined knowledge. In *AAAI*, 2013. 1

[17] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In *ICCV*, 2011. 1, 2, 5

[18] T. Lan, L. Sigal, and G. Mori. Social roles in hierarchical models for human activity recognition. In *CVPR*, 2012. 1

[19] T. Lan, Y. Wang, and G. Mori. Discriminative figure-centric models for joint action localization and recognition. In *ICCV*, 2011. 5, 6

[20] V. I. Morariu and L. S. Davis. Multi-agent event recognition in structured scenarios. In *CVPR*, 2011. 1

[21] J. C. Niebles, C.-W. Chen, and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *ECCV*, 2010. 7

[22] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis, et al. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR*, pages 3153–3160. IEEE, 2011. 1

[23] X. Peng, C. Zou, Y. Qiao, and Q. Peng. Action recognition with stacked fisher vectors. In *ECCV*, 2014. 1

[24] M. Raptis, I. Kokkinos, and S. Soatto. Discovering discriminative action parts from mid-level video representations. In *CVPR*, 2012. 2, 3, 4, 5, 6, 8

[25] M. Rodriguez, J. Ahmed, and M. Shah. Action mach: A spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008. 2, 5

[26] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In *CVPR*, 2012. 1

[27] C. Schüldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In *ICPR*, 2004. 5

[28] Y. Sheikh, O. Javed, and T. Kanade. Background subtraction for freely moving cameras. In *ICCV*, 2009. 5

[29] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, 2011. 1

[30] Y. Tian, R. Sukthankar, and M. Shah. Spatiotemporal deformable part models for action detection. In *CVPR*, 2013. 1, 5, 6, 7, 8

[31] C. Wang, Y. Wang, and A. L. Yuille. An approach to pose-based action recognition. In *CVPR*, 2013. 1

[32] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*, 2011. 4

[33] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, 2013. 1, 2, 3, 5, 6

[34] L. Wang, Y. Qiao, and X. Tang. Video action detection with relational dynamic-poselets. In *ECCV*, 2014. 1, 5, 6, 7, 8

[35] D. Weinland, R. Ronfard, and E. Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *CVIU*, 2011. 7

[36] R. Xu, P. Agarwal, S. Kumar, V. N. Krovi, and C. J. J. Combining skeletal pose with local motion for human activity recognition. In *Proceedings of VII Conference on Articulated Motion and Deformable Objects*, 2012. 1

[37] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011. 1

[38] B. Yao and S.-C. Zhu. Learning deformable action templates from cluttered videos. In *ICCV*, 2009. 7