Joint Angles Similiarities and HOG^2 for Action Recognition

Eshed Ohn-Bar and Mohan M. Trivedi
Computer Vision and Robotics Research Lab
University of California, San Diego
La Jolla, CA 92093-0434
{eohnbar,mtrivedi}@ucsd.edu

Abstract

We propose a set of features derived from skeleton tracking of the human body and depth maps for the purpose of action recognition. The descriptors proposed are easy to implement, produce relatively small-sized feature sets, and the multi-class classification scheme is fast and suitable for real-time applications. We intuitively characterize actions using pairwise affinities between view-invariant joint angles features over the performance of an action. Additionally, a new descriptor for spatio-temporal feature extraction from color and depth images is introduced. This descriptor involves an application of a modified histogram of oriented gradients (HOG) algorithm. The application produces a feature set at every frame, and these features are collected into a 2D array which then the same algorithm is applied to again (the approach is termed HOG^2). Both feature sets are evaluated in a bag-of-words scheme using a linear SVM, showing state-of-the-art results on public datasets from different domains of human-computer interaction.

1. Introduction

Action representation and classification is an active branch of computer vision and pattern recognition, with many potential applications to human-machine interactivity. In this paper we present three contributions. First, we introduce a spatio-temporal feature for depth maps, based on a modified histogram of oriented gradients (HOG) [4, 14]. In the case of full-body gestures, the proposed descriptor (termed HOG^2) involves applying the algorithm spatially at each frame in box regions around each joint given by the skeleton tracker, thereby producing a histogram descriptor set for each frame. Then, in order to capture temporal dynamics in the spatial feature sets, we re-apply the algorithm over time.

A second main contribution is by exploring pairwise skeleton-based features that have not been extensively studied in the gesture classification literature. Such features differ from other common techniques in the field in several ways. For instance, unlike a common technique to model joint trajectories as independent time-series', we experimentally show that a more powerful descriptor can be formed using pairwise affinities of joint trajectories along a gesture. For action recognition, such a descriptor works best when derived using simple distance functions, such as the Euclidean distance, as opposed to those allowing time-shifts and gaps (e.g. the Longest Common Subsequence distance (LCSS) or dynamic temporal warping (DTW)) but are often used in studying trajectory clustering.

Finally, the third contribution stems from the computational speed and relatively low-dimensional feature set compared to other techniques on the same public datasets achieving state-of-the-art. The two descriptors in the core of the framework are studied with a linear SVM, suitable for real-time applications.

2. Related Work

Recent efforts in the field of action recognition are surveyed in [6, 1]. We focus on efforts which are closest to this work, first in terms of the depth-based descriptor proposed, and then in terms of the skeleton-based features.

Common methods for depth-based feature extraction may use descriptors that were originally developed for color images on depth images. Yang *et al*. [21] employ a HOG [4] feature extraction after projecting the depth maps into three orthogonal planes and accumulating the depth maps throughout each gesture into a motion image. Occupancy-based features have been proposed in [20, 19]. These may be used around each joint location as provided by a skeleton tracker (similarly to our approach), or randomly sampled over the scene, as in [19]. In [16], a 4D histogram over depth, time, and spatial coordinates is used to encode the distribution of the surface normal orientation. The main difference between the aforementioned works and our paper is in the spatio-temporal extraction and modeling of the actions.

There have been several efforts for extending HOG to the

978-0-7695-4990-3/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPRW.2013.76

temporal domain [7, 13, 17, 9]. We show that using a modified (it is applied in sub-cells of the image with 50% overlap) HOG descriptor applied spatially in each frame around each joint provided by a skeleton tracker can be used to extract useful depth information. Next, a temporal descriptor is proposed-by collecting the spatial descriptors over time, and consequently applying the modified HOG algorithm again to extract a temporal HOG descriptor. Such an approach stems from a need to bridge object detection in images and temporal events (in particular, hand detection [2, 15]), and pose benefits that will be discussed in Section 4.

Motion models may employ joint trajectories as features. Since we are concerned with a multi-variate series modeling problem, DTW can be employed to perform action recognition by template matching [12]. Alternatively, other common motion models may utilize a hidden Markov model (HMM) [11], or a conditional random field (CRF) [5]. The jury is still out on whether using such high-level features, as opposed to only low-level image descriptors, is necessary. With the overarching goal of understanding naturalistic human activity-it's difficult to use such features on their own to capture subtle differences in actions performed, such as object-subject interactions. Therefore, skeleton-based features are integrated with a depth-based descriptor in this work.

Joint trajectories are transformed into joint angles to gain view invariance. Next, pairwise affinities are used in a bag-of-words approach. Relevant literature of such approaches is beyond the scope of this paper, but the main insight is in demonstrating that such features are more powerful than using the trajectories without this extra processing step. Recently used features leverage such information within a single frame or a small window in time of 2-3 frames [22, 3, 20]. Hence, these features are used to capture a static posture information that most resembles an action class. This is essentially different from our descriptor of the dynamics between angles along the *entire gesture*. The descriptor is shown to outperform some of these aforementioned methods with a significantly smaller sized feature set and a simpler framework. It is also immediately extendable to more complex activity analysis scenarios, such as multiple people interaction.

3. Joint Angles Affinity Clustering

The general scheme proposed in this paper is shown in Fig. 1. We observed that it was common to model joint trajectories or joint angles as independent trajectories. That is, template based approaches would involve comparison of the features of each joint separately (i.e. in classification, elbow joint trajectory in a gesture instance would be compared to a elbow joint trajectory in another instance). To explicitly distinguish our method, such features are categorized as *first-order*, and it will show that useful information for the classification of gestures can be derived using affinities within sequences of joint angles along the same gesture. This processing step produces a set of *second-order features*.

3.1. Angular Skeletal Representation (First-Order Features)

In the proposed feature extraction method, we use pairwise affinities between the joint angles along the gesture. Such features naturally arise from the hierarchical nature of the skeleton. Topologically, the human skeleton is an open directed graph that can be depicted with a particular joint as the root, and the other joints connected to the root in a hierarchical manner. In such a tree, every node has exactly one parent node (except for the root node). Nodes may have zero or more children nodes below it. Descendants nodes inherit a component of rotational and transnational transformations from their ancestors nodes in a relationship known as forward kinematics. We aim to effectively incorporate the relationship between the joints in the skeleton throughout the gesture into our feature set.

High-quality skeleton tracking data, such as the one outputted by the Kinect camera, allows us to map motion information in the scene into a smaller set of features-point trajectories. A depth-first tree traversal gives the relative azimuth and elevation angles of each joint with respect to its parent node. These will be referred to as first-order features, and may be used as input to classification tools [18]. For example, in order to calculate the first-order features at the left elbow (lE) joint, we translate the sensor coordinate system such that the origin is at the left shoulder (lS). Next, we construct a spherical coordinate system so that now the vector $\overrightarrow{(lS, lE)}$ is in terms of (r,θ,φ) with θ as the elevation angle (from the x-y plane), and φ as the azimuth angle (from the positive x-axis). We drop the radius so that any i^{th} joint is associated with two first-order features, $S_i = \{\theta_i, \varphi_i\}$. Hence, we derive a skeleton configuration $K^t = \bigcup_{i=1:p} S_i^t$ (where p is the number of joints tracked) at each time step t.

3.2. JAS-Joint Angles Pairwise Similarities (Second-Order Features)

The set of first-order features is transformed using a similarity measure into a set of second-order features. These are shown to outperform a first-order feature set through experimental validation in Section 5. Given a distance function $d : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ (where n is the number of frames in the gesture instance) the time series data is converted into a distance matrix of joint angle similarities between each of the angles along the entire action time series. The final set of features is therefore a $\frac{m(m-1)}{2}$ long feature vector, where

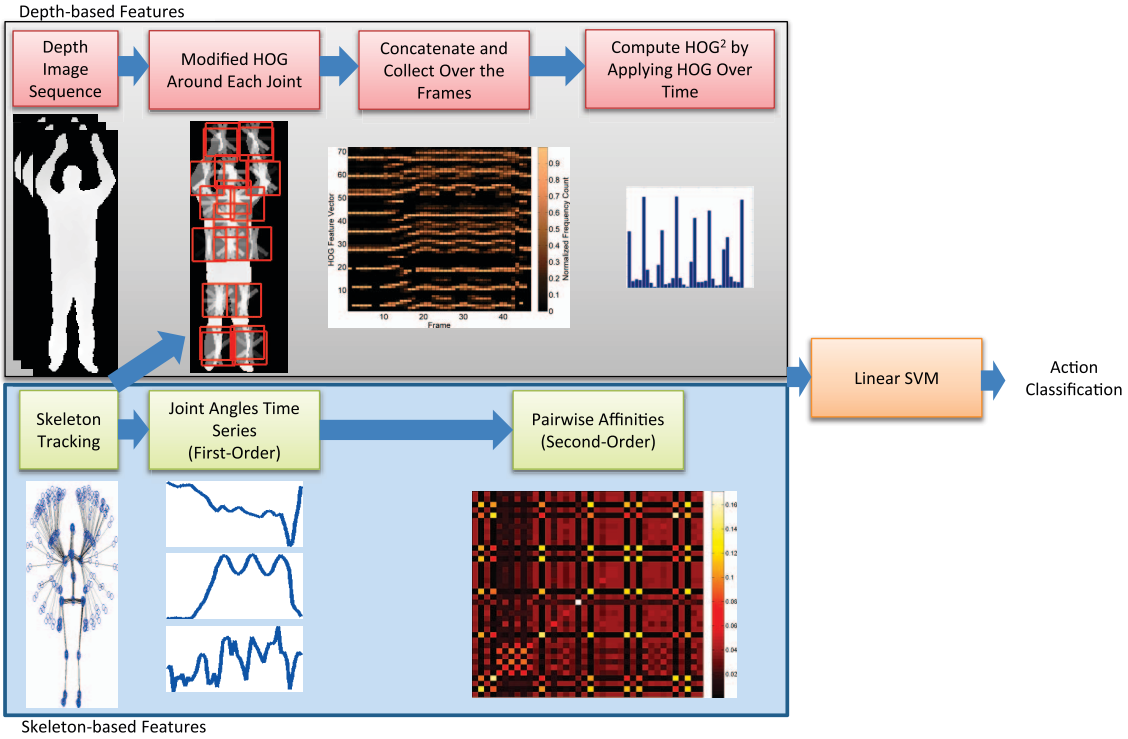


Figure 1: Overview of the feature extraction and action classification scheme proposed in this paper.

m is the number of angles in the angular skeleton representation.

After experimenting with a wide array of distance functions, simple ones were shown to perform best. In this work, three functions will be used for demonstration. These are compared against a first-order feature set, where the length of each time series is interpolated to 60. Given two vectors of joint angles over a gesture instance, $x_i, x_j \in K$, the *cosine* distance between them is defined as

$$d_{cos}(x_i, x_j) = \frac{x_i^T x_j}{\|x_i\|_2 \|x_j\|_2} \tag{1}$$

We also investigate another distance function, referred to as *weighted Euclidean*:

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

the weight term $w_i(t) \propto exp(-x_i(t)^2/2(c^2))$ provides higher weight to points in the middle of the gesture instance and lower near the beginning and end. $\lambda_{i,j}(t)$ is a penalty term introduced in order to avoid propagating errors in case of a noisy trajectory in a subset of the tracked joints. $\lambda_{i,j}(t) = 0$ or $\lambda_{i,j}(t) = a > 0$ depending on the noise in the time series. Using $c = 1$ was shown to improve results on the MSR-Action3D dataset as opposed to simply using the Euclidean distance.

The Euclidean or cosine distance functions are known to be intolerable to time shifts between time series as well as gaps. On the other hand, the *longest common subsequence* (LCSS) is robust to these. It is defined as,

$$d_{LCSS}(x_i, x_j) = 1 - \frac{LCSS(x_i, x_j)}{min(|x_i|, |x_j|)} \tag{3}$$

Where, given two sequences $A = (a_1, ..., a_{m_0})$ and $B = (b_1, ..., b_{n_0})$, the LCSS measure is defined as

$$\text{LCSS(A,B)} = \begin{cases} 0 \text{ if A or B is empty} \\ \\ 1 + LCSS(Head(A), Head(B)) \\ \text{if } |a_{m_0} - b_{n_0}| < \epsilon_1 \text{ and } |m_0 - n_0| < \theta_0 \text{ and} \\ \ |a_{m_0-1} - a_{m_0}| > \epsilon_2 \text{ and } |b_{n_0-1} - b_{n_0}| > \epsilon_2 \\ \\ Max\{LCSS(Head(A), B), \\ LCSS(A, Head(B))\} \\ \text{otherwise.} \end{cases} \tag{4}$$

Head(A) is the remaining sequence of A after removing the last element in the sequence, ϵ_1 determines if $a_{m_0} \in A$ and $b_{n_0} \in B$ are matched or not, ϵ_2 measures the similarity only when the joint angles are changing, and θ_0 tolerates time shifting between the two sequences.

The outputs of the distance function can be turned into affinities using

$$s_{ij} = \frac{exp[-d_{ij}/\sigma^2]}{\sum_{j=1:m}[exp(-d_{ij}/\sigma^2)} \quad (5)$$

yielding a $m \times m$ similarity matrix. We use $\sigma = 0.7$ in all of the experiments. Finally, adding a small set of descriptors of the maximum and minimum (referred to as **MaxMin** in Section 5) within each joint trajectory slightly improves classification accuracy. Finally, we solve for a mapping, ϕ, such that $\phi : \mathbb{R}^{m(m-1)/2} \rightarrow \chi$, where $\chi := \{1, ..., k\}$ is the multi-class label set using a linear SVM formulation and a one-against-one scheme.

4. Spatio-Temporal HOG^2 Descriptor from Depth Maps

Modified HOG Spatial Feature Extraction The modified HOG descriptor is created as follows: the gradient image of the image patch (using a centered mask $[-1, 0, 1]$) is divided into rectangular cells along the x- and y-directions. A 50% overlap between the cells is used. Within each cell, an orientation histogram is generated by quantizing the angles of each gradient vector into a pre-defined number of bins. These resulting histograms are concatenated to form the final spatial feature vector. For instance, a 2×2 grid of cells with 8 histogram bins on the image results in a 32D feature vector.

Spatio-Temporal Feature Extraction We collect the spatial descriptors over time to form a 2D array (shown in Fig. 1 for one joint). Changes in the feature vector correspond to changes in the shape and location of the joint. Consequently, the modified HOG algorithm is applied again to extract a temporal HOG descriptor. The approach is termed HOG^2, since it involves applying the same algorithm twice (once in the spatial domain, and then again in the temporal domain). The descriptor can be used on color or depth images.

Such an approach bridges the spatial and temporal feature extraction in a way that has implications to other fields in computer vision. It originated from a need to temporally extend a step of spatial application of HOG for object detection. By separating the two steps, fast spatial object detection approaches can first be used to detect a particular object in the scene. Next, if needed, the temporal extraction doesn't require as input the original image, which is of high dimensionality, but simply the representation of the image as a HOG descriptor. Representing changes in this descriptor, which is of lower dimensionality than the original image, allows for real-time gesture recognition. Finally, in order to produce the final feature vector we also append a vector of the average (over time) for each entry in the spatial feature vector.

Similar to in [4], we investigated several block normalization schemes. We may normalize the spatial descriptor differently from the temporal descriptor. If v is the un-normalized descriptor, spatial or temporal, it may be normalized using the *L2-norm*, $v \rightarrow v/\sqrt{\|(v)\|_2^2 + \epsilon}$, *L2-Hys*, which is an L2-norm followed by clipping (entries above a certain threshold) and re-normalizing, *L1-norm*, $v \rightarrow v/\|(v)\|_1 + \epsilon$, and the *L1-sqrt*, $v \rightarrow v/\sqrt{\|(v)\|_1 + \epsilon}$. There wasn't an observed benefit from different normalization schemes to different steps in deriving the HOG^2, and $L2 - norm$ and $L1 - norm$ performed equally well and slightly better than the other normalization schemes.

5. Experimental Evaluation

5.1. MSR-Action 3D Dataset

We perform evaluation of the proposed feature set on a recently introduced dataset containing both skeleton and depth information, the *MSR-Action3D* dataset [10]. It contains 20 actions, 10 subjects, and a total of 557 action samples. The dataset is challenging due to the small inter-class variations among actions, and the skeleton tracker fails often. Therefore, the tracked joint positions contain significant noise at times. We follow cross-subject test settings, where the first five actors are used in training and the rest for testing.

We note that skeleton-based methods alone generally performed below 70% on the dataset, yet JAS produced good results on its own using a 703D feature vector, a relatively small feature set (the distance measure is shown in parenthesis in Table 2). As mentioned befre, distance functions that don't allow for time-shifts or gaps better represent the unique motion in a particular action class. The parameters for the HOG^2 are optimizied together with the appended JAS feature set in. This is shown in Fig. 2 where the two parameters in the HOG^2 feature (a square block size used for binning spatially and temporally) for a fixed orientation quantization parameter of 9 bins were optimized. A box of size 60×60 pixels centered at each joint was shown to perform best. For the final results we chose a $3 \times 3 \times 8$ over space, and $4 \times 4 \times 9$ over time for a final feature set of size $(4 \times 4 \times 9 + 3 \times 3 \times 8) \times 20 = 4320$. Even lower-dimensional feature sets performed at state-of-the-art as shown in Fig. 2.

Table 1 shows the state-of-the-art scheme of HON4D+D_{disc} [16]. We include both the scheme with a discriminative learning refinement, and the performance of the descriptor itself to emphasize the little effort that was put into refinement of the raw feature set in our scheme (besides tweaking two parameters). Discriminative learning refinement is left for future work.

5.2. MSR-Hand Gesture Dataset

Introduced in [19], this is a depth-only dynamic hand gesture dataset containing 12 American sign language ges-

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%

Table 1: Existing results on the MSR-Action3D dataset.

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**%

Table 2: Performance comparison of our proposed descriptors on the MSR-Action3D dataset, with different distance functions for deriving the second-order feature set, a first-order feature set interpolated to length of 60 across the samples, and the depth-based descriptor, HOG^2.

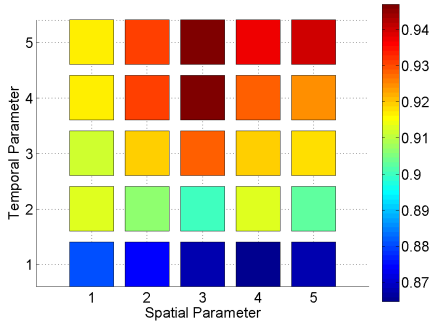


Figure 2: Accuracy of correct classification in cross-subject settings on the MSR-Action3D dataset for varying block size in the HOG^2 descriptor for a fixed orientation binning parameter of 9 bins. Results are shown after adding the best performing JAS feature from Table 2 in order to assure minimum redundancy.

tures. Wang *et al.* [19] proposed a Sparse coding technique to handle to challenging self-occlusion issues with this dataset. A total of 333 depth sequences performed by 10 subjects are tested in leave-one-subject-out cross validation. Results are shown in Table 3.

The HOG^2 descriptor is applied on the entire image using different parameter configurations. Fig. 3 shows the results on the dataset by choosing the same parameter across all cell sizes, both spatially and temporally, and a fixed binning parameter of 9. By varying this one parameter, we achieved good results on the dataset. The best results on this dataset were using a $8 \times 8 \times 9$ spatially and $7 \times 7 \times 9$ temporally at 91.81%, and $8 \times 8 \times 18$ and $8 \times 8 \times 18$ at 92.64% (shown in Table 3). As in the previous dataset, we compare the results of our descriptor to both schemes in [16] to emphasize the lack of refinement step in our algorithm and the little effort put into optimizing the descriptor's parameters, producing good results with a small sized feature vector.

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 [19])	92.45%
HOG^2	**92.64**%

Table 3: Results of our approach on the MSR-Hand Gesture 3D dataset compared to previously published results.

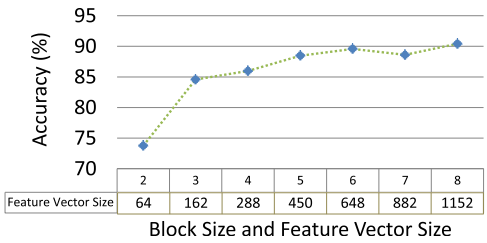
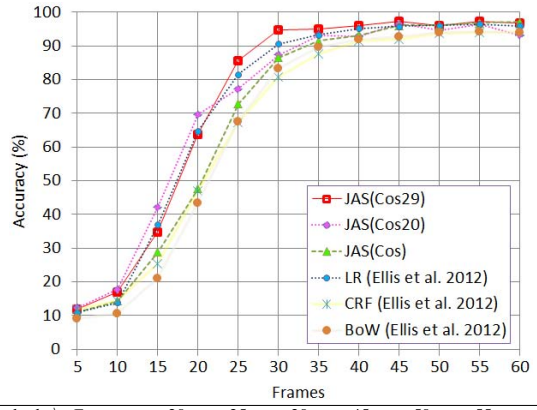


Figure 3: Accuracy of correct classification on the MSR-Hand Gesture 3D dataset for varying block size in the spatial and temporal stages of the HOG^2 descriptor for a fixed orientation binning parameter of 9 bins. The figure exhibits the strength of the descriptor even with a small sized feature set.

5.3. UCF-Kinect Dataset [3]

Finally, we evaluate the JAS scheme in terms of latency, to show its effectiveness even when partial gesture information is available. its own under a larger dataset with more reliable skeleton tracking. The dataset contains 16 actions suitable for a gaming environment, 16 subjects, and a total of 1280 actions samples. A 70/30 split is used for training/testing.

As before, incorporating a small number of first-order features was shown to slightly improve latency. These are the maximum and minimum azimuth and elevation angles in the elbow and knee joints along the gesture performance. Since there are 28 joint angles for body configurations in this dataset, with these additional 16 first-order features we get a total of 394 features for each gesture. This is quite small compared to the feature set used to evaluate this dataset in [3], which is of size 2776 and provides the best classification performance of 95.94%. JAS achieves im-



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

Figure 4: Results on the UCF-Kinect dataset [3] in terms of observational latency. The latency evaluation of the JAS descriptor using a cosine distance and a linear SVM.

proved latency performance compared to the two baseline methods in [3]-CRF and *bag of words* (BoW) model-from 15 frames into the performance of the gestures. Towards the end of the gesture (frame 45) our method outperforms in classification accuracy the method proposed in [3]. Expanding the JAS feature set to include pairwise affinities between the joint angles in the initial frame of the gesture instance and the angles in the 20^{th} or 29^{th} frame (or the latest frame if less than 29 frames occurred) for a total features set size of 1178D significantly outperforms in latency all of the other methods shown in Fig. 4.

6. Conclusion

Two descriptors were proposed in this work, one of joint angle similarities and another using a modified HOG algorithm which was used for a depth-based spatio-temporal feature extraction. The proposed features were tested for multiple applications of gesture recognition, both alone and combined, achieving state-of-the-art. The features are lightweight in size, implementation, and with a linear SVM are suitable for real-time gesture recognition.

References

[1] J. Aggarwal and M. Ryoo. Human activity analysis: A review. *ACM Comput. Surv.*, 2011.

[2] S. Y. Cheng and M. M. Trivedi. Vision-based infotainment user determination by hand recognition for driver assistance. *IEEE Trans. Intell. Transp. Syst.*, 11(3):759–764, Sep. 2010.

[3] E. Chris, M. Syed, T. Marshall, L. Joseph, and S. Rahul. Exploring the trade-off between accuracy and observational latency in action recognition. *IJCV*, 2012.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[5] L. Han, X. Wu, W. Liang, G. Hou, and Y. Jia. Discriminative human action recognition in the learned hierarchical manifold space. *Image and Vision Computing*, 2010.

[6] M. B. Holte, C. Tran, M. M. Trivedi, and T. B. Moeslund. Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments. *IEEE J. Sel. Topics Signal Process.*, 2012.

[7] M. Kaaniche and F. Bremond. Tracking HoG descriptors for gesture recognition. In *AVSS*, 2009.

[8] A. Klaser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3D-gradients. In *BMVC*, 2008.

[9] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[10] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3d points. In *CVPRW*, 2010.

[11] F. Lv and R. Nevatia. Recognition and segmentation of 3-D human action using hmm and multi-class adaboost. In *ECCV*. 2006.

[12] M. Müller and T. Röder. Motion templates for automatic classification and retrieval of motion capture data. In *ACM SIGGRAPH*, 2006.

[13] J. Niebles, C.-W. Chen, and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *ECCV*. 2010.

[14] E. Ohn-Bar, C. Tran, and M. Trivedi. Hand gesture-based visual user interface for infotainment. In *AutomotiveUI*, 2012.

[15] E. Ohn-Bar and M. M. Trivedi. In-vehicle hand localization using integration of regions. *IEEE Conf. Intell. Veh. Symp.*, 2013.

[16] O. Oreifej and Z. Liu. HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences. In *CVPR*, 2013.

[17] C. Thurau and V. Hlavac. Pose primitive based human action recognition in videos or still images. In *CVPR*, 2008.

[18] C. Tran and M. Trivedi. 3-D posture and gesture recognition for interactivity in smart spaces. *IEEE Trans. Ind. Informat.*, 2012.

[19] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu. Robust 3d action recognition with random occupancy patterns. In *ECCV*, 2012.

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

[21] Y. Xiaodong, C. Zhang, Y. Wu, and Y. Tian. Recognizing actions using depth motion maps-based histograms of oriented gradients. In *ACM Multimedia*, 2012.

[22] K. Yun, J. Honorio, D. Chattopadhyay, T. Berg, and D. Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *CVPRW*, 2012.