Human Action Segmentation with Hierarchical Supervoxel Consistency

Jiasen Lu[1], Ran Xu[1] Jason J. Corso[2]
[1]Department of Computer Science and Engineering, SUNY at Buffalo. [2]Department of EECS, University of Michigan.

Detailed analysis of human action, such as classification, detection and localization has received increasing attention from the community; datasets like J-HMDB [1] 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 introduce a hierarchical MRF model to automatically segment human action boundaries in videos "in-the-wild" (see Fig. 1).

We first propose a human motion saliency representation which incorporates two parts: foreground motion and human appearance information. For foreground motion estimation, we propose a new motion saliency feature by using long-term trajectories to build a camera motion model, and then measure the motion saliency via the deviation from the camera model. For human appearance information, we use a DPM 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.

Then, to segment the human action, we start by applying hierarchical graph-based video segmentation [2] to form a hierarchy of supervoxels. On this hierarchy, we define an MRF model, using our novel human motion saliency as the unary term. We consider the joint information of temporal connections in the direction of optical flow and human appearance-aware spatial neighbors as pairwise potential. We design an innovative high-order potential between different supervoxels on different levels of the hierarchy to alleviate leaks and sustain better semantic information. Given the graph structure $G = (\mathcal{X}, \mathcal{E})$ induced by the supervoxel hierarchy ($\mathcal{E}$ is the set of edges in the graph hiearchy). We introduce an energy function over $G = (\mathcal{X}, \mathcal{E})$ that enforces hierarchical supervoxel consistency through higher order potentials derived from supervoxel $\mathcal{V}$.

$$E(Y) = \sum_{i \in \mathcal{X}} \Phi_i(y_i) + \sum_{(i,j) \in \mathcal{E}} \Phi_{i,j}(y_i, y_j) + \sum_{v \in \mathcal{V}} \Phi_v(y_v) \quad (1)$$

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:

$$\Phi_i(y_i) = \gamma_M M_i(y_i) + \gamma_P P_i(y_i) + \gamma_S S_i(y_i) \quad (2)$$

where γ_M, γ_P and γ_S are weights for the unary terms. $M_i(y_i)$ reflects the motion evidence, $P_i(y_i)$ and $S_i(y_i)$ reflect the human evidence. **Pairwise potential:** we constrain the edge space with only two types of neighbors: temporal supervoxel neighbors and human-aware spatial neighbors, so we define the pairwise potential as:

$$\Phi_{i,j}(y_i, y_j) = \gamma_I I_{i,j}(y_i, y_j) + \gamma_K K_{i,j}(y_i, y_j)) \quad (3)$$

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 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. **Higher order potential:** We define the hierarchical supervoxel label consistency potential. We utilize the connection between different supervoxel hierarchical levels. In practice, we adopt the Robust P^n model [3] to define the potentials,

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

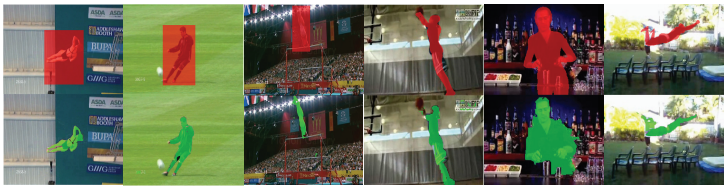

Figure 1: Segmentation visualization of UCF-Sports (first three column) and JHMDB data set. The red is the ground truth mask and the green is our result.

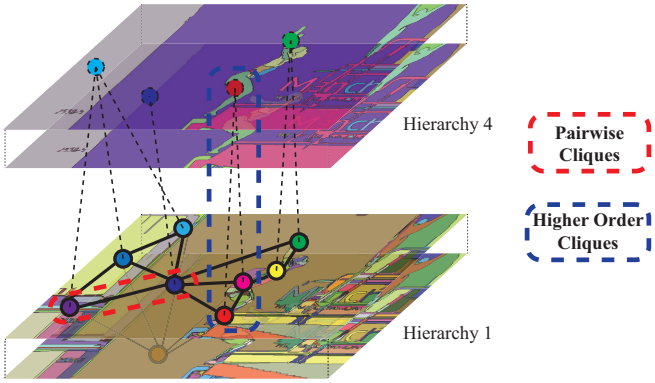

Figure 2: Edges corresponding to straight lines in the real world are detected (a) and the plumbline constraint is used to compute the distortion parameters, giving the rectified image (b)

where y_v denotes the labels of all the nodes corresponding to higher level supervoxel hierarchy $v \in \mathcal{V}$. Finally, we minimize the energy of the hierarchical MRF with α-expansion algorithm [3] and present a method to automatically learn the model parameters based on GMM estimation.

To fully evaluate our method, we report results on four tasks: actionness, action segmentation, action recognition and action localization. The evaluation verifies that our novel representation and hierarchical MRF model is effective, and our approach shows great potential as a weakly supervised video early processing tool for further video understanding. (see Fig. 1) To sumarize, 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] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *IEEE International Conference on Computer Vision*, December 2013.

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

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

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.