ACTION RECOGNITION USING UNDECIMATED DUAL TREE COMPLEX WAVELET TRANSFORM FROM DEPTH MOTION MAPS / DEPTH SEQUENCES

B. H. Shekar[1], Rathnakara Shetty P.[1]*, Sharmila Kumari M.[2], Leonid Mestetsky[3]

[1] Mangalore University, Mangalore, Karnataka, India - (bhshekar, rathnakar.sp)@gmail.com
[2] P A College of Engineering, Mangalore, Karnataka, India - sharmilabp@gmail.com
[3] Lomonosov Moscow State University, Moscow, Russia - mestlm@mail.ru

Commission II, WG II/10

KEY WORDS: Depth Maps, Wavelet Transform, Stridden Depth Motion Map, Action Recognition

ABSTRACT:

Accumulating the motion information from a video sequence is one of the highly challenging and significant phase in Human Action Recognition. To achieve this, several classical and compact representations are proposed by the research community with proven applicability. In this paper, we propose a compact Depth Motion Map based representation methodology with hastey striding, consisely accumulating the motion information. We extract Undecimated Dual Tree Complex Wavelet Transform features from the proposed DMM, to form an efficient feature descriptor. We designate a Sequential Extreme Learning Machine for classifying the human action secquences on benchmark datasets, MSR Action 3D dataset and DHA Dataset. We empirically prove the feasability of our method under standard protocols, achieving proven results.

1. INTRODUCTION

In the field of machine vision, Human Action Recognition (HAR) plays a significant role in many applications and hence it has been one of the active research area for the last two decades. HAR has wide applications and hence it is an inseparable part of video surveillance, scene understanding, health monitoring, fitness training, gait recognition and also human computer interaction. Due to the extensive research works in the domain of image/video processing, action recognition also attracted the research community. Initially, motion sequences were captured using traditional RGB video recording cameras and hence research was limited to the 2D information registered in the video frames. While dealing with recognition of more complex movements, such as movement of body part either nearer to a camera or away from the camera where there is no texture variation, the distance from body part (object surface) from the camera lens varies significantly. Owing to this, idea of depth sensors came into reality.

The invention of low price depth sensors addressed these limitations of 2D frames by registering the 3D depth (distance from the lens of camera) information. Recent research works on action recognition are exploring the feasibility of using depth information along with RGB (Aggarwal , Xia, 2014) data. Depth sensors are proven to be advantageous as the intensity based video frames are highly sensitive to the cluttering of background as well as lighting variance whereas depth frames are insensitive to lighting variance. Apart from this, depth frames are rich with discriminating, de facto depth and edge information.

1.1 Motivation and Contribution

In order to shuffle off the burden of accessing huge video files and high computationally intensive 3-layer RGB video frame processing, we need to explore the possibilities of action recognition solely based on depth frames. Further, to extract the textural features over the video frames, we need an efficient and concise representation of the entire depth frame sequence.

The proposed work aims to explore depth motion map concept to efficiently represent and classify human actions. The work date backs to the pioneering work carried out by (Bobick , Davis, 2001) where an accumulated foreground regions of human to track the shape changes using motion history images is addressed. Further, (Yang et al., 2012) extended this idea to represent video frames using depth motion map.

With this backdrop, we propose a compact and computationally efficient DMM representation, namely Stridden Depth Motion Map (S-DMM), which hastily generates an accumulated dense structure striding 4 frames of the video sequence at a time. This completes the DMM generation four times faster than the existing approach presented in (Chen et al., 2015b). Second, we explore a Undecimated Dual Tree Complex Wavelet Transform (UDTCWT) based feature descriptor for extracting the wavelet features from our S-DMMs. Third, using a Sequential Extreme Learning Machine (S-ELM) for classification, we compare our method with the state-of-the-art methods considering standard datasets.

The rest of paper is organized as follows. We briefly discuss significant related works in Section 2. The proposed methodology is detailed in Section 3, followed by experimental results in Section 4 and Section 5 concludes this paper.

2. RELATED WORKS

Invention of depth sensors and action based game controlling techniques have brought significant importance to HAR. Initial attempts of action recognition from depth sequences were focused on extracting local features. For example, (Li et al., 2010) presented collection of 3D points features wherein silhouettes of depth images were used to sample the 3D points. (Vieira et al., 2012) segregates 3D points onto 4D grids of equal size, encoding these grids as Spatio-Temporal Occupancy

*Corresponding author

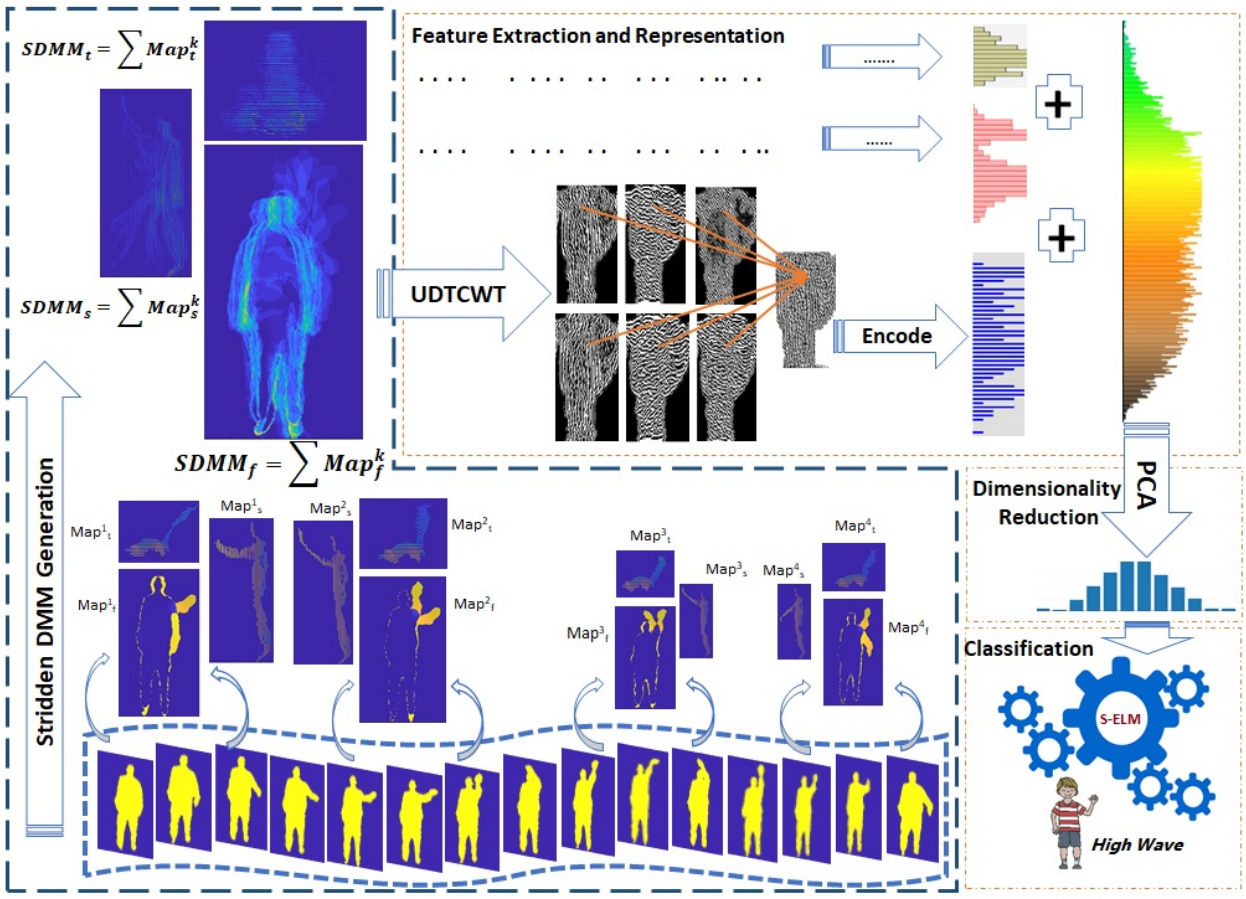

Figure 1. Proposed methodology for Stridden Depth Motion Map generation, extracting UDTCWT features and Action Recognition from Depth sequences

Patterns (STOP). (Wang et al., 2012) used sparse coding method to encode large sampling space of Random Occupancy Pattern (ROP) features.

Global feature representations generally use the differences between consecutive frames to accumulate the motion regions.(Bobick , Davis, 2001) proposed Motion History Image (MHI) from the temporal history of each point giving rise to the intensity of pixels in MHI. (Yang et al., 2012) proposed Depth Motion Map (DMM) representation wherein absolute difference between consecutive frames is calculated that results to a frontal depth map, which is projected onto the other two orthogonal planes (side and top). These projected maps are accumulated yielding the corresponding DMMs in three planes. Enhanced DMM based approaches are presented in (Chen et al., 2015b, Chen et al., 2016a), such as stacking of depth frames across an entire depth sequence on three orthogonal planes. (Chen et al., 2015b) exploited LBP feature on DMM and Kernal Extreme Learning Machine(K-ELM) for classification whereas (Chen et al., 2016a) uses vectorized DMM for feature representation and l_2-regularized collaborative representation classifier for classification. (Yang , Yang, 2014) trained Deep Convolutional Neural Networks (D-CNNs) on DMMs.

Various approaches of DMM representation and feature representations on DMMs are presented in (Zhang , Tian, 2013, Bulbul et al., 2015b, Bulbul et al., 2015c, Bulbul et al., 2015a, Chen et al., 2015a, Chen et al., 2017, Chen et al., 2016b). (Liu et al., 2018) attempted a recognizer using multi-scale energy based Global Ternary Image (GTI) modality on depth sequences which accounted the spatial-temporal discrimination and action speed variations in order to address the problems of distinguishing similar actions and identifying the actions with different speeds. DMM-based representations effectively transform the action recognition problem from 3D to 2D, with promising accuracies on the task of depth-based action recognition (Liu et al., 2018).

3. PROPOSED ACTION RECOGNITION METHODOLOGY

The proposed method has three phases namely, S-DMM based depth video representation, UDTCWT based feature extraction followed by classification using S-ELM technique. Different phases of our methodology such as S-DMM generation, UDTCWT feature representation, dimensionality reduction and classification are briefed in sections 3.1, 3.2, 3.3 and 3.4 respectively. The work flow is pictorially presented in Figure 1.

3.1 Stridden Depth Motion Maps

Aiming for computation efficiency and more compact representation, we traverse the depth video in strides (steps) of 4 frames per iteration during DMM generation. In addition, we compute the frame variation at an interval of two frames rather than subtracting the consecutive frames to find the energy

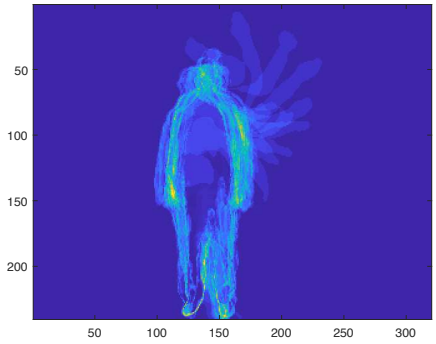

Figure 2. Stridden DMM generated for *high wave* action due to the proposed method.

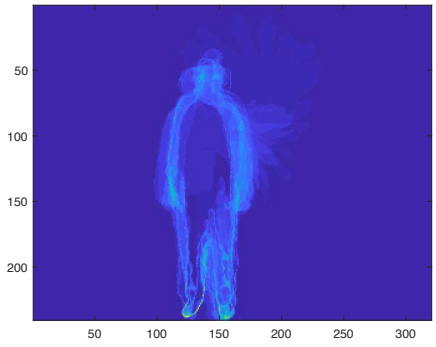

Figure 3. DMM generated for *high wave* action due to an existing method (Chen et al., 2015b).

difference. The summation of absolute differences between two alternative frames across the video sequence yields a single S-DMM. The process is mathematically summarized in equation 1.

$$SDMM_{\{f,s,t\}} = \sum_{k=3,k=k+4}^{N} \left| D^{k}_{\{f,s,t\}} - D^{k-2}_{\{f,s,t\}} \right| \quad (1)$$

where f, s, t are *front*, *side* and *top* projections respectively
$D^{k}_{\{f,s,t\}}$ is the *k*-th frame under the projection view
N is the number of frames in the depth sequence
$SDMM$ is the generated Stridden Depth Motion Map.

It shall be observed that the S-DMM generated for *high wave* action, using the proposed methodology (Figure 2) is more prominent than the DMM generated using the existing method (Figure 3) presented in (Chen et al., 2015b).

In our methodology, apart from the *front (f)* view available in the depth video frame we generate two more additional projection views namely, *side(s)* and *top(t)*.

The computation of *side* and *top* 2D projected views from the corresponding *front* view using equations 2 and 3 is given below,

$$D^{s}_{ik} = \begin{cases} j, & \text{if } D^{f}_{ij} \neq 0 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

$$D^{t}_{kj} = \begin{cases} i, & \text{if } D^{f}_{ij} \neq 0 \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where $i = 1, 2...M, j = 1, 2...N$ and $k = 1, 2...L$
L being the maximum depth value across *front* frames
D^{f} is the current *front* frame of size *MxN*
D^{s} and D^{t} are the generated *side* and *top* projections.

3.2 Undecimated Dual Tree Complex Wavelet Transform

During the second phase, each of the S-DMMs generated as described above are scanned with an Undecimated Dual Tree Complex Wavelet Transform (UDTCWT) function to produce an efficient feature descriptor. The traditional wavelet transforms such as DWT, DTCWT are having certain limitations when applied in digital image processing(Rajesh , Shekar, 2016). For example, the wavelet function produced from 2D DWT is not suitable for extracting diagonal edge details of an image due to its checker board pattern spectrum, whereas Dual Tree Complex Wavelet Transform (DTCWT) is free from such limitations. DTCWT wavelet coefficients are invariant to signal shift, free from oscillations around signal discontinuities such as edges and moreover do not suffer from aliasing.

An undecimated version of the DTCWT i.e UDTCWT is proposed in (Hill et al., 2012). It is more robust than DWT and DTCWT as downsampling is not employed, which helps in incorporating perfect one-to-one relationship between the co-located coefficients extracted at different scales. This helps to attain a flawless shift invariance in UDTCWT.

In our work, we form an higher dimensional global descriptor using UDTCWT similar to the method presented in (Rajesh , Shekar, 2016). This is achieved in four stages. At first, we compute UDTCWT at 6 different orientations (-15^0, -45^0, -75^0, 15^0, 45^0, 75^0) and 4 scales leading to 24 complex coefficient images. We separate the real and imaginary part of each coefficient image producing 48 complex coefficient images in total. During the second stage, we generate 8 Global UDTCWT Phase Pattern (GUPP) images from the above said 48 complex coefficient images. In this process, the values of 6 orientation coefficient images at each location (x,y) are concatenated forming a 6 element vector. The elements in vector are binarised (0 if value is greater than or equal to 0, 1 otherwise), to get a 6 bit vector which is in turn 0 padded at the two ends to form a vector byte. The value of this byte is the GUPP image value at (x,y). The process is shown in figure 4. GUPP image for a given scale is computed by computing the values in the similar manner for all (x,y) locations.

In the third stage, for each of the 48 complex coefficient images Local UDTCWT Phase Pattern (LUPP) images are generated. The LUPP image value at location (x,y) is computed from its eight neighbor locations as shown in Figure 5, forming a 8 bit vector. LUPP value at location p is computed using equation 4 as shown below,

$$LUPP(p) = \sum_{j=0}^{7} sgval(a_j * P) * 2^j \quad (4)$$

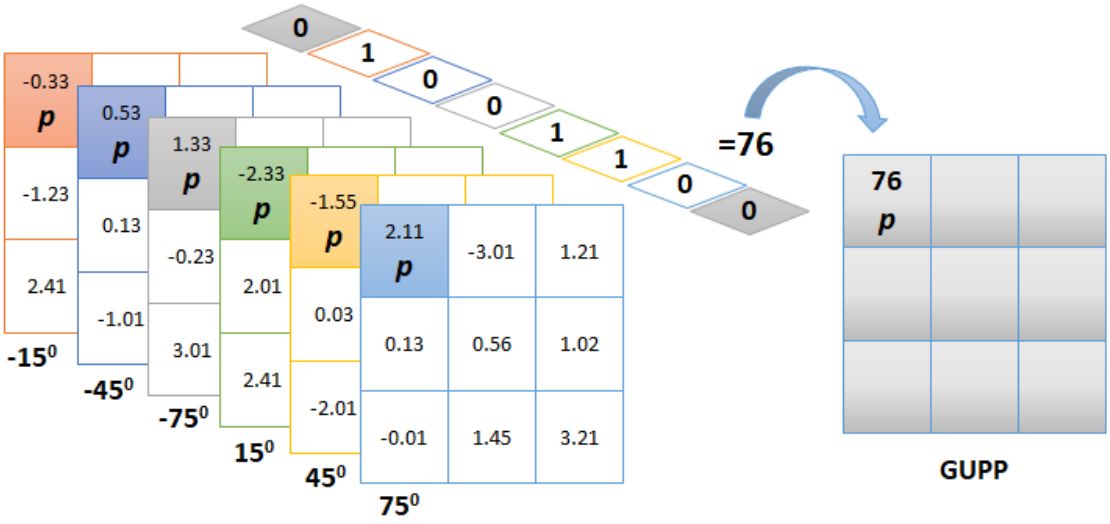

Figure 4. Computing GUPP value at location p from complex coefficient images at six different orientations.

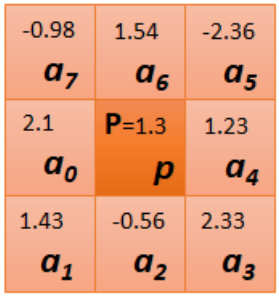

Figure 5. Eight neighbor locations for computation of LUPP value at p.

where sgval(product)=0 if product is +ve, =1 if product is -ve.
P is the value of UDTCWT at the center pixel *p*.
a_js are values of UDTCWT at the eight neighbors.

Finally, we divide GUPPs and LUPPs into blocks, compute spatial histogram in each block and concatenate these histograms to form a global descriptor for the given S-DMM. We as well extract similar features from *side* and *top* projections of S-DMMs.

3.3 Dimensionality Reduction

The global descriptor obtained due to GUPP and LUPP images is found to be robust. However, due to consecutive concatenation of spatial histogram of each blocks and concatenation of similar vectors from three projected DMMs, the dimension of feature vector turns to be very high. It is advisable to reduce the dimensionality in order to facilitate an efficient and smooth computation. We employ Principal Component Analysis (PCA) to reduce the dimension of feature vector, retaining the most discriminating features. PCA linearly maps the data to a new lower dimensional space, with the key objective of maximizing the variance of the data in the lower dimensional space.

3.4 Classification using S-ELM

We have used Sequential Extreme Learning Machine (Huang et al., 2005) for classification. It is a known fact that the single-hidden layer feed forward neural networks (SLFNs)(Huang et al., 2006) with randomly chosen weights between the input layer and the hidden layer and adequately chosen output weights are universal approximators for any bounded non-linear piecewise continuous function. In ELM, the input weights and bias matrix are randomly assigned. Considering an output neuron with a linear activation function, the SLFN network can be regarded as a linear regression model between the output vector of the hidden layer and the output of the SLFN. The sequential implementation of the ELM results in the application of recursive least-squares (RLS) to estimate the output weights vector. In brief, S-ELM is capable of randomly choosing the input weights and analytically determining the output weights of SLFNs.

4. EXPERIMENTAL RESULTS

In this section, we present experimental results conducted on two standard datasets namely, MSR Action 3D dataset and DHA dataset. The details of the datasets are presented in subsection 4.1. We conducted extensive experiments as per the standard protocols found in the existing literature and is detailed in subsection 4.2. A comparative analysis of our work and the results obtained are presented in subsection 4.3.

4.1 Dataset Description

We evaluate our approach with two of the popular benchmark datasets namely MSR Action 3D dataset(Li et al., 2010) and Depth-included Human Action (DHA) dataset (Lin et al., 2012). The MSR Action 3D dataset contains 557 depth video sequences, depicting 20 different actions, where each action is performed 2 or 3 times by 10 subjects facing the depth sensor. This dataset is challenging due to similarity of actions (e.g. draw x, draw tick, draw circle) and variations in the speed of actions. The DHA dataset is an extended dataset from the Weizmann dataset (Blank et al., 2005), containing 23 actions in total, each performed by 12 male and 9 female, 21 subjects in total.

4.2 Experimental Setup

For comparison with the standard published results, we have examined our method under two different experimental settings

Method		Action Subsets			Average (%)
		AS1	AS2	AS3	
Li et al.(Li et al., 2010)		72.9	71.9	79.2	74.7
DMM-HOG (Yang et al., 2012)		96.2	84.1	94.6	91.6
Chen et al. (Chen et al., 2016a)		96.2	83.2	92.0	90.5
HOJ3D (Xia et al., 2012)		88.0	85.5	63.6	79.0
STOP (Vieira et al., 2014)		91.7	72.2	**98.6**	87.5
DMM-LBP (Chen et al., 2015b)		**98.1**	92.0	94.6	94.9
Proposed	DMM-UDTCWT	95.6	**93.82**	96.6	95.34
	Stridden DMM-UDTCWT	96.52	**93.82**	96.92	**95.75**

Table 1. Average recognition accuracies (%) under Cross Subject tests on fixed subsets on MSR Action 3D dataset.

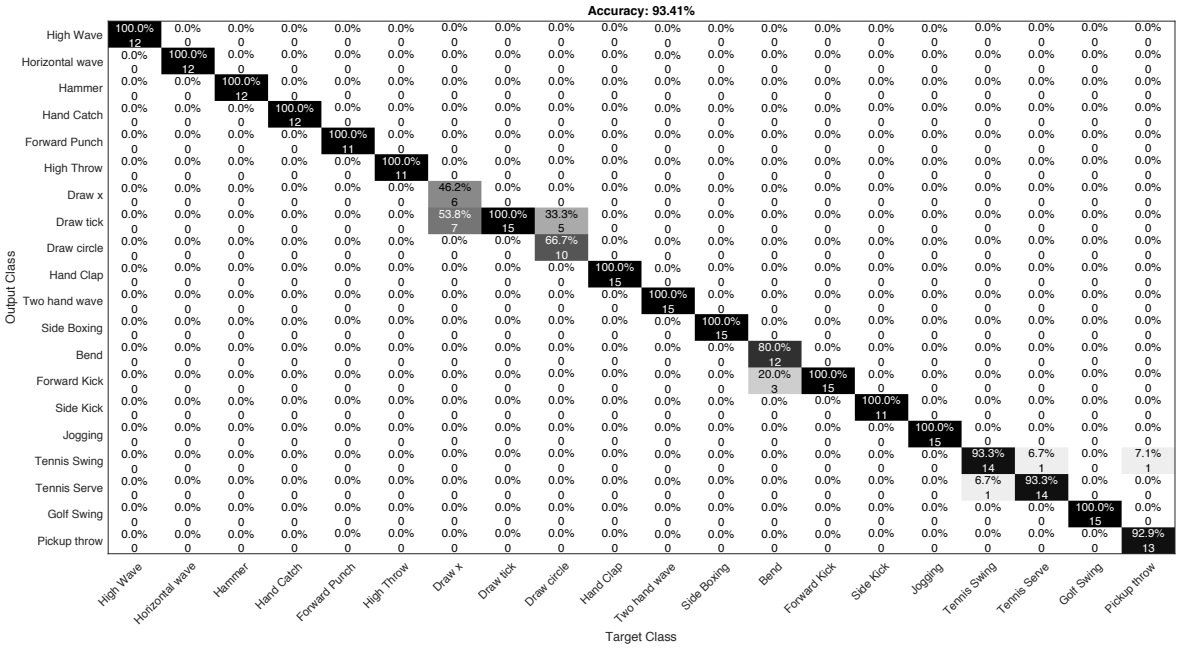

Figure 6. Confusion matrix under setting 2 on MSR Action 3D dataset demonstrating the class-wise recognition accuracy.

Method	Accuracy (%)
DMM-HOG(Yang et al., 2012)	85.5
ROP (Wang et al., 2012)	86.5
HON4D (Oreifej , Liu, 2013)	88.9
DMM-LBP (Chen et al., 2015b)	91.9
DMM-UDTCWT	92.67
Stridden DMM-UDTCWT	**93.41**

Table 2. Recognition accuracies under setting 2 on MSR Action 3D dataset .

Method	Accuracy (%)
D-STV/ASM (Lin et al., 2012)	86.8
SDM-BSM (Liu et al., 2015)	89.5
D-DMHI-PHOG (Gao et al., 2015)	92.4
D-STV (Gao et al., 2014)	86.8
DMM-UDTCWT	94.2
Stridden DMM-UDTCWT	**94.6**

Table 3. Average recognition accuracies under Leave One Subject Out test on DHA dataset.

available in the literature(Chen et al., 2015b). On MSR Action 3D dataset, under setting 1, the actions are divided into three subsets (AS1, AS2 and AS3) comprising of 8 actions each. The AS1 includes actions namely, *Horizontal wave, Hammer, Forward punch, High throw, Hand clap, Bend, Tennis serve* and *Pickup throw*. The AS2 comprises of action sequences such as *High wave, Hand catch, Draw x, Draw tick, Draw circle, Two hand wave, Forward kick* and *Side boxing* whereas AS3 is made up of actions like *High throw, Forward kick, Side kick, Jogging, Tennis swing, Tennis serve, Golf swing* and *Pickup throw*. On these three action subsets, we perform cross subject test, wherein one half of the subjects (1, 3, 5, 7, 9) were used for training and the rest for testing.

On Action 3D dataset, in setting 2, all the 20 actions are employed and one half of the subjects (1, 3, 5, 7, and 9) are used for training and the remaining subjects are used for testing. Thus, setting 2 is more challenging than setting 1, comprising more action classes. To evaluate our method on DHA dataset, we employ Leave One Subject Out (LOSO) testing protocol.

4.3 Discussion

We have made a comparative study with state-of-the-art action recognition methods, following their standard train/test protocol. Under setting 1 of MSR Action 3D dataset, our method achieves 95.75% of average accuracy. The results are tabulated in Table 1. Under setting 2 with cross subject test, experimental results are obtained and the results are tabulated in Table 2 along with other existing methods. In addition, we have also made a comparative study considering much more complex dataset such as DHA dataset with Leave One Subject Out strategy. The comparative results are presented in Table 3.

The confusion matrix is presented in Figure 6, that demonstrates the class-wise recognition accuracies of MSR Action 3D dataset under setting 2 and it is evident that seven out of thirteen *Draw x* actions are wrongly classified as *Draw tick* whereas five out of fifteen *Draw circle* actions are wrongly predicted to be *Draw tick*. This is due to the fact that there is strong interclass similarities among these three action sequences. Figure 7 shows action frames from *Draw x*, *Draw tick* and *Draw circle* actions respectively. Comparing the confusion matrix for Action 3D dataset of our proposed method to that of presented in (Chen et al., 2017, Chen et al., 2015b), it is evident and we conclude that our method effectively discriminates and better classifies *Draw x* action from *Horizontal wave* and *Hammer* actions as opposed to the methods presented in (Chen et al., 2017, Chen et al., 2015b).

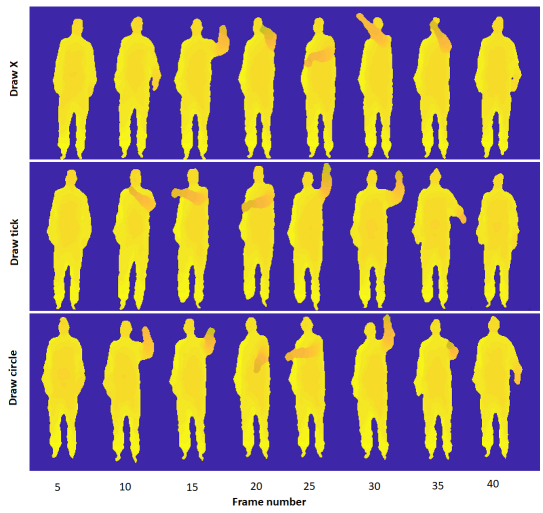

Figure 7. Interclass similarity across *Draw x*, *Draw tick* and *Draw circle* action sequences on MSR Action 3D dataset.

Considering various recognition accuracies on MSR action 3D dataset presented in Table 1, Table 2 and the recognition accuracies on DHA dataset presented in Table 3, we observe that there is a substantial difference between the recognition rate achieved by the proposed method and it is much more evident in the case of DHA dataset.

5. CONCLUSION

An accurate and efficient method for human action recognition using UDTCWT based feature descriptor considering newly proposed Depth Motion Map from depth sequences is presented. The proposed Stridden DMMs efficiently and quickly register the action cues and UDTCWT extracts the wavelet features from this compact DMM representation. A Sequential ELM is employed to classify the action sequences. The proposed method is extensively evaluated on two benchmark datasets, under standard protocols presented in the literature. Our experiments exhibit better results when compared to state-of-the-art methods. However, effective and discriminative representation to overcome the challenge of interclass similarities is to be addressed in our future works. Also, we intend to address the lack of DMMs in registering the speed variations of action sequences and to improve the computational speed of the proposed UDTCWT feature descriptor.

ACKNOWLEDGEMENT

This work is supported jointly by the Department of Science & Technology, Govt. of India and Russian Foundation for Basic Research, Russian Federation under the grant No. INT/RUS /RFBR /P-248.

REFERENCES

Aggarwal, Jake K, Xia, Lu, 2014. Human activity recognition from 3d data: A review. *Pattern Recognition Letters*, 48, 70–80.

Blank, Moshe, Gorelick, Lena, Shechtman, Eli, Irani, Michal, Basri, Ronen, 2005. Actions as space-time shapes. *null*, IEEE, 1395–1402.

Bobick, Aaron F., Davis, James W., 2001. The recognition of human movement using temporal templates. *IEEE Transactions on pattern analysis and machine intelligence*, 23, 257–267.

Bulbul, Mohammad Farhad, Jiang, Yunsheng, Ma, Jinwen, 2015a. DMMs-based multiple features fusion for human action recognition. *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, 6, 23–39.

Bulbul, Mohammad Farhad, Jiang, Yunsheng, Ma, Jinwen, 2015b. Human action recognition based on dmms, hogs and contourlet transform. *2015 IEEE International Conference on Multimedia Big Data*, IEEE, 389–394.

Bulbul, Mohammad Farhad, Jiang, Yunsheng, Ma, Jinwen, 2015c. Real-time human action recognition using dmms-based lbp and eoh features. *International Conference on Intelligent Computing*, Springer, 271–282.

Chen, Chen, Hou, Zhenjie, Zhang, Baochang, Jiang, Junjun, Yang, Yun, 2015a. Gradient local auto-correlations and extreme learning machine for depth-based activity recognition. *International Symposium on Visual Computing*, Springer, 613–623.

Chen, Chen, Jafari, Roozbeh, Kehtarnavaz, Nasser, 2015b. Action recognition from depth sequences using depth motion maps-based local binary patterns. *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, IEEE, 1092–1099.

Chen, Chen, Liu, Kui, Kehtarnavaz, Nasser, 2016a. Real-time human action recognition based on depth motion maps. *Journal of real-time image processing*, 12, 155–163.

Chen, Chen, Liu, Mengyuan, Zhang, Baochang, Han, Jungong, Jiang, Junjun, Liu, Hong, 2016b. 3d action recognition using multi-temporal depth motion maps and fisher vector. *IJCAI*, 3331–3337.

Chen, Chen, Zhang, Baochang, Hou, Zhenjie, Jiang, Junjun, Liu, Mengyuan, Yang, Yun, 2017. Action recognition from depth sequences using weighted fusion of 2D and 3D auto-correlation of gradients features. *Multimedia Tools and Applications*, 76, 4651–4669.

Gao, Z, Zhang, H, Xu, GP, Xue, YB, 2015. Multi-perspective and multi-modality joint representation and recognition model for 3D action recognition. *Neurocomputing*, 151, 554–564.

Gao, Zan, Song, Jian-ming, Zhang, Hua, Liu, An-An, Xue, Yan-bing, Xu, Guang-ping, 2014. Human action recognition via multi-modality information. *Journal of Electrical Engineering and Technology*, 9, 739–748.

Hill, Paul, Achim, Alin, Bull, David, 2012. The undecimated dual tree complex wavelet transform and its application to bivariate image denoising using a cauchy model. *Image Processing (ICIP), 2012 19th IEEE International Conference on*, IEEE, 1205–1208.

Huang, Guang-Bin, Chen, Lei, Siew, Chee Kheong et al., 2006. Universal approximation using incremental constructive feedforward networks with random hidden nodes. *IEEE Trans. Neural Networks*, 17, 879–892.

Huang, Guang-Bin, Liang, Nan-Ying, Rong, Hai-Jun, Saratchandran, Paramasivan, Sundararajan, Narasimhan, 2005. On-line sequential extreme learning machine. *Computational Intelligence*.

Li, Wanqing, Zhang, Zhengyou, Liu, Zicheng, 2010. Action recognition based on a bag of 3d points. *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, IEEE, 9–14.

Lin, Yan-Ching, Hu, Min-Chun, Cheng, Wen-Huang, Hsieh, Yung-Huan, Chen, Hong-Ming, 2012. Human action recognition and retrieval using sole depth information. *Proceedings of the 20th ACM international conference on Multimedia*, ACM, 1053–1056.

Liu, Hong, Tian, Lu, Liu, Mengyuan, Tang, Hao, 2015. Sdm-bsm: A fusing depth scheme for human action recognition. *2015 IEEE International Conference on Image Processing (ICIP)*, IEEE, 4674–4678.

Liu, Mengyuan, Liu, Hong, Chen, Chen, 2018. 3D action recognition using multiscale energy-based global ternary image. *IEEE Transactions on Circuits and Systems for Video Technology*, 28, 1824–1838.

Oreifej, Omar, Liu, Zicheng, 2013. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 716–723.

Rajesh, DS, Shekar, BH, 2016. Undecimated dual tree complex wavelet transform based face recognition. *Advances in Computing, Communications and Informatics (ICACCI), 2016 International Conference on*, IEEE, 720–726.

Selesnick, Ivan W, Baraniuk, Richard G, Kingsbury, Nicholas G, 2005. The dual-tree complex wavelet transform. *IEEE signal processing magazine*, 22, 123–151.

Vieira, Antonio W, Nascimento, Erickson R, Oliveira, Gabriel L, Liu, Zicheng, Campos, Mario FM, 2012. Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences. *Iberoamerican congress on pattern recognition*, Springer, 252–259.

Vieira, Antonio W, Nascimento, Erickson R, Oliveira, Gabriel L, Liu, Zicheng, Campos, Mario FM, 2014. On the improvement of human action recognition from depth map sequences using space–time occupancy patterns. *Pattern Recognition Letters*, 36, 221–227.

Wang, Jiang, Liu, Zicheng, Chorowski, Jan, Chen, Zhuoyuan, Wu, Ying, 2012. Robust 3d action recognition with random occupancy patterns. *Computer vision–ECCV 2012*, Springer, 872–885.

Xia, Lu, Chen, Chia-Chih, Aggarwal, Jake K, 2012. View invariant human action recognition using histograms of 3d joints. *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, 20–27.

Yang, Rui, Yang, Ruoyu, 2014. Dmm-pyramid based deep architectures for action recognition with depth cameras. *Asian Conference on Computer Vision*, Springer, 37–49.

Yang, Xiaodong, Zhang, Chenyang, Tian, YingLi, 2012. Recognizing actions using depth motion maps-based histograms of oriented gradients. *Proceedings of the 20th ACM international conference on Multimedia*, ACM, 1057–1060.

Zhang, Chenyang, Tian, Yingli, 2013. Edge enhanced depth motion map for dynamic hand gesture recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 500–505.

Revised April 2019