HUMAN ACTION RECOGNITION IN VIDEOS USING STABLE FEATURES

Mohib Ullah[1], Habib Ullah[2] and Ibrahim M. Alseadoon[2]

[1]Department of Computer Science,
Norwegian University of Science and Technology, Gjøvik, Norway
[2]Department of Computer Science and Software Engineering,
University of Ha'il, Ha'il, Saudi Arabia

ABSTRACT

Human action recognition is still a challenging problem and researchers are focusing to investigate this problem using different techniques. We propose a robust approach for human action recognition. This is achieved by extracting stable spatio-temporal features in terms of pairwise local binary pattern (P-LBP) and scale invariant feature transform (SIFT). These features are used to train an MLP neural network during the training stage, and the action classes are inferred from the test videos during the testing stage. The proposed features well match the motion of individuals and their consistency, and accuracy is higher using a challenging dataset. The experimental evaluation is conducted on a benchmark dataset commonly used for human action recognition. In addition, we show that our approach outperforms individual features i.e. considering only spatial and only temporal feature.

KEYWORDS

Neural Networks, Local Binary Pattern, Action Recognition, and Scale Invariant Feature Transform

1. INTRODUCTION

Human action recognition is a significant component in many applications including but not limited to video surveillance, ambient intelligence, human-computer interaction systems, and health-care. In spite of incredible research efforts and many encouraging advances in the last ten years, accurate recognition of the human actions is still a challenging problem.

To improve human action recognition performance, many techniques have been proposed and they rely on: Bag-of-Word [1] representations extracted from spatio-temporal interest points [2], dynamic time warping [3] algorithm derived from exemplar-based approaches, and eigenjoints [4] stem from skeleton-based approaches. However, designing effective features for human action recognition is difficult due to large intra-class variation arising from pose appearance and temporal variations. Therefore, it is crucial to design discriminative features. It is worth noticing that the combination of features could boost the discriminative power of a model. Considering combination of two features could provide much more information than observing occurrence of two features individually. We propose a robust approach to combine spatial and temporal features to design a unified model for human action recognition. For this purpose, we consider pairwise local binary pattern (P-LBP) [40] and scale invariant feature transform (SIFT) [41]. The P-LBP

DOI : 10.5121/sipij.2017.8601

and SIFT are spatial and temporal features, respectively. We then adopt an MLP neural network using the spatio-temporal features (P-LBP and SIFT) during the training stage. The action classes are inferred from the testing videos during the testing stage. The overall process of our proposed approach is presented in Fig. 1.

The rest of the paper is organized as follows: Section 2 presents the related work; Section 3 elaborates our proposed method of features extraction; Section 4 presents experimental evaluation; and Section 5 concludes this paper.

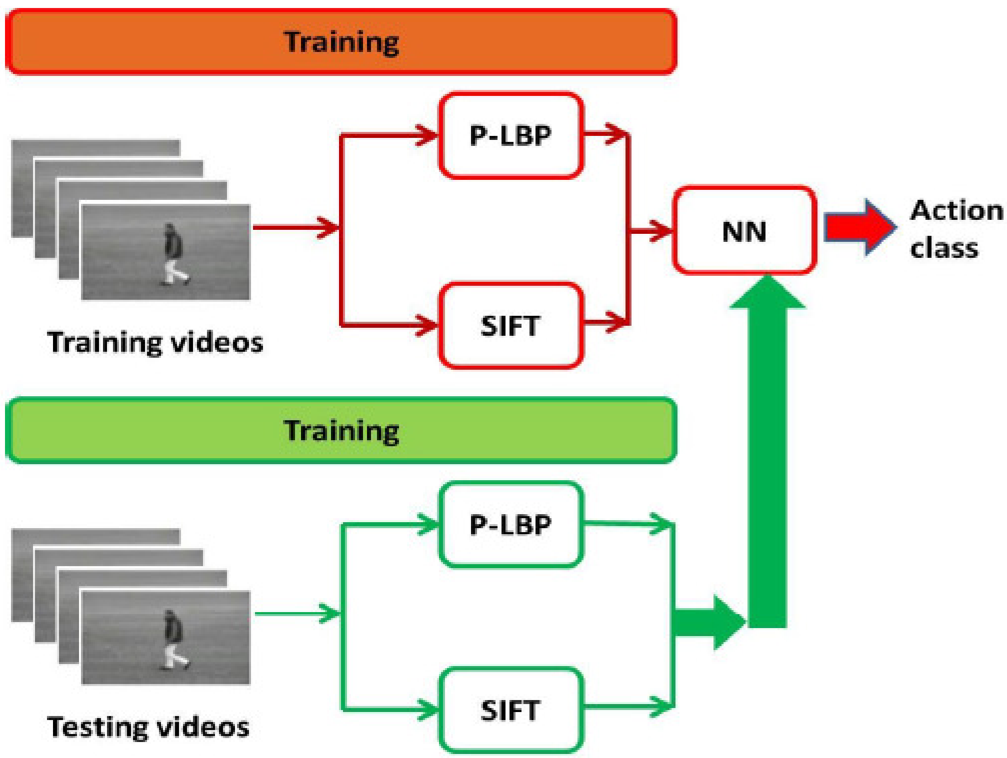

Figure 1. Flow diagram. The training videos are provided to train the neural network and testing videos are used to recognize action classes during the testing stage.

2. RELATED WORK

Herath et al. [5] and Dhulekar et al. [6] present comprehensive surveys on action recognition methods. Fernando et al. [7] propose a function-based temporal pooling method that explores the latent structure of the video sequence. For this purpose, they find out that how frame-level features evolve over time in a video. Rahmani et al. [8] present a robust non-linear knowledge transfer model for human action recognition from different views. The model is based on a deep neural network that transfers knowledge of human actions from any unprecedented view to a shared high-level virtual view by finding a set of non-linear transformations that combines the views. Idrees et al. [9] introduce the THUMOS challenge to serve as a benchmark for action recognition. In THUMOS, action recognition is promoted to a more practical level by introducing temporally untrimmed videos. These also include background videos which share similar scenes and backgrounds as action videos, but are devoid of the specific actions. Zhang et al. [10] present a descriptor called 3D histograms of texture to extract discriminant features from a sequence. The descriptor compactly characterizes the salient information of a specific action, on which texture features are calculated to represent the action. Yang et al. [11] propose a discriminative multi-instance multitask learning framework for human action recognition. For this purpose, they discover the intrinsic relationship between joint configurations and action classes. Wang et al. [12] recognize human action by extracting feature point matches between frames using SURF descriptors and dense optical flow. Han et al. [1] extract sparse geometric features based on the second generation Bandelet transformation. Liu et al. [14] select, characteristic frames using a martingale-based method, followed by the formation of the corresponding motion history through backtracking along the characteristic frames.

Deep learning approaches [15, 16] are also used for human action recognition. For using these approaches, the characteristics of deep networks can be exploited either by improving the connectivity of the network in time [17] or by using optical flow. The convolutional independent subspace analysis method [18] is a deep neural network that uses both visual appearance and motion information in an unsupervised way on video volumes instead of frames as input to the network. However, human actions last several seconds depicting spatio-temporal structure. The deep learning based methods use this structure and learn action representations at the level of a few video frames failing to model actions at their full temporal extent. To cope with this problem, Varol et al. [19] learn video representations using neural networks with longterm temporal convolutions. They demonstrate that this type of modeling with increased temporal extents improve the accuracy of action recognition.

Some researchers prefer depth sensors over color cameras due to their invariance to lightning and color conditions. These methods are generally based on a bag of 3D points [20], projected depth maps [21, 22, 23], spatio-temporal depth cuboid [24], occupancy patterns [25], surface normals [26, 27], and skeleton joints [28, 29, 30, 31]. Methods [32, 33, 34, 35, 36] considering only depth maps can deteriorate from noisy depth maps. To handle this problem, Tran et al. [37] propose a feature combination scheme. However, it can be costly if several features are used since this method needs one classifier and one weight coefficient for each feature, separately. Wang et al. [38] present the actionlet ensemble model (AEM) that exploits both the depth maps and skeleton joints. The AEM combines the relative 3D position of subgroups of skeleton joints and the local occupancy pattern descriptor. Additionally, to capture the temporal structure of actions, they use the short time Fourier transform on fused features to model the final feature vector.

3. PROPOSED METHOD

First we consider spatial feature P-LBP. In order to calculate pairwise rotation invariant local binary pattern (P-LBP), for each pixel on the video frame, a threshold is applied to its symmetric neighbor set on a circle of radius R and the result is considered as a binary number. The P-LBP is formulated in Eq. (1) and Eq. (2).

$$\psi(x) = \sum_{p=0}^{P-1} s_p(x) 2^p \qquad (1)$$

$$s_p(x) = \begin{cases} 1, & V(N_p(x)) \geq V(x) \\ 0, & V(N_p(x)) < V(x) \end{cases} \qquad (2)$$

Where x is a coordinate, $N_p(x)$ is the p^{th} neighbor of point x, and $V(x)$ is the pixel value of point x. It is worth noticing that the function $s_p(x)$ isn't affected by changes in mean luminance. Therefore, the P-LBP could achieve invariance. In order to achieve rotation invariance, Ojala et al. [39] formulates the rotation invariance in Eq. (3)

$$\Psi^{ri}(x) = \min\{ROR(\Psi(x), i) \mid i \in [0, P-1]\} \qquad (3)$$

Where ROR(x, i) calculates a circular bit-wise right shift for i times on P-bit number x. Ojala et al. [39] also investigate that patterns presenting limited spatial transitions indicate the fundamental properties of frame microstructure. For example, the pattern represented by the sequence 11110000 describes a local edge, and the pattern represented by the sequence 11111111 describes a flat region or a dark spot. To formally define these patterns, a uniformity measure is formulated in Eq. (4),

$$U(x) = \sum_{p=1}^{P} | s_p(x) - s_{p-1}(x) | \qquad (4)$$

where $s_P(x)$ is defined as $s_0(x)$. The uniform patterns are subject to the condition $U(x) \leq 2$.

We calculate temporal feature using SIFT. Scale invariant feature transform (SIFT) have different properties that make them suitable for extracting temporal information. In fact, SIFT features are invariant to scaling and rotation, and partially invariant to change in illumination and 3D camera viewpoint. They are well concentrated in both the spatial and frequency domains, reducing the chances of disruption by occlusion and other unprecedented noise. SIFT features are very distinctive, which allows a single feature to be uniquely identified, providing a basis for human action recognition.

For extracting the SIFT features, the scale space of a video frame is defined as a function L(x, y, σ) that is produced from the convolution of the input video frame with a variable scale Gaussian as formulated in Eq. (5)

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \qquad (5)$$

Where G(x, y, σ) is the variable scale Gaussian as formulated in Eq. (6)

$$G(x, y, \sigma) = \frac{1}{2\Pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \qquad (6)$$

In the next stage, the difference of two scales separated by a constant multiplicative factor k is computed according to Eq. (7)

$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y)$$
$$= L(x, y, k\sigma) - L(x, y, \sigma) \qquad (7)$$

To find the local maxima and minima of D(x, y, σ), each spot is compared to its eight neighbors in the frame under observation and nine neighbors in the scale above and below it. It is chosen only if it is larger or smaller than all the neighbors. An important aspect of SIFT approach is that it produces large numbers of features that densely cover the video frame over the full range of scales and locations. The quantity of features is particularly important for human action recognition, where the ability to consider objects exhibiting different actions is very crucial.

We extract both spatial and temporal features from a set of testing videos. We then adopt an MLP feed-forward neural network to learn the behavior of these features instead of considering the values of all the pixels. In fact, these features are exploited to learn different classes of human actions. The motivation for exploring MLP is in its substantial ability, through backpropagation, to resist to noise, and the dexterity to generalize. During the training stage, the weights W and biases b are updated so that the actual output y becomes closer to the desired output d. For this purpose, a cost function is defined as in Eq. (8).

$$E(W, b) = \frac{1}{2} \sum_{i=1}^{n_l} (d_i - y_i^L)^2 \qquad (8)$$

The cost function calculates the squared error between the desired and actual output vectors. Backpropagation algorithm requires the gradient of the cost function E(W ,b) with respect to the weights and biases in each iteration for optimizing the overall cost. According to the learning rate α, the parameters are updated as in Eq. (9) and Eq. (10).

$$w^{k+1} = w^k + \alpha \frac{\partial E(w^k, b)}{\partial w^k} \qquad (9)$$

$$b^{k+1} = b^k + \alpha \frac{\partial E(w, b^k)}{\partial b^k} \qquad (10)$$

4. EXPERIMENTAL EVALUATION

We extensively evaluated our approach on benchmark KTH dataset [2]. A few frames from the dataset are depicted in Fig. 2. The KTH dataset is very diverse since a set of actions are viewed in front of a uniform background. It consists of six human action classes: walking, jogging, running, boxing, waving and clapping. Each action is performed several times by 25 participants. Four different scenarios were considered for recording these sequences: outdoors, outdoors with scale variation, outdoors with different clothes, and indoors. In total, the dataset consists of 2391 video samples. Each sequence averages about 4 seconds in length.

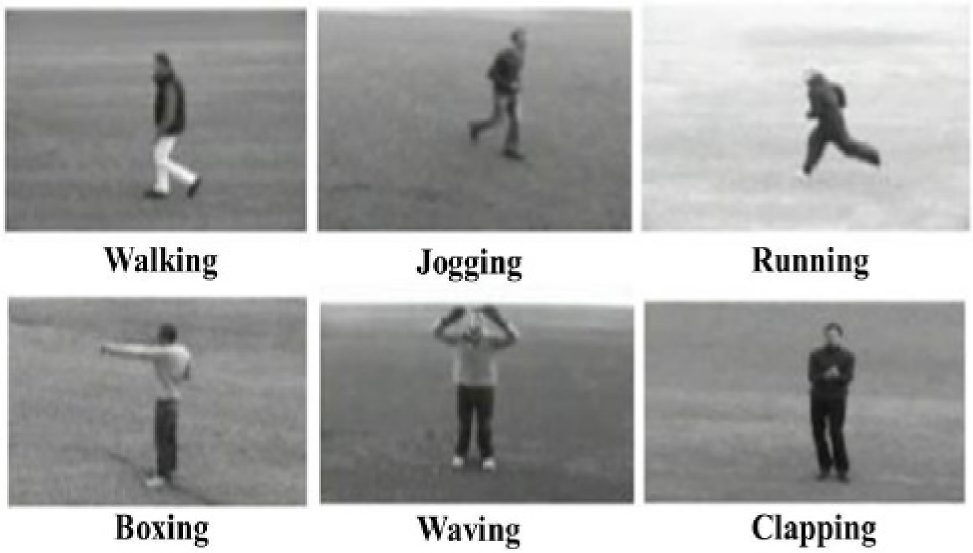

Figure 2. KTH dataset. There are six human action in this dataset including walking, jogging, running, boxing, waving, and clapping

The neural network has been configured using one input layer, two hidden layers and one output layer. The input layer consists of three neurons, each hidden layer consists of three neurons, and a single neuron is considered in the output layer. The adjustment of the neural network in terms of number of layers and number of neuron does not affect the performance significantly. We use an MLP neural network to learn six action classes based on the extracted spatio-temporal features.

The experimental results are presented in Table 1 in term of a confusion matrix. It can be seen that our proposed method shows very good performance on all action classes of KTH dataset. To further elaborate the effectiveness of our proposed method, we carried out experiments considering only spatial and only temporal features. These results are presented in Table 2 and Table 3. It can be seen that the performance considering either spatial or temporal features significantly declines. Thus it shows the robustness of our method.

Table 1. Our spatio-temporal model. Our proposed spatio-temporal model shows very good performance considering KTH dataset.

ACTION	WALK	JOG	RUN	BOX	CLAP	WAVE
WALK	.75	.21	.04	0	0	0
JOG	.07	.71	.20	.02	0	0
RUN	0	.29	.65	.06	0	0
BOX	0	0	.01	.80	.15	.04
CLAP	0	.03	.02	.12	.68	.15
WAVE	0	0	.02	.11	.10	.77

Table 2. Only spatial features. It can be seen that the performance significantly declines using only spatial features P-LBP.

ACTION	WALK	JOG	RUN	BOX	CLAP	WAVE
WALK	.22	.45	.10	.09	.10	.04
JOG	.10	.33	.30	.21	.04	.02
RUN	.15	.10	.10	.43	.20	.02
BOX	.11	.05	.21	.25	.27	.11
CLAP	.01	.02	.11	.33	.31	.22
WAVE	0	.01	.01	.15	.35	.48

Table 3. Only temporal features. It can be seen that the performance significantly declines also in this case considering only temporal features SIFT.

ACTION	WALK	JOG	RUN	BOX	CLAP	WAVE
WALK	.32	.25	.20	.09	.11	.03
JOG	.02	.43	.20	.20	.05	.10
RUN	.12	.13	.31	.22	.18	.04
BOX	.12	.06	.19	.35	.10	.18
CLAP	.02	.03	.19	.23	.35	.18
WAVE	.02	0	.02	.23	.24	.49

5. CONCLUSIONS

In this paper, we proposed an approach for human action recognition using spatio-temporal features and an MLP feed-forward neural network. We demonstrated the capability of our approach in capturing the dynamics of different classes by extracting these features. These features adopt the MLP neural network to learn six action classes. The main advantage of the proposed method is its simplicity and robustness.

REFERENCES

[1] Niebles, J. C., Wang, H., & Fei-Fei, L. Unsupervised learning of human action categories using spatial-temporal words. International journal of computer vision, 79(3), 299-318, 2008.

[2] Laptev, Ivan. "On space-time interest points." International journal of computer vision 64.2-3: 107-123, 2005.

[3] Rabiner, Lawrence R., and Biing-Hwang Juang. "Fundamentals of speech recognition." 1993.

[4] Yang, Xiaodong, and YingLi Tian. "Effective 3d action recognition using eigenjoints." Journal of Visual Communication and Image Representation 25.1: 2-11, 2014

[5] Herath, Samitha, Mehrtash Harandi, and Fatih Porikli. "Going deeper into action recognition: A survey." Image and Vision Computing 60: 4-21, 2017.

[6] Dhulekar, Pravin, et al. "Human Action Recognition: An Overview." Proceedings of the International Conference on Data Engineering and Communication Technology. Springer Singapore, 2017.

[7] Fernando, Basura, et al. "Rank pooling for action recognition." IEEE transactions on pattern analysis and machine intelligence 39.4: 773-787, 2017.

[8] Rahmani, Hossein, Ajmal Mian, and Mubarak Shah. "Learning a deep model for human action recognition from novel viewpoints." IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.

[9] Idrees, Haroon, et al. "The THUMOS challenge on action recognition for videos “in the wild”." Computer Vision and Image Understanding 155: 1-23, 2017.

[10] Zhang, Baochang, et al. "Action recognition using 3D histograms of texture and a multi-class boosting classifier." IEEE Transactions on Image Processing 26.10: 4648-4660, 2017.

[11] Yang, Yanhua, et al. "Discriminative Multi-instance Multitask Learning for 3D Action Recognition." IEEE Transactions on Multimedia 19.3: 519-529, 2017.

[12] Wang, H., Oneata, D., Verbeek, J., & Schmid, C. A robust and efficient video representation for action recognition. International Journal of Computer Vision, 119(3), 219-238, 2016.

[13] Han, H., and X. J. Li. "Human action recognition with sparse geometric features." The Imaging Science Journal 63, no. 1: 45-53, 2015.

[14] Liu, Xueping, Yibo Li, and Qing Shen. "Real-time action detection and temporal segmentation in continuous video." The Imaging Science Journal 65, no. 7: 418-427, 2017.

[15] Minhas, Rashid, Aryaz Baradarani, Sepideh Seifzadeh, and QM Jonathan Wu. "Human action recognition using extreme learning machine based on visual vocabularies." Neurocomputing 73, no. 10: 1906-1917, 2010

[16] Wang, Tingwei, Chuancai Liu, and Liantao Wang. "Action recognition by Latent Duration Model." Neurocomputing 273: 111-119, 2018.

[17] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, “Large-scale video classification with convolutional neural networks,” in CVPR, 2014.

[18] Z. Lan, D. Yao, M. Lin, S.-I. Yu, and A. Hauptmann, “The best of both worlds: Combining data-independent and data-driven approaches for action recognition,” arXiv preprint arXiv:1505.04427, 2015.

[19] Varol, Gul, Ivan Laptev, and Cordelia Schmid. "Long-term temporal convolutions for action recognition." IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.

[20] Wanqing L,ZhangZ, Liu Z (2010) Action recognition based on a bag of 3d points. In: IEEE computer society conference on Computer Vision and Pattern Recognition Workshops CVPRW, 2010.

[21] Chen C, Jafari R, Kehtarnavaz N. Action recognition from depth sequences using depth motion maps based local binary patterns. In: 2015 IEEE winter conference on Applications of Computer Vision WACV, 2015.

[22] Chen C, Liu K, Kehtarnavaz N. Real-time human action recognition based on depth motion maps. Journal of real-time image processing, p 1–9, 2013.

[23] Xiaodong Y, Zhang C, Tian YL. Recognizing actions using depth motion maps-based histograms of oriented gradients. Proceedings of the 20th ACM international conference on multimedia, 2012.

[24] Lu X, Aggarwal JK. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In: IEEE conference on Computer Vision and Pattern Recognition CVPR, 2013.

[25] Wang J, Liu Z, Chorowski J, Chen Z, Wu, Y. Robust 3d action recognition with random occupancy patterns. In: ECCV, p 872–885, 2012.

[26] Omar O, Liu Z (2013) Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In: IEEE conference on Computer Vision and Pattern Recognition CVPR, 2013.

[27] XiaodongY, TianYL. Super normal vector for activityrecognitionusingdepthsequences. 2014IEEE conference on Computer Vision and Pattern Recognition CVPR, 2014.

[28] Bin Liang and Lihong Zheng. A Survey on Human Action Recognition Using Depth Sensors. In International Conference on Digital Image Computing: Techniques and Applications (DICTA), pages 1–8, 2015.

[29] Georgios E, Singh G, Horaud R Skeletalquads: Human action recognition using joint quadruples. In: 2014 22nd international conference on Pattern Recognition ICPR, 2014.

[30] Raviteja V, Arrate F, Chellappa R. Human action recognition by representing 3d skeletonsas points in a lie group. In: IEEE conference on Computer Vision and Pattern Recognition CVPR, 2014.

[31] Xiaodong Y, Tian YL. Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In: IEEE computer society conference on Computer Vision and Pattern Recognition Workshops, 2012.

[32] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action Recognition Based on a Bag of 3D Points. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 9–14, 2010.

[33] C. Lu, J. Jia, and C. K. Tang. Range-Sample Depth Feature for Action Recognition. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 772–779, 2014.

[34] O. Oreifej and Zicheng Liu. HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 716– 723, June 2013.

[35] H. Rahmani, A. Mahmood, D. Huynh, and A. Mian. Histogram of Oriented Principal Components for Cross-View Action Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), PP(99):1–1, 2016.

[36] X. Yang and Y. Tian. Super Normal Vector for Activity Recognition Using Depth Sequences. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 804–811, 2014.

[37] Q. D. Tran and N. Q. Ly. An effective fusion scheme of spatio-temporal features for human action recognition in RGB-D video. In International Conference on Control, Automation and Information Sciences (ICCAIS), pages 246–251, 2013.

[38] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining Actionlet Ensemble for Action Recognition with Depth Cameras. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1290–1297, June 2012.

[39] Ojala, T., Pietikainen, M., & Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on pattern analysis and machine intelligence, 24(7), 971-987, 2002.

[40] Qi, Xianbiao, Rong Xiao, Chun-Guang Li, Yu Qiao, Jun Guo, and Xiaoou Tang. "Pairwise rotation invariant co-occurrence local binary pattern." IEEE Transactions on Pattern Analysis and Machine Intelligence 36, no. 11: 2199-2213, 2014.

[41] Lowe, David G. "Object recognition from local scale-invariant features." In Computer vision, 1999. The proceedings of the seventh IEEE international conference on, vol. 2, pp. 1150-1157. Ieee, 1999.

AUTHORS

Mohib Ullah did Master degree in Telecommunication engineering from the University of Trento, Italy, in 2015. He is currently pursuing Ph.D. degree in computer science from the Norwegian University of Science and Technology (NTNU), Gjøvik, Norway. His research interests include crowd analysis, object detection and tracking, and human action recognition.

Habib Ullah did PhD in 2015 from the University of Trento, Italy. He is currently working as Assistant Professor in the College of Computer Science and Engineering at the University of Ha'il, Ha'il, Saudi Arabia. His research interests include action recognition, crowd motion analysis, and machine learning.

Ibrahim M. Alsaadoen did PhD from the Queensland University of Technology, Brisbane, Australia. Currently, he is working as Assistant Professor and Dean of the College of Computer Science and Engineering at the University of Ha'il, Saudi Arabia. His research interests encompass information security, risk management, and computer vision.