Actionness-assisted Recognition of Actions

Ye Luo Loong-Fah Cheong An Tran
National University of Singapore

Abstract

We elicit from a fundamental definition of action low-level attributes that can reveal agency and intentionality. These descriptors are mainly trajectory-based, measuring sudden changes, temporal synchrony, and repetitiveness. The actionness map can be used to localize actions in a way that is generic across action and agent types. Furthermore, it also groups interacting regions into a useful unit of analysis, which is crucial for recognition of actions involving interactions. We then implement an actionness-driven pooling scheme to improve action recognition performance. Experimental results on various datasets show the advantages of our method on action detection and action recognition comparing with other state-of-the-art methods.

1. Introduction

In this paper, we wish to inquire what principles can be employed to extract actions in a way agnostic to the underlying agent as well as the type of action itself. In other words, we hope to discover low-level cues that are hallmarks of all biological motions. Oddly enough, little works in the action recognition area seek such precategorical cues. To us, it is an interesting question to ask, because it appears that human has special sensitivity to the perceptual category of biological motions in comparison to non-biological motions, even though these category boundaries are unlikely to be sharply defined. We also feel that the very notion of action is so intimately connected with action recognition research that it will be no digression to revisit the definition of action, with a view to discover some attributes which are capable of distinguishing actions.

1.1. The Definitions of Action

If one consults what philosophers commonly understand by action, one may elicit the following definition of action — an intentional movement made by an agent. While one may take issue with such a simplistic definition of action[1], we are more concerned about the relevance of this definition as far as discovering useful attributes of actions for computer vision is concerned. The first thing to note in this definition is that "movement" can be interpreted very loosely, so that it can include "movement" like standing fast and mental acts like computing. We do not plan to tackle these troublesome types of "movement" and so by movement, we restrict ourselves to the bodily types of movements. Even setting aside the non-bodily types of movements, we are still faced with difficulties as far as effecting some litmus test of intention is concerned. In general, there is no reason to expect that merely knowing some low-level trait in a video say the trajectory, one can tell whether or not a movement is intentioned. Consider these two action categories in

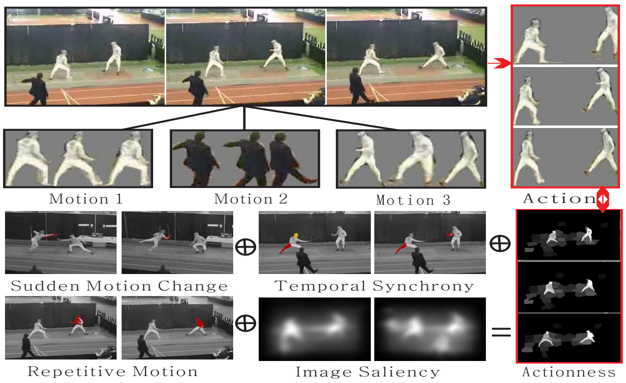

Figure 1. Top half illustrates the movements present in a fencing sequence and which ensemble of motions should be understood as the fencing action. The bottom half illustrates how the proposed actionness attributes contribute to the action extraction. From left to right, the trajectories highlighted in red show respectively those with significant sudden change in motion, repetitive motion and temporal synchrony, with the final panel on the right showing the fused actionness map. Note that the two fencers have been grouped together into a useful unit for further analysis. Best viewed in color.

[1] There are other philosophical difficulties that we do not consider in this paper; for instance, should we draw distinction between the primitive actions (e.g. throwing a stone) and the actions which we refer to as consequences (e.g. breaking a window), or should we puff out the primitive action to include the effect caused by the basic act? This difficulty is also inextricably bound up with the problems caused by the notion of causality. Interested readers are referred to [10] for a more detailed discussion. While considering the problem at this level of abstraction is an important philosophical project, thankfully for us, most of the action datasets in the computer vision community seem to deal with clearly primitive actions.

the HMDB51 dataset [19]: falling to the floor and smiling (other examples not in HMDB51 abound: coughing, knocking over someone, blinking, etc.). All the aforementioned acts might be done spontaneously or involuntarily; there is no simple way to know one from the other from low-level features. Indeed, with the aforementioned definition, falling to the floor is normally not an action (and thus should not be included in the HMDB51 dataset), unless it is done intentionally.

Despite the aforementioned exceptions and difficulties in imputing intention, we will adopt in this paper the above definition, and in what follows, we will show there exist good tests of agency. Before that, we must spend a few moments on the current state of action recognition research which has prompted our research.

1.2. Background and Motivation

Bag of features (BoF) together with its extension like hierarchical spatial pyramid structure is a general paradigm which is used in many classification problems. One of the well-documented drawbacks of such approach for action recognition is that simple spatial arrangement such as grids (or the complete lack thereof) does not contain information about the pertinent structure of the various actions to be recognized. In action recognition, indeed, the spatio-temporal location of the meaningful actions varies drastically within each category. Spatial locations are not inherently meaningful. This makes the subsequent mid-level representation, even after pooling, not distinct enough to classify actions correctly to categories. In a related development, there are several recent works [30] that questioned the meaning of the results obtained by many action recognition algorithms based on the aforementioned paradigm. Have we really learnt the essence of the action in these works, or are we learning the background? When faced with a more complex dataset like HMDB51 [19] with increasing varieties in the background, performance of many algorithms dropped significantly.

In an attempt to address this issue, a saliency-driven perceptual pooling method [2] has been proposed to capture structural properties of the actions, independently from their position in the image. While this has been shown to work better in complex dataset like HMDB51, we contend that there is still significant inadequacy in adopting these conventional notions of saliency for the domain of action recognition. The dominant paradigm in saliency research is still that of picture viewing in which the interpretation of the perceiver is supposed to be neutral. Even if we extend the saliency feature channel with optic flow (like in various saliency works dealing with video), it is still basically a series of optic flows between two frames strung together. The perceiver's interpretation will be significantly different when he or she is "thrown" (in the sense of Martin Heidegger [15]) into a natural dynamic settings. As an embodied agent situated in and engaged with the world, we are highly sensitive to various aspects of actions executed by other agents. At the most basic level, all animals are probably highly sensitive to the difference between animate and inanimate motions, given the importance of the distinction of agents versus non-agents to their survival. At a finer level, they would also be sensitive to the intent of these moving agents or meaning of these actions. In sum, current conspicuity-based saliency models lack explanatory power for the aforementioned dynamic and situated aspects of saliency. This in turn motivates us to look for cues that can capture these aspects.

1.3. Actionness

We now show that there exist low level attributes — generic over action classes — that can make salient actions stand out against the background and we term these the "actionness" attributes. These attributes are based on information residing in a single trajectory that is related to the distinction between agents and non-agents, as well as the degree of temporal synchrony between different trajectories that is characteristic of many animate motions; details will be provided in the following paragraphs about these intra- and inter-trajectory cues. Based on these actionness values, we can then create an actionness-driven pooling scheme that is more robust to space-time variance of actions, as the actionness-based content has a more intrinsic relationship with the semantics of the actions. Another significant advantage is that our temporal synchrony cue naturally highlights the spatiotemporal regions that belong together as a unit of analysis. This is important for many human actions that involve similar dynamics but have different meanings depending on the objects being manipulated. Being able to bind the object to the agent manipulating it is important in differentiating these closely-related actions. In many interaction scenarios, the different regions that form a useful unit of analysis might indeed be a mosaic of non-contiguous regions with varying characteristics and thus difficult to bind together, e.g. a forward and a defender (wearing different jerseys) in a one-on-one situation in a football game. Again, our actionness attributes will be able to pick out these players engaged in close interaction. In this sense, these actionness cues can be viewed as a kind of perceptual grouping cues, the difference with the classical perceptual grouping cues being that here the perceptual unit to be united together is not object per se, but action, and it takes place in time, not in space.

Let us consider a concrete example of action: that of handshaking. Note that to recognize the proper meaning of the extended hand, we need to discover the pertinent interacting units, i.e. who or what is the target of the extended hand? Only when we are able to discern the two hands extending towards one another as one perceptual unit, the meaning of the hand gesture can be properly interpreted. More generally, the meaning of an action appears not only in the form and motion of individual agent but in the spatial and temporal relationships between the agent and other agents or objects; i.e. these relationships form a higher-order structure which carries the meaning. Thus it is important to group the units involved in the action. Note that there may be large intervening background regions separating the two hands, i.e. the hands may not be close to one another initially, so the grouping cannot be based on proximity. As one of the actionness attributes mentioned above, temporal synchrony between the two extending hands can be used as a perceptual grouping cue to group them as a useful perceptual unit. Also note that it is not necessary for the entire object (the entire human body in this example) to be involved. Arguably, the key unit here should involve only the two extending hands approaching one another; other parts of the body might yield some supporting cues but they are not absolutely essential; they might in fact be involved in other composite actions (e.g. the two persons involved might be standing or walking towards one another, or even executing other composite actions like eating a snack), even though one can argue that these composite actions can also provide some useful contextual cues.

We now consider how the rather abstract defining characteristics of actions (such as agency and intention) quoted above can manifest themselves in low-level features. [13] gave a good summary of cues that have been considered in the visual psychophysics community: sudden direction and speed change, rational interactions with spatial contexts and other objects, apparent violations of Newtonian mechanics, and coordinated orientations. For this work, we wish to eschew the use of high-level semantics and non-visual cue such as gravity direction. Thus, we only consider the following cues as the desired characteristics of actionness:

1. Trajectory showing *sudden direction and speed change*; this is characteristic of ballistic movements involving impulsive propulsion of the limbs, such as striking and kicking.

2. Trajectory showing *repetitive motions* sustained over a period of time; this is characteristic of mass-spring movements of a more sustained character, like walking and running.

3. Trajectory exhibiting *temporal synchrony* with other trajectories. These well-synchronized movements can be between various body parts of the same person (e.g. in diving action), between different persons (e.g. interactions like embracing, shaking hands), or between person and object (e.g. hand and cup in drinking action). Such coordinated movements often indicate a sense of purpose and intention (e.g. in diving), and a sense of cause and effect (cup lifted up by hand in drinking action).

4. Trajectory that is associated with a *salient region* in the conventional static sense. This is because action is executed by agents and often on objects, not some amorphous background stuff (sky, ground, water etc).

Note that these criteria are not mutually exclusive. In the second and third criteria, we also favour trajectories that exhibit significant changes in directions, as this indicates self-propelledness rather than objects moving under external forces such as gravity. For instance, two features or regions moving in perfectly synchronous linear motion may not be that interesting. Fig. 1 summarizes the ideas of our framework, using the action fencing as an example.

Once the actionness map is extracted, it can be used to detect actions in a way that is generic across action and agent types. Furthermore, it also groups interacting regions together as a more useful unit for pooling; this is crucial for recognition of actions involving interactions, as will be demonstrated experimentally later.

2. Related Work

In this section, we review works that are related from various perspectives.

As remarked by [9], there has been a lack of attention paid to the very notion of action itself in the action recognition community. The definition of action in high level terms is not difficult to secure, as it has been discussed thoroughly in the philosophical community; for instance, [10] defined it as intentional biological movement. However, from a computer vision point of view, it is much more difficult to secure the concrete operational steps that can capture the essence of action from these high level notions. For instance, [9] defines actionness entirely in high-level terms and assumes that the notion of action can be defined implicitly by annotated examples in a dataset used to train an actionness classifier. Other recent action localization works [1, 6, 7, 12, 16, 18, 20, 26, 33, 37, 38] are similar in that they are ambivalent about what exactly defines action and require access to manually annotated training video set. In contrast, we explicitly posit that actionness can be defined in terms of low-level operations; the strength of our approach is corroborated by the much better action detection results obtained.

The discriminative parts discovery approach does not look for the action regions per se but any ensemble of regions that could hopefully discriminate actions into classes. Methods such as [20, 4, 36] try to learn a detector for these regions but they require accurately annotated sub-regions of every video (e.g. parts of the human bodies, a whole person, etc.), a tedious and troublesome process. Weakly supervised methods [5, 17, 29] reduce the workload of annotation, requiring only the action label for every frame or every video instead of every regions, but they usually need extra process to mine for the discriminative regions. In contrast, our action detection step is unsupervised. There are other unsupervised works [30, 2] but they can only extract foreground or salient regions which may not pertain in any way to the agent's action. [22] characterizes the interaction between all parts in terms of Granger causality; this is similar to our use of temporal synchrony cues to characterize motion in an unsupervised manner. However, the causality descriptors are not used for action localization in [22] but serve as an additional feature channel to augment a standard BoF recognition pipeline.

3. Actionness Attributes

To better characterize the long term motion cues that will lead to actionness descriptors, we first employ [8] to obtain the so-called temporal superpixels. We denote the i^{th} superpixel trajectory as a sequence of superpixel locations:

$$Tr_i = \{(x_i^k, y_i^k, t_i^k), k = t_i^s \cdots t_i^e\}, \quad i = 1 \cdots n, \tag{1}$$

where (x_i^k, y_i^k, t_i^k) is the spatio-temporal position of the centroid of the i^{th} superpixel R_i^k at frame k, t_i^s and t_i^e are the start and the end time indices of Tr_i, with $[t_i^s, t_i^e]$ being an interval inside $[1, T]$, and n is the number of detected trajectories in V.

Based on this temporal superpixel representation, we can now proceed to describe in details the various attributes used in our actionness descriptor.

3.1. Sudden Change

One of the key attributes for agency is sudden direction and speed change in the trajectory. In addition, we also model human's sensitivity to onset and offset (when a particular spatial region appears or disappears over time). Specifically, we look out for any sudden change in the size of a superpixel. For the former, we describe the displacement change as $\Delta R_{disp}^k = d\left(R_i^k, R_i^{k-1}\right)$, where $d()$ returns the Euclidean distance between the centroids of R_i^k and R_i^{k-1}. For the latter, we describe the size change of a superpixel i between two consecutive frames k and $k-1$ as $\Delta R_{sz}^k = abs\left(|R_i^k| - |R_i^{k-1}|\right)$, where $|R_i^k|$ is the cardinality of the superpixel R_i^k, and $abs()$ returns the absolute value. The "sudden change" (SC) attribute for the i^{th} trajectory at frame k (or equivalently, R_i^k) can then be estimated as follows, with both the ΔR_{sz}^k and ΔR_{disp}^k weighted equally with a suitable normalization:

$$SC(R_i^k) = \begin{cases} \frac{1}{2}\left(\frac{\Delta R_{sz}^k}{\Delta R_{sz}^{Max}} + \frac{\Delta R_{disp}^k}{\Delta R_{disp}^{Max}}\right) & t_i^s < k < t_i^e \\ 1 & k = t_i^s \text{ or } k = t_i^e \end{cases}. \tag{2}$$

Here ΔR_{sz}^{Max} and ΔR_{disp}^{Max} are the maximum size and displacement change over all the trajectories in the current video clip V. The second condition represents the instant when the i^{th} superpixel appears or disappears (onset and offset respectively), during which we give maximum value to the attribute. However we do not want to consider the appearance and disappearance of superpixels at the image boundary to be pertinent in that it is simply an artificial onset/offset caused by the image boundary. Furthermore, sudden change in speed or direction is also difficult to ascertain at the image boundary. Thus, in addition to the above, we also remove all those trajectories currently lying close to the image boundaries from consideration.

3.2. Temporal Synchrony

There are motions that might not be considered particularly meaningful individually, but when they exhibit temporal synchrony with other motions, they become highly indicative of agency. These well-synchronized movements might be between various body parts of the same person or even from different persons. At the coarsest level, they alert us to the presence of purposive behaviors and encode causality. At a more fine-grained level, it could signify something socially relevant and govern our interaction with others, or it could even be maneuvers perceived as threatening (either in real physical combats or in sports).

We use mutual information (MI) to measure the degree of synchrony between two trajectories $Tr_i = \{(x_i^k, y_i^k, t_i^k), k = t_i^s \cdots t_i^e\}$ and $Tr_j = \{(x_j^k, y_j^k, t_j^k), k = t_j^s \cdots t_j^e\}$ over the time interval during which they overlap. We denote this overlapping time interval between Tr_i and Tr_j by $[t^s, t^e] = [t_i^s, t_i^e] \cap [t_j^s, t_j^e]$, assuming $[t^s, t^e] \neq \emptyset$. For simplicity, we use the Gaussian distribution to model the probability of motion vectors from a trajectory. That is $(v_x^i, v_y^i) \sim N(\mu_i, \Sigma_i)$, where $\mu_i = [\mu_x^i \quad \mu_y^i]^T$ and $\Sigma_i = C_{ii} = diag(\sigma_x^i, \sigma_y^i)$. Similarly, for Tr_j, we have another Gaussian $N(\mu_j, \Sigma_j)$, where $\mu_j = [\mu_x^j \quad \mu_y^j]^T$ and $\Sigma_j = C_{jj} = diag(\sigma_x^j, \sigma_y^j)$. The mutual information between Tr_i and Tr_j can then be estimated as [3]:

$$MI(Tr_i, Tr_j) = \begin{cases} \frac{1}{2}\log\frac{|C_{ii}|\cdot|C_{jj}|}{|C|} & Tr_j \notin \mathcal{N}(Tr_i) \text{ and } |[t^s, t^e]| \geq 3 \\ 0 & \text{Otherwise} \end{cases}, \tag{3}$$

where $|.|$ is the determinant of a matrix, $C = \begin{bmatrix} C_{ii} & C_{ij} \\ C_{ji} & C_{jj} \end{bmatrix}$, and $C_{ij} = C_{ji}^T$ is the between-sets covariance matrix computed as $C_{ij} = \begin{bmatrix} cov(v_x^i, v_x^j) & cov(v_x^i, v_y^j) \\ cov(v_y^i, v_x^j) & cov(v_y^i, v_y^j) \end{bmatrix}$. $\mathcal{N}(Tr_i)$ in the first condition is the spatial-temporal neighborhood of Tr_i

used to enforce a mutual inhibition zone: the reason being that we should be allocating more attention only if the temporally synchronous trajectories are not originating from superpixels immediately adjacent to one another (immediately adjacent superpixels exhibiting synchrony would be less surprising). More specifically, $\mathcal{N}(Tr_i)$ is defined as all the trajectories which are spatially connected to Tr_i at some point in time. An example can be seen in Fig. 2, in which the spatial-temporal neighbors of Tr_5 originating from frames k and $k+1$ are illustrated, i.e. $\mathcal{N}(Tr_5) = \left\{ \cdots, \underbrace{Tr_1, Tr_2, Tr_3, Tr_6, Tr_7, Tr_8}_{\text{from frame } k}, \underbrace{Tr_9, Tr_{10}}_{\text{from frame } k+1}, \cdots \right\}$. The condition $|[t^s, t^e]| \geq 3$ aims to measure MI only for those trajectories which have temporal intersection of at least three frames.

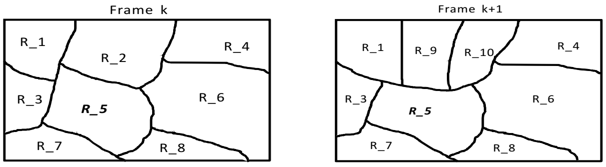

Figure 2. The spatial-temporal neighbors of Tr_5 at frame k and frame $k+1$.

From the MI computed between all pairwise trajectories, a mutual information matrix $G \in R^{n \times n}$ with $G(i,j) = MI(Tr_i, Tr_j)$ can be obtained between all trajectories. The temporal synchrony (TS) attribute of Tr_i should then be the maximum MI values in row i of G. However, we also want to put into context the value of this MI. For instance, the temporal synchrony exhibited between two ballet dancers involved in complex *pas de deux* sequence should have higher value than that between two parallel linear trajectories. Thus we use the entropy of motion vectors from Tr_i itself to weigh TS as:

$$TS(Tr_i) = max_j \left(G(i,j) \right) \times H_i, \tag{4}$$

where $H_i = \sum_{k=t_i^s}^{t_i^e} (-p_k \log(p_k))$ is the entropy of motion vectors of Tr_i, and p_k, the probability of the motion vector at frame k, can be obtained from $N(\mu_i, \Sigma_i)$. This attribute is defined at the level of trajectory; thus all superpixels on the trajectory Tr_i are assigned the same value.

3.3. Repetitive Motion

Many actions are rhythmic and repetitive in nature. To find such repeating patterns, we employ the MI measure in the preceding subsection but apply it to pairs of sub-segments in a trajectory. Formally, given a trajectory Tr_i, we split it into M sub-trajectory with equal time duration $Tr_i = \{Seg_1, \cdots, Seg_M\}$. Mutual information is then calculated for all pairs of sub-segments as $MI(Seg_m, Seg_n)$. Then the "repetitive motion" (RM) attribute is obtained as:

$$RM(Tr_i) = max_{m,n} MI(Seg_m, Seg_n). \tag{5}$$

By adjusting the length of the sub-segments, a scale-invariant way to find the RM of the trajectory is implemented. In all our experiments, five lengths (i.e. 3,6,9,12,15) are used and the one which has the largest RM is selected as the repetitiveness value of the trajectory.

3.4. Image Saliency

For static cues, we use conventional image saliency. Given a video clip V with T frames, for the $t^{th} (t \in [1,T])$ frame, we obtain its image saliency map S_I^t by the well-known GVBS algorithm [14]. Since GVBS is pixel based, we take the average saliency value within a superpixel as the saliency value of the superpixel Sa.

3.5. Fusion and others

Before fusion, Sa is frame-wise normalized by its maximum value, whereas SC, TS and RM are video-wise normalized by the respective maximum values. We now perform a simple weighted combination of the four normalized actionness attributes, with the weights equal to $\frac{1}{4}$:

$$Ac(R_i^k) = \frac{1}{4} \left(SC + TS + RM + Sa \right). \tag{6}$$

where $Ac(R_i^k)$ is the fused actionness score indexed by the i^{th} superpixel at the k^{th} frame.

Background motion induced by camera movement could significantly affect our actionness estimation. Thus, we first estimate the background motion with a simple homography model, using RANSAC to mark out the outliers (i.e. the objects of interest) on each frame. Background motion is then removed. For each trajectory, we now compute a confidence factor of it being a foreground; this is given by the fraction of the pixels inside the associated superpixel being counted as outliers and this statistic is averaged across all frames. Finally, the actionness value of the trajectory is weighted by this confidence factor.

4. Experiments

4.1. Performance on Action Detection

We first test how well our algorithm performs in terms of action detection. UCF-Sports [28] and HOHA datasets [21] are employed because the ground truth in terms of bounding box for each frame is provided. We compare our work with [9], which also detects actions in a precategorical manner. Fig. 3 shows the actionness maps of both methods. Qualitatively, it seems our method provides an actionness map that more accurately characterizes the key action regions. For a quantitative comparison, we adopt the evaluation protocol used in [9]. Basically, the actionness map is divided

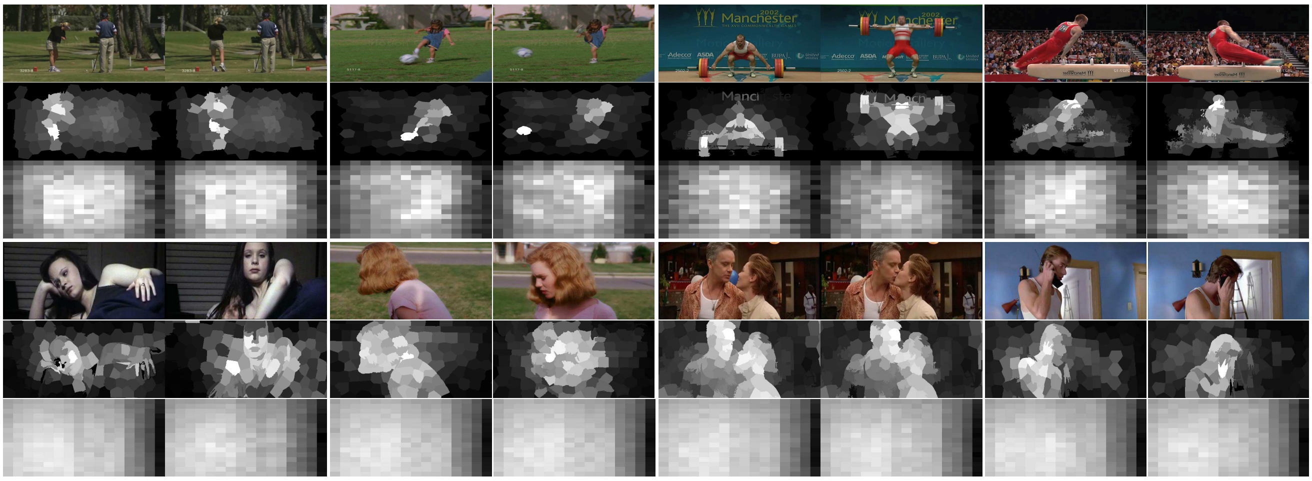

Figure 3. Actionness comparisons between our method (rows 2,5) and [9] (rows 3,6) on four UCF-Sports sequences (rows 1, 2, 3) and four HOHA sequences (rows 4, 5, 6).

into uniform patches, with each patch evaluated separately. Then, a precision-recall curve is obtained by continuously varying a threshold beyond which action is deemed to occur. At last, the statistics are averaged over all patches to obtain the mean average precision (mAP). Due to space limitation, readers are referred to [9] for details of the evaluation method. The mAP for all videos in UCF-Sports and HOHA dataset are shown in Table 1, with the results of [9, 11] directly cited from [9].

Table 1. Mean average precision (mAP) of action detection on the UCF-Sports and HOHA datasets.

	Our Method	L-CORF [9]	DPM [11]
UCF-Sports	**66.81**	60.8	54.9
HOHA	**70.16**	68.5	60.8

4.2. Performance on Action Recognition

Endowed with a relatively stable tracking of the actionness regions, we want to see how much an actionness-driven pooling scheme can improve action recognition. For this purpose, we adopt the action recognition pipeline in the dense trajectories approach [35], which together with its spatial-temporal pyramid variant [21], will serve as our baseline. The comparison was carried out over three public datasets: SSBD [31], HMDB51 [19] and UCF50 [27].

4.2.1 BoF Pipeline with Actionness Pooling

The BoF approach is a powerful and prevalent framework in action recognition [21, 35, 34, 2, 30]. In our experiments, we follow the BoF pipeline but utilize the actionness measures in the feature pooling stage. For BoF, we use similar settings as in [35] which yields the best results. We extract various feature descriptors including HOG, HOF, MBHx and MBHy, using the improved dense trajectories [35]. For each descriptor, we randomly sample 100,000 features from the training videos and train a codebook of 4000 words using the *k*-means algorithm. In the feature pooling stage, we divide a video into K spatio-temporal regions by clustering the actionnness values with *k*-means, and pool the features according to these K regions. Different descriptors in the K regions are then concatenated as one feature vector. Finally, a linear SVM is used for classification. For multi-class classification, we use a one-against-rest approach and select the class with the highest score. In the following, K is set as three unless otherwise stated.

Table 2. Comparisons of our actionness-pooling method with baseline methods. Average accuracy is reported.

Methods	SSBD	HMDB51	UCF50
BoF	76.0	51.74	88.35
BoF-STP	69.3	52.75	88.22
Ours	**77.33**	**56.38**	**89.35**

4.2.2 Comparisons to the Baselines

First, we compare our method (BoF-actionness-pooled) to two baseline methods: the improved version of global BoF [35] and BoF with spatial-temporal pyramid (BoF-STP) [21]. For fair comparisons, we employ in all methods the same features (i.e. improved dense trajectories) and feature encoding scheme (i.e. vector quantization). The results are shown in Table 2. In contrast to [21], we only use a single level $2 \times 2 \times 2$ spatial-temporal grid. It can be observed that our actionness pooling mechanism improves action recognition performance over the baselines on all the employed datasets.

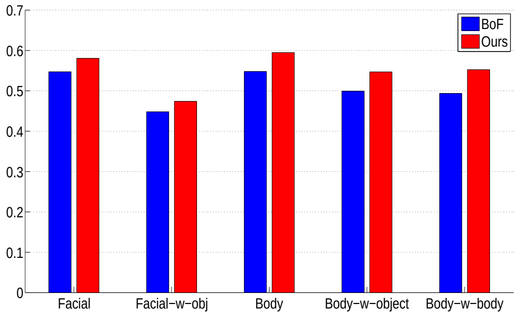

Figure 4. Average accuracy of five sub-categories of videos in HMDB51 dataset: general facial actions (Facial), facial actions with object manipulation (Facial-w-obj), general body movements (Body), body movements with object interaction (Body-w-obj) and body movements for human interaction (Body-w-body).

Among the three action recognition datasets, the HMDB51 dataset is the most challenging one. Delving into the performance gain in HMDB51, we find that our method achieves better accuracy in four of the five action types in the HMDB51 dataset (see Fig. 4). These tend to be those categories which involve interaction, because our method is able to group the interacting humans or objects into a unit. Some of the performance gains over BoF in individual class are substantial, such as in hug (81% vs. 74%), brush-hair (88% vs. 82%) and climb (82% vs. 73%). The hug and brush-hair results are also illustrated in Fig. 5. In many body movement sequences, our performance gain over BoF is also significant, such as cartwheel (54% vs. 47%) and flic-flac (82% vs. 69%). Cartwheel is also illustrated in Fig. 5; as can be seen, in many such sequences with fast motion and strong background clutter, often the camera motion compensation is not perfect. Actionness attributes, being sensitive to biological motion, serve to further suppress residual (induced) motion that remains in the background.

In the category "facial actions with object manipulation", our method's performance is less satisfactory. Looking at the examples of "drink" and "eat" in rows 3 and 4 of Fig. 5, the reason is not difficult to surmise. Our algorithm is able to pick up the hands, the cup/food, but the face, with little facial movements visible, are liable to be missed altogether. For future work, one can probably integrate a face detector in the saliency attribute (perhaps even a generic face that is not agent-specific), so that the face will feature more prominently in the actionness map. Note that in contrast, many sequences in the "facial actions" category tend to be taken from more close-up views, so the facial motions are more visible and thus explaining the better results.

4.2.3 Comparisons to the State-of-the-art

In the rest of the paper, unless otherwise stated, we use Fisher vector encoding for better performance; this encoding is also being employed in most state-of-the-arts methods.

Table 3. Comparisons to the state-of-the-art works. Average accuracy is reported.

SSBD		HMDB51		UCF50	
[31]	44.0	[34]	46.6	[34]	84.5
[25]	73.6	[5]	47.2		
		[2]	51.8	[2]	**92.8**
		[39]	54.0		
		[32]	58.8		
		[35]	57.2	[35]	91.2
		[22]	58.7	[22]	92.5
		[23]	61.1	[23]	92.3
		[24]	**66.7**		
Ours	**76.0**	Ours	60.41	Ours	92.48

Table 4. Results of our method on HMDB51 dataset based on the individual actionness attributes and all four fused together. Here, vector quantization is used for feature encoding.

%	SC	TS	RM	Sa	Fused
mAP	54.81	54.10	51.20	54.59	56.38

Table 5. Sensitivity analysis of parameter K on HMDB51 dataset.

%	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$
mAP	59.0	60.04	60.41	60.24	60.15

Table 3 compares our approach to the various state-of-the-arts works that are based on the BoF paradigm. On SSBD dataset, we achieve a new state-of-the-art result (76%). On HMDB51 dataset, we are one of the three works that have recognition rate over 60%. On UCF50 dataset, our performance is lower, as in these scenarios, background provides important information for recognizing the types of sport actions.

4.2.4 Others

We show in Table 4 the contributions of the individual attributes towards the performance in action recognition. The four attributes seem to complement each other well so that when all four are fused together, there is a significant improvement. We also analyze in Table 5 the sensitivity of our approach to the parameter K. As can be seen, the performance is not sensitive to the value of K, with the best result obtained when $K = 3$.

A note on the computational complexity. The computational cost consists of four main parts: 1) temporal superpixels extraction, 2) actionness attributes calculation, 3) actionness clustering, 4) trajectory features extraction. Excluding (1) and (4) which are implemented via public codes, (2) incurs 34.74 minutes and (3) 0.13 seconds per video (on UCF-Sports dataset) with Matlab implementation on an Intel (R) Xeon (R) workstation (CPU E5-2609 and 32G RAM).

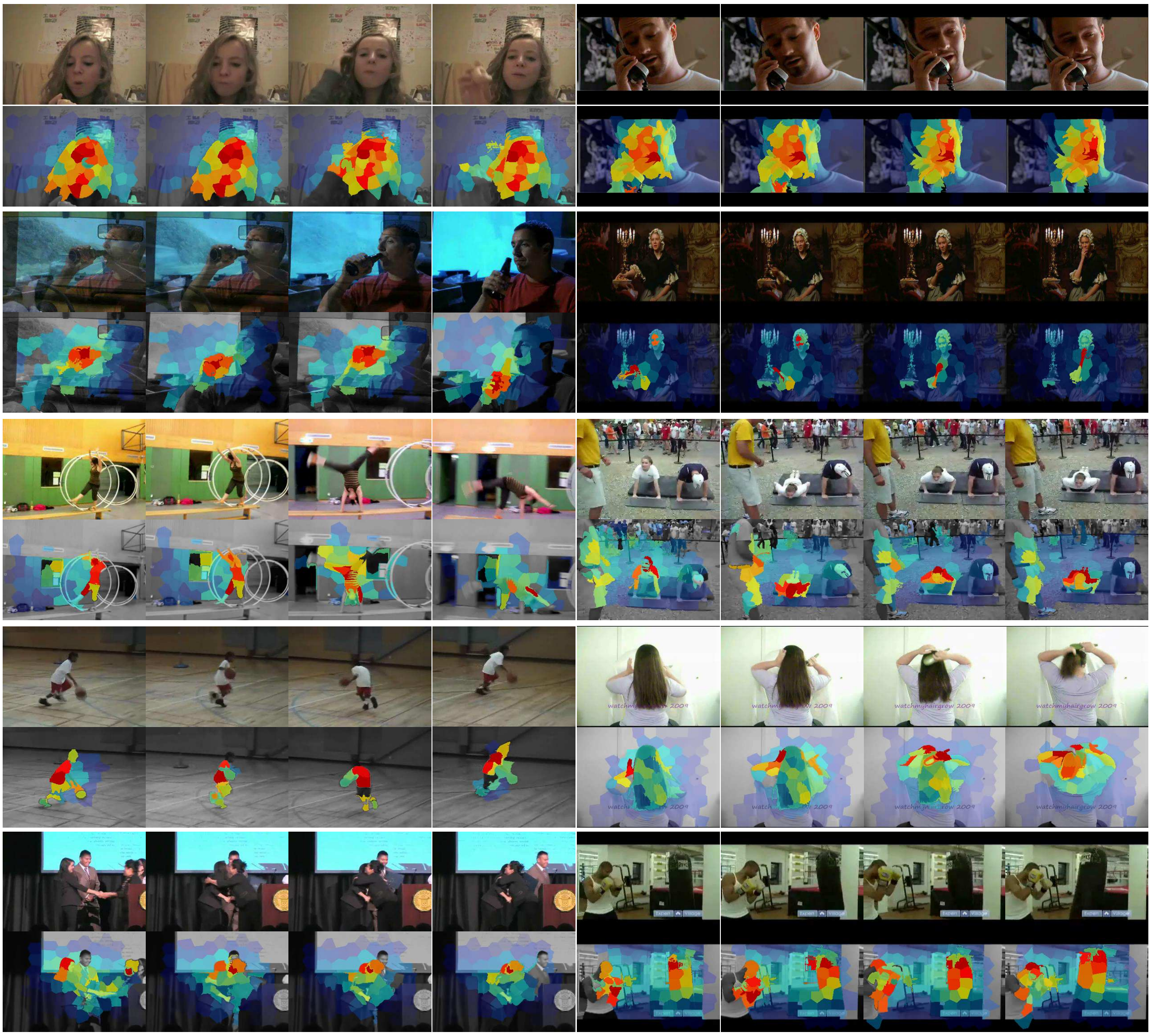

Figure 5. The estimated actionness maps on HMDB51 dataset, with examples from the five subcategories, each occupying two rows. From top to bottom: general facial actions, facial actions with object manipulation, general body movements, body movements with object interaction, and body movements for human interaction.

5. Conclusion

Our work shows that there exist reliable low level cues that can differentiate biological from nonbiological motions. We offer a clear account of how various abstract notions of action such as agency and intention manifest themselves in low level trajectory features such as sudden changes, repetition, and temporal synchrony. This ability to capture the action region allows us to better handle the difficulties besetting the pooling mechanisms proposed hitherto in capturing the pertinent structure of action, especially in actions involving interaction. Our advocated method has the advantage of dividing the scenes into distinct units of analysis that are related to the structure of the actions, and has a better chance of capture the interacting units of actions. For future work, we plan to improve the efficiency of the actionness attributes computation. For instance, we could use a hashing scheme to create an index in which correlated trajectories are mapped into the same hash bins with high probability. Then, one can locate the most correlated trajectories much faster.

Acknowledgement. This work was partially supported by the Singapore PSF grant 1321202075, the China NSFC grant 61572362 and the China NSFC grant 81571347.

References

[1] M. R. Amer and S. Todorovic. A chains model for localizing participants of group activities in videos. In *ICCV*, pages 786–793, 2011.

[2] N. Ballas, Y. Yang, Z.-Z. Lan, B. Delezoide, F. Preteux, and A. Hauptmann. Space-time robust representation for action recognition. In *ICCV*, 2013.

[3] M. Borga. *Learning Multidimensional Signal Processing*. PhD thesis, Linköping University, Sweden, 1998.

[4] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009.

[5] H. Boyraz, S. Masood, B. Liu, M. Tappen, and H. Foroosh. Action recognition by weakly-supervised discriminative region localization. In *BMVC*, September 2014.

[6] H. Boyraz, M. Tappen, and R. Sukthankar. Localizing actions through sequential 2d video projections. In *CVPR*, pages 34–39, 2011.

[7] L. Cao, Z. Liu, and T. Huang. Cross-dataset action detection. In *CVPR*, pages 1998–2005, 2010.

[8] J. Chang, D. Wei, and J. W. Fisher III. A video representation using temporal superpixels. In *CVPR*, pages 2051–2058, 2013.

[9] W. Chen, C. Xiong, R. Xu, and J. Corso. Actionness ranking with lattice conditional ordinal random fields. In *CVPR*, pages 748–755, 2014.

[10] D. Davidson. Actions, reasons, and causes. *Journal of Philosophy*, 60(23):685–700, 1963.

[11] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *TPAMI*, 32(9):1627–1645, 2010.

[12] A. Gaidon, Z. Harchaoui, and C. Schmid. Temporal localization of actions with actoms. *TPAMI*, 35(11):2782–2795, 2013.

[13] T. Gao and B. Scholl. Chasing vs. stalking: Interrupting the perception of animacy. *Journal of Experimental Psychology*, 37(3):669–684, 2011.

[14] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *NIPS*, pages 545–552, 2007.

[15] M. Heidegger. *Being and Time*. Blackwell Publishers Ltd., 1927. translated by John Macquarrie and Edward Robinson, Harper and Row, 1961.

[16] N. Ikizler-Cinbis and S. Sclaroff. Object, scene and actions: Combining multiple features for human action recognition. In *ECCV*, pages 494–507, 2010.

[17] A. Jain, A. Gupta, M. Rodriguez, and L. Davis. Representing videos using mid-level discriminative patches. In *CVPR*, pages 2571–2578, 2013.

[18] M. Jain, J. van Gemert, H. Jegou, P. Bouthemy, and C. Snoek. Action localization by tubelets from motion. In *CVPR*, pages 740–747, 2014.

[19] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *ICCV*, 2011.

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

[21] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, pages 1–8, 2008.

[22] S. Narayan and K. Ramakrishnan. A cause and effect analysis of motion trajectories for modeling actions. In *CVPR*, pages 2633–2640, 2014.

[23] X. Peng, L. Wang, X. Wang, and Y. Qiao. Bag of Visual Words and Fusion Methods for Action Recognition: Comprehensive Study and Good Practice. *ArXiv e-prints*, 2014.

[24] X. Peng, C. Zou, Y. Qiao, and Q. Peng. Action recognition with stacked fisher vectors. In *ECCV*, pages 581–595, 2014.

[25] S. S. Rajagopalan and R. Goecke. Detecting self-stimulatory behaviours for autism diagnosis. In *ICIP*, 2014.

[26] M. Raptis, I. Kokkinos, and S. Soatto. Discovering discriminative action parts from mid-level video representations. In *CVPR*, pages 1242 – 1249, 2012.

[27] K. Reddy and M. Shah. Recognizing 50 human action categories of web videos. *Machine Vision and Applications*, 24(5):971–981, 2013.

[28] M. Rodriguez, J. Ahmed, and M. Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, pages 1–8, 2008.

[29] N. Shapovalova, A. Vahdat, K. Cannons, T. Lan, and G. Mori. Similarity constrained latent support vector machine: An application to weakly supervised action classification. In *ECCV*, pages 55–68, 2012.

[30] W. Sultani and I. Saleemi. Human action recognition across datasets by foreground-weighted histogram decomposition. In *CVPR*, 2014.

[31] S. Sundar Rajagopalan, A. Dhall, and R. Goecke. Self-stimulatory behaviours in the wild for autism diagnosis. In *ICCV Workshops*, 2013.

[32] E. Taralova, F. De la Torre, and M. Hebert. Motion words for videos. In *ECCV*, 2014.

[33] D. Tran and J. Yuan. Max-margin structured output regression for spatio-temporal action localization. In *NIPS*, 2012.

[34] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *IJCV*, 103(1):60–79, 2013.

[35] H. Wang and C. Schmid. Action Recognition with Improved Trajectories. In *ICCV*, 2013.

[36] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*, pages 1290–1297, 2012.

[37] A. Yao, J. Gall, and L. Van Gool. A hough transform-based voting framework for action recognition. In *CVPR*, pages 2061–2068, 2010.

[38] J. Yuan, Z. Liu, and Y. Wu. Discriminative subvolume search for efficient action detection. In *CVPR*, pages 2442–2449, 2009.

[39] J. Zhu, B. Wang, X. Yang, W. Zhang, and Z. Tu. Action recognition with actons. In *ICCV*, pages 3559–3566, 2013.