RGBD Human Action Recognition using Multi-Features Combination and K-Nearest Neighbors Classification

Rawya Al-Akam
Active Vision Group
Institute for Computational Visualistics
University of Koblenz-Landau
Universitatsstr. 1, 56070 Koblenz, Germany

Dietrich Paulus
Active Vision Group
Institute for Computational Visualistics
University of Koblenz-Landau
Universitatsstr. 1, 56070 Koblenz, Germany

***Abstract*—In this paper, we present a novel system to analyze human body motions for action recognition task from two sets of features using RGBD videos. The Bag-of-Features approach is used for recognizing human action by extracting local spatial-temporal features and shape invariant features from all video frames. These feature vectors are computed in four steps: Firstly, detecting all interest keypoints from RGB video frames using Speed-Up Robust Features and filters motion points using Motion History Image and Optical Flow, then aligned these motion points to the depth frame sequences. Secondly, using a Histogram of orientation gradient descriptor for computing the features vector around these points from both RGB and depth channels, then combined these feature values in one RGBD feature vector. Thirdly, computing Hu-Moment shape features from RGBD frames, fourthly, combining the HOG features with Hu-moments features in one feature vector for each video action. Finally, the k-means clustering and the multi-class K-Nearest Neighbor is used for the classification task. This system is invariant to scale, rotation, translation, and illumination. All tested are utilized on a dataset that is available to the public and used often in the community. By using this new feature combination method improves performance on actions with low movement and reach recognition rates superior to other publications of the dataset.**

Keywords*—RGBD Videos; Feature Extraction; k-means Clustering; KNN (K-Nearest Neighbor)

I. Introduction

Human action recognition using cameras is a very active research topic and it has been widely studied in the computer vision and pattern recognition fields to characterize the behavior of persons. Also, it has been used in many applications fields like, video surveillance, robotics human-computer interaction, and a variety of systems that involve interactions between persons and computers [1]. Therefore, the ability to design a machine that is capable of interacting intelligently with a human-inhabited environment is important in recognizing humans and activities of people from the video frames [2].

In the last few years, research on human activity recognition essentially concentrated on recognizing human activities from videos captured by conventional visible light cameras [3]. But recently, the action recognition studies have entered a new phase by technological advances and the emergence of the low-cost depth sensor like Microsoft Kinect [4]. This depth sensor has many advantages over RGB cameras, like to provide 3D structural information as well as color image sequences in real time, and can even work in total darkness which makes it possible to explore the fundamental solution for traditional problems in human action classification [5][6]. Of course, the depth camera also has severe limitations which can be partially enhanced by fusion of RGB and Depth. But all these advantages make it interesting to incorporate the RGBD cameras into more challenging environments.

Overview of our approach: In this work, we combined two sets of features. For the local motion and appearance features, which are improved the method of [7][8] to categorize the body motions on RGBD videos instead of using only RGB video, according to how to represent the spatial and temporal structure of actions from color and depth data together and combining the motion features extracted from both channels in one feature vector for each video action. And the Hu-moments shape invariant which introduced by [9] are used for global spatial-temporal features. The overview of the proposed approach is illustrated in Fig. 1. In order to represent the human activity recognition from RGBD, the two different sets of features vector are extracted, the first set is represented as follow:

- Detect the important interest points by extracting visually distinctive points from the spatial domain using Speed-Up Robust Features (SURF). After that filter these SURF points using Motion History Image (MHI) [10][11] and Optical Flows (OF) [12] to extract only the essential motion points from the sequences.
- HOG descriptor is applied to describe the detected interest points. the HOG features is computed from the frames, MHI and OF channels and represented in one feature vector for each video action.

While the second feature set is computed as:

- Represent the spatial and temporal information about an action in a single image. In order to do this, MHI is used, where the pixel intensity is a function of the recency of action.
- Hu moments [9] are used as descriptors of the motion history image. We are using the seven translation,

scale and orientation invariant Hu moments to get each seven Hu moments from the motion history image.

After the two feature sets are computed, the feature vectors are combined and encoded into a single code by using the bag of features algorithm [13]. The unsupervised learning k-means clustering and supervised learning K-nearest neighbor (KNN) [2] are used for classification the different action from videos.

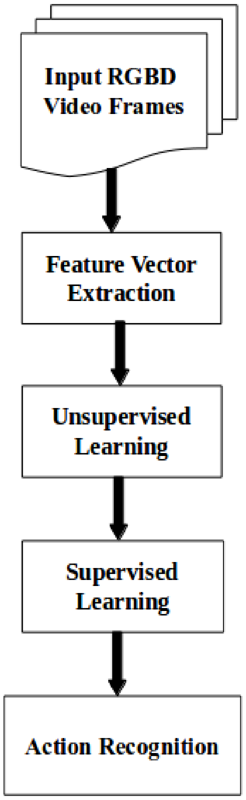

Fig. 1. General structure of our approach.

The rest of the paper is represented as follow, section II describes the related work done in this area. Section III explain in detail the system analysis of action recognition. Section IV represent the experimentation and results, and finally in Section V provides the conclusion.

II. Related Works

In this section, the state of the arts on human action recognition are summarized. During the last decades, several different approaches have been proposed to detection, representation and recognition, and understanding video events. Previous research on action recognition mainly focused on RGB videos, which yielded lots of feature extraction, action representation, and modeling methods.

In [14], the authors presented the human detection and simultaneous behavior recognition from RGB image sequences by using the action representation method depended on applying the clustering algorithm to the sequence of HOG descriptor of human motion images. Other people used a hierarchical filtered motion (HFM) method for recognizing the human action in crowded videos as in [7], they used 2D Harris corners for detection the motion interest points from motion history image (MHI) of the recent motion (i.e. locations with high intensities in MHI). Then applied a global spatial motion smoothing filter to the gradients of MHI to eliminate isolated unreliable or noisy motions. To characterize the spatial(appearance) and temporal(motion) features they used HOG descriptor in the intensity image and MHI respectively and the Gaussian Mixture Model (GMM) classifier for action recognition performance system. The work of [15] also used the invariant 7-Hu moments of MEI and MHI to estimate Gaussian Mixtures models of daily activities.

In the other hand, there are a lot of researchers presented action recognition depending on only depth data, like in [16], they recognized human action by projected to the depth maps onto three orthogonal levels and collect the global activities from entire video frames to compute the Depth Motion Maps (DMM), after that the Histograms of Oriented Gradients (HOG) is computed from DMM to represent an action video.

A lot of researchers improved the action recognition performance on RGBD data by computing a local spatial-temporal feature from RGB data, a skeleton joint feature, and a point cloud feature in-depth data, and combined all these features based on sparse coding features combination methods as in [17]. While in [2], presented a comparison of several well-known pattern recognition techniques. they used Motion History Images (MHI) to describe these activities in a qualitative way and computed Hu-moments. And the system was tested extracted features vectors with Support Vector Machines and K-Nearest Neighbours classifiers. Another method that was used for action recognition is based on features learned from 3D video data applying Independent Subspace Analysis (ISA) technique on data collected by RGBD cameras as in [18] and they followed the bag-of-visual-word model and an SVM classifier to recognize the activities. The other researchers considered a human's activity as composed of a set of sub-activities as in [19], they computed a set of features based on human poses and motion, as well as based on image and point-cloud information.

III. System Analysis of Action Recognition

In this section, we describe the steps for computing the feature vectors from each video action in details. Section III-A represents the Pre-processing to the input RGB and depth videos. Section III-B gives a brief description about Bag of Features Extraction. Section III-C explains the Bag of Words Generation and in section III-D explains the classification method used to compute the recognition accuracy. As shown in Fig. 2, the system scheme of action recognition is represented.

A. Preprocessing Input Data

The input dataset is color and depth videos were analyzed as a frame sequence to extract features presented in each frame. In this work, we choose to use a lower resolution of 320×240 in order to reduce the computational complexity of the system. The depth maps data captured by the Kinect camera are often noisy due to imperfections related to the Kinect infrared light reflections. For reducing noise and to eliminate the unmatched edges from the depth images, we used

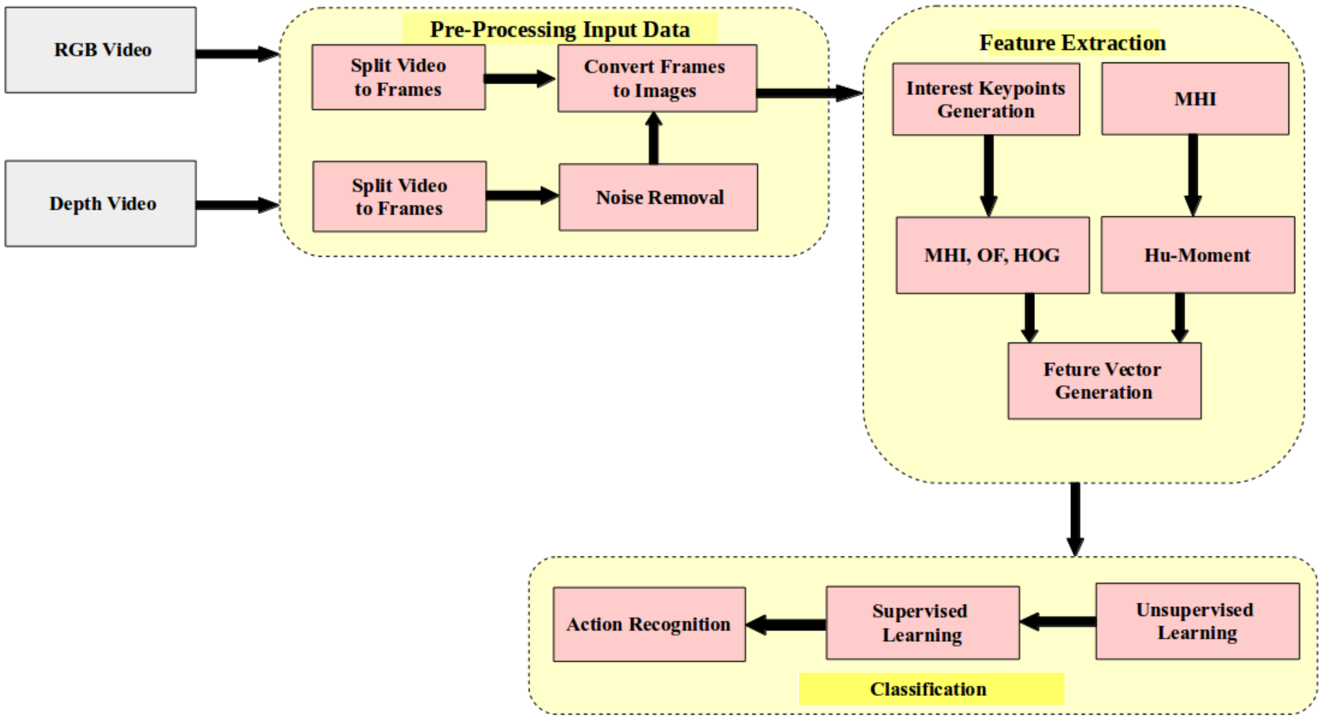

Fig. 2. System analysis schematics of action recognition. Using RGB and depth stream. Pre-processing to the input data; feature extraction; and classification.

a spatial-temporal bilateral filtering to smooth depth images. The joint-bilateral filtering proposed in [20] is formulated as in equation (1):

$$\hat{D}_{(P)} = \frac{1}{K_{(p)}} \sum_{q \in \Omega_p} f(p,q) g(\| \hat{D}_m(p) - \hat{D}_m(q) \|) \\ h(\| I_{(p)} - I_{(q)} \|) \tag{1}$$

where $f(p,q)$ refers to the domain term for measuring the closeness of the pixels, p, and q. the function $g(.)$ denotes a depth range term that computes the pixel similarity of the modeled depth map. $h(.)$ is function represent an intensity term to measure the intensity similarity. Moreover, Ω_p represents the spatial neighborhood of position p.

B. Bag of Features Extraction

For feature extraction, we flow the Bag-of-Features (BoFs) method, It is the most popular technique of feature representation for videos to learn and recognize the different human actions. The local features have been computed from the spatial-temporal domain by implementing the feature detector and descriptor methods on 3D data. The procedure for extracting features vectors include three steps: Interest keypoint generation, feature vector generation, and dictionary generation.

1) Interest Keypoints Generation: As the essential, we finding the motion interest points (keypoints) from RGB frame sequence using the Speed-Up Robust Features (SURF) detector [21] as a first step to extract visually distinctive keypoints from spatial domain. Then, these keypoints are filtered by using temporal (motion) template approach for detecting motion and computing its direction, this constraint from motion history images MHI that is generated by computing the difference between two adjacent frames as represented in [11][22]. Those points with larger intensities in MHI representing the moving object with more recent motion. After that compute optical flows of those keys preserved after MHI filtering using the Lucas-Kanade method [23]. To represent how motion the image is moving, form a motion-history image (MHI). In an MHI H , the pixel intensity, which is represent a function of the temporal motion history that point. The MHI shown in equation (2) is formally defined as in [11].

$$H_\tau(x,y,t) = \begin{cases} \tau, & \text{if D(x,y,t)=1} \\ \max(0, H_\tau(x,y,t-1) - 1) & \text{otherwise.} \end{cases} \tag{2}$$

where $D(x,y,t)$ is a binary image of differences between frames and τ is the maximum duration of motion. τ is the duration which decides the temporal extent of the movement (e.g., in terms of frames). After Computing the motion keypoints $P(x,y,t)$ from RGB images, this motion points are aligned to the related depth images $P_d(x,y,z,t)$, where (x,y,t) denote the coordinates and time of interest point p on RGB images and (x, y, z, t) refer to the 3D coordinate and time of interest point on depth images.

2) Feature Vector Generation: In order to represent the shape, appearance and motion information, we used two different descriptors. HOG features descriptor [7] is applied on both RGB and depth video frames and combined feature vector values to generate the BoFs. This descriptor is widely used in human detection [24] and action recognition [25]. For vector generation, the HOG descriptor was implemented around each keypoints in video frames of RGBD images, MHI and OF channel and also, can be well adapted to characterize local shape information from image channel and local motion information from MHI channel by computing distributions of local gradients. Seven Hu-moment shape features are extracted from MHI that computed above in equation (2). For two-dimensional $(M \times M)$ images that has MHI function $f(x, y); x, y = 0, 1, ..., M - 1$, geometric moment m_{pq} of $f(x, y)$ is computed as follow [26]:

$$m_{pq} = \sum_{x=0}^{x=M-1} \sum_{y=0}^{y=M-1} (x)^p.(y)^q f(x, y), \tag{3}$$

for $p, q = 0, 1, 2, 3, ..,,$ where p, q are positive integers and $(p + q)th$ is called the order of the moment of a density distribution function $f(x, y)$.

The moments value of $f(x, y)$ are translated by a quantity (a, b), which is computed as:

$$\mu_{pq} = \sum_x \sum_y (x + a)^p.(y + b)^q f(x, y), \tag{4}$$

Then, to make these moments invariant to translation, the central moment μ_{pq} can be defined from equation (4) as follow, by changing the values $a = -\bar{x}$, and $b = -\bar{y}$:

$$\mu_{pq} = \sum_x \sum_y (x + \bar{x})^p.(y + \bar{y})^q f(x, y), \tag{5}$$

where,

$$\bar{x} = \frac{m_{10}}{m_{00}}, \bar{y} = \frac{m_{01}}{m_{00}}$$

And the scaling invariance of central moment can be computed by normalizing the moments of the scaled image by the scaled energy of the original image to become invariant to scale change, which can be defined as stated below.

$$\eta_{pq} = \frac{\mu_{pq}}{\mu^{\gamma}{}_{00}}, \gamma = \frac{p + q}{2} + 1 \tag{6}$$

where γ is the value of normalization factor.

The values of η_{pq} represented a set nonlinear function that calculated by normalizing central moments, which are invariant to object rotation, translation, position and scale change. The seven Hu-moments is derived as in equation (7) [26][27]:

$$\begin{aligned}
M_1 &= \eta_{20} + \eta_{02}, \\
M_2 &= (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2, \\
M_3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2, \\
M_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2, \\
M_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\
&\quad +(3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[(\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2], \\
M_6 &= (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + \\
&\qquad 4\mu_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}), \\
M_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})] + \\
&\qquad (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} - \eta_{03}^2]
\end{aligned} \tag{7}$$

Where the numerical values of M_1 to M_6 are very small. To avoid precision problems the logarithms of the absolute values of these six functions, i.e. $log|M_i|$; where, $i = 1, .., 6$, are selected as features representing the action among video frames.

Finally, the feature vectors are generated by combining the hu-moment features with the HOG features to represent the action information from each RGBD video.

3) Dictionary Generation: After extracting features information from all RGBD video depending on the detector and descriptor strategy, the dictionary is generated from these feature vectors – this is the important step on (BoFs) method. The Dictionary was generated by clustering using the k-means algorithm as represented in Fig. 3. The size of the dictionary is important for the recognition process because if the size of the dictionary is set too small then the BoF model cannot express all the keypoints and if it is set too high then it might lead to over-fitting and increasing the complexity of the system [28]. The k-means clustering was applied on all BoF from training videos, the k is represents the dictionary size. The centroids of each cluster are combined to make a dictionary. In this method, we got the best result with a value of $k = 400$ as a dictionary size.

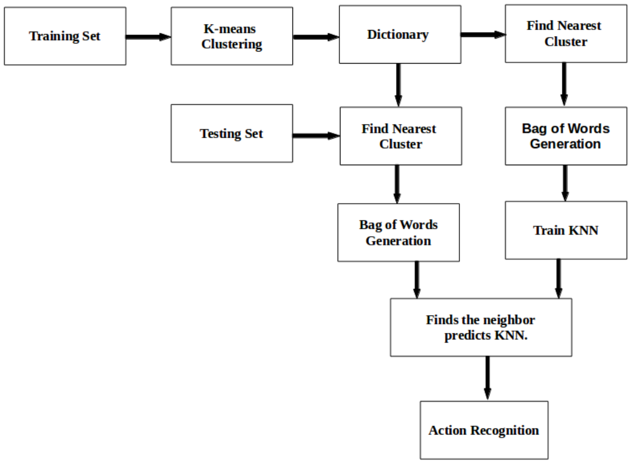

Fig. 3. Dictionary Generation from Feature vector for classified Action.

C. Bag of Words Generation

In order to generate the Bag-of-Words (BoWs), each feature description of the video frame is compared with each centroid of the cluster in the dictionary using Euclidean distance measure e as formulated in equation (8) [29].

$$e = \sum_{j=1}^{k} \sum_{i=1}^{n} \|X_i{}^{(j)} - C_j\|^2 \tag{8}$$

where $\|X_i{}^{(j)} - C_j\|^2$ is the selected distance measure between the feature vector point and the clustering center C_j. C_j is the clustering center length and n is the feature vector size. Then, we check the difference e, if the difference is small or features values is close to a certain cluster, the count of that index is increased. Similarly, the other feature description of video frames are also compared and the counts of the respective indices are increased of which the feature description values are closest to as in [28]. These BoWs vectors are computed for all the videos for training and testing dataset.

D. Action Classification

In order to make performance comparison for our system, a K-nearest neighbor (KNN) [2] is used. KNN is the simplest and mostly used classifier. It is assigned an object to a class according to the vote of its K-nearest neighbors, i.e. KNN is to classify unlabeled observations by assigning them to the class of the most similar labeled examples. Characteristics of observations are collected for both training and test dataset. K is an integer value and typically small and varied by the amount of test class. If K=1, the object is directly assigned to the class of its nearest neighbor.

In this work, the Bag of words vectors for all the videos is computed in training stage and labels are appended according to the class. This bag of words vectors are fed into the multi-class KNN in order to train the model that is further used in testing stage for human action recognition as shown in Fig. 3.

IV. Experimentation and Evaluation

In this section, we present the two types of datasets used and the experimental results on them using our approach.

A. Dataset

To evaluate the performance of our system approach, we conducted experiments on the MSR-Daily Activity 3D Dataset [1] and Online RGBD Action dataset (ORGBD) [2].

1) MSR-Daily Activity 3D dataset: The MSR-Daily Activity 3D Dataset is a daily activity dataset captured by a Kinect device and it is designed to cover humans daily activities in the living room [30]. This dataset contains 16 action and 10 subjects; each subject performs each activity in two different poses:*drinking, eating, read a book, call cell phone, writing on a paper, using laptop, using vacuum cleaner, cheer up, sitting, still, tossing paper, playing game, laying down on sofa, walking, playing guitar, stand up, and sit down.* see Fig. 4.

2) Online RGBD action dataset: The Online RGBD Action dataset (ORGBD) [31] are captured by the Kinect device. Each action was performed by 16 subjects for two times. This dataset contains seven types of actions which recorded in the living room: *drinking, eating, using a laptop, picking up a phone, reading phone (sending SMS), reading a book, and using a remote.* as shown in Fig. 5. We compare our approach with the state-of-the-art methods on the same environment test setting, where half of the subjects are used as training data and the rest of the subjects are used as test data.

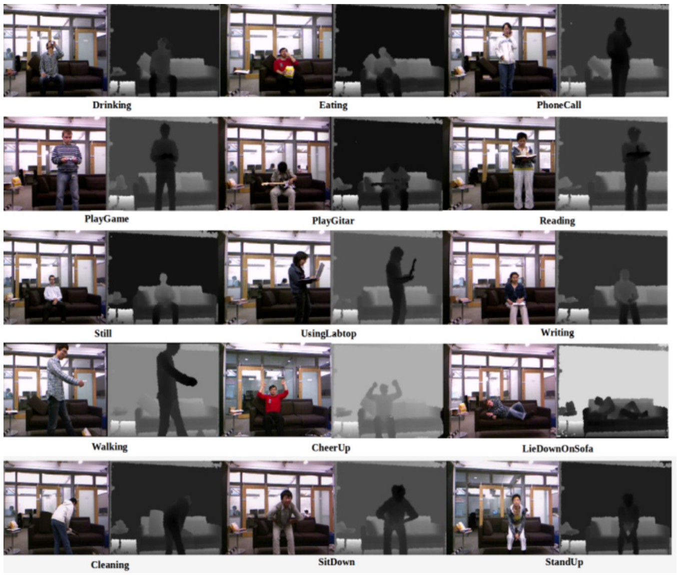

Fig. 4. Sample frames of MSR-Daily Activity 3D Dataset.

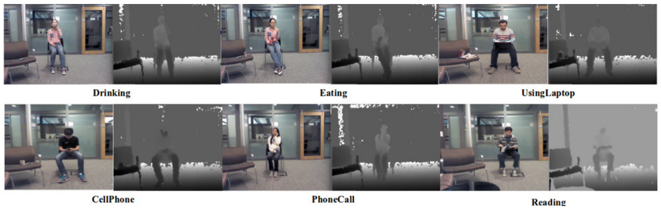

Fig. 5. Sample frames of Online RGBD Action Dataset.

B. Experimental results

In our experiments, we are combined two different feature descriptor information. The local features which encode information regarding all the available modalities and the shape invariant moment. The local features are extracted as follow: from RGB videos, the SURF detector is used on spatial domains and filtered these points by MHI and OF on temporal domains to extract the motion points all video frames and then aligned these points to the depth sequences to get the RGBD interest motion points as in Fig. 6, which shows the position of interesting motion points in the video frames. After that, the local appearance and motion features are characterized by grids of the histogram of orientation gradient (HOG) [7] around the motion interest points. Normalized histograms of all the patches are concatenated into HOG (for appearance features in the intensity image), HOG-MHI (for motion features in the MHI) and HOG-OF (for motion features in the OF) descriptor

[1] http://www.uow.edu.au/w̃anqing/#MSRAction3DDatasets
[2] https://sites.google.com/site/skicyyu/orgbd

TABLE I. COMPRISION OF RECOGNITION ACCURACY WITH OTHER METHODS ON MSR-DAILY ACTIVITY 3D DATASET

Methods	Accuracy
CHAR [32]	54.7%
Discriminative Orderlet [31]	60.1%
Relative Trajectories 3D [33]	72.00%
Moving Pose [34]	73.80%
Proposed Method	**100%**

TABLE II. COMPRISION OF RECOGNITION ACCURACY WITH OTHER METHODS ON ONLINE RGBD (ORGBD) DATASET

Methods	Accuracy
HOSM [35]	49.5%
Orderlet+SVM [31]	68.7%
Orderlet+ boosting [31]	71.4%
Human-Object Interaction[36]	75.8%
Proposed Method	**85.71%**

vectors as the input of the classifier for action recognition. In this test, we set x and y equal to 3 and use 6 bins for HOG in the intensity image, HOG-MHI, and HOG-OF. These selected values are applied on RGB and Depth channels.

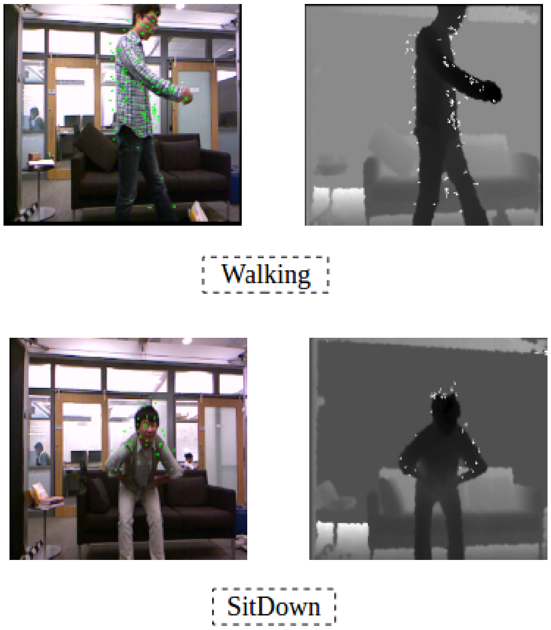

Fig. 6. Motion points in RGB and depth frames of different action represented by green points on RGB frame and white points on depth frames..

The Hu-moment features are computed from MHI channel on both RGB and depth video frames to compute the seven invariant features from each frame in video. The last step in computing features vector is combined the local and hu-moment feature to represent the feature vector. All testing results of the experiment are described on Table I and Table II, which shows the comparison results of recognition rate of our system test and the other state of the art using different methods of the MSR-Daily Activity 3D and ORGBD dataset respectively.

V. CONCLUSION AND FUTURE WORKS

In this paper, a human action recognition on 3D video (RGB and Depth data) is proposed. Our system starts from processing, removing the noise from the input depth data and aligning the RGB with the depth frames. We proposed two sets of feature information, which are represented by the local feature vector by extracting these features from 3D video data using SURF, MHI, and OF for detecting motion interest points, and for the appearance and motion features, the HOG descriptor is applied on image, MHI and OF of each RGB and depth video of all actions. and the other feature set is extracted using global Hu-moments shape descriptor from MHI, then combined all motion, shape and appearance vectors into one vector for each RGBD video action. These feature vector values are tested depending on the Bag-of-words method (BoWs) by using k-means clustering and KNN classifier. The presented approach is highly efficient and invariant to cluttered backgrounds, illumination changes, rotation, translation and scale.

The Experiment results showed that the proposed scheme can effectively recognize the similar action with high movement rate as walking, cleaning, etc., and improves performance on actions with low movement rate like: reading, using laptop, etc. It gives a 100% on MSR-Daily activity 3D dataset and 85.71% on ORGBD dataset recognition rates. From this method on RGBD dataset demonstrate that our approach significantly outperforms the existing state-of-the-art methods. The best performance is achieved because interest points are extracted solely from the RGB channel and aligned to the depth, then combined the RGB and depth based descriptors values depending on this detected motion points. For the future works, we will combine a new feature vector values like local binary pattern (LBP). Also for the classification task, the convolution neural networks (CNN), and random forest will be use.

ACKNOWLEDGEMENTS

This work was partially supported by Ministry of Higher Education and Scientific Research (MHESR), Iraq, and University of Koblenz Landau, Germany.

REFERENCES

[1] G. Chen, F. Zhang, M. Giuliani, C. Buckl, and A. Knoll, “Unsupervised Learning Spatio-temporal Features for Human Activity Recognition from RGB-D Video Data,” in *Social Robotics*. Springer International Publishing, 2013, pp. 341–350.

[2] M. N. ADEM KARAHOCA, “Human motion analysis and action recognition,” in *1st WSEAS International Conference on Multivariate Analysis and its Application in Science and Engineering*, 2008, pp. 156–161.

[3] X. Yang and Y. Tian, “Super Normal Vector for Human Activity Recognition with Depth Cameras,” *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELIGENCE*, pp. 1–12, 2016.

[4] M. G. Soares Beleboni, “A brief overview of Microsoft Kinect and its applications,” *Interactive Multimedia Conference, University of Southampton, UK*, 2014.

[5] D. Kim, W.-h. Yun, H.-s. Yoon, and J. Kim, “Action Recogntion with Depth Maps Using HOG Descriptors of Multi-view Motion Appearance and History,” *The Eighth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM)*, no. c, pp. 126–130, 2014.

[6] W. Li, Z. Zhang, and Z. Liu, "Action Recognition Based on A Bag of 3D Points," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pp. 9–14, 2010.

[7] Y. Tian, L. Cao, Z. Liu, and Z. Zhang, "Hierarchical filtered motion for action recognition in crowded videos," *IEEE Transactions on Systems, Man and Cybernetics Part C (Applications and Reviews)*, vol. 42, no. 3, pp. 313–323, 2012.

[8] X. Yang, Y. Tian, C. Yi, and L. Cao, "MediaCCNY at TRECVID 2012: Surveillance event detection," *NIST TRECVID, Workshop*, 2012. [Online]. Available: http://www-nlpir.nist.gov/projects/tvpubs/tv12.papers/mediaccny.pdf

[9] M.-K. Hu, "Visual pattern recognition by moment invariants," *IRE Transactions on Information Theory*, vol. 8, no. 2, pp. 179–187, 1962.

[10] M. A. R. Ahad, "Motion History Images, Chapter 3," in *Motion History Images for Action Recognition and Understanding*. Springer London, 2013. [Online]. Available: http://link.springer.com/10.1007/978-1-4471-4730-5

[11] A. F. Bobick and J. W. Davis, "The Recognition of Human Movement Using Temporal Templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[12] D.-M. Tsai, W.-Y. Chiu, and M.-H. Lee, "Optical flow-motion history image (OF-MHI) for action recognition," *Signal, Image and Video Processing*, vol. 9, no. 8, pp. 1897–1906, 2015. [Online]. Available: https://doi.org/10.1007/s11760-014-0677-9

[13] G. Somasundaram, A. Cherian, V. Morellas, and N. Papanikolopoulos, "Action recognition using global spatio-temporal features derived from sparse representations," *Computer Vision and Image Understanding*, vol. 123, pp. 1–13, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.cviu.2014.01.002

[14] C. Thurau, "Behavior Histograms for Action Recognition and Human Detection," in *Proceedings of the 2nd Conference on Human Motion: Understanding, Modeling, Capture and Animation*. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 299–312. [Online]. Available: http://dl.acm.org/citation.cfm?id=1785357.1785381

[15] R. Rosales and S. Sclaroff, "3D Trajectory Recovery for Tracking Multiple Objects and Trajectory Guided Recognition of Actions," in *In Proceedings IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, no. 10, 1999, pp. 117–123.

[16] 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*. New York, NY, USA: ACM, 2012, pp. 1057–1060. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2393347.2396382

[17] Y. Song and Y. Lin, "Combining RGB and Depth Features for Action Recognition based on Sparse Representation," in *Proceedings of the 7th International Conference on Internet Multimedia Computing and Service*, 2015. [Online]. Available: http://doi.acm.org/10.1145/2808492.2808541

[18] N. Nguyen, "Feature Learning for Interaction Activity Recognition in RGBD Videos," *CoRR*, vol. abs/1508.02246, 2015. [Online]. Available: http://arxiv.org/abs/1508.02246

[19] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Human Activity Detection from RGBD Images," in *Proceedings of the 16th AAAI Conference on Plan, Activity, and Intent Recognition*, ser. AAAIWS'11-16. AAAI Press, 2011, pp. 47–55. [Online]. Available: http://dl.acm.org/citation.cfm?id=2908772.2908779

[20] Y. Zhao, Z. Liu, L. Yang, and H. Cheng, "Combing RGB and Depth Map Features for Human Activity Recognition," in *Proceedings of The Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, 2012, pp. 1–4.

[21] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-Up Robust Features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[22] R. Hendaoui, M. Abdellaoui, and A. Douik, "Synthesis of spatio-temporal interest point detectors: Harris 3D, MoSIFT and SURF-MHI," *1st International Conference on Advanced Technologies for Signal and Image Processing, ATSIP*, pp. 89–94, 2014.

[23] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 1981, pp. 674–679.

[24] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, June 2005, pp. 886–893 vol. 1.

[25] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.

[26] M. Rizon, H. Yazid, P. Saad, A. Y. Shakaff, A. R. Saad, M. R. Mamat, S. Yaacob, H. Desa, and M. Karthigayan, "Object Detection using Geometric Invariant Moment," *American Journal of Applied Sciences*, vol. 2, no. 6, pp. 1876–1878, 2006.

[27] A. Khotanzad and J.-H. Lu, "Classification of invariant image representations using a neural network," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 6, pp. 1028–1038, 1990.

[28] Z. Zafar and K. Berns, "Recognizing Hand Gestures for Human-Robot Interaction," *Proceedings of the 9th International Conference on Advances in Computer-Human Interactions (ACHI)*, pp. 333–338, 2016.

[29] M. K. Kundu, D. P. Mohapatra, A. Konar, and A. Chakraborty, *Advanced Computing , Networking and Informatics - Volume 2 Wireless Networks and Security Proceedings of the Second International Conference on Advanced Computing, Networking and Informatics (ICACNI)*. Springer International Publishing, 2014. [Online]. Available: http://www.springer.com/series/8767

[30] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Learning Actionlet Ensemble for 3D Human Action Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, pp. 914–927, 2014.

[31] G. Yu, Z. Liu, and J. Yuan, "Discriminative Orderlet Mining for Real-Time Recognition of Human-Object Interaction," in *Computer Vision – ACCV 2014*. Springer International Publishing, 2015, pp. 50–65.

[32] G. Zhu, L. Zhang, P. Shen, and J. Song, "An Online Continuous Human Action Recognition Algorithm Based on the Kinect Sensor," *MDPI, Sensors (Basel)*, vol. 16, no. 2, pp. 1–18, 2016. [Online]. Available: http://www.mdpi.com/1424-8220/16/2/161/htm

[33] M. Koperski, P. Bilinski, and F. Bremond, "3D Trajectories for Action Recognition," in *The 21st IEEE International Conference on Image Processing (ICIP)*. Paris, France: IEEE, 2014, pp. 4176–4180. [Online]. Available: https://hal.inria.fr/hal-01054949

[34] M. Zanfir, M. Leordeanu, and C. Sminchisescu, "The Moving Pose: An Efficient 3D Kinematics Descriptor for Low-Latency Action Recognition and Detection," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2752–2759, 2013.

[35] W. Ding, K. Liu, F. Cheng, and J. Zhang, "Learning Hierarchical Spatio-temporal Pattern for Human Activity Prediction," *Journal of Visual Communication and Image Representation*, vol. 35, pp. 103–111, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.jvcir.2015.12.006

[36] Meng Meng, H. Drira, M. Daoudi, and J. Boonaert, "Human-object interaction recognition by learning the distances between the object and the skeleton joints," *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1–6, 2015. [Online]. Available: http://ieeexplore.ieee.org/document/7284883/