Integrating Depth-HOG and Spatio-Temporal Joints Data for Action Recognition

Noopur Arora
Indian Institute of Technology, Delhi
Hauz Khas, New Delhi-110016, India
mcs142128@cse.iitd.ac.in

Parul Shukla
Indian Institute of Technology, Delhi
Hauz Khas, New Delhi-110016, India
parul@cse.iitd.ac.in

Kanad K. Biswas
Indian Institute of Technology, Delhi
Hauz Khas, New Delhi-110016, India
kkb@cse.iitd.ernet.in

ABSTRACT

In this paper, we propose an approach for human activity recognition using gradient orientation of depth maps and spatio-temporal features from body-joints data. Our approach is based on an amalgamation of key local and global feature descriptors such as spatial pose, temporal variation in 'joints' position and spatio-temporal gradient orientation of depth maps. Additionally, we obtain a motion-induced global shape feature describing the motion dynamics during an action. Feature selection is carried out to select a relevant subset of features for action recognition. The resultant features are evaluated using SVM classifier. We validate our proposed method on our own dataset consisting of 11 classes and a total of 287 videos. We also compare the effectiveness of our method on the MSR-Action3D dataset.

Keywords

Action Recognition, Depth-HOG, Kinect, Body-Joints Data

1 INTRODUCTION

Human action recognition has been an active area of research for over a decade. With the proliferation of online videos and personalized cameras, the task of human action recognition for applications such as content-based video retrieval, surveillance, human-computer interaction has attained newer meanings. Further, the introduction of depth sensors such as Microsoft Kinect has added a new dimension. The depth data available from Kinect consists of depth maps and body-joints data. A number of ways have been used in the literature for action recognition from depth data [16], [4], [18], [8], [9], [21], [10]. Broadly, these could be categorized as methods that are based on data from depth maps and those, which use joints data.

Li et al.[8] use action graph to model the dynamics of action from depth maps sequences. They use a bag of 3D points to characterize a set of salient postures corresponding to nodes in action graph. Ni et al.[9] use depth-layered multi-channel representation based on spatio-temporal interest points. They propose a multi-modal fusion scheme, developed from spatio-temporal interest points and motion history images, to combine color and depth information. In [21], the average difference between the depth frames is computed and summarized in a single Depth Motion Maps (DMM), from which Histogram of Oriented Gradients features (HOG) are extracted. Oreifej and Liu [10] construct an activity descriptor called Histogram of Oriented 4D surface normal analogous to the histogram of gradients in color sequences. Jetley and Cuzzolin [3] divide the video into temporally overlapping blocks and generate motion history template (MHT) and binary shape template (BST) for each block. Gradient analysis is performed on MHT and BST to describe motion and shape respectively.

Amongst approaches driven by body-joints data, Sung et al.[16] use features extracted from estimated skeleton and use a two-layered Maximum-Entropy Markov Model (MEMM) where the top layer represents activites and the mid-layer represents sub-activities connected to the corresponding activites in top layer. In [4], the authors propose an encoding scheme to convert skeleton data into symbolic representation and use longest common subsequence for activity recognition. Wang et al.[18] use skeleton data and depth maps to construct novel Local Occupancy Pattern (LOP) feature wherein, each 3D joint is associated with a LOP feature which can be treated as depth appearance of a joint. They further propose fourier temporal pyramid and use these features in a mining approach to obtain a

subset of joints or an actionlet. In [14], the authors extract features in spherical coordinate system from body-joints data. The features are represented using bag-of-joint-features (BoJF) model for each joint. To incorporate temporal variations of an action, a hierarchical-temporal histogram (HT-hist) model is used. A new relational geometric feature called Trisarea has been proposed in [17]. It is a pose-based feature defined as the area of trianlge formed by three joints. An approach for reducing pose data over time to histograms of relative location, velocity, and their correlations has been presented in [2]. Subsequently, the partial least squares have been used to learn a compact and discriminative representation for an action sample.

The use of depth maps has the advantage that cues such as shape and geometry are better represented. Body-joints data, on the other hand, provides pose information which has been known to facilitate action recognition as humans tend to recognize actions easily from a sequence of poses. In this paper, we exploit both the data streams by learning a model based on features extracted from depth maps as well as body-joints data. The features extracted can be categorized as local or global depending on whether the feature descriptors are defined over a local region or the entire video volume. In this paper we propose a novel scheme by integrating both the depth maps and joints data. We estimate Gradient Orientation from depth maps (*depthHOG*) and motion-induced shape (*MIS*) features from depth maps. Further, we augment these features with Relative Joint Distance (*RJD*) and Temporal Joint Distance (*TJD*) features obtained from body-joints data.

The rest of the paper is organized as follows: Section 2 presents the proposed approach. In section 3, we present the experiments and results. Finally, in section 4 we discuss the conclusion and future extensions.

2 PROPOSED APPROACH

In this section, we present our proposed approach based on fusion of key local and global attributes such as pose, temporal joint distance, orientation of gradient and motion information.

2.1 Local Attributes

2.1.1 Spatial Features

It has been widely acknowledged that humans tend to recognize actions easily from a sequence of poses. We use this idea to extract spatial pose-based features, Relative Joint Distance (*RJD*), by computing mean of joint positions in each frame. Let it be denoted by μ^f. Subsequently, in each frame f we compute a Relative Joint Distance (*RJD*) R_j^f of a joint j from the mean as follows:

$$R_j^f = ||p_j^f - \mu^f|| \quad (1)$$

where $p_j^f(x,y,z)$ is the 3D position of a joint j in frame f and μ^f is the mean position of all the given joints in a frame f. We normalize the *RJD* with respect to the height(H) of a person as follows:

$$\hat{R}_j^f = R_j^f / H \quad (2)$$

The *RJD* of each joint over all the frames is concatenated to yield the final spatial descriptor from body-joints data. In particular, we have a 20-dimensional *RJD* feature vector corresponding to the 20 body-joints in a frame. Further, since the execution speeds of an action may vary for different actors, we select N number of frames with a step size of n_f/N and compute *RJD* in these frames only, where n_f is the number of frames in a video. The resultant $N*20$ features capture spatial pose information. However, if an action involves movements such as circular motion of an arm or waving of hands, there will not be significant change in pose. Therefore, there is a need to augment spatial pose features with information from other sources as well.

2.1.2 Temporal Features

We propose to augment spatial pose features with Temporal Joint Distance (*TJD*) features extracted from body-joints data. As with the spatial pose features, we first select N frames from a video sequence of n_f frames. We then compute *TJD* for the selected frames as follows:

$$T_j^f = ||p_j^f - p_j^{f+1}|| \quad (3)$$

Since there are N selected frames, the resultant *TJD* consists of $(N-1)*20$ features.

2.1.3 Spatio-Temporal Features

The *RJD* and *TJD* features are extracted from body-joints data. Additionally, we use depth map sequence to exploit cues such as shape, which are better represented in depth maps. We obtain gradient based spatio-temporal features, henceforth referred to as *depthHOG*. Use of histogram of gradients(HOG) for action recognition has been reported earlier in the literature for RGB data [13], [5], [11], [7]. In [5], the authors compute gradients in spatio-temporal pyramid and use regular polyhedrons for quantization of 3D orientations. In [11], the authors combine histogram of gradients into orientation tensors per frame.

As a pre-processing step, we normalize the input depth map by performing histogram equalization of intensity values within a person mask on each frame. The normalization step results in the depth values of person being covered over the entire intensity range. We then compute gradient (G_x, G_y, G_t) of the depth map sequence along the x, y and t directions. Let $D(i,j,f)$ denote the depth value at pixel (i,j) and frame f. The gradients are computed using the following:

$$G_x(i,j,f) = D(i,j+1,f) - D(i,j-1,f) \quad (4)$$

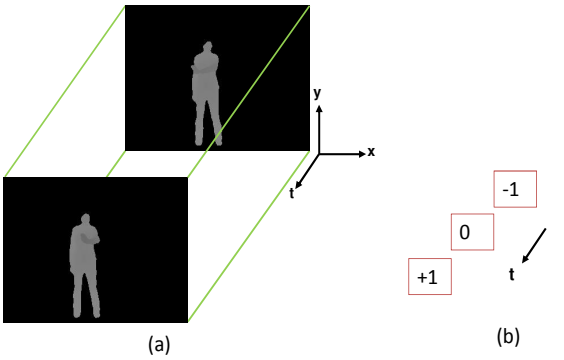

Figure 1: (a)Depth map sequence. (b)Gradient mask for a pixel along temporal domain.

$$G_y(i,j,f) = D(i+1,j,f) - D(i-1,j,f) \quad (5)$$

$$G_t(i,j,f) = D(i,j,f+1) - D(i,j,f-1) \quad (6)$$

Figure 1(a) shows a sample depth map sequence for 'hand wave' action. Figure 1(b) shows the gradient mask across temporal domain. We use the computed gradients (G_x, G_y, G_t) to find local 3D orientations in depth maps. Let $G_x(i,j,f)$ denote the gradient at pixel (i,j) and frame f computed along x direction. Similarly $G_y(i,j,f)$ and $G_t(i,j,f)$ denote the gradients computed along y and t directions respectively. In order to find the local 3D orientation of depth gradients, we convert G_x, G_y, G_t values into spherical coordinates. This results in a gradient magnitude $M(i,j,f)$ and angles $\theta(i,j,f)$ and $\phi(i,j,f)$.

$$M = \sqrt{G_x^2 + G_y^2 + G_t^2}\ ,\ M \geq 0 \quad (7)$$

$$\phi = \arccos\left(G_t/M\right)\ ,\ 0 \leq \phi \leq \pi \quad (8)$$

$$\theta = \arctan\left(G_y/G_x\right)\ ,\ 0 \leq \theta < 2\pi \quad (9)$$

Although, $\tan(\theta)$ is defined for $-\pi/2 \leq \theta \leq \pi/2$, we map the values in the range $0 \leq \theta < 2\pi$. It may be noted that there is a slight variation from the formulation in [7], in that, their formulation is for RGB data whereas ours is on depth maps. Secondly, in our case, ϕ signifies the orientation of gradient vector with respect to the temporal axis whereas in [7], ϕ is the angle that the gradient vector makes with its projection on the x-y plane.

The aggregation of the orientation values over the depth map sequence is done by dividing the depth map sequence into a spatio-temporal grid. In order to construct such a grid, we consider a Region of Interest (ROI) for a depth map sequence by finding a maximum of all possible bounding boxes (a bounding box contains a person) in a depth map sequence. We then divide this region

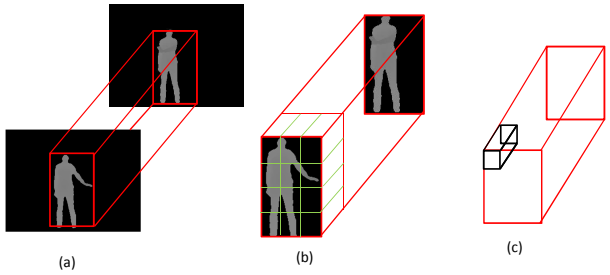

Figure 2: (a)Maximum bounding box for a depth map sequence. (b)Spatio-Temporal grid. (c)Cell in a Spatio-Temporal grid.

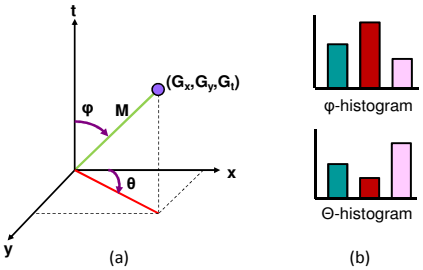

Figure 3: (a)Spherical coordinates for gradient of a pixel. (b)*depthHOG* in a cell.

into a grid consisting of $n_x * n_y$ cells in the spatial domain and n_t cells in the temporal domain. For aggregating the gradient orientations in a cell, we quantize the θ and ϕ angles into n_θ and n_ϕ bins respectively and the bins are weighted according to the gradient magnitude.

Figure 2 illustrates the process of cell creation. Figure 3 illustrates the conversion of pixel gradient into spherical coordiante system and the *depthHOG* as two 1D histograms, namely $\theta - histogram$ and $\phi - histogram$. Each histogram is normalized within a cell. Figure 4(a) and 4(b) illustrates the process of creating angular bins for ϕ and θ. Figure 4(c) and 4(d) illustrate sample histograms in a cell.

The histograms from all the cells are concatenated to give the final *depthHOG* features. The *depthHOG* features are obtained by concatenating $n_x * n_y * n_t$ histograms for both n_θ and n_ϕ bins. A typical choice of the parameters for creating spatio-temporal grid and gradient orientation bins is given as $n_x = 5$, $n_y = 8$, $n_t = 6$, $n_\theta = 12$, $n_\phi = 6$. This would result in 4320 *depthHOG* features.

2.2 Global Attributes

Recent research [21], [3], suggests that additional body shape and motion information from projections of depth map onto three orthogonal planes can be used to enhance performance of action recognition systems. We use this idea to define a Motion-Induced-Shape (*MIS*) feature. Yang et al. [21] obtain three 2D maps corresponding to top, front and side views for each

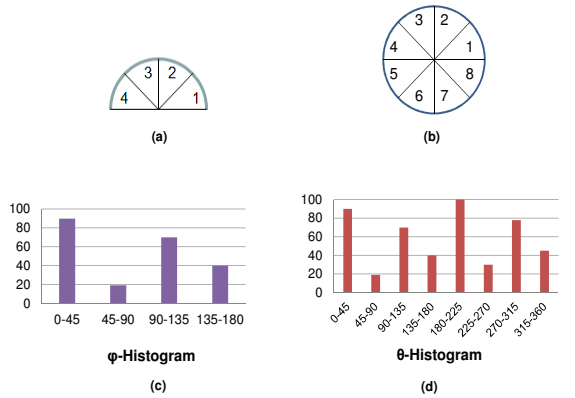

Figure 4: (a)Illustrative example showing 4 angular bins for ϕ. (b)Illustartive example showing 8 angular bins for θ. (c)-(d)Sample $\phi - Histogram$ and $\theta - Histogram$ for a cell.

depth frame. And for each projected map, obtain motion energy by computing and thresholding the difference between two consecutive maps. This, however, requires one to empirically set a threshold value. We modify this by extracting binary projections along the three directions. In particular, given a depth frame k, we obtain three masks B_k^f, B_k^s and B_k^t corresponding to the three views as:

- *Front view*: $B_k^f(i,j) = 1$, if $D(i,j,k) = z$ and $z > 0$
- *Side view*: $B_k^s(z,j) = 1$, if $D(i,j,k) = z$ and $z > 0$
- *Top view*: $B_k^t(i,z) = 1$, if $D(i,j,k) = z$ and $z > 0$

In all other cases, resultant pixel value will be 0. It may be noted that this procedure is applied only on human silhouette. Obtaining depth information of only human body has been greatly facilitated with devices such as Kinect.

We now aggregate the difference between consecutive binary masks as:

$$S_f(i,j) = \sum_{k=1}^{n_f-1} |B_k^f(i,j) - B_{k+1}^f(i,j)| \qquad (10)$$

$$S_s(i,j) = \sum_{k=1}^{n_f-1} |B_k^s(i,j) - B_{k+1}^s(i,j)| \qquad (11)$$

$$S_t(i,j) = \sum_{k=1}^{n_f-1} |B_k^t(i,j) - B_{k+1}^t(i,j)| \qquad (12)$$

where, $B_k^f(i,j)$ $B_k^s(i,j)$ and $B_k^t(i,j)$ are binary masks corresponding to front, side and top view of depth frame k for pixel (i,j), respectively. Next, we normalize the obtained motion maps as follows:

$$\hat{S}_f(i,j) = \frac{S_f(i,j) - min_f}{max_f - min_f} \qquad (13)$$

where, min_f and max_f are the minimum and maximum pixel values of S_f respectively. Similarly, we normalize S_t and S_s to obtain $\hat{S}_t$ and $\hat{S}_s$. Figure 5 illustrates the normalized motion maps for the 'High arm wave' action.

2.2.1 *Motion-Induced-Shape features*

We obtain *MIS* features by extracting HOG descriptor from the motion maps $\hat{S}_f$, $\hat{S}_t$, $\hat{S}_s$ corresponding to the three views. A typical choice of cell size is $c_x * c_y$ with number of orientation bins as $n_o = 9$ and a block size of $2*2$. c_x and c_y varies for different datasets.

The number of *MIS* features obtained from a single view (say front view) is given as $N_{MIS}^f = n_b * \delta_b * n_o$ where $n_b = n_b^x * n_b^y$ is the number of blocks, $\delta_b = b_x * b_y$ is the block size. Typical value of $b_x = b_y = 2$ indicates that a block consists of $2*2$ cells. If the image is of size $W * H$, then the number of blocks is given as:

$$n_b = \lfloor (\frac{\frac{W}{c_x} - b_x}{(b_x - b_o^x)} + 1) \rfloor * \lfloor (\frac{\frac{H}{c_y} - b_y}{b_y - b_y^o} + 1) \rfloor \qquad (14)$$

where $b_o^x * b_o^y$ denote the block overlap. Typically, $b_o^x = b_o^y = 1$. Likewise, N_{MIS}^s and N_{MIS}^t can be computed from $\hat{S}_s$ and $\hat{S}_t$ for side and top views respectively. Finally, the concatenated MIS descriptors from each of the three views constitute the final *MIS*.

2.3 Classification

The *RJD*, *TJD*, *depthHOG* and *MIS* features from a video are concatenated to form the final feature vector for the corresponding video. We perform classification on the features using SVM with RBF kernel. The resultant feature vector may contain some redundant or irrelevant features leading to large computational load on the classifier. We propose to obtain the most relevant set of features using a feature selection (FS) approach such as RELIEFF [6], [12]. It gives the relative importance of attributes or predictors by keeping into account k nearest neighbors in a class (called as nearest hits) and k nearest neighbors from each of the other classes (called as nearest misses). Prior probability of a class is taken into account while estimating the quality of an attribute.

Using RELIEFF we obtain a ranking order of all the features. From the entire set of *ranked* α features, we select a subset of $\hat{\alpha}$ *top* ranked features. We perform classification on the top ranked $\hat{\alpha}$ features using SVM with RBF Kernel. In section 3, we discuss the performance of proposed approach in relation to the number of top ranked features.

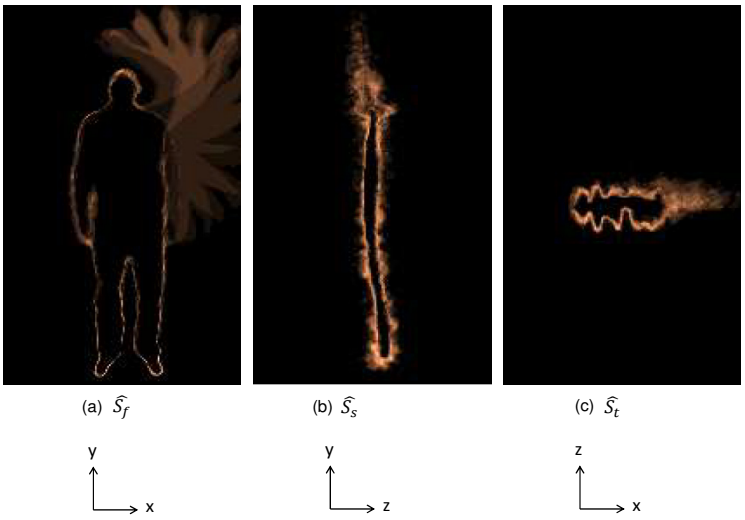

Figure 5: Normalized Motion Maps for High Arm Wave action. (a)Front View (b)Side View (c)Top View

3 EXPERIMENTS

In this section, we evaluate the proposed method. We tested our method on the MSR-Action3D dataset [8] and a dataset created by us.

3.1 MSR-Action3D

The MSR-Action3D dataset [8] consists of 20 actions namely 'high arm wave', 'horizontal arm wave', 'hammer', 'hand catch', 'forward punch', 'high throw', 'draw x', 'draw tick', 'draw circle', 'hand clap', 'two hand wave', 'side-boxing', 'bend', 'forward kick', 'side kick', 'jogging', 'tennis swing', 'tennis serve', 'golf swing', 'pick up and throw'. Each action is performed by 10 actors and has a total of 567 depth map sequences as well as body-joints data.

Li et al. [8] divide the 20 actions into three subsets, each having 8 actions as listed in Table 1. The AS1 and AS2 group similar actions with similar movements, while AS3 consists of complex actions. We used the same divisions as well for testing our method. The performance of entire feature set has been compared with that of reduced feature set obtained using RELIEFF in Tables 2 and 3 under 2 scenarios: 'cross-subject'[8] and 'five-fold cross validation'. "Without FS" column refers to the accuracy obtained when the entire set of α features is used. "With FS" column refers to the accuracy obtained using top $\hat{\alpha}$ features. From a total of $\alpha = 11512$, we selected $\hat{\alpha} = 2000$ top ranked features.

In 'cross-subject'[8] setting, half of the subjects are used for training and the remaining are used for testing. In Table 2, we report the accuracy obtained using cross-subject test scenario. We observed an increase in the overall accuracy from 91.28% to 94.61% using feature selection. In 'five-fold cross-validation' the entire dataset is split into five folds and training is done on four folds and tested on remaining fold. This is repeated so that each fold is tested once. The results of the same are reported in Table 3. The accuracy reported is the average over all the folds. We observed an increase in the overall accuracy from 93.73% to 95.92% using feature selection.

Action Set 1 (AS1)	Action Set 2 (AS2)	Action Set 3 (AS3)
Horz. arm wave	High arm wave	High throw
Hammer	Hand catch	Forward kick
Forward punch	Draw x	Side kick
High throw	Draw tick	Jogging
Hand clap	Draw circle	Tennis swing
Bend	Two hand wave	Tennis serve
Tennis serve	Forward kick	Golf swing
Pickup-throw	Side boxing	Pickup-throw

Table 1: The three subsets of actions in MSR Action3D dataset

Figure 6 illustrates the confusion matrix for AS1, AS2 and AS3 under the 'cross-subject' scenario. It may be observed from fig 6(b) that misclassification occurs mostly for the first five actions since 'draw x', 'draw circle', 'hand catch' involve similar movement of hands. We compare the performance of proposed method ('With FS') with the state-of-the-arts in Table 4.

We also tested our approach on the MSR-Action3D dataset in another scenario wherein the data is not divided into action sets i.e. all the 20 classes were used for evaluation. We obtained an accuracy of 85.09% without feature selection and an accuracy of 87.64% with feature selection in cross-subject test scenario.

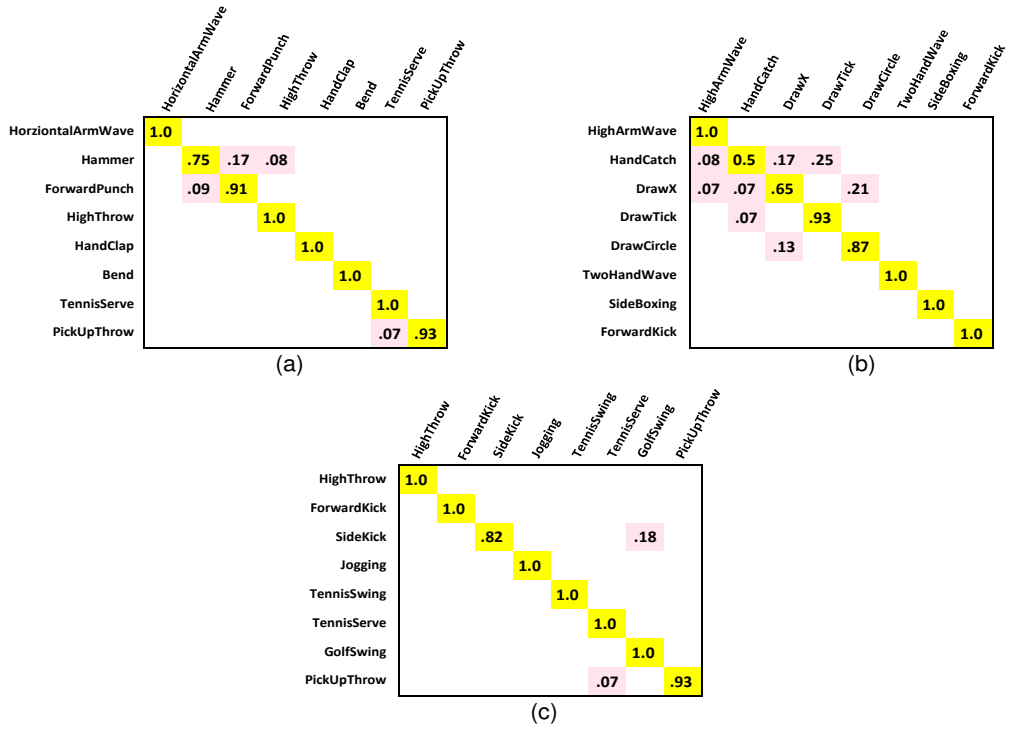

Figure 6: Confusion matrix for MSR Action3D Dataset. (a)AS1 (b)AS2 (c)AS3

	Without FS	**With FS**
AS1	92.45%	96.23%
AS2	84.96%	89.38%
AS3	96.43%	98.21%
Overall	91.28%	**94.61**%

Table 2: Cross-subject accuracy on MSR-Action3D dataset

	Without FS	**With FS**
AS1	93.36%	96.9%
AS2	90.04%	92.64%
AS3	97.78%	98.23%
Overall	93.73%	95.92%

Table 3: Five-fold cross-validation accuracy on MSR-Action3D dataset

Method	**Accuracy**
BOP[8]	74.7%
HOJ3D[19]	79.0%
EigenJoints[20]	82.3%
MHT+BST[3]	83.8%
BoJFH[14]	84.5%
GRMD[15]	86.21%
DMM-HOG[21]	91.63%
Ours	**94.61**%

Table 4: Comparative results on MSR-Action3D dataset in cross-subject scenario

3.2 Our Dataset

We created a dataset of depth maps and joints data using Microsoft Kinect to test our proposed approach. The dataset consists of 11 actions namely 'bending', 'clapping', 'drinking water', 'hand washing', 'jumping', 'kicking', 'left hand wave', 'right hand wave', 'punching', 'standing', 'stretching'. The data set consists of 287 videos where various actions were performed by 13 actors. Figure 7 shows a few sample frames from our dataset.

The total number of features(α) from each video turns out to be 16192 from which we select top 2000 features($\hat{\alpha}$). Table 5 shows the accuracy for 2 testing scenarios: five-fold cross-validation (FFCV) and New Subject(NS). In FFCV scenario, the entire dataset is divided into five folds and training is done on four folds and tested on remaining fold. This is repeated so that each fold is tested once. In 'NS' Test scenario six subjects were chosen for training and the remaining for testing. We observed that the accuracy increased by selecting $\hat{\alpha}$ top ranked feature.

Figure 8 illustrates the confusion matrix obtained in 'NS' scenario. Figure 9 shows the performance variation with respect to the number of selected top ranked features for MSR Action3D and our dataset. The horizontal axis indicates the number of selected *top* ranked features and the vertical axis indicates the accuracy obtained using the selected features.

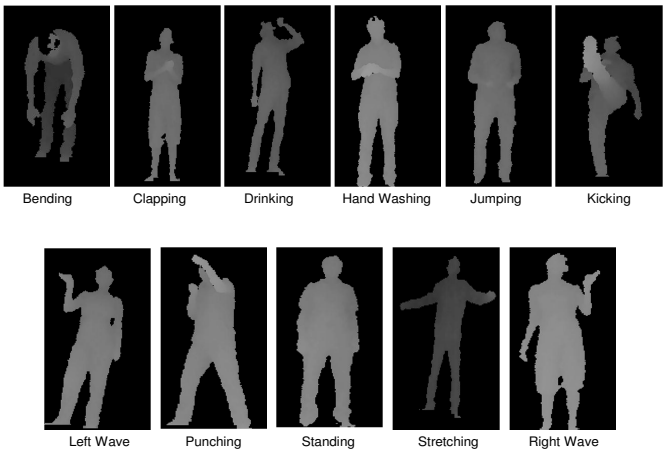

Figure 7: Sample frames from Our Dataset.

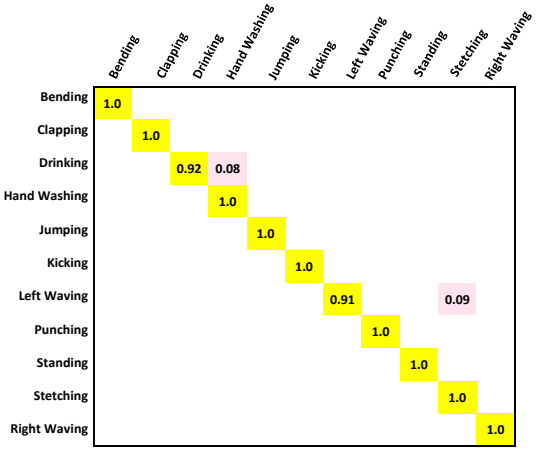

Figure 8: Confusion matrix for our Dataset.

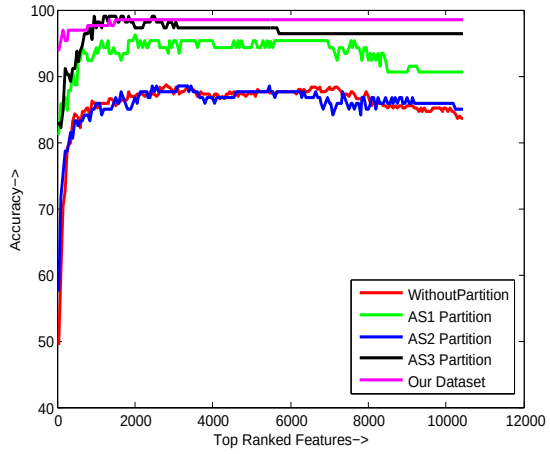

Figure 9: Recognition accuracies using different number of top ranked features.

	Without FS	With FS
Five-Fold CV	98.6%	99.3%
New Subject	97.67%	98.45%

Table 5: Results on Our dataset

4 CONCLUSION

In this paper, we have presented a new approach for action recognition based on fusion of local and global features from depth maps and body-joints data. We have proposed a novel gradient based spatio-temporal feature called as *depthHOG* and a motion-induced shape (*MIS*) feature, both extracted from depth maps. Further, we have augmented these features with Relative Joint Distance (*RJD*) and Temporal Joint Distance (*TJD*) feature obtained from body-joints data. We have used RELIEFF to obatin a small but more relevant subset of features from the entire feature pool. Experimental study reveals that the classification accuracy improves when relevant features are used. This further reduces the computational complexity of classification process.

5 REFERENCES

[1] C. C. Chang and C. J. Lin. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.*, 2(3), 2011.

[2] A. Eweiwi, M. S. Cheema, C. Bauckhage, and J. Gall. Efficient Pose-Based Action Recognition. In *Computer Vision – ACCV 2014: 12th Asian Conference on Computer Vision*, Singapore, November 1-5, 2014, Revised Selected Papers, Part V, pages 428–443, 2015.

[3] S. Jetley, F. Cuzzolin. 3D Activity Recognition Using Motion History and Binary Shape Templates. In *Computer Vision - ACCV 2014 Work-*

shops, volume 9008 of *Lecture Notes in Computer Science*, pages 129–144, 2014.

[4] S. Y. Jin and H. J. Choi. Essential body-joint and atomic action detection for human activity recognition using longest common subsequence algorithm. In *Computer Vision - ACCV 2012 Workshops*, volume 7729 of *Lecture Notes in Computer Science*, pages 148–159, 2012.

[5] A. Kläser, M. Marszałek, and Cordelia Schmid. A Spatio-Temporal Descriptor Based on 3D-Gradients. In *British Machine Vision Conference*, pages 995–1004, 2008.

[6] I. Kononenko, E. Simec, and M.R. Sikonja. Overcoming the myopia of inductive learning algorithms with RELIEFF. *Applied Intelligence*, Volume 7, pages 39–55, 1997.

[7] Y. Li, T. Sun, and X. Jiang. Human Action Recognition Based on Oriented Gradient Histogram of Slide Blocks on Spatio-Temporal Silhouette. In *International Journal of Signal Processing, Image Processing and Pattern Recognition*, Vol. 5, No. 3, September, 2012.

[8] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3D points. In *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010.

[9] B. Ni, G. Wang, and P. Moulin. RGBD-HuDaAct: A color-depth video database for human daily activity recognition. In *ICCV Workshops*, pages 1147–1153. IEEE, 2011.

[10] O. Oreifej and Z. Liu. HON4D: histogram of oriented 4d normals for activity recognition from depth sequences. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pages 716–723, 2013.

[11] E. A. Perez,V. F. Mota, L. M. Maciel, D. Sad, and M. B. Vieira. Combining gradient histograms using orientation tensors for human action recognition. In *21st IEEE International Conference on Pattern Recognition (ICPR)*, pages 3460–3463, 2012.

[12] M. Robnik-Sikonja, I. Kononenko. Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning* , 53, 23–69, 2003

[13] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th International Conference on Multimedia*, MULTIMEDIA '07, pages 357–360, 2007.

[14] P. Shukla, K.K. Biswas, P.K. Kalra. Bag-of-Features based Activity Classification using Body-joints Data. In *VISAPP 2015 - Proceedings of the 10th International Conference on Computer Vision Theory and Applications*, Volume 1, pages 314–322, Berlin, Germany, 11-14 March, 2015.

[15] R. Slama, H. Wannous, M. Daoudi. Grassmannian representation of motion depth for 3d human gesture and action recognition. In *22nd International Conference on Pattern Recognition, ICPR 2014, Stockholm, Sweden, August 24-28, 2014*, pages 3499–3504, 2014.

[16] J. Sung, C. Ponce, B. Selman, and A. Saxena. Human activity detection from RGBD images. In *Association for the Advancement of Artificial Intelligence (AAAI) workshop on Pattern, Activity and Intent Recognition*, 2011.

[17] M. Vinagre, J. Aranda, , and A. Casals. A New Relational Geometric Feature for Human Action Recognition. In *Informatics in Control, Automation and Robotics: 10th International Conference, ICINCO 2013 Reykjavík, Iceland, July 29-31, 2013 Revised Selected Papers*, pages 263–278, 2015.

[18] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '12, pages 1290–1297, 2012.

[19] L. Xia, C. Chen, and J. Aggarwal. View Invariant Human Action Recognition Using Histograms of 3D Joints. In *CVPR Workshop*, 2012.

[20] X. Yang and Y. Tian. EigenJoints based Action Recognition Using Naive Bayes Nearest Neighbor. In *CVPR Workshop*, 2012.

[21] X. Yang, C. Zhang, and Y. Tian. Recognizing actions using depth motion maps-based histograms of oriented gradients. In *Proceedings of the 20th ACM International Conference on Multimedia*, MM '12, pages 1057–1060, 2012.