Using Discrete Cosine Transform Based Features for Human Action Recognition

Tasweer Ahmad and Junaid Rafique
Electrical Engineering Department, Government College University, Lahore, Pakistan
Email: tasveer.ahmad@gcu.edu.pk, junaidumtee70@gmail.com

Hassam Muazzam
Electrical Engineering Department, University of Punjab, Lahore, Pakistan
Email: hassammuazzam@hotmail.com

Tahir Rizvi
Dipartimento di Automatica e Informatica, Politecnico di Torino, Turin, Italy
Email: Syed.rizvi@polito.it

***Abstract*—Recognizing human action in complex video sequences has always been challenging for researchers due to articulated movements, occlusion, background clutter, and illumination variation. Human action recognition has wide range of applications in surveillance, human computer interaction, video indexing and video annotation. In this paper, a discrete cosine transform based features have been exploited for action recognition. First, motion history image is computed for a sequence of images and then blocked-based truncated discrete cosine transform is computed for motion history image. Finally, K-Nearest Neighbor (K-NN) classifier is used for classification. This technique exhibits promising results for KTH and Weizmann dataset. Moreover, the proposed model appears to be computationally efficient and immune to illumination variations; however, this model is prone to viewpoint variations.**

***Index Terms*—motion history image, discrete cosine transform, K-nearest neighbor, human computer interaction, video indexing, video annotation**

I. Introduction

The task of Human Action recognition has always been challenging and fascinating for computer vision scientists and researchers within last two decades years. Human Action recognition has found numerous applications in video surveillance, motion tracking, scene modelling and behavior understanding [1]. Intelligent and effective Human Action recognition has received a lot of attention and funding due to rapidly increasing security concerns and effective surveillance of public places such as airports, bus stations, railway stations, shopping malls etc. [1]. Human Action recognition systems can also be deployed at health-care centers, day-care centers, and old homes for monitoring and for fall detection. Human Computer Interaction (HCI), using action recognition, finds ample of applications in interactive and gaming environment [2]. R. T. Collins *et al.* in 2000 [3] suggested that video surveillance can be widely categorized as human detection and tracking, human motion analysis and activity recognition. At that time, they further suggested that "...activity analysis will be the most important area of future research in video surveillance." Now, this projection seems true as a large number of research articles have been published in this domain over the last decade. Although surveillance cameras and monitoring systems are quite prevalent and affordable, but still it is very challenging to devise a robust surveillance systems due to human factors like fatigue and boredom.

It is highly desirable to devise such an intelligent system that can recognize common human actions with remarkable accuracy, multi-scale resolution and minimal computational complexity. A lot of efforts have been made by computer vision researchers to overcome these challenges. A survey by [4] highlights the importance and applications of Intelligent Video Systems and Analytics (IVA). In this survey, both system analytics and theoretical analytics have been targeted. Video system hardware is being developed at faster rate due to digital signal processors and VLSI Design, but still hardware-oriented issues are unresolved due to system scalability, compatibility and real-time performance [5]. Theoretical Analytics deal with more robust and computationally efficient algorithms.

Another breakthrough came in human action recognition by the introduction of multiple cameras for rendering Multi-View Videos for pose estimation and activity recognition. The performance of such systems drastically ameliorated when videos were accessed from multiple cameras [6]. The price paid for multi-channel video was computational complexity; certainly there must be compromise between performance and complexity of the system.

Now-a-days, Infra-Red (IR) Sensor based monocular cameras are widely spread for video gaming and human

Manuscript received February 3, 2015; revised August 25, 2015.

doi: 10.18178/joig.3.2.96-101

pose estimation. Microsoft Kinect Sensor is quite ubiquitous among robotics and computer vision researchers for hand gesture recognition. This new horizon of human action recognition using RGB-Depth Videos is quite popular among research community and, in true sense; this concept has surpassed the performance of systems many-fold.

The statistics shown in Fig. 1 vividly highlights the increasing trend in the area of Human action recognition. The remaining paper is organized as follows. Section II is a brief review of recent techniques for action recognition. Section III renders a basic concept and understanding of Motion History Image, Section IV is brief discussion about Discrete Cosine Transform. Section V elaborates experimental results and compares performance with other techniques. Finally, Section VI is about conclusion, limitations and direction for future work.

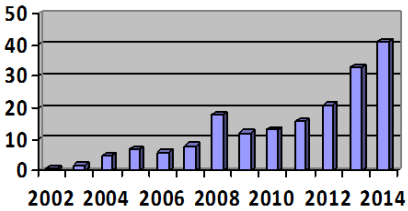

Figure 1. Frequency of research articles published in the domain of human action recognition.

II. Literature Review

Cedras and Shah in 1995 [7] illustrated the significance of Moving Light Display (MLD) for action recognition, MLD includes only 2D information without any structural information. Gavrila in 1999 [8], furnished a survey emphasizing on 2D approaches using shape models and without shape models. Aggarwal and Cai in [9], invigorate action recognition techniques by involving segmentation of low-level body parts. In the context of Human Action recognition, literature review can be categorized into two main approaches.

A. 2D Approaches

This approach exclusively incorporates 2D image data collected through either single camera or multiple cameras. This approach covers simple pointing gestures and complex human actions, e.g. dancing, fighting etc. Moreover, this line of work is used to figure out coarse details of body movement to fine details of hand gesture recognition. Ahmad *et al.* [10] involved motion features by computing Principal Component Analysis (PCA) of optical flow velocity and body shape information. Then, they represented each human action as a set of multi-dimensional discrete Hidden Markov Models (HMM) for each action and view point [11]. Cherala *et al.* [12] depicted some promising results using view invariant recognition that was carried out with the help of data fusion of orthogonal views. Lv *et al.* [13] used Synthetic Training Data to train their model and classified key human poses. Cross-View Recognition method is also very famous among computer vision researchers; several authors have explored this topic [11]. This method is considered to be complex due to training of model on one view and testing on another entirely different view (e.g. side view for training and frontal view for testing in IXMAS). Computer vision researchers have also shown promising results by using some other techniques like: metric learning [14], feature-tree [15] or 3-D Histogram of Oriented Gradients [16].

B. 3D Approaches

This approach deals with feature extraction and description from 3D data for human action recognition. 3D approaches involve both model based representation and non-model based representation of human body and its motion [11]. Ankerst *et al.* [17] used shape features and introduced 3D shape histogram as powerful similarity model for 3D objects. Huang *et al.* [18] combined shape descriptors with self-similarities and made a comparison with 3D shape histogram. Some authors first capture the temporal details of descriptors i.e. shape and pose changes over time, and then add temporal information for reliable action recognition [19]-[21]. Kilner *et al.* [19] used this concept for sports events by applying shape histogram and similarity measure for action matching. Cohen *et al.* [20] computed cylindrical histogram for 3D body shapes and then applied Support Vector Machine (SVM) for classification of view invariant body postures. Huang and Treivedi [21] rendered gesture analysis using volumetric data based on 3D cylindrical shape context. However, this study was view-dependent; while training the subject was asked to rotate [11]. Initially, Mikic *et al.* [22] proposed multi-view 3D human pose tracking using volumetric data. In this study, first they used a hierarchical procedure by figuring out head location due to its specific shape and size and then growing to other body parts. By using this technique, they exhibited very promising results even for some complex body actions but one disadvantage associated with this method was high computational cost. Cheng and Trivedi [23] succeeded to accurately track both body and hand models from volume data by using Gaussian Mixture model framework. This technique exhibited quite remarkable results but it always required manual initialization, therefore was unable to work in real-time scenarios. It is quite intuitive to have a compromise between detailed information of human body pose and computational cost [11]. 3D motion features are also used by authors [24]-[26] for action recognition. For detailed motion information, they had to devise 3D descriptors like 3D motion context (3D-MC) [24] and harmonic motion context (HMC) [24]. 3D-MC captures motion information by including 3D optical flow embedded with 3D histogram of optical flow (3D-HOF) [25], [26]. The drawback of view-point variation in 3D-MC is circumvented in HMC descriptor by computing harmonic basis functions for 3D-MC [11].

III. Motion History Algorithm

Motion History Image (MHI) is used for human action recognition along with Motion Energy Image (MEI). MEI coarsely defines motion energy distribution in spatial

domain for a specific view of a given action, while MHI depicts pixel intensity as a function of motion history in a sequence of images. Action classification is carried out by feature-based statistical framework. MEIs can promptly point out any region where any sort of motion has occurred using consecutive image differencing. First, square of consecutive image differences is computed and then summation of all these images is carried out to engender spatial distribution of signal. Due to lower computational cost for image differencing, it allows real-time acquisition of MEIs. The above concept of MEIs has been illustrated in Fig. 2 below.

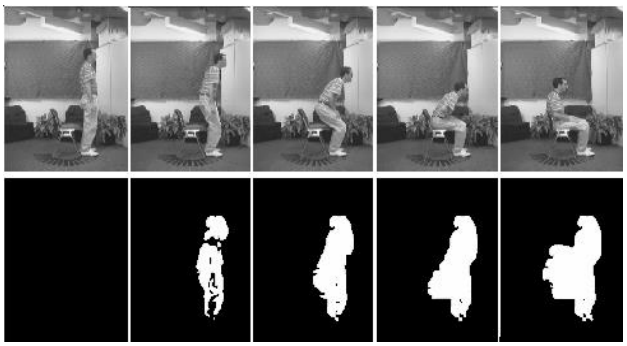

Figure 2. Row 1: Frames from sitting sequence, Row 2: Cumulative binary motion energy image sequence.

MHI is used to represent how motion occurred, where pixel intensity is a function of motion history at that location. Bright pixel values depict more recent motion while dark values correspond to delayed motion. A simple model can be designed by linear discrete decay operator to represent pixel intensity variation. The concept of MHI is illustrated in Fig. 3 below for three actions (sit-down, arms-raise, and crouch-down). It can be vividly perceived that the most recent motion locations are brighter in MHIs.

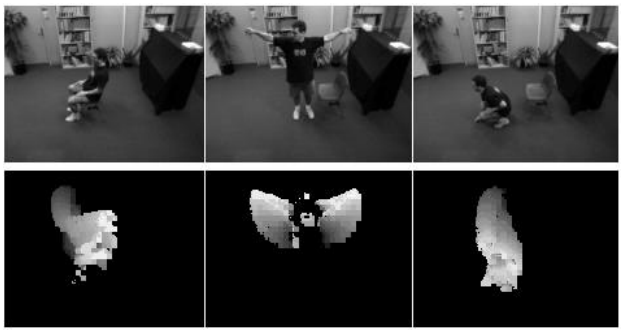

Figure 3. MHIs for sit-down, arms-raise and crouch-down sequences.

The potential advantages associated with MHI are invariance to linear changes in speed and can work for real-time scenarios using standard platforms. The limitation of MHI is that its performance severely mitigates when applied for dense (local) motion flow.

IV. Discrete Cosine Transform

Discrete Cosine Transform (DCT) expresses a sequence of data in terms of summation of cosine functions of varying frequencies. DCT is extensively used for compression of audio and video data and for spectral analysis using partial differential equations. In DCT, the cosine function has superseded the use of sine function because lesser cosine functions are required for a typical signal approximation. One significant advantage of DCT is that its smoothly varying basis vectors resemble the intensity variations of most natural images, such that image energy is matched to a few coefficients [26]. Moreover, more efficient and fast DCT algorithms made it pragmatic choice for most of image compression and dimensionality reduction applications.

A slightly variant of DCT to two dimensional DCT (2D-DCT) is used to for images. 2D-DCT is a separable process and can be implemented using two one-dimensional DCTs: one for horizontal direction, and other for vertical direction.

$$F(u,v)=\left(\frac{2}{M}\right)^{\frac{1}{2}}\left(\frac{2}{N}\right)^{\frac{1}{2}}\sum_{i=0}^{N-1}\sum_{j=0}^{M-1}\Lambda(i).\Lambda(j).\cos\left[\frac{\pi.u}{2.N}(2i+1)\right].$$
$$=\cos\left[\frac{\pi.v}{2.M}(2j+1)\right].f(i,j) \quad (1)$$

where:

$$\Lambda(\xi)=\begin{cases}\frac{1}{\sqrt{2}}, & \text{for } \xi=0\\ 1, & \text{otherwise}\end{cases} \quad (2)$$

In practice, M=N=8, block based 2D-DCT is computed e.g. 8×8=64 pixels, resulting in 64 transform coefficients. The choice of such 8×8 block size DCT is a compromise between compression efficiency and blocking artefacts. The Fig. 4 below shows a 2D-DCT for an 8×8 block of pixels. In this figure, for each step we move horizontally or vertically, the frequency increases by half-cycle. For instance, moving to right one step from top-left results in horizontal frequency increase by half-cycle. A further move to right results in two half-cycle increase. A move down at this stage, engender two half-cycle horizontal and one half-cycle vertical increase.

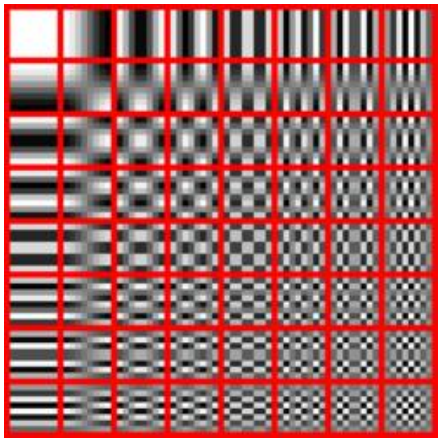

Figure 4. Two-Dimensional DCT frequencies for 8×8 pixel block.

One powerful feature of DCT over other transforms e.g. Discrete Fourier Transform (DFT), is its energy compactness. This concept has been revealed in Fig. 5, where narrower DCT histogram contains more energy compactness as compared to the wider DFT histogram.

This energy compactness of DCT is a significant tool for dimensionality reduction.

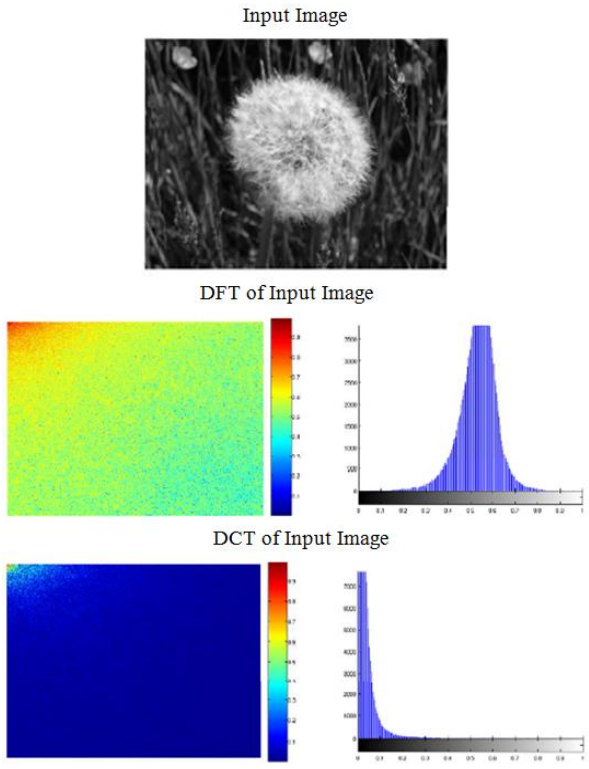

Figure 5. Energy compactness of DCT over DFT.

V. Experimental Results

In this research work, the task of human action recognition was carried out using Motion History Image and Discrete Cosine Transform. K-Nearest Neighbor (KNN) is finally used for classification purpose. The bench marks used for research are KTH dataset and Weizmann dataset for basic human actions. The performance of our algorithm was compared with [2] and found to be improved in terms of accuracy and number of actions. Algorithm in [2] was applied for seven human actions with success rate of 73%, while our algorithm works for 10 actions and success rate of 92.25%. While success rate during training phase was 94.6% and during testing phase was 89.9%. The recall rate for training data was 92.3%, while for testing data was 88.7%. The recall rate also known as sensitivity is the fraction of only relevant instances that are retrieved.

Both precision and recall measure the relevance of result. More intuitively, it can be perceived that high precision means more relevant results are obtained than irrelevant results by the algorithm. Whereas, high recall points out that algorithm returns most of the relevant results.

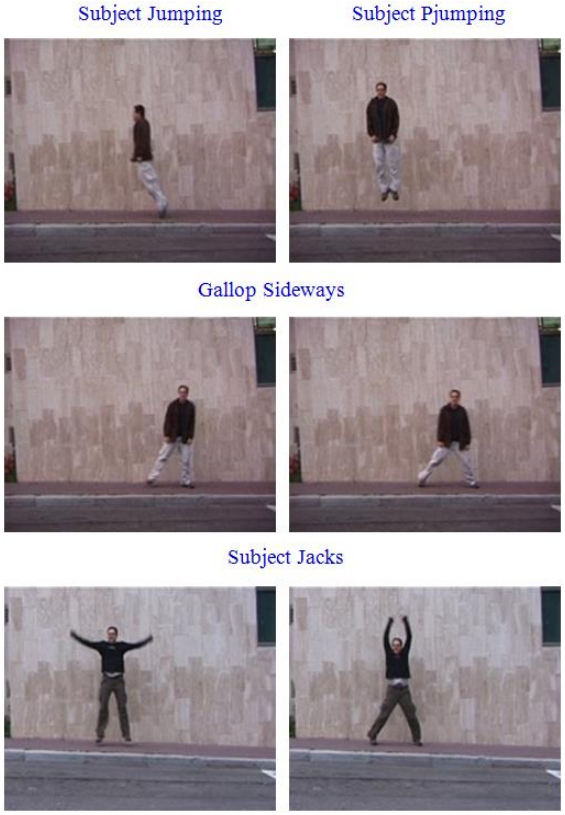

Figure 6. Illustration of different human actions.

TABLE I. Confusion Matrix for Human Action Recognition

Accuracy	Bending	Running	Sliding	Standing	Walking	Gallop	Jump	Jack	Skip	Pjump	Precision %
Bending(12)	12	0	0	0	0	0	0	0	0	0	100
Running(18)	1	17	0	0	0	0	0	0	0	0	94.4
Sliding(45)	0	0	45	0	0	0	0	0	0	0	100
Standing(27)	0	0	0	22	0	2	0	3	0	0	81.5
Walking(14)	0	0	0	2	8	0	1	0	2	1	57
Gallop(23)	0	0	0	1	0	22	0	0	0	0	95.6
Jump(28)	0	0	0	0	0	0	25	0	1	2	89.3
Jack(10)	0	0	0	0	1	0	0	9	0	0	90
Skip(22)	0	0	0	0	0	0	0	0	21	1	95.5
Pjump(19)	0	0	0	0	0	0	0	0	4	15	70

In Table I, twelve bending feature vectors were used. Each feature vector contained motion history of last ten-frames, so in actual bending video sequence contained 120 frames. Same methodology was followed for other video sequences. The difference between Jump and Pjump sequence is that in Jump the subject jumps ahead while in Pjump the subject jumps only at its own location. For Gallop sideway, the subject moves side-ways while jumping; a clear illustration of this action is presented in Fig. 6. The subject stretches and contracts his both legs and arm at its own position for the Jack action.

This algorithm first computes motion history of input sequence, then 8×8-block based 2D-DCT is computed for motion history image. A truncated version of 2D-DCT is used to give less weight to higher frequencies. For every action, each feature vector was a row vector of dimension 1×533. Then, a non-parametric algorithm K-Nearest Neighbor (K-NN) classifier was used for classification purpose. K-NN is dependent on the value of k; higher value of k is suitable for noise suppression but results in lesser distinct boundary between neighbors. In our case, the algorithm showed optimum results when a value k=10 was selected. However, the processing time of algorithm was surpassed as compared to case when default value of k was selected.

VI. Conclusion and Future Work

It is quite evident from the Confusion Matrix shown in Table I, that performance of the proposed algorithm is remarkable. This technique involves quite simple methodologies like MHI, DCT and K-NN. A lot of improved and fast algorithms are readily available for implementation of MHI and DCT techniques. One constraint of this technique is that it is view-dependent; view-point variation severely mitigates the performance of this algorithm. In this scheme of work, MATLAB7 R13, was used for computing MHI and for 2D-DCT feature vector extraction. Then, RapidMiner was used for KNN classification on hp 2.4GHz corei3machine.

Acknowledgment

We acknowledge to Department of Electrical Engineering, Government College University, Lahore Pakistan for mentoring and rending us state-of-the art lab facilities.

References

[1] O. P. Popoola and K. J. Wang, "Video-Based abnormal human behavior recognition—A review," *IEEE Trans. Systems, Man, and Cybernetics*, vol. 42, no. 6, pp. 865-878, Nov. 2012.
[2] M. Hassan, T. Ahmad, and M. Ahsan, "Human activity recognition using motion history algorithm," *International Journal of Scientific and Engineering Research*, vol. 5, no. 8, Aug. 2014.
[3] R. T. Collins, A. J. Lipton, and T. Kanade, "Introduction to the special section on video surveillance," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 745-746, Aug. 2000.
[4] H. H. Liu, S. Y. Chen, and N. Kubota, "Intelligent video systems and analytics: A survey," *IEEE Trans. Industrial Informatics*, vol. 9, no. 3, pp. 1222-1233, Aug. 2013.
[5] X. Q. Zhang, W. M. Hu, Q. Wei, and S. Maybank, "Multiple object tracking via species-based particle swarm optimization," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 20, no. 11, pp. 1590-1602, Nov. 2010.
[6] M. B. Holte, C. Tran, M. M. Trivedi, and T. B. Moeslund, "human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 5, pp. 538-552, Sept. 2012.
[7] C. Cedras and M. Shah, "Motion-Based recognition: A survey," *Image Vision Comput.*, vol. 13, pp. 129-155, 1995.
[8] D. M. Gavrila, "Visual analysis of human movement: A survey," *Comput. Vis. Image Understanding*, vol. 73, pp. 82-98, 1999.
[9] J. K. Aggarwal and Q. Cai, "Human motion analysis: A review," in *Proc. Nonrigid and Articulated Motion Workshop*, 1997, pp. 90-102.
[10] M. Ahmad and S. W. Lee, "HMM-Based human action recognition using multiview image sequences," in *Proc. ICPR 2006. 18th International Conference on Pattern Recognition*, 2006, vol. 1, pp. 263-266.
[11] S. Cherla, K. Kulkarni, A. Kale and V. Ramasubramanian, "Towards fast, view-invariant human action recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Anchorage, 23-28 Jun. 2008, pp. 1-8.
[12] F. J. Lv and R. Nevatia, "Single view human action recognition using key pose matching and viterbi path searching," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, 17-22 Jun. 2007, pp. 1-8.
[13] D. Tran and A. Sorokin, "Human activity recognition with metric learning," in *Proc. 10th European Conference on Computer Vision*, Marseille, 2008, pp. 548-561.
[14] K. Reddy, J. Liu, and M. Shah, "Incremental action recognition using feature-tree," in *Proc. 12th International Conference on Computer Vision*, Kyoto, 2009, pp. 1010-1017.
[15] D. Weinland, M. Ozuysal, and P. Fua, "Making action recognition robust to occulusions and viewpoint changes," in *Proc. 11th European Conference on Computer Vision*, Greece, 2010, pp. 635-648.
[16] M. Ankerst, G. Kastenmuller, H. P. Kriegel, and T. Seidl, "3D shape histograms for similarity search and classification in spatial databases," in *Proc. 6th International Symposium Spatial Databases*, Hong Kong, 1999, pp. 207-226.
[17] P. Huang, A. Hilton, and J. Starck, "Shape similarity for 3D video sequence of people," *International Journal of Computer Vision*, vol. 89, no. 2-3, pp. 362-381, 2010.
[18] J. Kilner, J. Y. Guillemaut, and A. Hilton, "3D action matching with key-pose detection," in *Proc. 2009 IEEE 12th International Conference on Computer Vision Workshop*, Kyoto, 2009, pp. 1-8.
[19] I. Cohen and H. Li, "Inference of human postures by classification of 3D human body shape," in *Proc. IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, 17 Oct. 2003, pp. 74-81.
[20] K. Huang and M. Trivedi, "3D shape context based gesture analysis integrated with tracking using omni video array," in *Proc. Comput. Vis. Pattern Recognit. Workshops*, 2005.
[21] I. Mikic, M. M. Trivedi, E. Hunter, and P. Cosman, "Human body model acquisition and tracking using voxel data," *Int. J. Comput. Vis.*, vol. 53, no. 3, pp. 199-223, 2003.
[22] S. Y. Cheng and M. M. Trivedi, "Articulated human body pose inference from voxel data using a kinematically constrained Gaussian mixture model," in *Proc. Computer Comput. Vis. Pattern Recognit. Workshops*, 2007.
[23] M. B. Holte, T. B. Moeslund, N. Nikolaidis, and I. Pitas, "3D human action recognition for multi-view camera systems," in *Proc. International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT)*, Hangzhou, 16-19 May 2011, pp. 342-349.
[24] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509-522, Apr. 2002.
[25] M. Kortgen, M. Novotni, and R. Klein, "3D shape matching with 3D shape contexts," in *Proc. 7th Central European Seminar on Computer Graphics*, 2003.
[26] M. Ghanbari, "Principles of video compression," in *Standard Codecs: Image Compression to Advanced Video Coding*, IET Press, 2003, ch. 3, pp. 26-30.

Tasweer Ahmad received his B. Engg. from University of Engineering and Technology, Taxila, Pakistan and M.S.c Engg. from University of Engineering and Technology, Lahore, Pakistan. His area of research includes Image Processing, Computer Vision, and Machine Learning.

Hassam Muazzam did his B.Sc. Electrical Engineering from University of the Punjab, Lahore Pakistan and pursuing M.S.c Electrical Engineering from Govt. College University, Lahore, Pakistan. His research interests include Image Processing and the application of soft computing techniques to problems in control.