Human Action Segmentation with Hierarchical Supervoxel Consistency

Jiasen Lu[1], Ran Xu[1] and Jason J. Corso[2]

[1] Computer Science and Engineering, SUNY at Buffalo
[2] Electrical Engineering and Computer Science, University of Michigan
{jiasenlu,rxu2}@buffalo.edu jjcorso@eecs.umich.edu

Abstract

Detailed analysis of human action, such as action classification, detection and localization has received increasing attention from the community; datasets like JHMDB have made it plausible to conduct studies analyzing the impact that such deeper information has on the greater action understanding problem. However, detailed automatic segmentation of human action has comparatively been unexplored. In this paper, we take a step in that direction and propose a hierarchical MRF model to bridge low-level video fragments with high-level human motion and appearance; novel higher-order potentials connect different levels of the supervoxel hierarchy to enforce the consistency of the human segmentation by pulling from different segment-scales. Our single layer model significantly outperforms the current state-of-the-art on actionness, and our full model improves upon the single layer baselines in action segmentation.

1. Introduction

In recent years, a great emphasis in video understanding have been action recognition [33, 28, 17] on large datasets like UCF101 [30] and HMDB51 [20]. To classify a video, a number of representations has been proposed, from low-level features that leverage point trajectories and local appearance/motion information [33, 34], to high-level features that create a high-dimension action space [28], leverage human pose [32], or even unsupervised features learned from deep neural networks [17].

Despite the progress, however, these methods remain limited in their ability to supply any deeper information than the video-wise action label. In reaction, sub-communities have begun to focus on aspects of the broader video understanding problem such as classification of group activities [22] and human-object interactions [11]. However, these foci still remain at a coarse granularity and are not suitable for many applications, such as autonomous driving [23] and robotic surgery [4], that require precise action boundaries in space-time.

In contrast, action localization [6, 21] and action detection [31] directly emphasize the finer "when" and "where" a particular action of interest occurs in a video. Not only do these finer action inferences enable a broader set of application, Jhuang et al. [16] also recently showed that precise human silhouette boundaries can impact action classification itself. They have recently evaluated the impact that different ground-truth scenarios have on action classification; they thoroughly annotated data of human actions, and find that a human "puppet" provides significant help towards better action classification compared with whole video or even bounding box constraints of a human action. They, however, do not pose a solution to automatically localizing and segmenting the silhouette or action automatically.

One line of work in this direction of automatic action segmentation follows a template-matching paradigm in which templates, which are normally in the form of a sequence of bounding boxes or a bounding volume, are used to localize actions in space-time. The templates are either rigid, manually chosen [8, 6] or deformable with flexible properties [31, 35]. Another line of work is driven by low-level segmentation, such as manually cropped human segments [18], tubelets [14] which merges supervoxels [9, 36] using motion, and space-time segments [24] that seek human-like segments with color, shape and motion cues. The latest results show that both directions [35, 14] achieve state-of-the-art performance while [35] use stronger supervision, such as pose annotation for training, which is non-trivial to acquire.

Leveraging low-level segmentation such as supervoxels [9, 36] as a precursor to action segmentation, localization and classification is a promising direction. It could potentially relax the amount of supervision needed and provides a general representational framework. However, some challenges need to be addressed. First, according to the study from Jhuang et al. [16], whole-human segmentation is better than a bounding box comprised of pieces of segments,

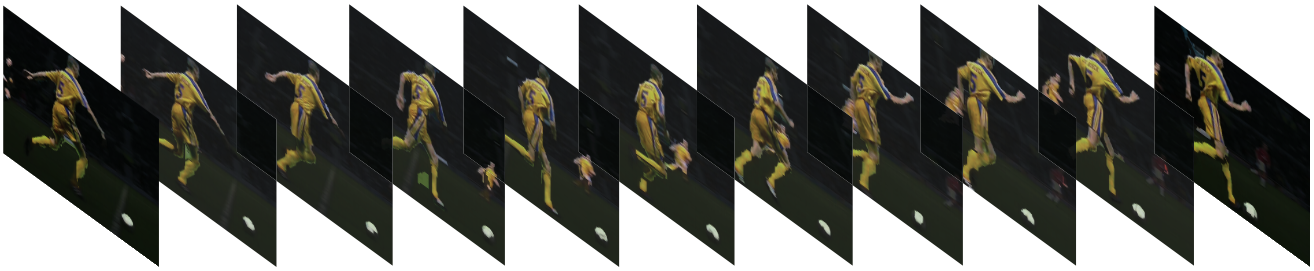


Figure 1. Our system output of human action segmentation, the example is from J-HMDB data set. The segmented regions are rendered using the original RGB pixels for foreground and the background regions are covered with a transparent black mask.

e.g. [24]. Second, segmentation methods that depend only on motion, e.g. [14], tends to miss objects or human parts that are static in the video. For example, it is unclear whether a person sitting at a desk waving hand will be fully or partially segmented as foreground. Third, segmentation quality is vital for further video understanding, thus making the selection of segmentation granularity a non-trivial problem [37, 26] shows.

To these ends, our paper proposes a hierarchical MRF model for human action segmentation that satisfies the following goals.

- Automatically segments the whole human action silhouette, as Fig. 1 shows thus further enabling deeper video understanding tasks, i.e., action classification and localization.
- Bridges low-level segmentation and a high-level human prior to recover both static body parts and difficult, articulating body parts.
- Improves the segmentation quality by enforcing supervoxel consistency between different scales (levels) in the hierarchy.

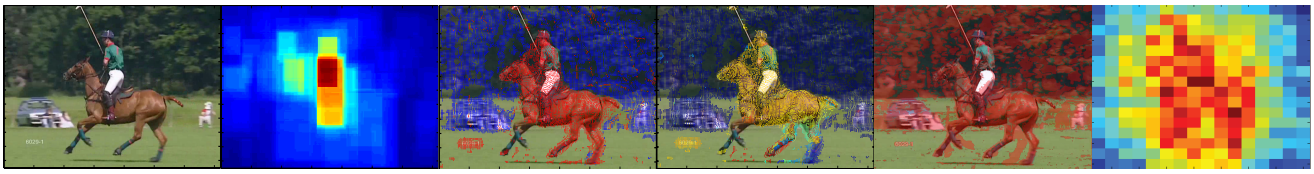
1.1. Overview of our Method

We first propose a human motion saliency representation that is able to account for camera motion and balance human motion and human appearance cues automatically. A similar concept called "actionness" was proposed by Chen et al. [5]: it produces a rank ordering of video regions according to the degree to which they contain an action, but the regions to be ranked are small 3D cuboid-volumes (see Fig. 2f) and the ground-truth is agnostic to the human action boundaries. Our human motion saliency, on the contrary, has the human action silhouette naturally incorporated. To be specific, we estimate foreground motion by forming a camera model via long term trajectories [3], and obtain a human prior by DPM-based person models [7]. We compare our human motion saliency with an optical flow based camera motion estimation method [25] and actionness [5], and find a $+16\%$ relative improvement in actionness ranking. (See Sec. 4.1 for more details.)

Then, to segment the human action, we start by applying hierarchical graph-based video segmentation [38] to form a hierarchy of supervoxels. On this hierarchy, we define an MRF model, using our novel human motion saliency as the unary term. As Jain et al. [15] noted, supervoxels with vast temporal-extent variance can make the graph over all supervoxels brittle. So, instead of choosing frame based superpixel graph, which may lose the human action boundary, we alternatively design a pairwise potential based on neighboring supervoxels connected only in the direction of optical flow. Hence, we only consider pairwise supervoxels temporally. On the other hand, to address the problem of static human body parts, we extract a shape prior from the learned parts of a person-DPM [7] and consider connections between supervoxels and the shape prior in the pairwise potential.

In this hierarchical MRF, we design an innovative high-order potential between different supervoxels on different levels of the hierarchy. Intuitively, supervoxels in higher levels of the hierarchy carry better human or human parts semantic meaning, but are more vulnerable to leaks correspondingly. Similarly, supervoxels in lower levels have less leaks but also carry less semantic meaning. Most existing approaches manually choose the hierarchy level based on visual inspection and the optimal choice of hierarchy level is different based on different videos, Oneata et al. [26] determine hierarchy level of supervoxels by finding best localization scores in the training set. Xu et al. [37] propose a flattening process which select supervoxels from different hierarchies of the segmentation by using the *uniform entropy slice* criterion. To alleviate leaks and sustain better semantic information, our high-order potential favors supervoxels from higher levels with constraints from motion and appearance cues. The strategy is tailored for the human motion segmentation problem, and we find it effective in both our quantitative and qualitative experiments in Sec. 4.2.

Finally, we minimize the energy of the hierarchical MRF by the α-expansion algorithm [1, 19] and present a method



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

Figure 2. Motion saliency and human saliency feature. (a) Original image. (b) Visualization of human saliency response. (c) Initial 2 clusters of trajectories from GMM (red are foreground and blue are background). (d) Visualization of our motion saliency response (Note that misclassified trajectories from (c) have low response). (e) Visualization of foreground motion estimated with optical flow, used by [25, 14, 13]. (f) Actionness ranking [5].

to automatically learn the model parameters based on GMM estimation.

The remaining sections are organized as follows. Sections 2 and 3 formulate the problem and presents our model in detail. Section 4 presents quantitative and qualitative evaluations on our method. Section 5 concludes the paper and discusses future works.

2. Human Motion Saliency for Human Action Segmentation

Our approach inputs a video clip containing human action and outputs a space-time segmentation that labels all the human-action pixels as foreground and otherwise as background. We make no assumptions about the scene context or the level of articulation in the humans; we hence use datasets such as JHMDB [16], UCF-sports [27], Penn Action [39] because they contain human action with large variation, gross deformation and strong camera motion. Detailed annotations such as human puppet or pose joints are available for detailed evaluation.

We begin the discussion of our method with a new approach to automatically measure human motion saliency. This new feature can be directly used to localize and rank human action, which we directly evaluate (Sec. 4), and it can be used as an action feature for later modeling, as we do in Sec. 3.

Our human motion saliency incorporate two parts: foreground motion and human appearance information. For foreground motion estimation, building a camera model is a natural choice. Current methods for foreground estimation generally fall into two categories, 1) use optical flow and RANSAC to find the dominate motion as background motion [29], and 2) do spectral clustering with long term trajectories and find dominant trajectory group [2]. Our early experiments show optical flow could be unreliable, see Fig. 2(e) for an example; and clustering of trajectories, though robust, sometimes could also leave outliers, such as the red dot in background in Fig. 2(c).

We hence combine these two schemes and propose a new motion saliency feature, we use the long term trajectories to build a camera motion model, and then measure the motion saliency via the deviation from the camera model. We use a 2D parametric affine motion model for the camera motion.

Concretely, given a trajectory set Tr of a video clip with L frames. The velocity difference between two trajectories tr^i and tr^j at time t is

$$d_t(tr^i, tr^j) = \frac{1}{T}(u_t^i - u_t^j)^2 + (v_t^i - v_t^j)^2 \qquad (1)$$

where u_t^i and v_t^j denote the motion of tr^i aggregated over T frames. We measure tr^i using the median value of the velocity distances between tr^i and all the others and further fit a 1D Gaussian Mixture Model to get two clusters as in [10]. However, without the structure information, the background trajectories are susceptible to being grouped in foreground clusters. To alleviate the false clustering, we compute the affine motion model and fit the trajectory points with the robust penalty on the background cluster as

$$\widehat{\theta} = \arg\min_{\theta} \sum_{p \in \Omega} \rho(r_\theta(p, t)) \qquad (2)$$

where θ is the affine motion model parameters, $\rho(\cdot)$ is defined as robust Tukey function [12] and $r_\theta(p, t)$ is the displaced frame difference at trajectory point p on frame t. We further reweight the motion saliency of all the trajectories by calculating the mean deviation through the clip. As can be seen in Fig. 2(d), the misclassified trajectories in (c) are re-classified to background.

For human appearance information, we use a DPM [7] person detector trained on PASCAL VOC 2007 and construct a saliency map by averaging the normalized detection score of all the scale and all components, as shown in Fig. 2(b). We further encode our foreground motion and human saliency in supervoxel as presented in Sec. 3.1.

3. Hierarchical MRF Graph Structure

In this section, we introduce our hierarchical Markov Random Field (MRF) model on the hierarchy of supervoxels. We denote a graph G consisting of nodes $\mathcal{X}$ and edges $\mathcal{E}$, where $\mathcal{X}$ is the set of supervoxels over the entire video volume and $\mathcal{E}$ is the edge set as shown in Fig. 3. As discussed in Sec. 1, building a graph with all supervoxels from

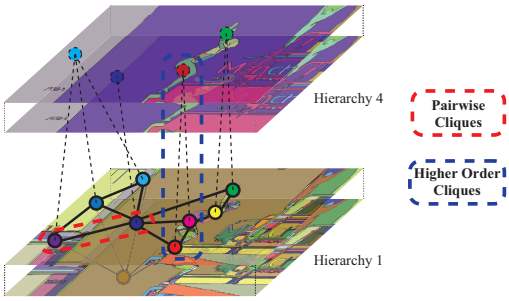


Figure 3. Proposed hierarchical MRF graph. Nodes are supervoxels, candidate edges exist if two supervoxels are neighbors, but are confirmed only if they are connected by supervoxel optical flow or overlapping person detectors. Higher-order cliques are defined by corresponding supervoxels among higher-level supervoxel levels. Only a small subset of nodes and connections are depicted for simplicity.

a video may lead to a brittle graph due to the large size of neighbors. Unlike [15], who build the graph with superpixels from each frame and propagate them to subsequent frames, we devise a mechanism to constrain the number of edges in our graph. Intuitively, we only build an edge between two supervoxels (x_i, x_j) when 1) they are neighbors in the direction of optical flow, which finds two supervoxels that are neighbors temporally, or 2) two supervoxels are both overlapping with a person detection that has confidence value larger than a certain threshold, which largely finds two supervoxels that are neighbors spatially but constrained by human appearance.

The nodes inside the red dash line in Fig. 3 represents a confirmed edge. We refer the reader to Sec. 3.1 for a more detailed presentation. We use $\mathcal{V}$ to denote the set of supervoxels in higher layer of the hierarchy. Each element $v \in \mathcal{V}$ represents a higher-order clique (with blue dashed line representing the correspondence in Fig. 3). And, y_v denotes the set of labels assigned to the supervoxel nodes belonging to the supervoxel v. We associate a random variable $y_i \in \{+1, -1\}$ with every node to represent the label it may take, which can be either human-with-action $(+1)$ or background (-1). Our goal is to label all the supervoxels $\mathcal{X} = \{x_i\}_{i=1}^{N}$ over the entire video.

3.1. Energy function

Given the graph structure $G = (\mathcal{X}, \mathcal{E})$ induced by the supervoxel hierarchy ($\mathcal{E}$ is the set of edges in the graph hierarchy). We introduce an energy function over $G = (\mathcal{X}, \mathcal{E})$ that enforces hierarchical supervoxel consistency through higher order potentials derived from supervoxel $\mathcal{V}$.

$$E(Y) = \sum_{i \in \mathcal{X}} \Phi_i(y_i) + \sum_{(i,j) \in \mathcal{E}} \Phi_{i,j}(y_i, y_j) + \sum_{v \in \mathcal{V}} \Phi_v(y_v) \tag{3}$$

where $\Phi_i(y_i)$ denotes unary potential for a supervoxel with index i, $\Phi_{i,j}(y_i, y_j)$ denotes pairwise potential between two supervoxels with edge, and $\Phi_v(y_v)$ denotes high order potential of supervoxels between two layers.

Unary potential: We encode the motion saliency and human saliency feature into supervoxels to get the unary potential components:

$$\Phi_i(y_i) = \gamma_M M_i(y_i) + \gamma_P P_i(y_i) + \gamma_S S_i(y_i) \tag{4}$$

where γ_M, γ_P and γ_S are weights for the unary terms. $M_i(y_i)$ reflect the motion evidence, $P_i(y_i)$ and $S_i(y_i)$ reflect the human evidence respectively. $M_i(y_i)$ can be calculated as:

$$M_i(y_i) = \exp\left(-\frac{\lambda_m}{|x_i|} \sum_{tr^j \in x_i} w^M(tr^j)\right) \tag{5}$$

where λ_m is the scale parameter, $w^M(tr^j)$ is the motion saliency weight for trajectory tr^j and $|x_i|$ is the size of supervoxel x_i.

The human saliency $P_i(y_i)$ can be formed as:

$$P_i(y_i) = \frac{1}{1 - \exp(-\frac{\lambda_p}{|x_i|} \sum_{px^j \in x_i} w^P(px^j))} \tag{6}$$

where $w^P(px^j)$ is the human saliency weight for pixel j. We use a DPM detection with a person-model trained on PASCAL VOC 2007 for human detection and form the saliency map with average of normalized detection score of all scales and components.

Following [24], background objects with straight boundaries are common in man-made scenes, but human body boundaries contain fewer points of zero curvature. So we also compute the curvature at all boundary points of each supervoxel as the shape saliency feature.

$$S_i(y_i) = \exp\left(-\frac{\lambda_s}{|x_i|} \sum_{px^j px^k \in B_i} w^S(px^j, px^k)\right) \tag{7}$$

where B_i is the set of boundary point in supervoxel x_i, (px^j, px^k) are two nearby pixels and w^S is the curvature.

Pairwise potential: As described in Sec. 3, we constrain the edge space by only two types of neighbors: temporal supervoxel neighbors and human appearance-aware spatial neighbors, so we define the pairwise potential as:

$$\Phi_{i,j}(y_i, y_j) = \gamma_I I_{i,j}(y_i, y_j) + \gamma_K K_{i,j}(y_i, y_j)) \tag{8}$$

where γ_I and γ_K are pairwise potential weights. $I_{i,j}(y_i, y_j)$ is the cost between supervoxel i and supervoxel j with human detection constraints, which ensures the smoothness

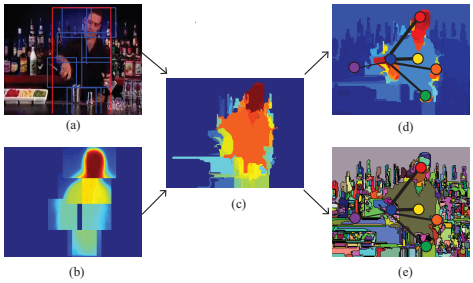


Figure 4. Human with action representation (a) DPM human detection with root and part bounding box. (b) Corresponding DPM part mask extracted from PASACAL VOC. (c) supervoxel response for the part masks. (d) and (e) Supervoxel pairwise connections of motion saliency map and segmentation respectively, bold line represent strong connections.

spatially. Note that i and j could be determined as neighbors without pixel-level connection. $K_{i,j}(y_i, y_j)$ is the virtual dissimilarity which ensures the smoothness temporally.

The mixture components of a part-based model typically reflect a number of body parts of human. We exploit this by defining new potentials which utilize a shape prior for each part. We incorporate this information in our model by encouraging the supervoxels in the same human detection with stronger consensus. Therefore, we define r_i^t to be the response of supervoxel i given detection on frame t:

$$r_i^t = s_r^t \mu_r^t \max(\mu_p^t) \tag{9}$$

where s_r^t is the root detection score. μ_r^t is percentage that a superpixel (we use the 2D plane from supervoxel) lies in the bounding box and μ_p^t is the percentage that the superpixel lies in the part mask extracted from PASCAL VOC, see Fig 4(b) for an example. The human connection score for two supervoxel i, and j at frame t is $S_{i,j}^t = \min(r_i^t, r_j^t)$. Thus we define

$$I_{i,j}(y_i, y_j) = \delta(y_i \neq y_j) \exp(-\beta_I \sum_{t \in T} S_{i,j}) \tag{10}$$

where β_I is the scale parameter. Fig. 4 depicts this process.

The temporal smoothness term $K_{i,j}(y_i, y_j)$ is defined as

$$K_{i,j}(y_i, y_j) = \delta(y_i \neq y_j) \exp(-\beta_D D(x_i, x_j)) \tag{11}$$

where $D(x_i, x_j)$ is the χ^2 distance between the histogram feature of two supervoxels. For each supervoxel, we compute two features: 1) an RGB color histogram with 33 bins (11 bins per channel), and 2) a histogram of optical flow with 9 bins.

Higher order potential: We define the hierarchical supervoxel label consistency potential. Different from [15], which use the higher order potential as temporal smoothness of superpixels, we utilize the connection between different supervoxel hierarchical levels. In practice, we adopt the Robust P^n model [19] to define the potentials

$$\Phi_v(y_v) = \begin{cases} N(y_v)\frac{1}{Q}\gamma_{\max}(v) & \text{if } N(y_v) \leqslant Q \\ \gamma_{\max}(v) & \text{otherwise} \end{cases}$$

where y_v denotes the labels of all the nodes corresponding to higher-level supervoxel hierarchy $v \in \mathcal{V}$. $N(y_v)$ is the number of nodes within v that do not take the dominant label. Q is a truncation parameter that controls how rigidly we want to enforce the consistency within the supervoxels in two layers.

The penalty $\gamma_{\max}(v)$ is a function considering the size, color and motion diversity of supervoxels. If the foreground supervoxel and background supervoxel are inappropriately merged in v, $\gamma_{\max}(v)$ should be large and less penalty is paid for label inconsistencies. Specifically, $\gamma_{\max}(v) = |y_v| \exp(-\eta(\sigma_v^c + \sigma_v^m))$ where σ_v^c and σ_v^m is the variance of RGB color as well as motion in supervoxel v.

3.2. Energy minimization and parameters

The energy function defined in Eqn. 3 can be efficiently minimized using the α-expansion algorithm [19]. Local color prior is a strong evidence for segmentation. Since we do not have the initial labeled frame, we use the output of our first segmentation result and further refine our output by learning a Gaussian mixture model (GMM) on RGB color space.

The value of γ_M, γ_P and γ_S reflects the weight for motion and human cue. For most videos, we set the unary weights at reasonable value, $\gamma_M = 6$, $\gamma_P = 4$ and $\gamma_S = 3$ in our experiment. However, for videos with little motion, such as golf, we want the human detection feature to dominate the unary term. Thus, we automatically determine γ_M and γ_P by comparing the mean of estimated gaussian centers and the mean of all individual distance terms. We empirically found that GMM estimation performs better. But GMM estimation can fail if μ_2 is located on an outlier. In this case, the values of the two estimations are significantly different. If the difference is larger than a threshold, we set $\gamma_M = 3$, $\gamma_P = 7$.

We set the weight parameter $\gamma_I = \gamma_K = 0.1$, $\gamma_H = 2$. The scale parameter λ_m, λ_p, β_I, β_D and η are all set automatically as the inverse of the mean of all individual distance terms. The truncation parameter $Q = 0.3|y_v|$. We use the same parameter setting in all dataset.

4. Experiments

To fully evaluate our method, we report results on three tasks: first, with our human motion saliency feature, we

evaluate actionness ranking and compare with state-of-the-art methods [5, 25] in Sec. 4.1; second, we evaluate our human action segmentation with baseline methods that without pairwise or high order terms, in Sec. 4.2; third, with our segmentation output, we evaluate action classification and localization with other state-of-the-art methods, in Sec. 4.3.

Dataset As presented in previous sections, our approach makes no assumptions about scene context, camera motion or the level of articulation in the humans, so general action recognition or localization dataset "in-the-wild" is suitable. In addition, due to the segmentation nature of our method, we favor data set with ground truth segmentation, e.g., JHMDB [16], or at least with bounding box ground truth, e.g. UCF-Sports [27] and Penn Action [39].

UCF-Sports contains 150 videos over 10 classes with large human action variation, gross deformation and strong camera motion, we fully evaluate actionness ranking, action segmentation, classification and localization on this data set because it is widely used and evaluated in the community. We use training/testing split from [5] for these tasks.

JHMDB is a subset of HMDB and contains 928 clips comprising 21 action categories, all frames are annotated with a "human puppet", which we take as our human action segmentation ground truth for evaluation. We evaluate our action segmentation and action recognition with this data set.

Penn Action contains 15 actions and 2326 video clips, the annotations consist of 2D keypoint positions, in which we recover ground truth bounding box. We evaluate action segmentation and recognition with this data set.

Evaluation protocol For actionness evaluation, we follow the protocol of [5]: rank our human motion saliency map and get mean AP. For action segmentation, with our output binary mask, we evaluate Intersection Over Union (IOU) value, precision (IOU over our segmentation area) and recall (IOU over ground truth area) with ground truth. Note that ground truth is segmentation mask in JHMDB and bounding box in UCF-sports and Penn Action data set. For action recognition, we extract dense trajectory feature [33] and evaluate a number of settings with segmentations, we follow the standard bag-of-visual-words procedure to obtain codebooks with five channels of dense trajectory and train χ^2 kernel SVM for action classification. For action localization, we evaluate mean IOU of our segmentation masks only when it is correctly classified, and set the IOU as 0 when the mask is misclassified.

4.1. Evaluate Actionness

Following [5], we measure mean average precision (mAP) of actionness ranking of our motion saliency map and joint human motion saliency map, described in Sec. 2, and compare with [5] which use a ranking-CRF to rank the actionness of 3D cuboid volumes, and Motion 2D [25] which generate foreground motion saliency map with optical flow and RANSAC. From Table. 1, we can see our human motion saliency map shows a significant improvement of more than 10% mAP gain over state-of-the-art method [5]. In addition, our method needs very limited supervision (only image-based person models trained with PASCAL VOC), which makes a good potential for further detailed action understanding tasks. We also observe that the score for whole UCF-Sports data set is much lower than testing set, which could indicate that the test set has more human action cues and less noise compared with training set.

	subset of video				all video		
	[5]	[25]	Motion	Joint	[25]	Motion	Joint
dive	58.7	63.9	**69.5**	66.4	58.1	**66.7**	64.1
golf	61.8	41.9	66.6	**66.8**	35.2	63.7	**69.1**
kick	68.7	71.1	**73.0**	61.3	70.7	**67.9**	60.7
lift	**86.7**	61.5	76.4	77.1	67.2	**76.1**	75.4
ride	18.8	52.4	51.3	**52.5**	33.1	35.7	**42.3**
run	48.4	48.2	**61.6**	61.0	51.5	58.4	**59.1**
skate	57.6	65.4	**81.1**	63.8	55.4	57.7	**66.6**
sw-b	80.1	**89.8**	87.3	84.4	**85.7**	80.9	77.3
sw-s	54.0	64.9	78.9	**79.5**	61.8	**72.7**	67.0
walk	50.4	55.3	70.3	**79.1**	40.5	57.1	**69.1**
Avg.	60.8	61.9	**71.8**	68.6	56.0	63.9	**65.4**

Table 1. mAP of our actionness with proposed motion and joint feature against actionness [5] on UCF sport dataset

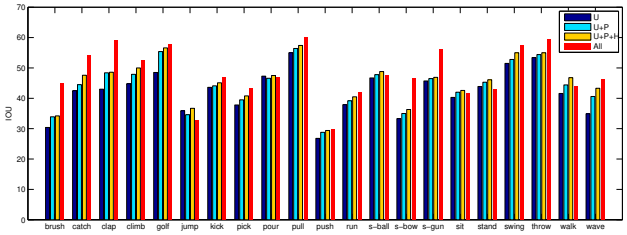


Figure 5. Action segmentation results with IOU measurement on JHMDB data set. We report three baselines and our full model on all 21 classes.

4.2. Human Action Segmentation

We thoroughly evaluate our human action segmentation methods with all three data sets and all baselines, we denote U as only using unary potential, $U + P$ as using unary and pairwise potential, $U+P+H$ as using additional high order potential and All as our full model. Table. 2 summarizes mean IOU, mean precision and mean recall, it demonstrates pairwise potential generally contribute $1\% - 3\%$ gain over the unary potential, and with high order term we observe another $1\% - 3\%$ gain over $U + P$. Surprisingly, a second path of inference with additional color prior from $U+P+H$ shows another $1\% - 4\%$ gain. The score demonstrate the effectiveness our model. Qualitatively, Fig. 6 illustrates the ground truth, unary segmentation mask and full model

	IOU				Precision				Recall			
	U	U+P	U+P+H	All	U	U+P	U+P+H	All	U	U+P	U+P+H	All
UCF-Sport	40.0	42.1	44.2	**47.6**	67.8	70.4	**73.8**	72.8	54.4	55.2	56.3	**60.6**
J-HMDB	42.7	44.6	45.8	**48.8**	57.2	60.5	**63.0**	62.7	70.0	68.0	67.7	**72.0**
Penn Action	48.9	49.0	49.9	**51.5**	66.1	67.0	69.6	**69.8**	68.1	67.7	65.8	**69.4**

Table 2. Table shows mean IOU (Intersection over union), mean Precision and mean Recall of UCF-Sports, J-HMDB and Penn Action data sets, with only unary potential(U), unary and pairwise potential (U+P), with high order (U+P+H) and our full model (All).

	UCF-Sport	JHMDB	Penn Action
Baseline DT	83.0	54.4	94.5
Seg DT*	**93.6**	**58.6**	95.0
GT bbox DT	89.0	55.5	**95.4**
GT puppet mask DT	-	56.2	-

Table 3. Action Recognition results

	subset of frames			all frame	
	[21]	[24]	Ours	[24]	Ours
dive	43.4	46.7	**48.0**	44.3	**48.3**
golf	37.1	**51.3**	49.6	**50.5**	50.0
kick	36.8	**50.6**	36.0	**48.3**	35.5
lift	**68.8**	55.0	57.2	51.4	**57.1**
ride	21.9	29.5	**29.8**	**30.6**	29.8
run	20.1	**34.3**	33.9	33.1	**33.7**
skate	13.0	40.0	**46.1**	38.5	**45.9**
swing-b	32.7	54.8	**62.6**	54.3	**62.3**
swing-s	16.4	19.3	**53.9**	20.6	**54.9**
walk	28.3	39.5	**58.1**	39	**58.1**
Avg.	31.8	42.1	**48.1**	41.0	**48.0**

Table 4. Action localization results measured as average IOU (in %) on the UCF sports dataset

mask, the full model can effectively remove mis-segmented regions with unary segmentation, e.g. "Diving", "Riding-Horse" and "Running", due to the pairwise and high order consistency. In addition, our full model may also complete missing segments with unary segmentation, such as "Golf" and "Kicking", possibly because of our human-aware edges in pairwise potential.

In Fig. 5, we show class-wise segmentation results in JHMDB data set, actions with full upright human pose, such as "golf" and "pull" perform relatively better than actions with large body motion articulations, e.g. "jump", or the human motion is more from small child instead of adult, e.g. "push". Qualitatively, in Fig. 6, our automatically generated masks align very well with ground truth masks. Some of our masks have leaks due to supervoxel segmentation such as "pick", "push" and "sit", but notably, as the ground truth is annotated with deforming "puppet", the alignment may not be perfect when the human body cannot fit the puppet such as "wave" and "push". In these cases our method generate masks that cover whole human body.

4.3. Action Recognition and Localization

After getting human action segmentation, we further evaluate how it will impact action recognition and localization. First, we extract dense trajectory features inside 1) whole video, 2) our segmentation mask, 3) ground truth bounding box and 4) ground truth puppet mask (only for JHMDB). We find using bounding box or segmentation mask can generally improve classification accuracy over whole video, and specifically, using segmentation mask generally get better results than bounding box. The finding is consistent with evaluations from [16] and demonstrating our method is a plausible tool for better video understanding. Notably, our UCF-Sports classification accuracy achieves $+10\%$ gain over recent segmentation-based video understanding papers [24] (with 81.7%) and [14] (with 80.24%).

In addition, in JHMDB data set, we find using our segmentation mask achieves even better accuracy than ground truth puppet mask, which echoes the finding in Sec. 4.2 that our mask covers more complete human body.

For action localization, we compare average IOU scores with [21] and [24], although it is not tailored for localization, we still achieve $6-7\%$ gain in mean IOU.

5. Conclusion

In this paper, we introduce a hierarchical MRF model to automatically segment human action boundaries in videos "in-the-wild". We make several contributions, including a strong human motion saliency feature and a novel higher-order potentials that connect different granularities of video segments, to empower accurate action segmentation, and achieve promising results in several important video understanding tasks such as action recognition, localization and actionness ranking. Our approach needs minimum supervision compared with many existing methods, and show potential for more complex tasks such as human pose estimation. In future work, we plan to extend this model and learn better human action representation for action detection and human pose estimation in the video. Code for our method and segmentation results on three datasets are available from the authors' website.

Acknowledgments. This work was partially supported by the National Science Foundation CAREER grant (IIS-0845282), the Army Research Office (W911NF-11-1-0090) and the DARPA Mind's Eye program (W911NF-10-2-0062).

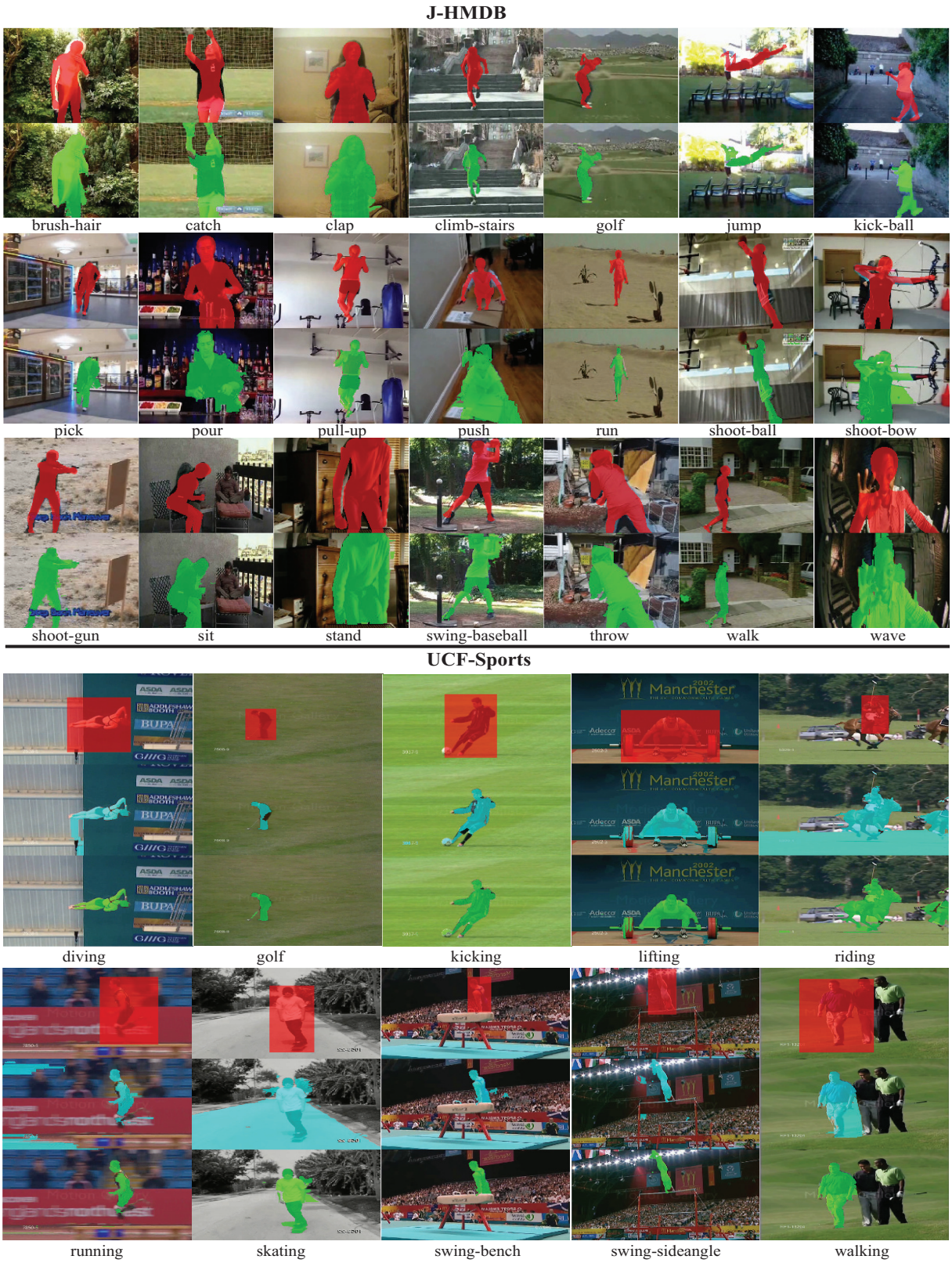


Figure 6. Segmentation visualization of JHMDB and UCF-Sports data set of all the classes. The red is the ground truth mask, the cyan in UCF-Sports is the unary result and the green is our result with full model.

References

[1] Y. Boykov, O. Veksler, and R. Zabih. Fast Approximate Energy Minimization via Graph Cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.

[2] T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. In *European Conference on Computer Vision*. 2010.

[3] T. Brox and J. Malik. Large displacement optical flow: Descriptor matching in variational motion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):500–513, 2011.

[4] D. Burschka, J. J. Corso, M. Dewan, W. Lau, M. Li, H. Lin, P. Marayong, N. Ramey, G. D. Hager, B. Hoffman, D. Larkin, and C. Hasser. Navigating Inner Space: 3-D Assistance for Minimally Invasive Surgery. *Robotics and Autonomous System*, 2005.

[5] W. Chen, C. Xiong, R. Xu, and J. Corso. Actionness ranking with lattice conditional ordinal random fields. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[6] K. G. Derpanis, M. Sizintsev, K. Cannons, and R. P. Wildes. Efficient action spotting based on a spacetime oriented structure representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[7] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–45, 2010.

[8] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253, 2007.

[9] M. Grundmann, V. Kwatra, M. Han, and I. Essa. Efficient hierarchical graph-based video segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[10] J. Guo, Z. Li, L.-F. Cheong, and S. Zhou. Video co-segmentation for meaningful action extraction. In *IEEE International Conference on Computer Vision*, Dec 2013.

[11] A. Gupta, A. Kembhavi, and L. S. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31, 2009.

[12] P. Huber. *Robust statistics*. Wiley, New York, 1981.

[13] M. Jain, H. Jegou, and P. Bouthemy. Better exploiting motion for better action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[14] M. Jain, J. C. van Gemert, H. Jégou, P. Bouthemy, and C. G. M. Snoek. Action localization by tubelets from motion. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[15] S. Jain and K. Grauman. Supervoxel-consistent foreground propagation in video. In *European Conference on Computer Vision*. 2014.

[16] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *IEEE International Conference on Computer Vision*, Dec. 2013.

[17] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[18] Y. Ke, R. Sukthankar, and M. Hebert. Spatio-temporal shape and flow correlation for action recognition. In *IEEE International Conference on Computer Vision*, 2007.

[19] P. Kohli, P. H. Torr, et al. Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision*, 82(3):302–324, 2009.

[20] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In *IEEE International Conference on Computer Vision*, 2011.

[21] T. Lan, Y. Wang, and G. Mori. Discriminative figure-centric models for joint action localization and recognition. In *IEEE International Conference on Computer Vision*, Nov 2011.

[22] T. Lan, Y. Wang, W. Yang, and G. Mori. Beyond actions: Discriminative models for contextual group activities. In *Proceedings of Advance in Neutral Information Processing*, 2010.

[23] J. Leonard, J. How, S. Teller, M. Berger, S. Campbell, G. Fiore, L. Fletcher, E. Frazzoli, A. Huang, and S. Karaman. A perception-driven autonomous urban vehicle. *Journal of Field Robotics*, 25(10):727–774, 2008.

[24] S. Ma, J. Zhang, N. Ikizler-Cinbis, and S. Sclaroff. Action recognition and localization by hierarchical space-time segments. In *IEEE International Conference on Computer Vision*, 2013.

[25] J.-M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. In *Journal of Visual Communication and Image Representation*, 1995.

[26] D. Oneata, J. Revaud, J. Verbeek, and C. Schmid. Spatio-temporal object detection proposals. In *European Conference on Computer Vision*, 2014.

[27] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach: A spatio-temporal maximum average correlation height filter for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[28] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[29] Y. Sheikh, O. Javed, and T. Kanade. Background subtraction for freely moving cameras. In *IEEE International Conference on Computer Vision*, 2009.

[30] K. Soomro, A. R. Zamir, and M. Shah. A dataset of 101 human action classes from videos in the wild. Technical report, University of Central Florida, Center for Research in Computer Vision, 2012.

[31] Y. Tian, R. Sukthankar, and M. Shah. Spatiotemporal deformable part models for action detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[32] C. Wang, Y. Wang, and A. L. Yuille. An approach to pose-based action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[33] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action Recognition by Dense Trajectories. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[34] H. Wang and C. Schmid. Action recognition with improved trajectories. In *IEEE International Conference on Computer Vision*, 2013.

[35] L. Wang, Y. Qiao, and X. Tang. Video action detection with relational dynamic-poselets. In *European Conference on Computer Vision*, 2014.

[36] C. Xu and J. J. Corso. Evaluation of super-voxel methods for early video processing. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[37] C. Xu, S. Whitt, and J. Corso. Flattening supervoxel hierarchies by the uniform entropy slice. In *IEEE Conference on Computer Vision and Pattern Recognition*, Dec 2013.

[38] C. Xu, C. Xiong, and J. J. Corso. Streaming hierarchical video segmentation. In *Proceedings of European Conference on Computer Vision*, 2012.

[39] W. Zhang, M. Zhu, and K. G. Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *IEEE International Conference on Computer Vision*, 2013.