Motionlets: Mid-Level 3D Parts for Human Motion Recognition

LiMin Wang[1,2], Yu Qiao[2], and Xiaoou Tang[1,2]

[1]Department of Information Engineering, The Chinese University of Hong Kong
[2]Shenzhen Institutes of Advanced Technology, CAS, China

Introduction

- **Goal**: Design a “motion part” based representation for human motion recognition.
- **Existing works**:
  - **Low level local spatio-temporal features**: HOG, HOF, HOG3D etc. These features share *local* and *repeatable* properties.
  - **High level representations or models**: Motion Energy and History Image, Action Bank etc. The features share *global* and *discriminative* properties.
- **Our idea**: To preserve the advantages of low level features and global templates, we propose a mid-level 3D (spatio-temporal) part, called *motionlet*. It corresponds to the moving process of parts, objects, visual phrase etc.
- **Properties**:
  - **High motion saliency**: it is able to capture the part with strong motion cues.
  - **Multiple scales**: it is a balance between local features and global template.
  - **Representative and discriminative**: it can provide rich information for classifying motions.

Low Level Features

- **Spatio-temporal Orientation Energy [1,2]**:
  - **3D orientation filter**: $E_{\hat{\theta}}(\mathbf{x}) = \sum_{\mathbf{x}'\in\Omega(\mathbf{x})} (G^3_{\hat{\theta}} * V)^2$.
  - **Marginalization**: $\widetilde{E}_{\hat{\mathbf{n}}}(\mathbf{x}) = \sum_{i=0}^{N} E_{\hat{\theta}_i(\hat{\mathbf{n}})}(\mathbf{x})$.
  - **Substraction**: $\overline{E}_i = \max(\widetilde{E}_i - \widetilde{E}_s - \widetilde{E}_o, 0), \quad \forall i \in \text{All} - \{s, o\}$.
  - **Normalization**: $\overline{E}_i = \frac{\overline{E}_i}{\sum_{j=1}^{M} \overline{E}_j}$.
- **Dense HOG and HOE:**

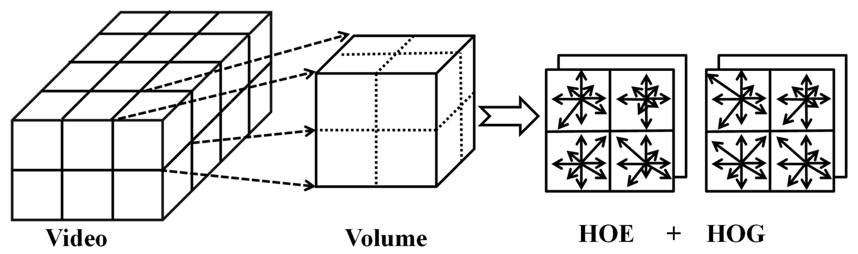

Figure 1: Illustration for dense HOE and HOG.

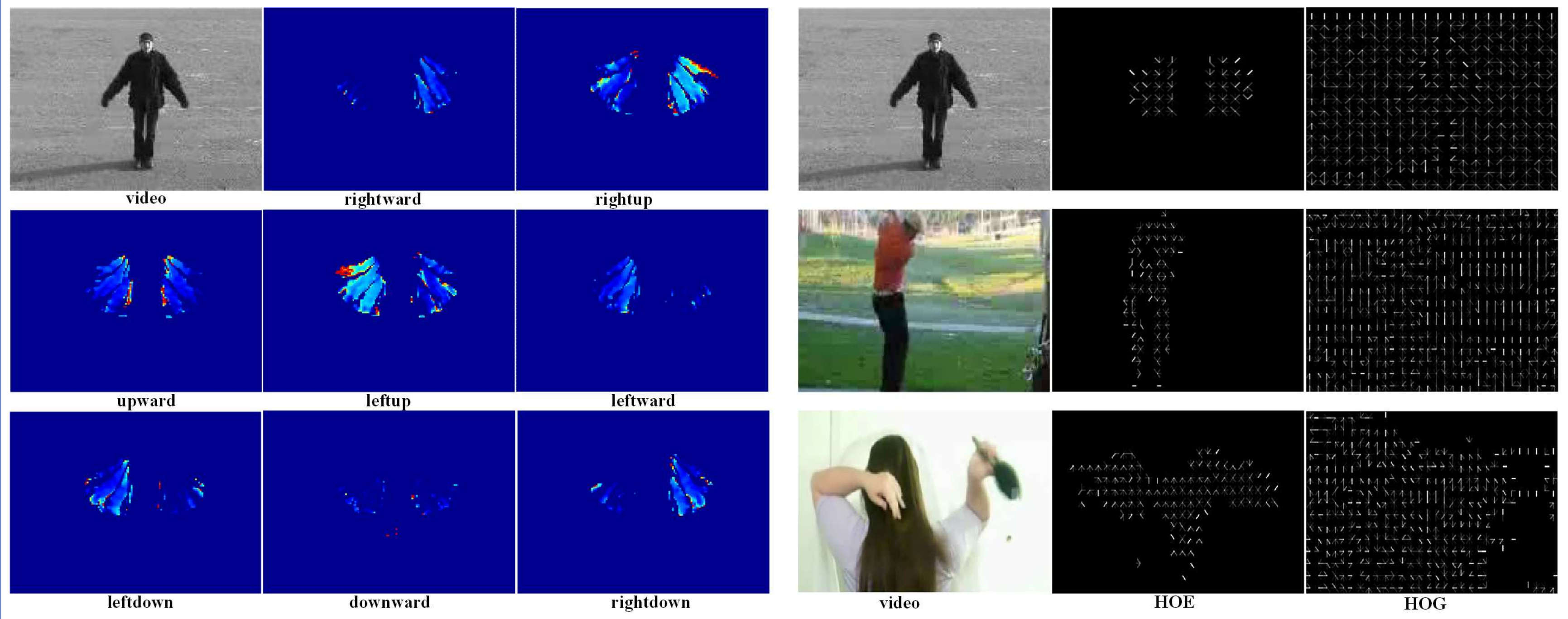

Figure 2: Low level motion saliency and features.

Motionlet Extraction & Video Representation

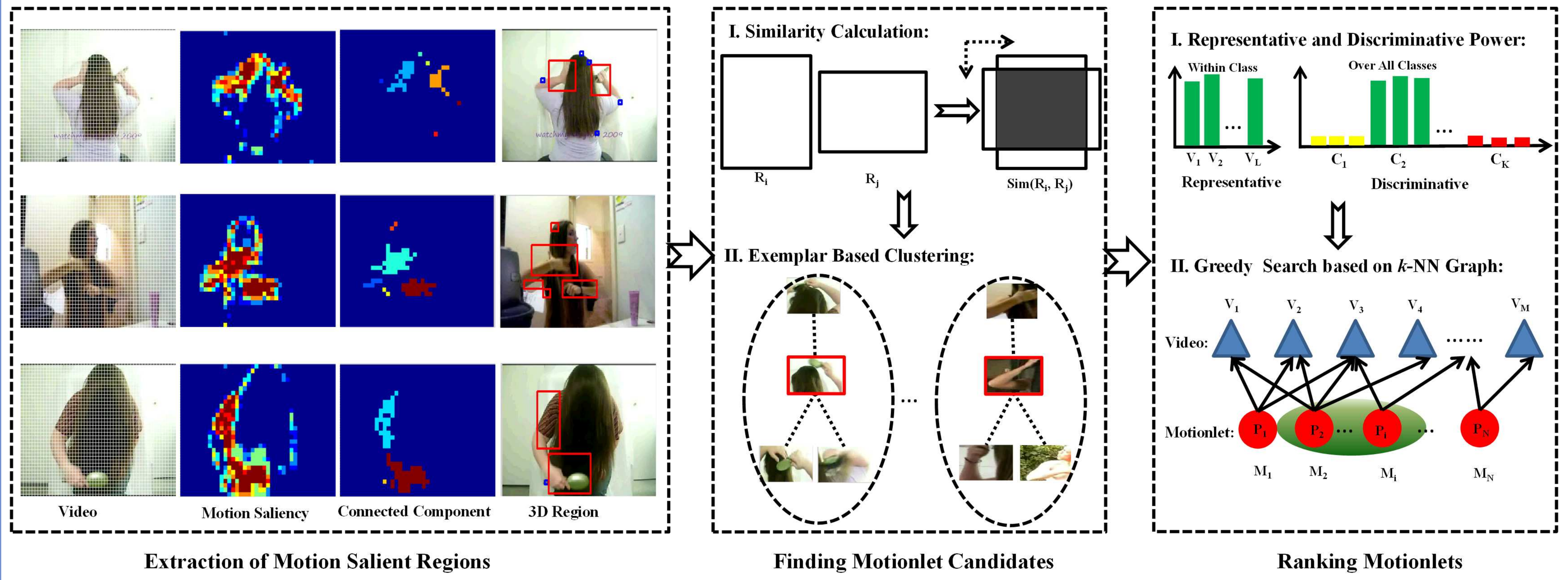

Figure 3: Pipeline of motionlet Construction.

- **Motionlet Construction**:
  - **Extraction of Motion Salient Regions**:
    1. We obtain the motion salience map and its binarization:
       $s(\Omega) = \sum_{\mathbf{x}\in\Omega} \sum_{i\in \text{All}-\{s,o\}} \overline{E}_i(\mathbf{x}) \quad \Rightarrow \quad \mathcal{B}(\Omega) = \begin{cases} 1 & \text{if } s(\Omega) > \alpha, \\ 0 & \text{otherwis.} \end{cases}$
    2. We can obtain a large pool of 3D regions with different sizes by connected component analysis: $\{\mathcal{R}_1, \cdots, \mathcal{R}_M\}$.
  - **Finding Motionlet Candidates**:
    1. We define the similarity between two subregions:
       $\text{Sim}(\mathcal{R}_i, \mathcal{R}_j) = \max_{\mathbf{x}} \{\sum_{\mathbf{u}} m(\mathcal{R}_i(\mathbf{x}+\mathbf{u}), \mathcal{R}_j(\mathbf{u}))\}$.
    2. With similarity measures, we use Affinity Propagation to cluster 3D regions.
  - **Ranking Motionlet**:
    1. We define the representative and discriminative measure of $\mathcal{M}_j$ using matching score s:
       $P_j = \frac{\sum_{k=1}^{K} N_k(\bar{s}^j_k - \bar{s}^j)^2}{\sum_{k=1}^{K}\sum_{\mathcal{V}_i\in C_k}(s^j_i - \bar{s}^j_k)^2}, \quad \bar{s}^j_k = \frac{1}{N_k}\sum_{\mathcal{V}_i\in C_k} s^j_i, \quad \bar{s}^j = \frac{1}{\sum_{k=1}^{K} N_k}\sum_{k=1}^{K} N_k \bar{s}^j_k$
    2. Considering the correlation of motionlets, we design a greedy algorithm to select effective motionlets:

```
Input  : Representative and Discriminative power: P.
         Coverage table: T. Selecting number l.
Output: Selected motionlets: S

Init: coverage counter C ← 0, selected set S ← ∅;
for i ← 1 to l do
    1. videoset ← FindLeastCoverage(C);
    2. motionletset ← FindActive(T,videoset);
    3. bestmot ← FindBest(P,S,motionletset);
    4. Update(S,C,T,bestmot);
end
```

- **Video Representation**:
  - **Motionlet activation vector**: we represent an action video using max pooling for the matching score of motionlets.
  - **Spatio-temporal pyramid**: three layers $1\times1\times1$, $2\times2\times2$, and $1\times1\times4$.
  - **Classification**: LibSVM and one vs. all for multi-class classification.

Experiment Results

- **Settings**: We conduct experiments on three datasets: KTH, HMDB51, UCF50.
- **Examples**:

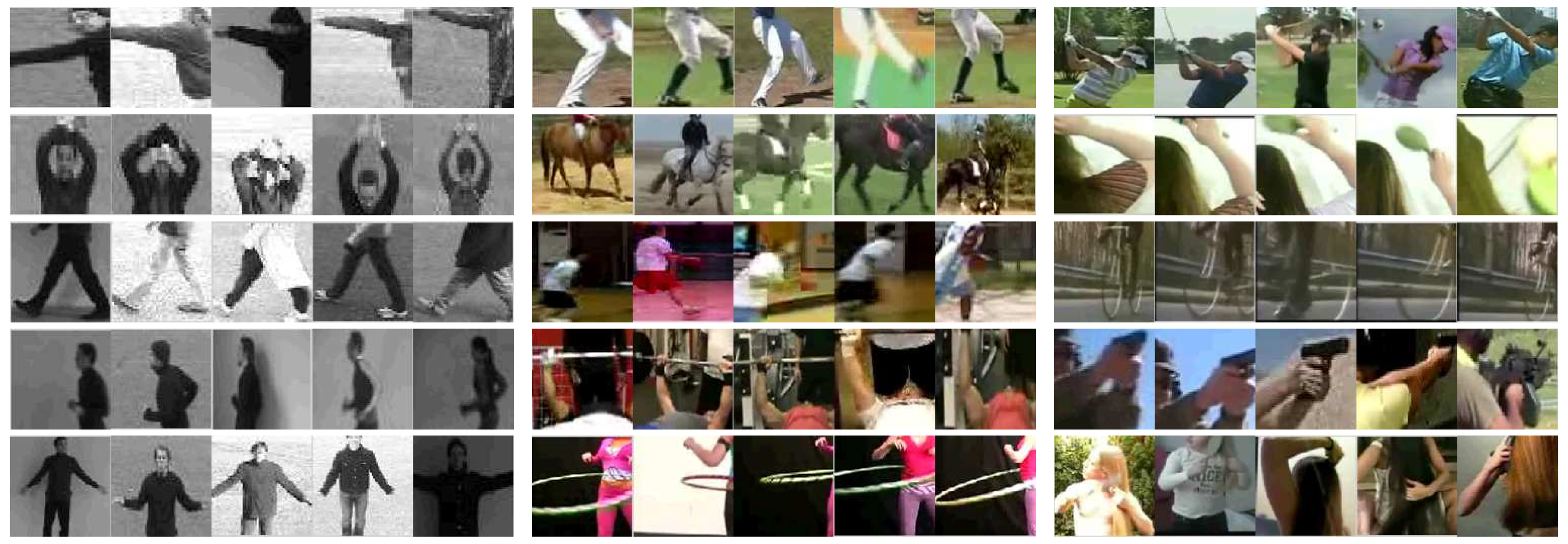

Figure 4: Examples of motionlet from three datasets: KTH, UCF50, and HMDB51.

- **Results and Comparisons**:

Method	Accuracy
Harris3D + HOG/HOF	91.8
Cuboids + HOF3D	90.0
Dense + HOF	88.0
Hessian + ESURF	81.4
HMAX(C2)	91.7
3D CNN	90.2
GRBM	90.0
ISA (dense sampling)	91.4
ISA (norm thresholding)	93.9
ActionBank	**98.0**
Motionlet (1000)	92.1
Motionlet (3000)	93.3

Method	Accuracy
Gist	13.4
Harris3D + HOG/HOF	20.2
HMAX(C2)	23.2
Motion Interchange Pattern	29.2
Action Bank	26.9
Motionlet (1000)	32.1
Motionlet (3000)	**33.7**

Method	GV	LOGO
Gist	38.8	-
Harris3D + HOG/HOF	47.9	-
Motion Interchange Pattern	68.5	72.7
Action Bank	57.9	-
Motionlet (1000)	67.9	70.2
Motionlet (3000)	**71.7**	**73.9**

Table 1: Results on three datasets: KTH, HMDB51, and UCF50.

- **Combined with other representations**:

Method	HMDB51	UCF50
Combined with Harris3D + HOG/HOF	35.5	73.6
Combined with Action Bank	39.0	74.0
Combine All	**42.1**	**78.4**

Table 2: Recognition accuracy of combined representation.

- **Varying number of motionlets**:

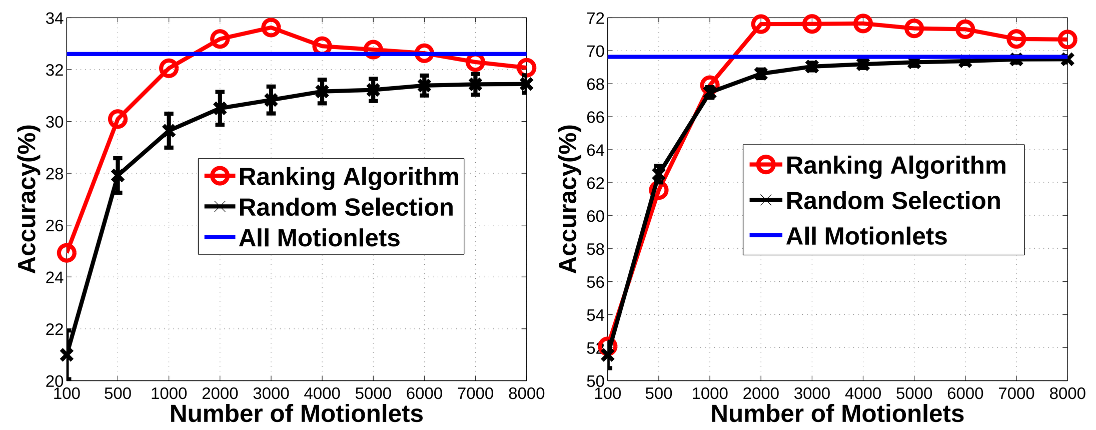

Figure 5: Results of varying motionlet sizes and compare ranking algorithm with random selection, Left: HMDB51 and Right: UCF50.

References

1. S. Sandanand and J. J. Corso. Action bank: a high level representation of activity in video, in CVPR 2012.
2. K. G. Derpanis, M. Sizintsev, K. J. Cannons, and R. P. Wildes. Efficient action spotting based on a spacetime oriented structure representation. In CVPR, 2010.
3. H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In BMVC, 2009.