©2016 Society for Imaging Science and Technology
DOI: 10.2352/ISSN.2470-1173.2016.3.VSTIA-514

# A Real-Time System for Shoppers' Action Recognition

***Srenivas Varadarajan* [a] and *Shahrokh Shahidzadeh* [b]**

[a] *Intel Labs, Intel Corporation, Bangalore, India 560103*
[b] *Intel Labs, Intel Corporation Hillsboro, OR 97124*
*srenivas.varadarajan@intel.com, shahrokh.shahidzadeh@intel.edu*

## Abstract

*The paper proposes a pose-based real-time system for inferring the engagement of a shopper with a retail shelf by recognizing some atomic actions of shelf-interaction. These actions include examining the shelf, reaching for an object, taking an object, reading a product's label and placing it on a cart for check-out. A novel pose-representation that is robust to large intra-class variations while performing these retail actions, is proposed in this work. The paper also extends the framework to do real-time action segmentation, abnormal action detection and configurable privacy protection of shoppers. The abnormality detection also offers a scope for learning new un-modelled actions through crowdsourcing. Though the system currently relies on a Kinect sensor (RGBD) for computing the joints of the human body, the system can work with a combination of RGB surveillance camera and any 2D video-based pose-tracking algorithm. The system has an accuracy of 90% in recognizing the 5 actions considered in this work and exhibits a latency of about 1 sec w.r.t. real world action. This can have a huge potential in optimizing store resources and in improving the shopping experience of the customer.*

**Index Terms—** Action Recognition, Pose Estimation, Hidden Markov Model, Abnormal Actions Detection, Crowdsourcing

## Introduction

Understanding human actions is very important for the normal and efficient functioning of various ambiences like retail-shops, hospitals or airports. The accurate and timely action recognition can include benefits like efficient resource management, enhanced user-experience and higher levels of safety and security of the stake-holders. The existing approaches for action recognition can be classified as single layer approaches and multi-layer approaches as mentioned in [1]. Single layer approaches directly infer the action from the video sequence or from the features extracted from the video sequence. In hierarchical approaches, the lower layers infer atomic action units like the movement of a limb while the higher layers infer the action and activities from the results of the lower layers. The single layer approaches can be further classified as approaches based on describing the space-time volume of the human silhouette ([2], [3]), methods based on tracking human body joints ([4],[5]) and those based on local patch descriptors around the interest points ([6],[7]). The single layer approaches also include characterizing action as a sequence of features modelled through exemplar videos [8] or as a state-space model ([9], [10]) in which each state corresponds to a pose. Multi-layer approaches include the statistical methods based on layered state space models [11], syntactic approaches based on a sequence of action grammars [12] and description based approaches that attributes a time interval to the action grammar [13].

**Table 1: Inference from recognizing customer's interactions with a retail shelf**

| Action | Engage-ment Level | Inference |
|---|---|---|
| Examining without touching products | Least | 1.Products or the price is un-compelling<br>2.Non-serious shopper is looking around |
| Reaching Object, taking Objects, reading labels but not buying | Medium | 1.The Shopper is serious about buying but not getting exactly what he\she wants due to brand, pricing, size or some other reason |
| Placing in the shopping cart | Highest | 1.Customer is satisfied with the product<br>2.On recognizing the product a smart checkout is possible |

The action recognition involves representing the action compactly in terms of feature descriptors and classifying them. Based on the action representation, the approaches can be divided as global representation based approaches and local representation based approaches [14]. The global approaches start with the extraction of human silhouette of an action from the video frames and form feature descriptors of the 3D space-time volume present in the silhouette. On the other hand, the local representation approaches start with a set of spatially and temporally significant interest points, create patches around the interest points in each frame and characterize the action by correlating the patch descriptors of successive frames. The global methods are proven to be more reliable than the local methods under controlled environmental conditions like lighting, occlusion, clutter etc. Pose-based approaches like [15] describe action as a sequences of human poses. They have proven to be more effective in recognizing human actions as pose is agnostic to image properties like skin color, dress color etc. Hence we take such an approach in this work. We propose an alternative to the Histogram of Joints representation [15] and our work was more robust to intra-class variations while performing the shelf-actions.

The work aims in recognizing the shopper actions which are of great business impact to the retail store. In particular, we endeavor to measure the amount of customer engagement with a shelf. This could vary from as low as examining the shelf to as high as picking a product for check-out. The level engagement can reflect on the shopper's needs and the popularity of a product or a category of products as tabulated in Table 1. The detection of placing a product in the cart can also pave the way to a smart checkout use-case where in shoppers can avoid long times at checkout counters for scanning

DOI: 10.2352/ISSN.2470-1173.2016.3.VSTIA-514

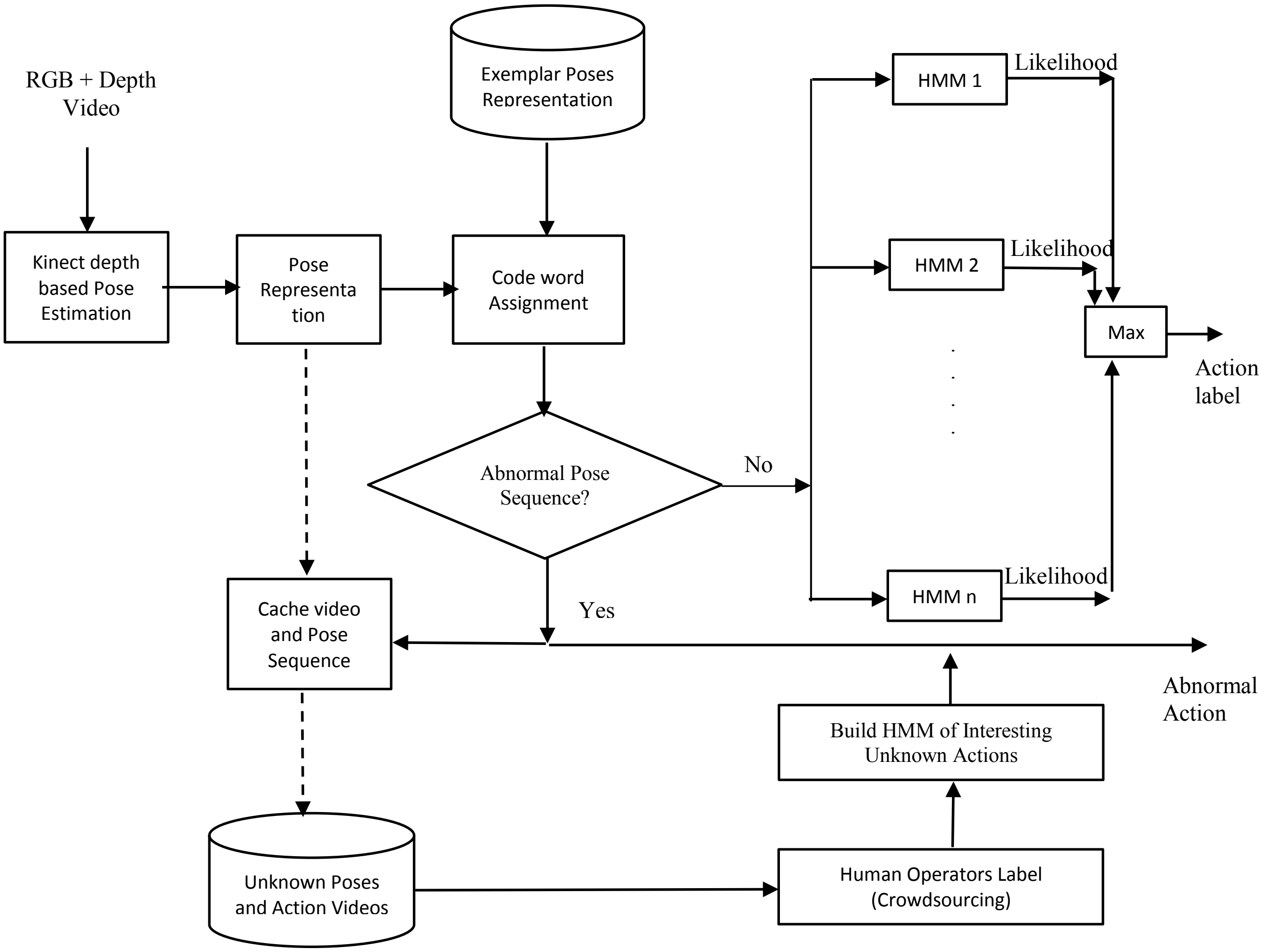

RGB + Depth Video
Exemplar Poses Representation
Kinect depth based Pose Estimation
Pose Representa tion
Code word Assignment
Abnormal Pose Sequence?
No
Yes
HMM 1
HMM 2
HMM n
Likelihood
Likelihood
Likelihood
Max
Action label
Cache video and Pose Sequence
Abnormal Action
Build HMM of Interesting Unknown Actions
Unknown Poses and Action Videos
Human Operators Label (Crowdsourcing)

***Figure 1.*** *A frame-work for Human Action Recognition and Actions Learning*

the products. A work aiming at measuring customer engagement and people counting in a store, using RGBD cameras was proposed in [16]. But the work in [16] addresses only a subset of actions covered in this work and does not model or incorporate human poses for action recognition as proposed in this work. In addition to these normal actions our system is capable of detecting abnormal actions like falling down of a shopper, drop and pick up of an object etc.

This paper is organized as follows. The proposed action recognition framework is described in Section 2. Performance results are presented in Section 3 followed by a conclusion in Section 4.

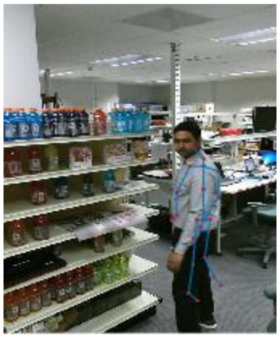
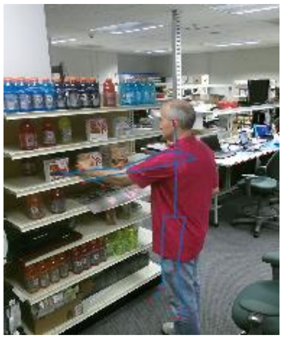
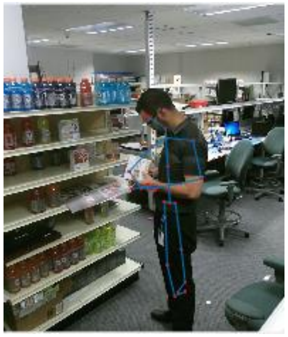
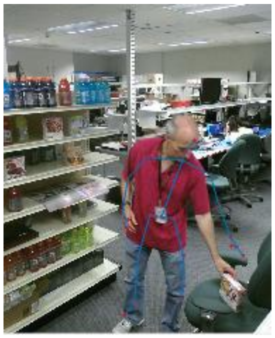
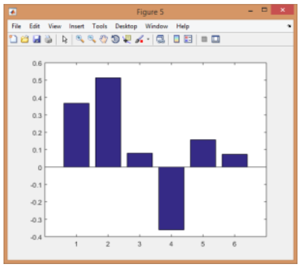
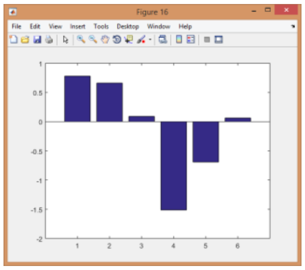
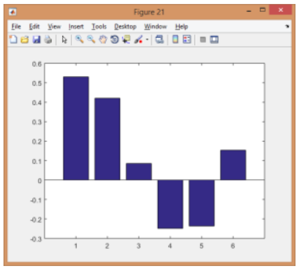
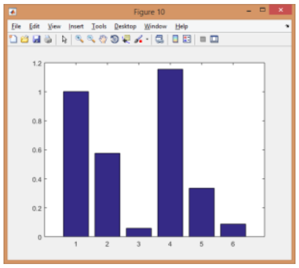
## Proposed Method

The proposed frame-work for action recognition and learning novel actions is depicted in Figure 1. The system employs a Kinect sensor which uses a depth camera to detect the human joints and skeleton. This current method of estimating human poses can be replaced by any algorithm which does joint tracking from RGB or RGBD videos. The pose is represented through the relative displacements of the most discriminative body joints for the actions under consideration. In our work, we consider the X,Y,Z displacements of the wrist and elbow joints from the hip as these are most discriminative in detecting the shelf-interactions of the shoppers. In order to account for only the actions of the active hand and form a pose representation that is agnostic to the location of the inactive hand, we consider the maximum of the right and left hand's displacements of the wrist and elbow joints. This reduced pose representation is robust to a large variety of intra-class variations and hence effective in recognizing the shelf actions. For example, a shopper might be speaking in phone or holding on to a product in one hand while fetching for another object using his other hand. The displacements are normalized by the torso length of the individuals to make the system robust to shopper's height and the depth of the shopper from the image plane of the camera. The torso is measured as the distance between the neck and hip center.

The incoming frame's pose is mapped onto one of the exemplar poses in a pose-dictionary using a nearest neighbor mapping. The exemplar poses in the pose-dictionary are chosen such that they are well-defined and most discriminative in symbolizing a sequence of poses in a considered ambience. For example, the pose-dictionary in a retail ambience will be different from that used in a hospital

DOI: 10.2352/ISSN.2470-1173.2016.3.VSTIA-514

**Table 2. Novel Pose Representation using the signed max displacements (X, Y, Z) of elbow and wrist joints from hip**

| Hands down | Hands on Shelf | Hands holding Product | Hands on Cart |
|---|---|---|---|
|  |  |  |  |
|  |  |  |  |

ambience. Examples of the 4 poses used in the retail ambience pose-dictionary and the corresponding pose features are shown in rows 2 and 3 of Table 2 respectively. The poses include hands-down posture, hand on a shelf posture, reading a product's label and placing the object on a cart for checkout. A set of frames is thus mapped to a set of pose symbols. Based on the deviation of the observed pose sequence from an expected set of postures in a given ambience abnormal action is detected as explained below. If the pose sequence is normal, then it is input to a set of HMMs, each of which is modelled to recognize a single action. Each HMM gives out the likelihood of the considered pose sequence generate by the corresponding action class. The maximum amongst these likelihoods gives the action label for the sequence.

If the pose sequence is abnormal, then the corresponding RGB frames along with the pose-data is cached in to an unknown actions database. Depending on the business logic, these actions can be annotated on a periodic basis through crowdsourcing where human observers assign action labels to these frames. When an un-modelled action occurs sufficient number of times, it is modelled by training a new HMM and promoted as a newly learnt known action in the given ambience. Once this happens, it can be recognized as another normal action. The actions learning frame-work proposed in this work is not yet implemented but can be accomplished with minimal effort.

## Action Segmentation

Since actions happen seamlessly one after the other, the system a priori has no clue about which set of frames characterize an action. So a sliding window based approach is used where in the pose representations of not only the past frames but also the future frames is used for assigning an action label to the frame under consideration. A HMM based Action Label (AL) assignment is done for N frames in the local neighborhood of the current frame as shown in Figure 2, and the maximum occurring action label amongst the N frames is assigned to the current frame. This helps in segmenting the contiguous video into different action segments as shown in Figure 2. The sliding window creates an N frame delay in the system between the times an action was performed to the time it was detected. The value of N was 15 in our system which corresponded to a 1 sec delay.

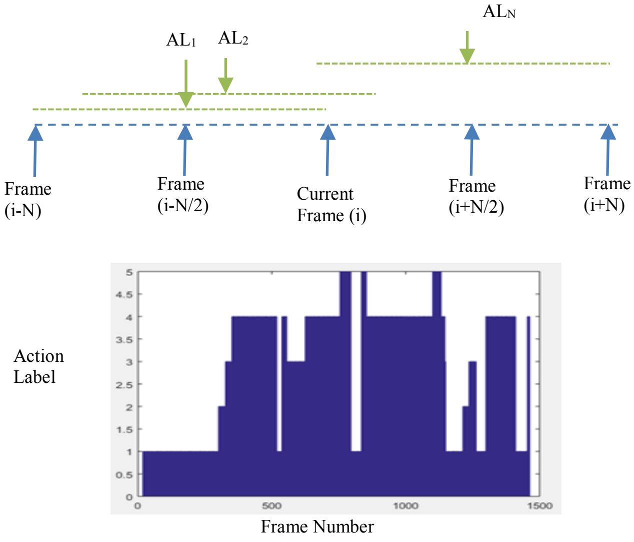


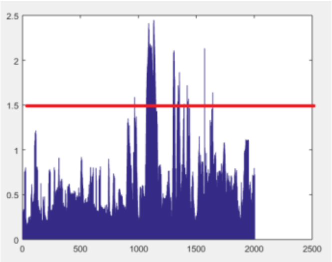
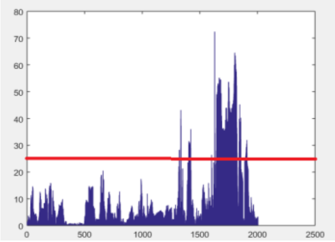
***Figure 2.** HMM based action segmentation*

## Abnormal Action Detection

There are 2 ways of detecting an abnormal action in a retail ambience as mentioned in Table 3. The first type of abnormality (Type I) occurs if the poses deviate a lot from any of the modelled poses in the pose-dictionary. Also, since shoppers are normally

©2016 Society for Imaging Science and Technology
DOI: 10.2352/ISSN.2470-1173.2016.3.VSTIA-514

**Table 3. Types of Abnormal Poses in Retail Ambience**

| | |
|---|---|
| **Type I**. When wrist and elbow poses deviate a lot from the dictionary of exemplar poses |  |
| **Type II**. When the torso and the thighs deviate a lot from vertical axis. (Assuming Shoppers are upright normally in a store) |  |

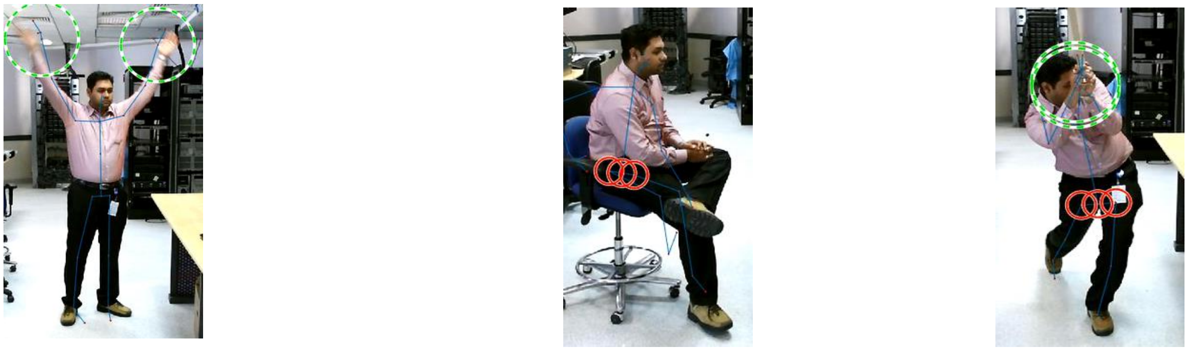
expected to be standing erect in a shop, a large angular deviation of the torso, left and right thighs from the vertical axis acts as a strong clue for abnormality detection and this defines the second type of abnormality (Type II). Whenever the pose-deviation or the angular deviation exceeds a corresponding threshold an abnormal pose is detected. The frame-wise pose and angular deviations and the corresponding detection thresholds (indicated by the red colored bars) are shown in column 2 of Table 3. Figure 3 illustrates Type I and Type II abnormalities with large yellow-green circles and small red circles respectively. As shown in Figure 3(a), Type I abnormality happens due to the pose-deviation of elbow and wrist joints from modelled poses. In Figure 3(b), the wrist and elbow joints look similar to a “Hands-Down” pose but the angles which the torso and things make with horizontal deviate a lot from that of a standing shopper and hence an abnormality is detected. In Figure 3(c), the shopper plays a cricketing shot which is caught by both types of abnormalities.

## Privacy Protection

Automatic surveillance systems should be capable of detecting interesting events without compromising on the privacy of the individuals. The proposed system is capable of protecting the privacy of shoppers by blurring the skeletal joints selectively as shown in Figure 4. The head is unaltered in the examples shown in Figure 4, while other body parts are masked. This is done by replacing all the pixels in the local neighborhood of a skeletal joints with the local average. A rectangular neighborhood is chosen centered around the joint locations. The spread of the rectangular region can be determined by the foreground mask or from the height of the person and anthropometric ratios. The head can also be replaced by the local average if a completely anonymous setting is needed.

## Simulation Results

The accuracy of the proposed algorithm in detecting shopper’s actions in front of a shelf is tested by the Leave Out One Cross Validate (LOOCV) technique on 74 hand-clipped retail actions videos, each exclusively representing one of 5 actions classes. The actions were performed by 6 subjects which included both left and right handed subjects, male and female in the age group of 30-50 years and belonging to 5 different demographics. In this approach one of the action video is picked up at random as a test sequence and the rest are used for training the HMMs of different action classes.



(a) Type I

(b) Type II

(c) Both Types I and II

***Figure 3**. Illustration of Type I and Type II abnormalities*

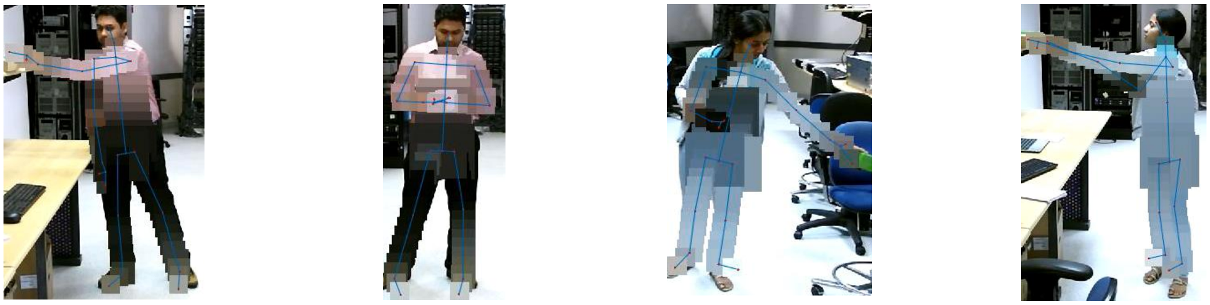


***Figure 4.** Privacy Protection of people*

DOI: 10.2352/ISSN.2470-1173.2016.3.VSTIA-514

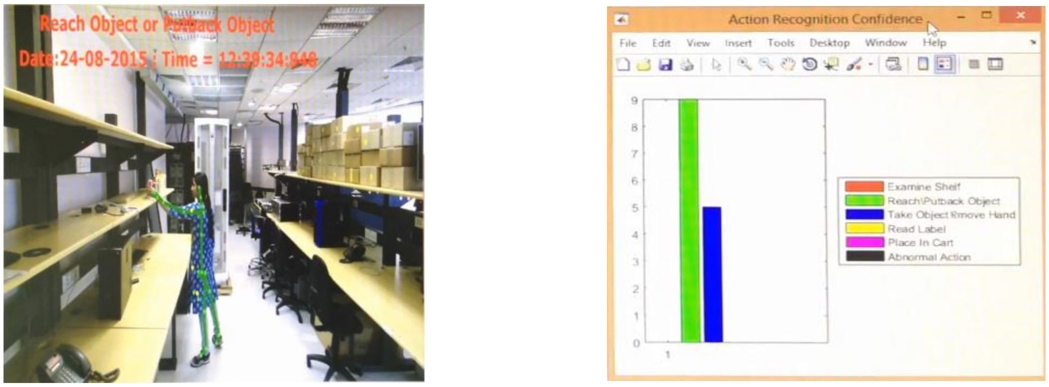


***Figure 5**. Real-Time Action Recognition: (a) RGB data with action label and time overlaid on a frame; (b) Confidence in action classification for that frame.*

The trained HMMs are used for classifying the randomly picked test pose sequence and the classification result is recorded. The process is repeated for all the 74 test sequences in the database and action recognition accuracy is averaged over all sequences of an action class through a confusion matrix shown in Table 4. Each row of the table indicates the real action that was performed and each column indicates the classification result. The confusion matrix quantifies the expectation of classifying one action as another. While a diagonal matrix of all ones along the main diagonal indicates a perfect classification, the confusion matrix in Table 4 indicates that our proposed system has an accuracy of 90% in classifying the 5 shelf actions under consideration. The ability of our system to perform real-time action recognition is shown in Fig. 5. As shown in Fig. 5(a), when the shopper reaches for an object on the shelf, that action is recognized and overlaid on top of the image. Along with the action label, the date and time of performing the action is also overlaid on top of the corresponding frame. The system also shows the confidence levels of recognizing the action through a dynamic bar graph as shown in Fig. 5(b). This is obtained from the relative votes for different action classes in the sliding window surrounding the frame under consideration. As shown in Fig. 5(b), the confidence level of reach object is highest while that action is performed. Since the sliding window not only depends on past frames but also future frames for action recognition, it has a delay of about 15 frames with respect to the real world. This corresponds to a delay of 1 sec as the Kinect sensor's capture rate was around 15 fps.

## Conclusion and Future Work

An action recognition system capable of recognizing retail interactions of a shopper with a shelf and its products is proposed in this work. A pose-representation that is more robust to intra-class variations and more discriminative for the actions under consideration is also proposed. The suggested approach achieves up to 90% accuracy in detecting retail actions that help in understanding the engagement of a shopper with the shelf. The proposed system is also capable of real-time action recognition, exhibiting a delay of less than 1 sec between the action performance and recognition. The work also proposes a methodology for abnormal action detection in a given ambience and the potential of crowdsourcing and building new models to detect those actions. Finally, a configurable privacy protection engine in which body parts can be selectively masked based on the skeletal data, is also enclosed. The proposed framework for retail action recognition can be extended to any other ambience like hospital or airport by modelling the corresponding actions and poses. This work is a part of Ambient Intelligence Research (AIR) which aims at inferring the states of objects, environment and humans in a given ambience. The accuracy in inferring the human actions can be further improved by contextual clues derived from a better understanding of the 3D structure of the environment and that of the interacting objects.

**Table 4. Confusion Matrix for Retail Action Recognition**

| | Shelf Examine | Reach Object | Take Object | Read Label | Place InCart |
|---|---|---|---|---|---|
| ShelfExamine | 1 | 0 | 0 | 0 | 0 |
| ReachObject | 0 | 0.875 | 0 | 0.125 | 0 |
| TakeObject | 0 | 0 | 1 | 0 | 0 |
| ReadLabel | 0.1429 | 0 | 0 | 0.8571 | 0 |
| PlaceInCart | 0.1818 | 0 | 0 | 0 | 0.8182 |

## References

[1] J.K. Aggarwal and M.S. Ryoo, "Human activity analysis: A review." *ACM Computational Survey* Vol.43, No. 3, 16.1 – 16.43, 2011.

[2] A. Bobick and J. Davis, "The recognition of human movement using temporal templates" *IEEE Trans. Pattern Analysis Machine Intelligence vol. 23*, no. 3, pp. 257–267, 2001.

[3] M. D. Rodriguez, J. Ahmed and M. Shah, "Action MACH: A spatio-temporal maximum average correlation height filter for action recognition." *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2008.

[4] Y. Sheikh, M. Sheikh, and M. Shah, "Exploring the space of a human action." *IEEE International Conference on Computer Vision (ICCV)*. Vol. 1, pp.144–149, 2005.

[5] A. Yilmaz M. Shah, "Recognizing human actions in videos acquired by un-calibrated moving cameras*",* ICCV, 2005.

[6] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features**"**. *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*. 2005.

[7] I. Laptev and T. Lindberg, "Space-time interest points" *IEEE International Conference on Computer Vision (ICCV), pp.* 432-439, 2003.

[8] A. Efros, A. Berg, G. Mori and J. Malik, "Recognizing action at a distance**"**. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Vol. 2, pp. 726–733, 2003.

DOI: 10.2352/ISSN.2470-1173.2016.3.VSTIA-514

[9] S.Park and J.K. Aggarwal "A hierarchical Bayesian network for event recognition of human actions and interactions". *Multimedia Syst. 10*, 2, 164–179. 2004.

[10] P. Natarajan and R. Nevatia, "Coupled hidden semi-Markov models for activity recognition". *IEEE Workshop on Motion and Video Computing (WMVC),* 2007.

[11] D. Zhang, D. Gatica-Perez, S. Bengio, and I. Mccown, "Modeling individual and group actions in meetings with layered HMMs". *IEEE Trans. Multimedia 8*, vol. 3, pp.509–520, 2006.

[12] S.-W, Joo and R. Chellappa,"Attribute grammar-based event recognition and anomaly detection". *IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW),* 2006.

[13] M. S. Ryoo and J.K. Aggarwal, "Recognition of composite human activities through context-free grammar based representation". In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, Los Alamitos, CA, pp. 1709–1718, 2006.

[14] Ronald Poppe, "A survey on vision-based human action recognition," Image and Vision Computing, Vol. 28, No. 6, pp. 976-990, 2010.

[15] L. Xia, C. Chen and J-K. Aggarwal, "View invariant human action recognition using histograms of 3D joints," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp.20-27, 2012.

[16] E.Frontoni, P. Raspa, A. Mancini, P. Zingaretti, and V. Placidi, "Customers' Activity Recognition in Intelligent Retail Environments," *International Workshops on New Trends in Image Analysis and Processing,* Vol. 8158. pp. 509-516, 2013.

## Author Biography

*Srenivas Varadarajan received his B.E. degree in Electronics and Communications Engineering from PSG College of Technology, Coimbatore, India in 2003 and the M.S and PhD degrees in Electrical Engineering from Arizona State University in 2009 and 2014, respectively. Currently he is working as a Research Scientist at Intel Labs in field of Image Processing and Computer Vision. He has about 10 years of Industrial experience in Image and video processing in companies including Texas Instruments, Qualcomm Technologies and Intel Corporation.*

*Shahrokh is a Senior Principal Technologist leading the Ambient Intelligence Research (AIR), a multimodal application-specific system for modeling the scene and registering events of interest and analytics to enhance efficiency and quality of workflows in various segments including healthcare, retail, transportation, enterprise and intelligence community security and safety. For the last 24 years Shahrokh has served in a variety of leadership roles including CPU architecture & design and management (P6 and WMT:2000), Principal Engineer and Technical Marketing Manager of P4P Platform Application Engineering (DPG-PAE 2005), Sr. PE of Assembly and Test Manufacturing and (ATM and ATTD(: 2009) and the Sr. PE In Software & Services focused on platform trust elevation and policy orchestration (SSG:2014). He has chaired and led number of industry initiatives including iNEMI Consumer/Portable Products, OASIS Electronic Identity Credential Trust Elevation Methods (Trust Elevation) TC, EPCglobal-GS1, and founding board member of the RAIN alliance helping with worldwide adoption of Electronic Product Code.*