Joint Angles Similarities and HOG² for Action Recognition

Eshed Ohn-Bar and Mohan M. Trivedi
Computer Vision and Robotics Research Laboratory
Electrical and Computer Engineering Dept.
University of California, San Diego

June 24, 2013

International Workshop on Human Activity Understanding from 3D Data
Computer Vision and Pattern Recognition Conference 2013

1

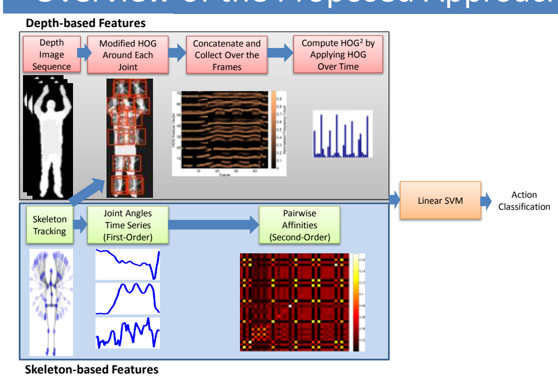
Contribution

1) Characterize actions using **pairwise affinities** between view-invariant joint angles features over the performance of an action.

2) A **new spatio-temporal feature** for RGB and depth images, based on a modified HOG, termed **HOG²** involves applying the algorithm over space, and then re-applying over time.

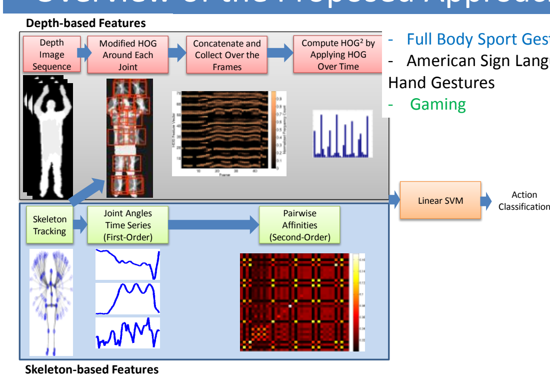
2

Overview of the Proposed Approach



3

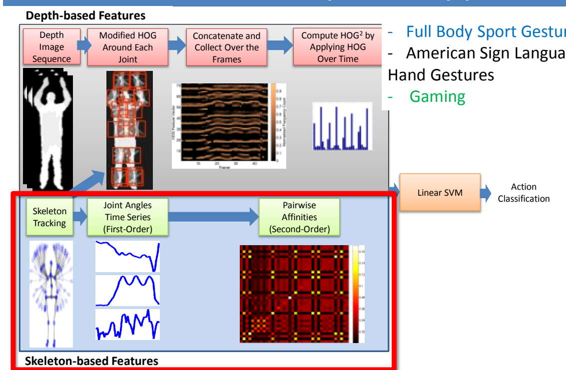
Overview of the Proposed Approach



- Full Body Sport Gestures
- American Sign Language Hand Gestures
- Gaming

4

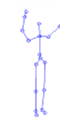
Overview of the Proposed Approach



- Full Body Sport Gestures
- American Sign Language Hand Gestures
- Gaming

Joint Angles Affinity Clustering – Previous Work

- Model joint trajectories or joint angles as independent trajectories (**First-order features**).



$$\textbf{sample}\ [x^1_{LeftElbow}\ \ x^2_{LeftElbow} \cdots x^n_{LeftElbow}]$$
compared to
$$\textbf{template}\ [x^1_{LeftElbow}\ \ x^2_{LeftElbow} \cdots x^n_{LeftElbow}]$$

- Pairwise similarities – recent works leverage such information within a single frame or a small window in time of 2-3 frames (Ellis *et al., IJCV* 2012, *Wang et al.* CVPR 2012, Yun *et al. HAU3D* 2012).

$[x^1_{LeftElbow}\ \ x^2_{LeftElbow} \cdots x^n_{LeftElbow}]$
$[x^1_{RightElbow}\ \ x^2_{RightElbow} \cdots x^n_{RightElbow}]$

Similarity Measure as Feature

This work: pairwise similarities of angles along the *entire gesture* (**Second-order features**). **How to define similarity? Why not use first-order features directly?**

Angular Skeleton Representation (First-Order Features)

- A depth-first tree traversal gives the relative azimuth and elevation angles of each joint with respect to its parent node.

$$S_i = \{\theta_i, \phi_i\}$$
$$K^t = \bigcup_{t=1:p} S_i^t$$

Skeleton configuration at time t, where p is number of joints

JAS – Joint Angles Pairwise Similarities (Second-Order Features)

- Transform first-order => second-order using
$$d: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$$

Where n is number of frames in the gesture instance.

- Produces $\frac{m(m-1)}{2}$ feature set

Which d works well?

Good: Simple distance functions, **Euclidean**

Bad: Allowing time-shifts and gaps (e.g. the Longest Common Subsequence distance (**LCSS**) or dynamic temporal warping (**DTW**))

Which d works well?

- Given two vectors of joint angles over a gesture instance $x_i, x_j \in K$

$$d_{cosine}\left(x_i, x_j\right) = \frac{x_i^T x_j}{\|x_i\|_2 \|x_j\|_2}$$

Which d works well?

- Given two vectors of joint angles over a gesture instance $x_i, x_j \in K$

$$d_{cosine}\left(x_i, x_j\right) = \frac{x_i^T x_j}{\|x_i\|_2 \|x_j\|_2}$$

$$d_{weigthedEuc}\left(x_i, x_j\right) = \sum_{t=1:n} w_i(t)\left\|x_i(t) - x_j(t)\right\|_2^2 (1 + \lambda_{i,j}(t))$$

$$w_i(t) \propto \exp(-\frac{x_i(t)^2}{2(c^2)})$$

Which d works well?

- Given two vectors of joint angles over a gesture instance $x_i, x_j \in K$

$$d_{cosine}\left(x_i, x_j\right) = \frac{x_i^T x_j}{\|x_i\|_2 \|x_j\|_2}$$

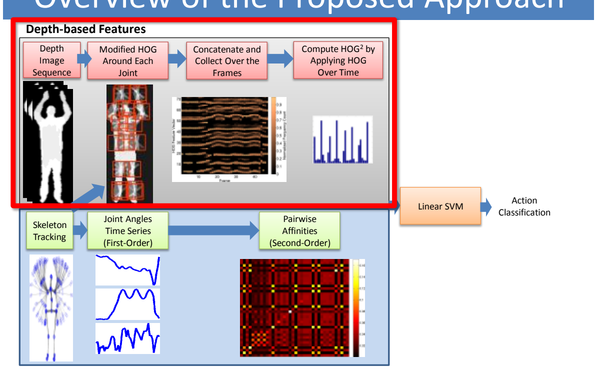
$$d_{weigthedEuc}\left(x_i, x_j\right) = \sum_{t=1:n} w_i(t)\left\|x_i(t) - x_j(t)\right\|_2^2 (1 + \lambda_{i,j}(t))$$

$$w_i(t) \propto \exp(-\frac{x_i(t)^2}{2(c^2)})$$

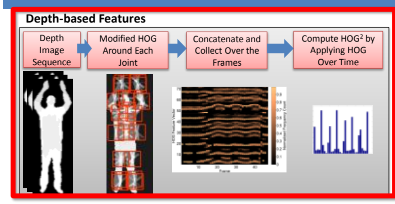
$$\boxed{s_{ij} = \frac{\exp(-d_{ij}/\sigma^2)}{\sum_{j=1:m} \exp(-d_{ij}/\sigma^2)}}$$

Final JAS Feature Set

Overview of the Proposed Approach



Spatio-Temporal HOG^2 Descriptor from Color or Depth Images



Modified HOG:

1) **Parameters** are the **number of blocks** in the x and y direction, and **orientation bins**

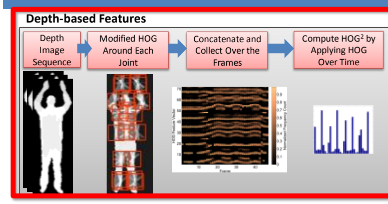
2) Gradient image => cells, 50% overlap => Orientation histogram for each cell => Concatenate

Example:

2 × 2 grid of cell with **8 histogram bins** results in a **32D feature vector**.

$$h_t = mHOG(I) = [h^1 \ldots h^{M \cdot N}]$$

Spatio-Temporal HOG^2 Descriptor from Color or Depth Images



Spatio-Temporal Feature Extraction:

$$\phi(I_1, \ldots, I_t) = mHOG\left(\begin{bmatrix} h_1 \\ \vdots \\ h_t \end{bmatrix}\right)$$

Block Normalization of the spatial and temporal histograms:

1) L2-norm: $\phi \rightarrow \phi / \sqrt{\|\phi\|_2^2 + \epsilon}$

2) L2-Hys: L2-norm followed by clipping and renormalization

3) L1-norm

4) L1-sqrt

L2-norm and L1-norm performed best

Experimental Evaluation – MSR-Action 3D Dataset

- Dataset Statistics:

skeleton and depth
20 actions
557 action samples
Cross-subject Testing

- Challenges: Joint position tracking is noisy, small inter-class variation

Existing results

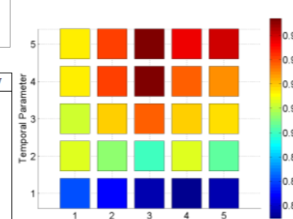
Method	Accuracy
DMM-HOG (Yang *et al.* [21])	85.52%
HON4D (Oreife and Liu [16])	85.8%
Random Occupancy Patterns (Wang *et al.* [19])	86.50%
Actionlet Ensemble (Wang *et al.* [20])	88.20%
HON4D + D_{disc} (Oreife and Liu [16])	88.89%

Experimental Evaluation – MSR-Action 3D Dataset

Method	Accuracy
DMM-HOG (Yang *et al.* [21])	85.52%
HON4D (Oreife and Liu [16])	85.8%
Random Occupancy Patterns (Wang *et al.* [19])	86.50%
Actionlet Ensemble (Wang *et al.* [20])	88.20%
HON4D + D_{disc} (Oreife and Liu [16])	88.89%

Method	Accuracy
JAS (LCSS)	53.95%
SVM on Joint Angles	80.29%
JAS (Cosine)	81.37%
SVM on Joint Angles+MaxMin	81.63%
JAS (Weighted Euclidean)	82.20%
JAS (Cosine)+MaxMin	83.53%
HOG^2+SVM on Joint Angles	91.72%
HOG^2	91.81%
JAS (Weighted Euclidean)+HOG^2	92.96%
JAS (Cosine)+HOG^2	93.66%
JAS (Cosine)+MaxMin+HOG^2	**94.84%**

Performance comparison of the proposed descriptors.



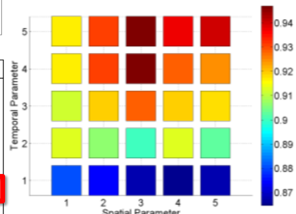
Accuracy of correct classification for **varying block size** in the HOG2 descriptor for a **fixed orientation** binning parameter of 9 bins. Results are shown after adding the best performing JAS feature.

Experimental Evaluation – MSR-Action 3D Dataset

Method	Accuracy
DMM-HOG (Yang *et al.* [21])	85.52%
HON4D (Oreife and Liu [16])	85.8%
Random Occupancy Patterns (Wang *et al.* [19])	86.50%
Actionlet Ensemble (Wang *et al.* [20])	88.20%
HON4D + D_{disc} (Oreife and Liu [16])	88.89%

Method	Accuracy
JAS (LCSS)	53.95%
SVM on Joint Angles	80.29%
JAS (Cosine)	81.37%
SVM on Joint Angles+MaxMin	81.63%
JAS (Weighted Euclidean)	82.20%
JAS (Cosine)+MaxMin	83.53%
HOG^2+SVM on Joint Angles	91.72%
HOG^2	91.81%
JAS (Weighted Euclidean)+HOG^2	92.96%
JAS (Cosine)+HOG^2	93.66%
JAS (Cosine)+MaxMin+HOG^2	**94.84%**

Performance comparison of the proposed descriptors.



Accuracy of correct classification for **varying block size** in the HOG2 descriptor for a **fixed orientation** binning parameter of 9 bins. Results are shown after adding the best performing JAS feature.

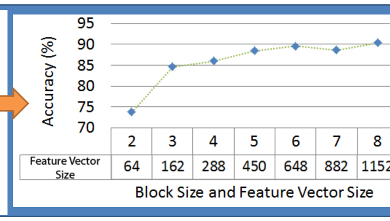
Experimental Evaluation – MSR-Hand Gesture Dataset

- Dataset Statistics:

Depth only,
12 gestures
333 sequences
leave-one-subject-out cross validation.

Method	Accuracy
HOG 3D (Klaser *et al.* [8])	85.23%
HON4D (Oreife and Liu [16])	87.29%
Random Occupancy Patterns (Wang *et al.* [19])	88.5 %
DMM-HOG (Yang *et al.* [21])	89.20%
HON4D + D_{disc} (Oreife and Liu [16])	92.45%
HOG^2	**92.64%**

Varying block size in the spatial and temporal stages of the HOG2 descriptor for a fixed orientation binning parameter of 9 bins. The figure exhibits the strength of the descriptor even with a **small sized feature set**.



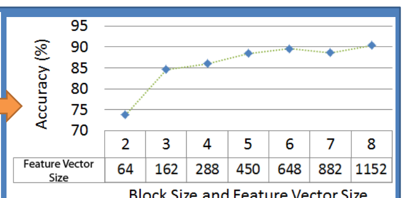
Experimental Evaluation – MSR-Hand Gesture Dataset

- Dataset Statistics:

Depth only,
12 gestures
333 sequences
leave-one-subject-out cross validation.

Method	Accuracy
HOG 3D (Klaser *et al.* [8])	85.23%
HON4D (Oreife and Liu [16])	87.29%
Random Occupancy Patterns (Wang *et al.* [19])	88.5 %
DMM-HOG (Yang *et al.* [21])	89.20%
HON4D + D_{disc} (Oreife and Liu [16])	92.45%
HOG^2	**92.64%**

Varying block size in the spatial and temporal stages of the HOG2 descriptor for a fixed orientation binning parameter of 9 bins. The figure exhibits the strength of the descriptor even with a **small sized feature set**.



Observational Latency Evaluation – UCF-Kinect Dataset

- **Latency**: JAS using partial gesture information?

- Dataset Statistics:

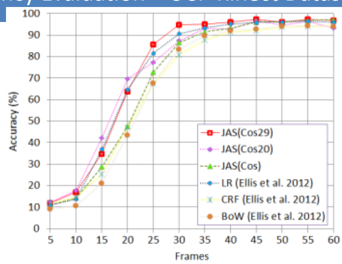
16 actions
Gaming applications
1280 gesture instances
High-quality skeleton tracking

Baseline: Ellis et al. *IJCV* 2012. Logistic Regression model, **2776-D** feature set. Best performance: **95.94%.**

JAS: **394-D** feature set. Best performance: **97.07%.**

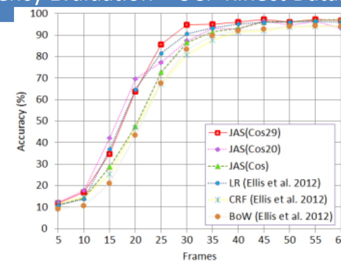
20

Latency Evaluation – UCF-Kinect Dataset



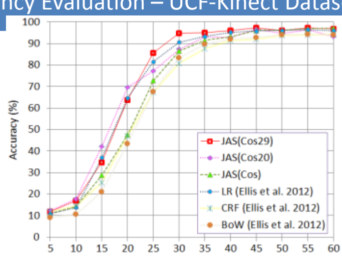
Method \ *Frames*	20	25	30	45	50	55	60
JAS (Cos29)	63.85	**85.45**	**94.72**	**97.37**	**96.05**	**97.37**	96.83
JAS (Cos20)	**69.66**	77.25	87.34	96.58	94.47	96.32	93.09
JAS (Cos)	47.63	72.75	86.54	96.05	**96.05**	96.84	**97.07**
LR (Ellis *et al.* [3])	64.77	81.56	90.55	95.78	96.1	96.48	95.94
CRF (Ellis *et al.* [3])	46.88	67.27	80.7	91.81	93.75	93.98	94.29
BoW (Ellis *et al.* [3])	43.52	67.58	83.2	92.73	93.98	94.22	94.06

Latency Evaluation – UCF-Kinect Dataset



Method \ *Frames*	20	25	30	45	50	55	60
JAS (Cos29)	63.85	**85.45**	**94.72**	**97.37**	**96.05**	**97.37**	96.83
JAS (Cos20)	**69.66**	77.25	87.34	96.58	94.47	96.32	93.09
JAS (Cos)	47.63	72.75	86.54	96.05	**96.05**	96.84	**97.07**
LR (Ellis *et al.* [3])	64.77	81.56	90.55	95.78	96.1	96.48	95.94
CRF (Ellis *et al.* [3])	46.88	67.27	80.7	91.81	93.75	93.98	94.29
BoW (Ellis *et al.* [3])	43.52	67.58	83.2	92.73	93.98	94.22	94.06

Latency Evaluation – UCF-Kinect Dataset



Method \ *Frames*	20	25	30	45	50	55	60
JAS (Cos29)	63.85	**85.45**	**94.72**	**97.37**	**96.05**	**97.37**	96.83
JAS (Cos20)	**69.66**	77.25	87.34	96.58	94.47	96.32	93.09
JAS (Cos)	47.63	72.75	86.54	96.05	**96.05**	96.84	**97.07**
LR (Ellis *et al.* [3])	64.77	81.56	90.55	95.78	96.1	96.48	95.94
CRF (Ellis *et al.* [3])	46.88	67.27	80.7	91.81	93.75	93.98	94.29
BoW (Ellis *et al.* [3])	43.52	67.58	83.2	92.73	93.98	94.22	94.06

Conclusions

Contributions

- **HOG²**: Proposed a new spatio-temporal descriptor based on a modified HOG, applied at every frame, collected into a 2D array, and then applied again.
- **JAS**: Experimentally validated that characterizing gestures using angle affinities with distance functions not allowing for time-shifts and gaps is a good idea.
- **Evaluation** on three datasets in different domains of human-machine interaction, and in terms of classification accuracy and latency.
- Relatively low-dimensional feature set and a Linear SVM suitable for real-time applications.

Future Work

- Discriminative choice of features.
- Multiple people interaction.

24