Local Feature Extraction from RGB and Depth Videos for Human Action Recognition

Rawya Al-Akam and Dietrich Paulus

Abstract—**In this paper, we present a novel system to analyze human body motions (actions) for recognizing human actions by using 3D videos (RGB and depth data). We apply the Bag-of-Features techniques for recognizing human actions by extracting local-spatial temporal features from all video frames. Feature vectors are computed in two steps: The first step consists of detecting all interest keypoints from RGB video frames by using Speed-Up Robust Features detector; then the motion points are filtered by using Motion History Image and Optical Flow, and these important motion points are aligned to the depth frame sequences. In the second step, the feature vectors are computed by using a Histogram of Orientation Gradient descriptor, this descriptor is applied around these motion points from both RGB and depth channels, then the feature vector values are combined in one RGBD feature vector. Finally, the k-means clustering and multi-class Support Vector Machines are used for the action classification task. Our system is invariant to scale, rotation and illumination. All tested results are computed from a dataset that is available to the public and often used in the community. This new features combination method is help to reach recognition rates superior to other publications on the dataset.**

Index Terms—**RGBD videos, feature extraction, K-means clustering, SVM classification.**

I. Introduction

Human action recognition by using the cameras is a very active research topic and it has been widely used in the pattern recognition and computer vision studies to characterize the behavior of persons. Also, it has been widely used in many applications fields like, video analysis, video surveillance, robotics human-computer interaction, robotic and a variety of systems that involve interactions between persons and the computers [1]. Also, the ability to design a machine that is can able to interact intelligently with a human-inhabited environment is important for recognizing human actions from different video frames with different activities of people [2]. In the last decade, the research on human activity recognition is concentrated on recognizing human activities from videos captured by conventional visible light cameras [3]. But recently, action recognition studies have entered a new phase by technological advances and the emergence of the low-cost depth sensor like Microsoft Kinect and ASUS Xtion pro [4]. This depth sensor has many advantages over RGB cameras as it can work in total darkness which makes it possible to explore the fundamental solution for traditional problems in solution for human action classification and it provides 3D structural information as well as color image sequences in real time, and [5], [6]. However, the depth camera also has a limitation which can be partially enhanced by incorporation of RGB and Depth. But all these advantages make it interesting to incorporate the RGBD cameras into more challenging environments.

In this work, we improved the method of [7], [8] to categorize the body motions on RGBD videos rather than 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 into one feature vector for each RGBD video actions. There are two reasons to combine depth information to the RGB features: **firstly**, the RGB data can be strictly impacted by the variation of lighting condition, alongside with the variation of subjects clothing and viewing angle both may influence the recognition results. In compare with depth data, this data contains more intrinsic information for action representation. And **secondly**, because of both physical bodies and movements are presenting as four dimensions, (x, y, z, t), in the real world. which is, human activities involve not only spatial-temporal axes but also the depth axis, that are constraints of the 3D scenes and activities can be directly transposed into image/video contents. Therefore, we should base on the color and depth map images for human action recognition. Our approach represents the human activity recognition in **four** steps: The **First** is the detection of 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) [9] and Optical Flows (OF) [10]. The **Second** is to describe the detected interest points using Histogram of Oriented Gradients (HOG) descriptor. The HOG descriptor is applied on images, MHI and OF channels to extract the feature vector from each video action. The **Third** is the bag of features algorithm, which encodes all the descriptors extracted from each video into a single code. And Finally, in the **Fourth**, we use unsupervised learning like K-means clustering and supervised learning like support vector machine SVM for classification the different action from videos, as shown in Fig. 1 that represents the general structure for our action recognition method.

II. Related Works

In this section, we summarize the state of the art on human action recognition. During the last few years, there are several approaches that proposed to detection, representation and recognition, and understanding video events. Previous research on action recognition mainly focused on RGB

Manuscript received March 1, 2018; revised May 12, 2018.

The authors are with the Faculty of Computer Science, Institute of Computational Visualistics, Active Vision Group (AGAS), University of Koblenz-Landau, Germany, Universitätsstr. 1, 56070 Koblenz (e-mail:rawya@ uni-koblenz.de, paulus@ uni-koblenz.de).

doi: 10.18178/ijmlc.2018.8.3.699

videos, which yielded lots of feature extraction, action representation, and modeling methods using the color images. As in [11], the human detection and simultaneous behavior recognition are presented from RGB image sequences by applying the action representation methods depended on using the clustering methods to the sequence of HOG features descriptor on human motion images. In the other research a hierarchical filtered motion (HFM) method are used to recognize the human action on crowded videos [7]; the 2D Harris corners is used to detect the interest motion points from motion history image (MHI) of the recent motion (i.e. locations with high intensities in MHI). After that, a global spatial motion smoothing filter is applied to the gradients of MHI to eliminate isolated unreliable or noisy motions. To characterize the appearance and temporal motion features, they used HOG descriptor in the intensity image and MHI and for action recognition performance, the Gaussian Mixture Model (GMM) classifier is implemented on these feature vector values.

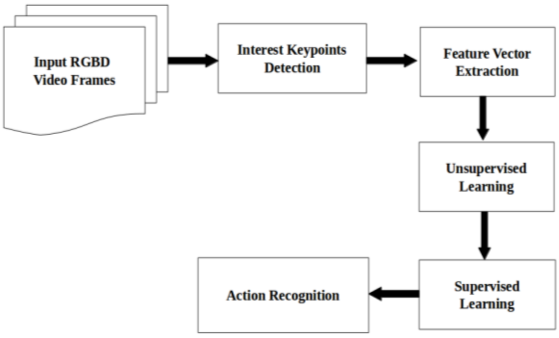


Fig. 1. General structure of our approach.

In the other hand, there are a lot of researchers who improved the human action recognition methods depending on only depth data, as in [12]: they improved the human action recognition by projected depth maps onto three orthogonal planes and collect the global activities through entire video sequences to generate the Depth Motion Maps (DMM). Histograms of Oriented Gradients (HOG) are then computed from DMM as the representation of an action video. . 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 camera as in [13] and the bag-of-visual-word and an SVM classifier are used to recognized the human activities from the 3D data.

In the last few years, a lot of researchers are improved the action recognition performance by using RGBD data, which computing a local spatio-temporal feature based on RGB data, a skeleton joint feature, and a point cloud feature in-depth data, and all these features are combined using sparse coding features combination methods which represented in [14], or by considered a human's activity as composed of a set of sub activities by computing a set of features based on human poses and motion as well as based on image and point cloud information [15]. In [2], the authors are represented a comparison of several well-known pattern recognition methods; which they used Motion History Images (MHI) to describe human activities in a qualitative way and the extracted feature vectors are classified by using machine learning methods such as support vector machine and K-nearest neighbors to recognize the human activities.

III. Overview of Action Recognition System

In this section, we illustrate our steps for computing the feature vectors from each video action in details. In section ***A***, we represent pre-processing to the input RGB and depth videos. In section ***B***, we give a brief description of the Bag of Features extraction. In section ***C***, we explain the Bag of Words generation and in section ***D***, which explain the classification method used to compute the recognition accuracy. In Fig. 2 shows our system scheme.

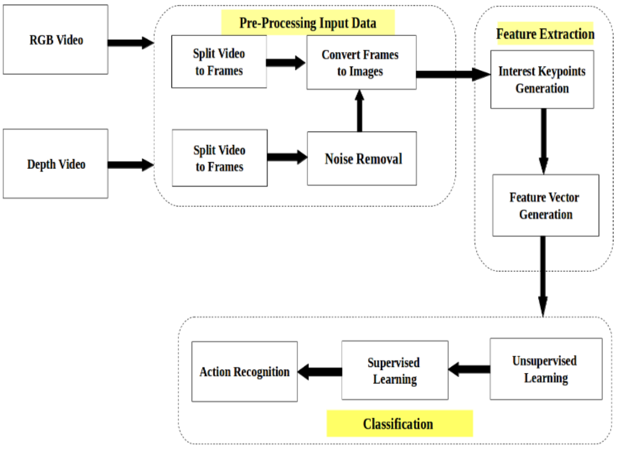


Fig. 2. System analysis schematics of action recognition. Using RGB and depth stream. Pre-processing to the input data; feature extraction; and classification.

A. *Pre-Processing Input Data*

The input dataset consisting of color (RGB) and depth videos was analyzed as a sequence of frames to extract spatio-temporal features which presented in each video action. In this work, we used a lower resolution of (320 × 240) in order to reduce the computational complexity of the system. The depth maps information captured by the Kinect camera are often noisy due to imperfections related to the Kinect infrared light reflections. To reduce like this noise and to eliminate the unmatched edges from the depth images, we used a spatial-temporal bilateral filtering to smooth depth images. The joint-bilateral smoothing filtering is proposed in [16] which is used as in equation (1)

$$\widehat{D}_{(P)} = \frac{1}{K_{(p)}} \sum\nolimits_{\in\Omega_p} \begin{array}{r} f(p,q)g\left(\left\|\widehat{D}_m(p) - \widehat{D}_m(q)\right\|\right) \\ h\left(\left\|I_{(p)} - I_{(q)}\right\|\right) \end{array} \qquad (1)$$

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

B. *Bag of Features Extraction*

For the feature vector extraction, we used the Bag-of-Features (BoFs) technique. It is the most popular technique of feature representation from videos to learn and recognize different human actions. Local features have been computed from the spatio-temporal domain by applying the feature detector and descriptor methods on 3D data(RGBD). The features vectors are extracted and computed into three steps: Interest keypoint generation, feature vector generation, and dictionary generation.

1) Interest keypoints generation

The core of our approach is to find motion interest points (keypoints) from RGB frame sequence by using the Speeded Up Robust Features (SURF) detector [17] to extract more visually characteristic keypoints from spatial domain. Then, these keypoints are filtered by using a temporal (motion) template approach to detect motion and to compute its direction; this constraint by motion history images (MHI), which is generated by computing the difference between two adjacent frames [18], [19]. Those points with larger intensities in MHI represent the moving object with more recent motion. After that, the optical flow is computed for those keys preserved after MHI filtering by using the Lucas-Kanade method [20]. To present motion in the image, a motion-history image (MHI) is used. In a MHI H_τ , pixel intensity is a function of the temporal history of motion at that point. The MHI shown in equation (2) is formally defined as in [18].

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

where $D(x, y, t)$ is a binary image of differences between frames and τ is the maximum duration of motion. The duration τ 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, these motion points are aligned from RGB 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 from all RGBD videos actions, we used the HOG features descriptor [7] on both RGB and depth video sequences and these feature vector values are combined to generate the BoFs. This descriptor is widely used in human detection [21] and action recognition [22]. To generate feature vectors, the descriptor was applied around each motion interest keypoints in video frames of RGBD videos images, MHI, and OF channel; it can be well adapted to represent local shape information from image channel and local motion information from MHI channel by computing distributions of local gradients.

3) Dictionary generation

After extracting the local features information from all RGBD video action by using the detector and descriptor technique, a dictionary from these feature vectors is generated– this is the important step on (BoFs) context. The Dictionary was generated using k-means clustering algorithm as shown 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 important keypoints and if it is set too high then it might lead to over-fitting and increasing the complexity of the system [23].

k-means clustering was applied on all BoF at the learning stage from training videos, the k is represented the dictionary size. The centroid of each cluster is combined to make a dictionary. In our method, a dictionary size is set as k=100 and we have got the best result with it.

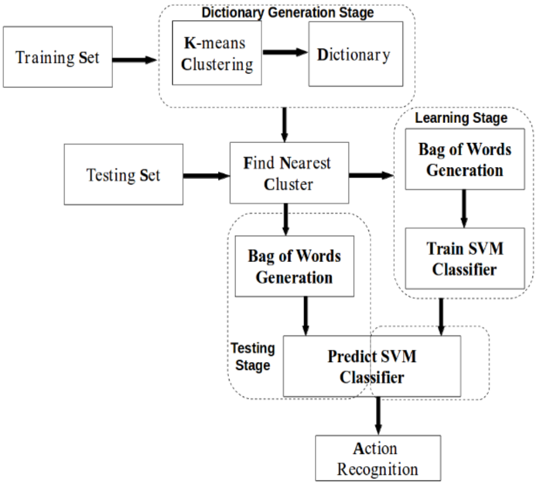


Fig. 3. Dictionary Generation from Feature vector for classified Action.

C. Bag of Words Generation

In order to generate the Bag-of-Words (BoWs), we followed the method in [24]. For each feature vector of the video frames is compared to each centroid of the cluster in the dictionary using Euclidean distance measure e as formulated in equation (3):

$$e=\sum_{j=1}^{m}\sum_{i=1}^{n}\left\|X_i^{(j)}-C_j\right\|^2 \tag{3}$$

where $\left\|X_i^{(j)}-C_j\right\|^2$ is the selected distance measure between the feature vector point and the clustering center C_j . And m is the clustering center length, and n is the feature vector size. Then, the difference e is checked, if this difference is small or features values is closed to a certain cluster, the count of that index is increased. Similarly, the other feature description of the video frames is also compared and the counts of the respective indices are increased to which the feature description values are closest to as in [23]. These BoWs vectors are computed for all the videos for training and testing dataset.

D. Action Classification

In order to make performance comparison of our system, we use a Support Vector Machines (SVM). It uses hyper-planes in high dimensional space to divide training data with the largest margin of the points and it is powerful and widely used in a Bag-of-Words context [25]. In this work, we used multi-class Support Vector Machines (SVM) with RBF (Radial Basis Function) kernel. An RBF is maps the data into an infinite dimensional Hilbert space. The RBF is a Gaussian distribution, calculated as in [26]:

$$k(x,z)=e^{-(x-z)/2\sigma^2} \tag{4}$$

where $K(.)$ is the kernel function, x and z are the input vectors. The Bag of words vectors for all the videos are computed in training stage and labels are appended according to the class. This bag of words vectors is fed into the multi-class SVM learning stage, 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 and the experimental results on them using our approach.

A. Dataset

To evaluate the performance of our system approach, we test our system using two types of datasets, the MSR Daily Activity 3D[1] and Online RGBD Action dataset (ORGBD)[2] Datasets.

1) MSR-DailyActivity3D dataset

The MSR Daily Activity 3D Dataset is a daily activity dataset recorded by the Kinect device and it is designed to represents the human's daily activities in the living room [27]. This dataset contains 16 actions and 10 subjects; each subject performs each activity twice (i.e. in two different poses) like: 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, [24].

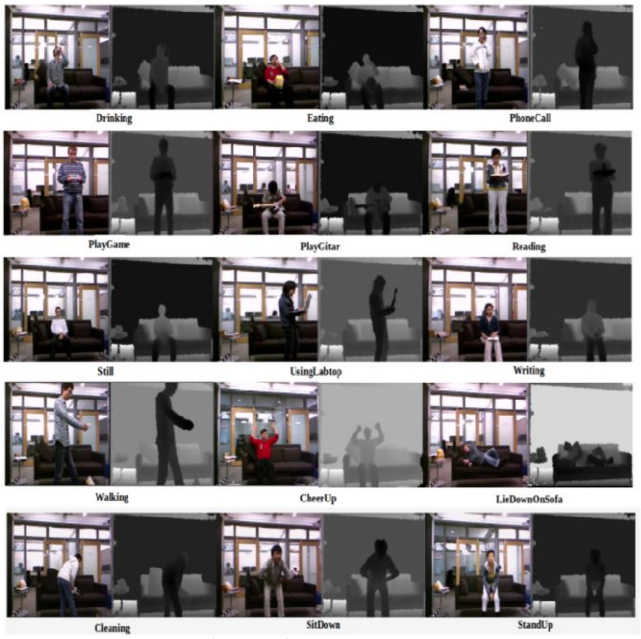


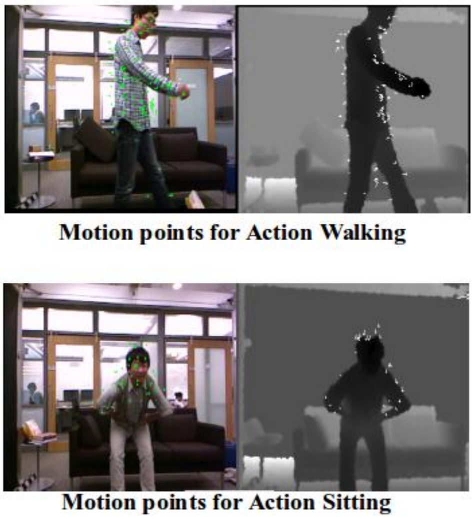
Fig. 4. Sample frames of MSR-Daily Action 3D Dataset.

2) Online RGBD action dataset

The RGB and Depth videos in Online RGBD action dataset (ORGBD) [28] are captured by the Kinect device (RGBD sensor) and designed for three environments tasks: same-environments, cross environments and continues action recognition. Each action was performed by 16 subjects for two times. This dataset contains seven actions categories which are recorded in the living room such as: 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, [24]. In this work, the comparison results is done 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.

B. Experimental Results

In our experiments, the local features extraction methods is used to encode the information which regarding all the available modalities. This feature is extracted from two video channels RGB and Depth channels for each video sequences. As we used SURF on spatial domains of RGB videos to find the interest points in each frame and filtered these points by MHI and OF on temporal domains to extract the motion points from all video frames and then aligned these points to the depth sequences from the depth channel to detect 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 shape and motion features are represented by grids of the histogram of orientation gradient (HOG) [7] which applied around the motion interest points. Normalized histograms of all the patches are combined into HOG (for shape features in the intensity image), HOG-MHI (for motion features in the MHI) and HOG-OF (for motion features in the OF) descriptor vectors as the input to the SVM classifier for action recognition. In our 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.



Motion points for Action Walking

Motion points for Action Sitting

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

TABLE I: Comparison of Recognition Accuracy with Other Methods on MSR-Daily Activity 3D Dataset

Methods	Accuracy
CHAR [29]	54.7%
Discriminative Orderlet [28]	60.1%
Feature Covariance [30]	65.00%
Moving Pose [31]	73.80%
Unsupervised Training [32]	86.9%
Proposed Method	**91.11%**

All the testing results 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

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

methods of the MSR-Daily at 3D datasets and ORGBD Dataset respectively.

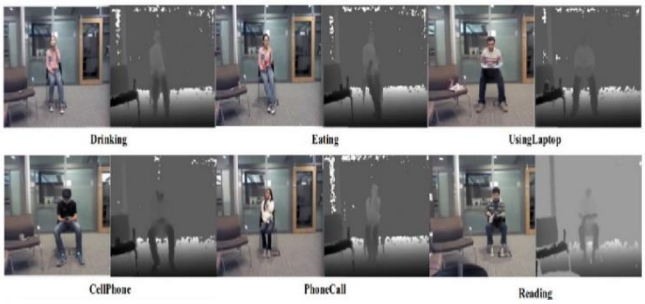


Fig. 6. Motion points in RGB and depth frames of different action represented by green points on RGB frame and white points on depth frames.

TABLE II: COMPARISON OF RECOGNITION ACCURACY WITH OTHER METHODS ON ONLINE RGBD (ORGBD) DATASET

Methods	Accuracy
HOSM [33]	49.5%
Orderlet+SVM [28]	68.7%
Orderlet+ boosting [28]	71.4%
Human-Object Interaction [34]	75.8%
Proposed Method	**92.86%**

V. CONCLUSION AND FUTURE WORKS

In this paper, we improved a human action recognition on 3D video (RGB and Depth data). Our method starts from preprocessing to the input data by removing the noise from the depth data with different filtering and smoothing methods, feature detection and align the RGB motion points with the related depth frames, feature description. The local feature vectors are computed by extracting these features from 3D video data using SURF, MHI, and OF for detecting motion interest points. For the appearance and motion features, the HOG descriptor is used and implemented on the image, MHI, and OF from both RGB and depth video of all different actions. These feature vector values are tested depending on the Bag-of-words method (BoWs) by using k-means clustering and SVM classifier. The presented approach is highly efficient and invariant to cluttered backgrounds, illumination changes, rotation, and scale.

The Experiment results of this method showed that the proposed scheme can effectively recognize the similar action of eating, walking, drinking and so on from different actions. And the recognition accuracy reached to 91.11% on 3D MSR Daily activity dataset and 92.86% on ORGBD dataset. From this accuracy results, we demonstrate that our approach significantly outperforms existing state-of- the-art methods on the same RGBD datasets, because of the motion interest points are computed and extracted solely from the RGB and then aligned to the all depth channels of video frames. After that the RGB and depth based descriptors values are combined depending on this points. For the future works, we will combine a new feature vector values like local binary pattern (LBP), and 3D Trajectory. Also for the classification task, we will use deep learning with convolution neural networks (CNN), and random forest.

ACKNOWLEDGMENT

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. pp. 341–350, 2013.

[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,” in *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELIGENC*, 2016, pp. 1-12.

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

[5] D. Kim, W.-H. Yun, H.-S. Yoon, and J. Kim, “Action recognition with depth maps using hog descriptors of multi-view motion appearance and history,” in *Proc. the Eighth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies*, pp. 126–130, 2014.

[6] W. Li, Z. Zhang, and Z. Liu, “Action recognition based on a bag of 3d points.pdf,” in *Proc. Conf. on Comput. Vis. Pattern Recognition. Workshop. (CVPRW)*, pp. 9–14, 2010.

[7] Y. Tian, L. Cao, Z. Liu, and Z. Zhang, “Hierarchical filtered motion for action recognition in crowded videos,” *IEEE Trans. Syst. Man 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 Surveillance event detection,” *NIST TRECVID*, 2012.

[9] M. A. R. Ahad, “Motion history images”, Chapter 3, in Motion history images for action recognition and understanding,” *Springer Briefs in Computer Science*, 2013.

[10] 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*, pp. 1897–1906, 2015.

[11] C. Thurau, “Behavior histograms for action recognition and human detection,” *in Proceedings of the 2nd Conference on Human Motion: Understanding, Modeling, Capture and Animation, Springer-Verlag Heidelberg, Germany*, pp. 299–312, 2007.

[12] X. Yang, C. Zhang, and Y. Tian, “Recognizing actions using depth motion maps-based histograms of oriented gradients,” in *Proc. Of 20th International. Conference of. Multimedia ACM*, 2012.

[13] N. Nguyen, “Feature learning for interaction activity recognition in rgbd videos,” {CoRR}, arXiv150802246N, 2015.

[14] Y. Song and Y. Lin, “Combining rgb and depth features for action recognition based on sparse representation,” in *Proc. ICIMCS’ 15*, ACM, p.1–49. 2015.

[15] 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*, pp. 47–55, 2011.

[16] Y. Zhao, Z. Liu, L. Yang, and H. Cheng, “Combing rgb and depth map features for human activity recognition,” in *Proc. the Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, 2012.

[17] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, “Speeded-up robust features,” *Computer Vision and Image Understanding, vol. 110*, no.3. pp. 346–359, 2008.

[18] A. F. Bobick and J. W. Davis, “The recognition of human movement using temporal templates,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, 2001.

[19] R. Hendaoui, M. Abdellaoui, and A. Douik, “Synthesis of spatio-temporal interest point detectors: Harris 3D, MoSIFT and SURF-MHI,” in *Proc. 1st Int. Conf. Adv. Technol. Signal Image Process*, pp. 89–94, 2014.

[20] 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), pp. 674–679, 1981.

[21] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 886–893, 2005.

[22] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” in *Proc. 26th IEEE Conf. Comput. Vis. Pattern Recognition*, CVPR, 2008.

[23] Z. Zafar and K. Berns, “Recognizing hand gestures for human-robot interaction,” in *Proc. the 9th International Conference on Advances in Computer-Human Interactions*, pp. 333–338, 2016.

[24] R. Al-Akam and D. Paulus, "RGBD human action recognition using multi-features combination and k-nearest neighbors classification," *(IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 8, no. 10, pp. 383–389, 2017.

[25] J. Uijlings, I. C. Duta, E. Sangineto, and N. Sebe, "Video classification with Densely extracted HOG/HOF/MBH features: An evaluation of the accuracy/computational efficiency trade-off," *Int. J. Multimed. Inf. Retr*., vol. 4, no. 1, pp. 33–44, 2015.

[26] Y. Mahesh, Dr. M. Shivakumar and Dr. H. S. Mohana, "Classification of human actions using non-linear svm by extracting spatio temporal hog features with custom dataset," *International Journal of Research in Science & Engineering*, 2015.

[27] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Learning actionlet ensemble for 3D human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 914–927, 2014.

[28] G. Yu, Z. Liu, and J. Yuan, "Discriminative orderlet mining for real-time recognition of human-object interaction," in Computer Vision ACCV, Lect. *Notes Comput. Sci.*, pp. 50–65, 2015.

[29] 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, p. 1–18, 2016.

[30] A. Perez, H. Tabia, D. Declercq, and A. Zanotti, "Feature covariance for human action recognition," in *Proc. 6th Int. Conf. Image Process. Theory*, 2016.

[31] M. Zanfir, M. Leordeanu, and C. Sminchisescu, "The moving pose: an efficient 3d kinematics descriptor for low-latency action recognition and detection," in *Proc. IEEE Int. Conf. Comput. Vis,* pp. 2752–2759., 2013.

[32] Z. Luo, B. Peng, D.-A. Huang, A. Alahi, and L. Fei-Fei, "Unsupervised learning of long-term motion dynamics for videos," in Proc. CVPR 2017.

[33] W. Ding, K. Liu, F. Cheng, and J. Zhang, "Learning hierarchical spatio-temporal pattern for human activity prediction," *J. Vis. Commun. Image Represent*, vol. 35, pp. 103–111, 2016.

[34] M. Meng, H. Drira, M. Daoudi, and J. Boonaert, "Human-object interaction recognition by learning the distances between the object and the skeleton joints," in *Proc. 11th IEEE Int. Conf. Work. Autom. Face Gesture Recognit*, vol. 1–6, 2015.

Rawya Al-Akam obtained her bachelor degree in 2004 and master degree in 2007 in computer science from Al-Nahrain University, Baghdad, Iraq. She is a PhD student at the Faculty of Computer Science, Institute of Computational Visualistics, University of Koblenz-Landau. Her primary interests are computer vision and robot vision, image and video processing, and pattern recognition.

Ing Dietrich Paulus obtained a bachelor degree in computer science from University of Western Ontario, London, Canada, followed by a diploma (Dipl.Inf.) in computer Science and a PhD (Dr. Ing) from Friedrich-Alexander-University-Erlangen-Nuremberg, Germany. He obtained his habilitation in Erlangen in 2001. Since 2001 he is at the institute for Computational Visualistics at the University Koblenz-Landau, Germany where he became a full professor in 2002. He is head of the Active Vision Group (AGAS). His primary interests are computer vision and robot vision.