Mehran University Research Journal of Engineering & Technology
Vol. 37, No. 4, 545-558 October 2018
p-ISSN: 0254-7821, e-ISSN: 2413-7219
DOI: 10.22581/muet1982.1804.09

First Person Vision for Activity Prediction Using Probabilistic Modeling

SHAHEENA NOOR*, AND VALI UDDIN*

RECEIVED ON 20.07.2017 ACCEPTED ON 13.11.2017

ABSTRACT

Identifying activities of daily living is an important area of research with applications in smart-homes and healthcare for elderly people. It is challenging due to reasons like human self-occlusion, complex natural environment and the human behavior when performing a complicated task. From psychological studies, we know that human gaze is closely linked with the thought process and we tend to "look" at the objects before acting on them. Hence, we have used the object information present in gaze images as the context and formed the basis for activity prediction.

Our system is based on HMM (Hidden Markov Models) and trained using ANN (Artificial Neural Network). We begin with extracting motion information from TPV (Third Person Vision) streams and object information from FPV (First Person Vision) cameras. The advantage of having FPV is that the object information forms the context of the scene. When context is included as input to the HMM for activity recognition, the precision increases. For testing, we used two standard datasets from TUM (Technische Universitaet Muenchen) and GTEA Gaze+ (Georgia Tech Egocentric Activities). In the first round, we trained our ANNs only with activity information and in the second round added the object information as well. We saw a significant increase in the precision (and accuracy) of predicted activities from 55.21% (respectively 85.25%) to 77.61% (respectively 93.5%). This confirmed our initial hypothesis that including the focus of attention of the actor in the form of object seen in FPV can help in predicting activities better.

Key Words: First Person Vision, Activity Recognition, Activity Prediction, Context-Aware System, Hidden Markov Models, Artificial Neural Networks.

1. INTRODUCTION

Much work is done on ADL (Activities of Daily Living) for smart homes in which their main focus is to improve the QoL (Quality of Life) for elderly people, health care system and for those people who need supervision. Despite the fact that a lot of research efforts have been directed to human activity recognition and significant results have been achieved, there are still many challenges w.r.t. to the human self-occlusion, complex natural environment and the human behavior when performing a complicated task [1]. It is

Authors E-Mail: (shanoor@ssuet.edu.pk, vali.uddin@hamdard.edu.pk)
* Department of Computer Engineering, Hamdard University, Karachi.

highly desirable for such systems to be proactive, hence it is important not only to detect and recognize the activities currently happening, but also to predict the next activity and possibly the overall scenario. Ni et. al. [2] proposeda 3-layered context aware architecture for improving recognition of ADL by using different sensors and then combining it with context ontology for the abstract level representation. In this work we use FPV to utilize the object information in gaze data to predict the next activity in an overall scenario. We exploit the fact that gaze is strongly linked to human actions and thought process and thus provides a strong cue of what is going in the mind w.r.t. goal accomplishment.

This paper is organized as follows: We begin by covering an account of existing literature in Section 2. Next, in Section 3, we lay the foundation for our activity recognition and prediction model. We present the dataset, the experiments, results and observations in section 4 and give conclusions in Section 5.

2. LITERATURE REVIEW

HAR (Human Activity Recognition) is a huge field of research with lots of constituent steps, data collection methods and learning approaches. Hence, each research area addressing HAR focuses on typically one aspect like ma-chine learning method e.g. HMM and ANN, devices used e.g. wearable and smartphones, type of sensors like depth or 2D (Two-Dimensional) images, using contextual information like objects and location etc. In the remainder of this section, we look at these different areas and cover the important and recent research in each.

MLS (Machine Learning Solutions): HAR is essentially a machine learning problem [3] with a goal of identifying human activities in a real setup [4], comprising of feature extraction followed by generating and training a model for classification. A number of machine learning solutions have been explored like Naive Bayes [5], SVM (Support Vector Machines) [6], Decision Trees [7] and HMM [1,3]. Also, much literature can be found on a comparison on these approaches e.g. [9-10]. A number of variations of HMM have been used for activity classification.San-Segundo et. al. [3] used inertial signals from smart phone to train an HMM to identify activities like walking, walking upstairs, walking downstairs, sitting, standing and lying down. However, such an approach is not applicable in identifying finer motion activities like performing a cooking task, where most actions are differentiated by hand motions and do not depend on full body. Till recently, HMMs were widely used for HAR because they offer dynamic time warping, have clear semantics and are robust i.e. can be trained on one person and tested on another. With the uprise of deep learning networks [11], attentionhas been shifted back to the use of ANN for classification in general and activity recognition in particular [12].

Data Collection: In [13], the authors have applied sensors on the subjects' bodies and gathered a number of motion signals using accelerometers. Next, they reduced the system complexity by projecting the signals to a lower dimensional space followed by action classification and subject identification. They have considered USC Human Activity Dataset [14] with 12 activities e.g. Walking Forward, Walking Left, Walking Right, Walking Upstairs, Walking Downstairs, Running Forward, Jumping Up, Sitting, Standing, Sleeping, Elevator Up and Elevator Down. The authors have also provided a detailed review of state of the art research in the field of sensor-based activity recognition and compared the techniques presented and performance achieved.

The widespread availability of depth cameras and images has facilitated a variety of object and activity recognition tasks [15]. In [16], the authors proposed a human action recognition algorithm exploiting the skeleton provided by Microsoft Kinect and discussed its application to AAL (Active and Assisted Living) scenarios. They begin with a skeleton model and compute posture features. Then they identified the most informative postures and form a feature vector using them. This vector is used to train a multiclass SVM for activity identification. Another attempt to using depth information for activity recognition is by Kamal et. al. [1], where they extracted spatial depth shape features and temporal joints features and used a Modified HMM (M-HMM) for activity classification. Similarly, Kumar et. al. [17] used kinect to extract a 3D skeleton and after dimensionality reduction with PCA (Principal Component Analysis), trained an ANN for recognizing activities. In another work Mo et. al. [18] also used skeleton information from kinect sensor and trained a deep learning model for activity recognition.

As smart devices like smartphones and watches etc. are becoming more popular, researchers are exploring them to not only collect more data but also to propose simple and light-weight solutions to activity recognition. For example, a deep convolution neural network (convnet) is proposed [19] to perform activity recognition using smartphone sensors by exploiting the inherent characteristics of activities and 1D time-series signals for activities like walking (straight, upstairs and downstairs), sitting, standing and laying. In another work, Sefen et. al. [20] used sensor data of smartphones and smart-watches to recognize in realtime, human activities like walking, jogging, stair climbing, rope jumping, pushups and crunches.

Context-Awareness: According to Ronao et. al. [19] human activitiesare inherently hierarchical in structures and are thus very prone to small variations at the input level. An earlier work by Kim et. al. [4] involved composite activities where they showed that each higher level activity is actually a combination of lower level "poses", which helps in overall activity data pattern discovery.

Most of the work in the literature [21-22] focuses on using devices (wearable or mounted) to extract motion information of humans for activity recognition. On the other hand, attention or context information, which corresponds to the objects in the scene can greatly enhance the quality of activity recognition. In fact, in this paper, we have shown that the contextual information cannot only be used to detect the current activity, but also predict the next one. In addition, we have shown that it can be used to identify the scenarios as a whole.

Some work has been done using contextual information combined with motion information for activity recognition [23-24]. Identifying context is in itself hard because it needs complex recognition systems and/or manual annotation. Hasan et. al. [25] formulated a continuous learning framework using CRF (Conditional Random Field) model and identified the most informative query instances using the system entropy for manual labeling.

Attention plays an important role in scenario recognition and noting gaze is widely used as a reliable way to identify attention. Neuroscience studies have shown that incorporating gaze view with third person perspective positively influences human activity and behavior recognition [26]. Attention identification and in turn scenario recognition is important for a number of applications, Das et. al. [27] proposed a system for robot to initiate interaction with a human depending on his scenario and level of attention. In their system, the robot observes the human gaze patterns to detect the interest of the human in a specific activity and starts communication if necessary. The features are extracted from gaze images and then used to train a mutliclass SVM.

3. MODEL GENERATION FOR ACTIVITY PREDICTION

In this section, we form the mathematical foundation of our model. Any activity recognition system begins with extracting features from the input data (set of images or video stream). These features are then used to build the model and train the classifier. We have used SIFT (Scale-Invariant Feature Transform) [28-29] as feature because it is invariant to scale and robust to changes in viewpoint, rotation and illumination. We begin with presenting a brief account of feature extraction using SIFT. Next, we describe our model for context-aware activity recognition. Our hypothesis is that human gaze is closely linked with the thought process and humans tend to "look" at the objects before acting on them. Hence, knowledge of the object-seen increases the recognition precision of the activity in progress. We have used the object information present in gaze images as the context and formed the basis for activity prediction: we take the above hypothesis to the next level and show that using the contextual information i.e. object-focused-in-gaze can be combined with information of current activity to actually predict the next one. We conducted experiments and showed that this prediction of next activity can even be done reasonably well using object information only.

Fig. 1 which gives an overview to the two-fold activity prediction system using first person vision information. In the following subsections, we briefly touch upon our feature extraction and then give details of our activity recognition and prediction models.

3.1 Feature Extraction

The first step to activity recognition is to identify segments in the image stream causing the action and extract the representative features. Some earlier work [30-32] proposed to use raw pixels for an effective initial feature representation for learning. However, this requires the overall number of pixels to be small, which is rarely the case in current imaging scenarios. e.g. in our dataset, each image is 1920x1080 and the activity segments range between 10 and 650 frames. Including the multi-camera views, the overall number of pixels become over whelmingly large and hence computationally expensive. We adopted the alternate approach to extract representative features for compact and effective activity-segment representation. A number of such approaches exist e.g. STIP (Space-Time Interest Points) [33] and Spatial pyramid [34] etc. each based on a primitive feature detector e.g. SIFT [28], Harris corner detector [35], or HOG (Histogram of Oriented Gradients) [36] computed over time. We decided to use SIFT for our activity-segment representation, because it offers a number of advantages inour scenario e.g. scale, rotation and view point in variance and robustness to illumination changes.

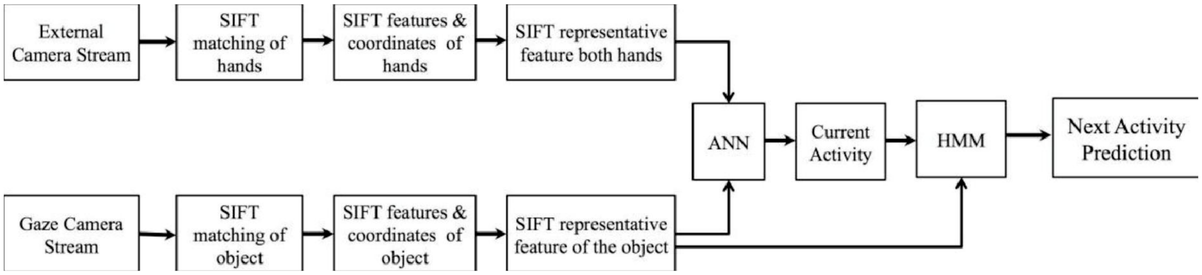

External Camera Stream
SIFT matching of hands
SIFT features & coordinates of hands
SIFT representative feature both hands
ANN
Current Activity
HMM
Next Activity Prediction
Gaze Camera Stream
SIFT matching of object
SIFT features & coordinates of object
SIFT representative feature of the object

FIG. 1. OVERVIEW OF ACTIVITY PREDICTION SYSTEM

We begin with extracting SIFT features of the start (f_s) and end (f_e) frames of the activity segment and identify the left (H_l) and right (H_r) hands in each. Let SIFT = <$SIFT_1$, $SIFT_2$, …,$SIFT_n$> be the extracted SIFT features with LOC = <loc_1, loc_2, … loc_n> the corresponding locations of each SIFT point. We compute the representative SIFT point and its location given by S = Avg (SIFT) and L = Avg (LOC), where S is 128D feature vector and L:(x,y). The goal is to reduce the size of the feature vector without losing the distinctiveness. For the external cameras, this is computed per hand per frame and thus each activity segment is denoted by $AS_{fs}^{fe} =< S_{Hl}, L_{Hl}, S_{Hr}, L_{Hr} >$. Similarly the object is extractedfrom gaze-directed camera as O_f i.e. Object(s) seen in frame f. AS_{fs}^{fe} and O_f are then fed to the neural network the for prediction.

3.2 Training the System and Activity Model Generation

In order to learn features for activity recognition, we employed supervised learning and used an ANN. ANNs are inspired by human brains and thus adopt the terminology and high-level structure of their biological counterpart. A general ANN model comprises of one input layer with an input feature vector length of n, one output layer with m output variables and one hidden layer with k neurons. A generic back propagation learning algorithm is given in Algorithm-1 [37].

ALGORITHM-1. ALGORITHM FOR NEURAL NETWORK

```
Class Training
Initialize W = [$W^1$, $W^2$, $b^1$, $b^2$]
Repeat for i = 1 : m
Perform feedforward pass:
Compute $\hat{x}^i$.
Perform backpropagation:
Compute gradients: $\Delta_W J_a(W)$.
Compute weight change: $\Delta W$.
Update weight W.
Class Feature Encoding
Compute: $\tilde{x}^i$ = f(W $^i x^i$ + $b^1$)
```

3.3 Using Gaze and Current Activity for Next Activity Prediction

As per psychology dictionary [38], goal-directed behavior implies behavior oriented towards achieving a particular goal. This implies that given an end-goal, humans follow a set of activities to achieve that goal. While the overall sequence may be long and unpredictable, at any given state there exists a finite set of possibilities for the next state. This can very well be captured by HMM, which forms the activity model by observing the effects of an activity. HMM is a generative probabilistic model used for generating hidden states from observable data [39]. Mathematically, the goal is to determine the sequence of hidden state (y_1, y_2, … y_t) corresponding to the sequence of observed outputs (x_1, x_2, … x_t).

For tractable inference, HMM needs two independence assumptions [39]:

The 1st order Markov assumption of transition as given in Equation (1) which states that the next state depends only on the current state, not on past states. In other words, the hidden variable at time t, y_t, depends only on the previous hidden variable y_{t-1}.

$$P\ (y_t|y_1,\ y_2,\ y_{3\ldots,}\ y_{t-1}) = P\ (y_t|y_{t-1}) \qquad (1)$$

Conditional independence of observation parameters as indicated in Equation (2) i.e. the observable variable at time t, x_t, depends only on the current hidden state y_t. In other words, the probability of observing x while in hidden state y is independent of all other observable variables and past states.

$$P\ (x_t|y_t,\ x_1,\ x_{2\ldots,}\ x_{t-1},\ y_1,\ y_2,\ y_{3\ldots,}\ y_{t-1}) = P\ (x_t|y_t) \qquad (2)$$

To find the most probable hidden state sequence from an observed output sequence, HMM finds a state sequence which maximizes a joint probability p(x,y) of the transition probability and the observation probability (that is the probability that outcome x_t is observed in state y_t [39] as shown in Equation (3).

$$Tp(x,y) = \prod_{t=1}^{T} p(y_t \mid y_{t-1})\, p(x_t \mid y_t) \tag{3}$$

When an HMM is used for activity recognition, activities are the hidden states and observable output is sensor data, which is hand and objectinformation in our case. Consider Fig. 2, that shows the state transition probabilities for activity recognition for sandwich making scenario.

Given A_N: $a_1, a_2, \ldots a_n$ Set of next possible activities and A_C Current Activity: $A_N \alpha A_C$, or in other words represented as Equation (4):

$$A_N = kA_C \tag{4}$$

where k is the proportionality constant.

As a function of time, the same may be given by Equation (5):

$$A(t+1) = kA(t) \tag{5}$$

Next, considering O: $O_1, O_2, \ldots O_n$ Set of objects currently seen , A_N can be enriched as A_N^O to indicate the next activity given the current object information. Thus, given: $A_N \equiv P(A_N) = kP(A_C)$ and $A_N^O \equiv P() = kP\ (A_C|O)$; we show that Equation (6) holds true.

$$S(A(t+1|O))\ \ S(A(t+1)) \tag{6}$$

where S(.) is the probability score of next activity.

The state transition model for the sandwich making scenario of Fig. 2 has been redrawn in Fig. 3 with object information included.

The training of ANN is depicted in Algorithm-2. We begin with extracting hand motion information and computing current activities from external cameras. We also extract the object-seen information from gaze-directed camera. Next, we train two neural networks independently with and without object information and compare the results for prediction of next activity.

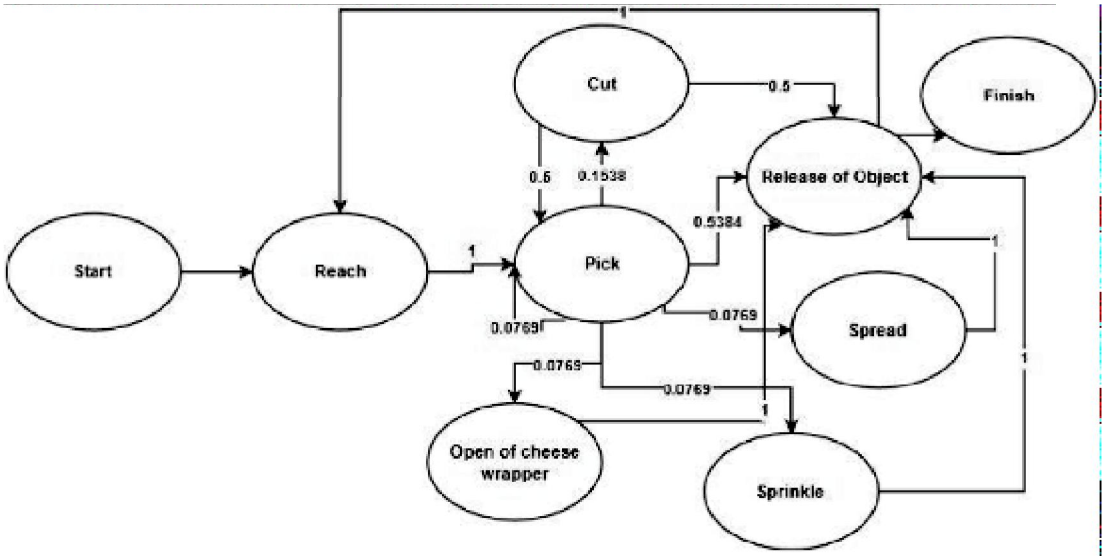


FIG. 2. STATE TRANSITION GRAPH FOR ACTIVITIES IN SANDWICH MAKING

4. EXPERIMENT AND RESULTS

In this section, we describe the experiments conducted and the results obtained. We begin by explaining the datasets and then talk about the method of analysis. The details are given below:

4.1 Data Set

We considered two independent datasets to conduct experiments for our work: TUM Kitchen dataset available at: http://web.ics.ei.tum.de/~karinne/Dataset/dataset.html) and GTEA Gaze+ dataset available at:(http://ai.stanford.edu/~alireza/GTEA_Gaze_Website/GTEA_Gaze+.html)

These datasets are independent datasets captured at the two universities and are used for activity recognition and other perception scenarios including first person vision. The details are given below:

TUM Kitchen Dataset: The data is captured from five different cameras including first and third person perspectives. The (stationary) external and (mobile) gaze-directed cameras are placed at different locations having different views of the scene, we need to combine the information from these cameras so as to allow inter-camera information exchange. Frame rate of gaze-directed camera is 25FPS and that of external cameras is 60FPS. Both have been temporally aligned by selecting the appropriate frames from each stream and stitching them together. The experiments comprise of a number of kitchen activities. The dataset comprises of pancake and sandwich making activities with a single actor repeating 10 sets for pancake

ALGORITHM-2. ALGORITHM FOR IDENTIFYING NEXT ACTIVITY

```
Input: $A_e$: $a_{e1}$, $a_{e2}$, ... $a_{en}$ ≡ activities from external camera
$O_G$: $o_{g1}$, $o_{g2}$, ...$o_{gn}$ ≡ set of objects seen in gaze (Given only in second round)
Output: {A|P}: Possible next activity with probability
Class ANN Modeling
Generic ANN Training Algorithm
```

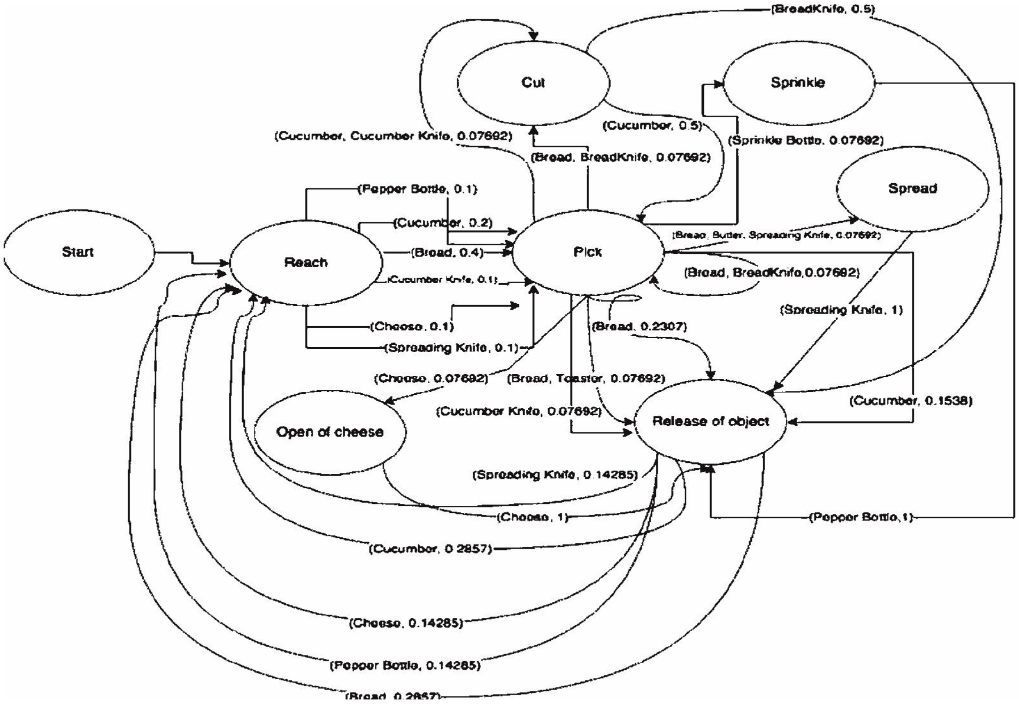


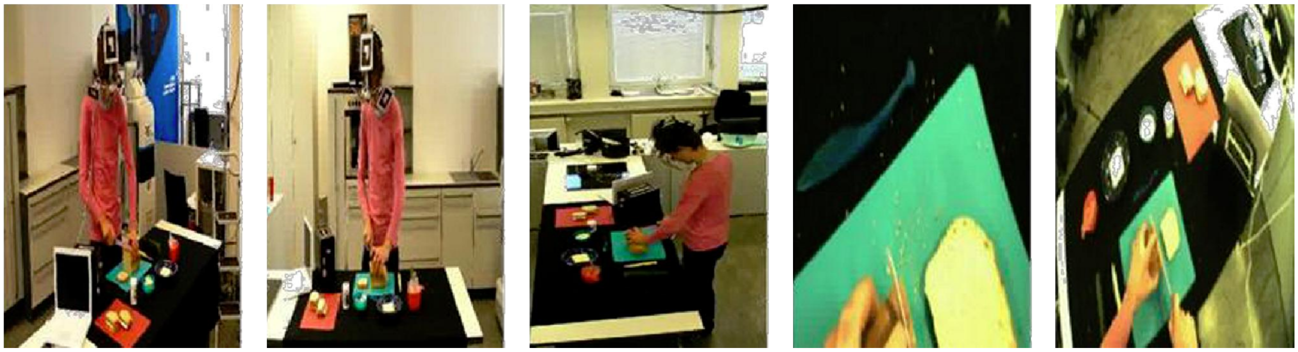
FIG. 3. STATE TRANSITION DIAGRAM FOR ACTIVITIES OF SANDWICH MAKING SCENARIO WITH OBJECT INFORMATION

making and 10 subjects with 18 iterations each for sandwich making. Each set lasts for approximately 3 minutes which implies a video of 570 minutes. Fig. 4(a-e) shows the images from different viewpoints of the TUM dataset.

Gaze+ Dataset: GTEA Gaze+ dataset is collected by SMI eye-tracking glasses [8]. In this dataset 7 meal-preparation activities were considered, each performed by 10 subjects. The subjects were given cooking recipes for American Breakfast, Pizza, Snack, Greek Salad, Pasta Salad, Turkey Sandwich and Cheese Burger. High Definition videos were recorded at 24 frames per second. Fig. 5(a-d) shows a subset of images from the GTEA Gaze+ dataset.

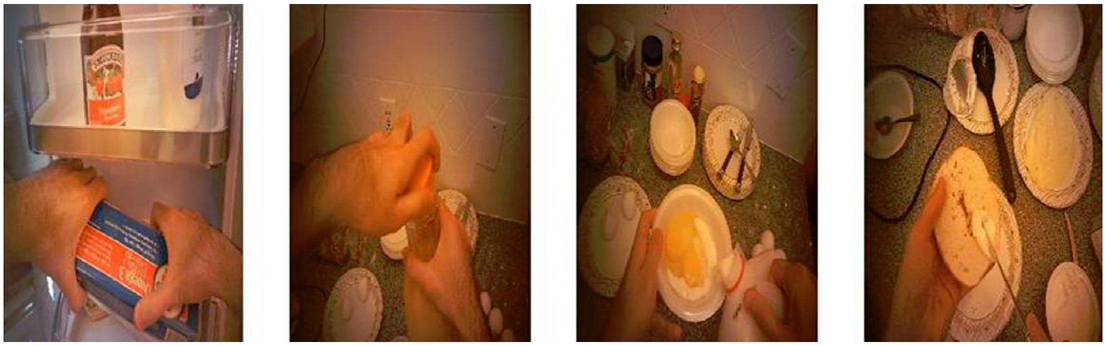
4.2 Method of Analysis

In this work, we considered the activities encountered in day-to-day life in cooking scenarios like reach, pick, release, cut, spread, sprinkle, open of cheese, open/close of lid, and pour. We begin with identifying motion segments in the TPV images using optical flow, followed by identifying left and right hands and their coordinates using SIFT. The subsequent steps depend on this information because the NN is trained using this hand information. First, we extract SIFT features of desired frames alongwith those in hand model. Next, we match the features using nearest neighbor and identify the left and right hands individually. SIFT being a dense descriptor returns multiple features for each hand so we compute a representative feature for each hand and identify its coordinates. This representative feature is used in subsequent steps for training the neural network in the first pass to recognize the current activity in progress. Next, we used the recognition results from the first step and combine with object information from internal camera (i.e. FPV) to predict the next activity and their



(a) LEFT (b) FRONT (c) RIGHT (d) GAZE (e)HEAD-MOUNTED

FIG. 4. TUM DATASET - DIFFERENT VIEWPOINTS FOR CUTTING A BREAD IN SANDWICH MAKING SCENARIO



(a) PICK (b) OPEN/CLOSE (c) POUR (d) SPREAD

FIG. 5. GTEA GAZE+DATASET

probabilities. Combining these two pieces of information improves the precision for predicting the next activity. For this we trained ANNs with and without object information and compared the recognition precision and accuracy.

4.3 Results and Observations

4.3.1 Model for Current Activity Recognition

The motion information helps us in identifying the start and end frames for the various activities. The hand coordinates from these frames are used to train the network. We used the Easy NN [8] to build our ANN, which is a fast and simple data analytics tool. Fig. 6 shows a sample dataset in the tool.

We trained the network multiple times with a varying number of hidden layers (0-3) and found that adding hidden neurons mostly decreases the precision and takes longer for training to complete. These findings are also published in our earlier work in [12]. Fig. 7 shows the generated network for TUM dataset. The final network in this case is a 3 layer network with input layer having 17 neurons each node representing one input feature, output layer with 8 neurons corresponding to the output categories and one hidden layer. A close-up of one each of the input and output neurons is shown in Fig. 8(a-c). The connections between the nodes are weighted. The color denotes the polarity i.e. green for +ve and red for -ve values. The thickness represents weight magnitude.

It took 73 cycles for the network to complete the learning for TUM dataset. According to our criteria, network training continues as long as average error is greater than the threshold 0.01. At the completion of training cycle, our maximum error is 0.0414, minimum error is 0.0004 and average error is 0.0099. We trained our network with a learning rate of 0.7 and momentum of 0.8.

The network was regenerated and trained with object information extracted from FPV. As expected, including the object information improved the recognition rate

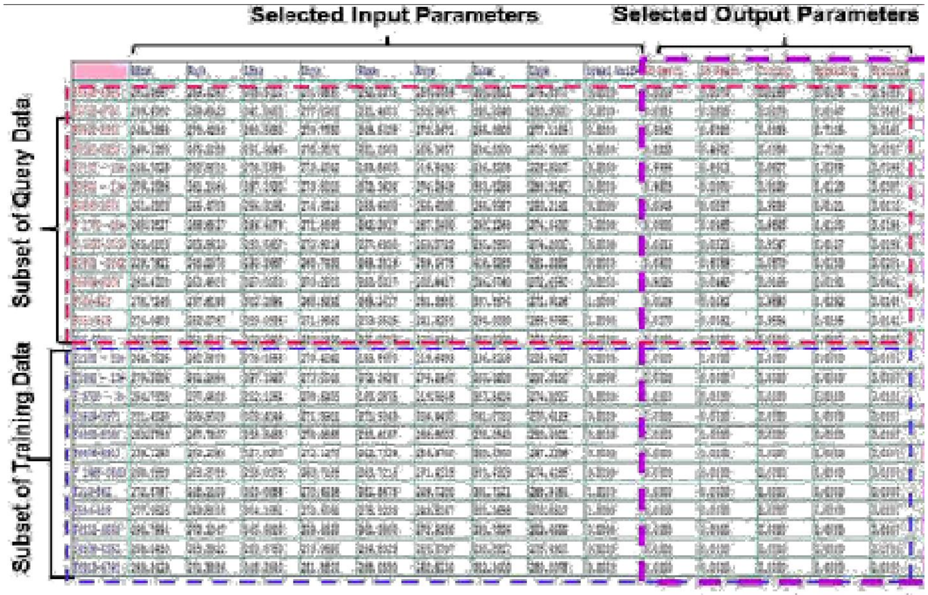


FIG. 6. SAMPLE PARAMETERS & DATASET FOR ANN LEARNING

significantly - 95.38% against 90.77%. This confirms our initial hypothesis that the gaze precedes the action and including the focus of attention of the actor can help in better action recognition.

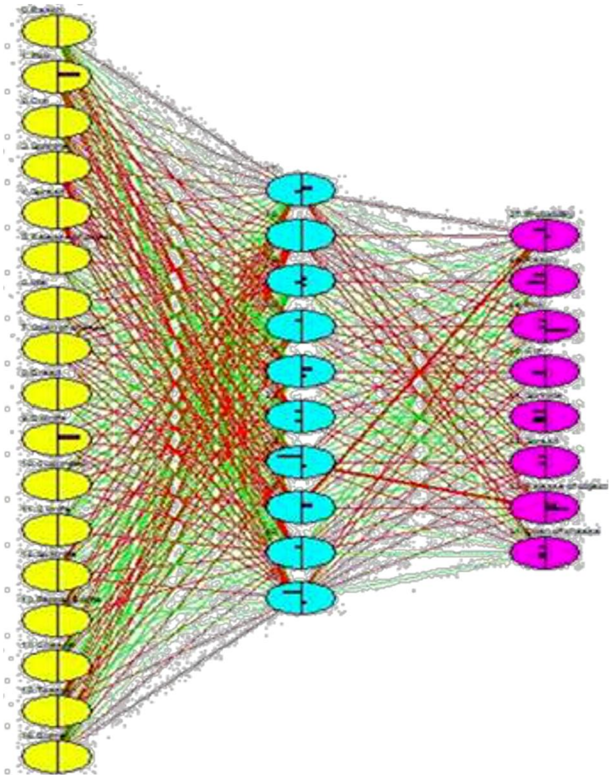
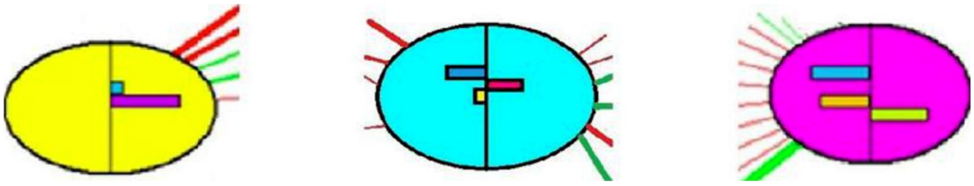


FIG. 7. NEURAL NETWORK FOR ACTIVITY RECOGNITION WITH OBJECTS FROM FPV

4.3.2 Next Activity Prediction

For prediction, we ran experiments on the two standard datasets (TUM and GTEA Gaze+) described above. The recognition results are detailed in **Table 1**. We show that the precision and accuracy for predicting next activity is higher when objects are considered. This is true for both TUM dataset where precision (respectively accuracy) increased to 82.5% (respectively 96.1%) from 57.1% (respectively 86.7%) and GTEA+ where precision (respectively accuracy) increased to 72.7% (respectively 90.1%) from 53.8% (respectively 83.8%). This confirms our initial hypothesis that the gaze precedes the action and including the focus of attention of the actor can help in predicting better action recognition. It is also important to note that the learning for ANN takes too long in case of training-without-objects > 707383 cycles, while it completes in a reasonable time when training-with-objects 148 cycles). We repeated the experiments using objects-observed only and were able to predict the next activity with an accuracy of 71.4%. The advantage of this approach is that it avoids all the setup costs and efforts of externally mounted cameras.

Fig. 9(a-e) shows the confusion matrices for predicting different activities with and without objects in focus.

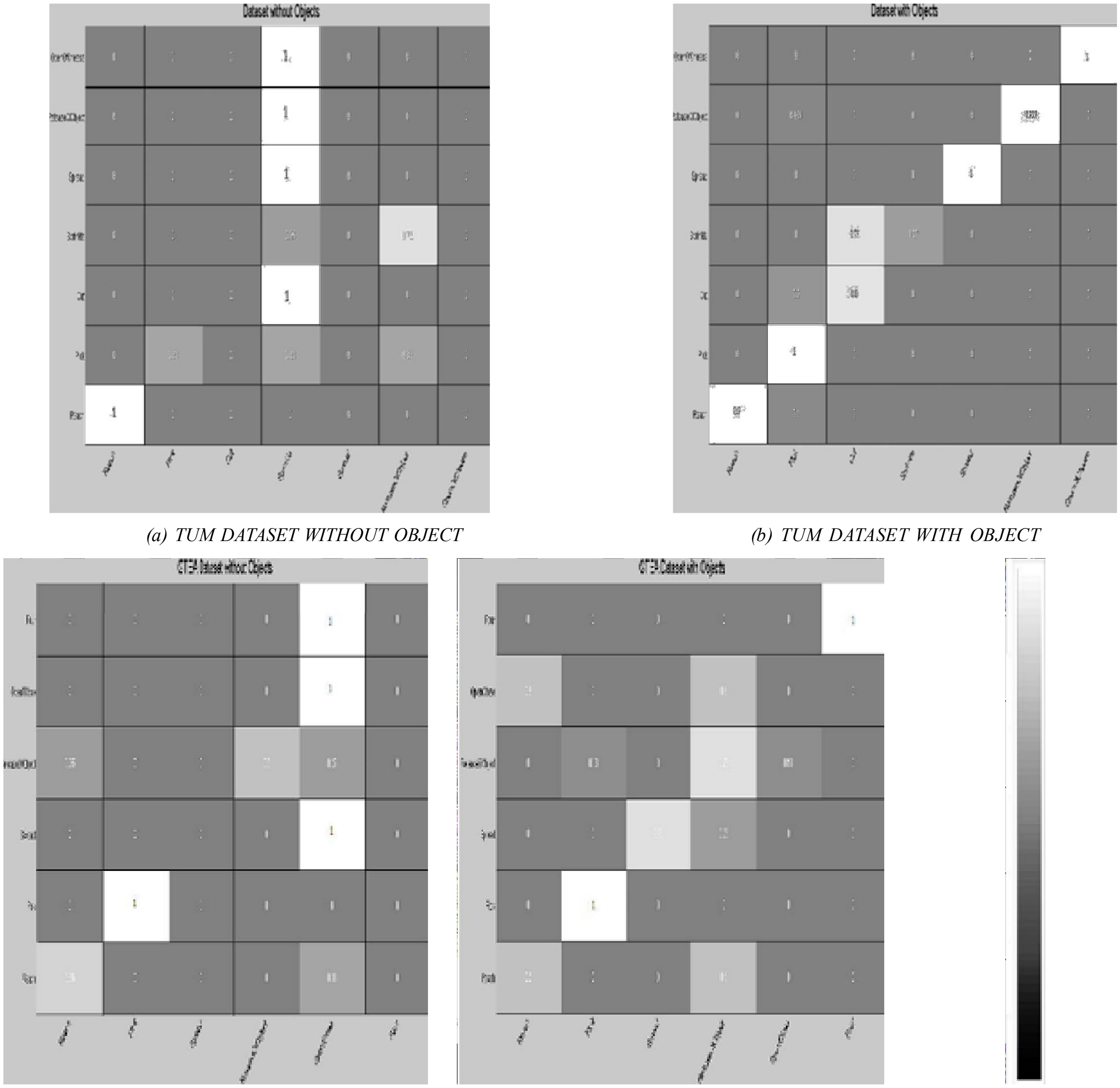


(a) INPUT NEURON (b) HIDDEN NEURON (c) OUTPUT NEURON

FIG. 8. INTERNAL PARAMETERS OF NEURONS

TABLE 1. ACTIVITY RECOGNITION RESULTS

Datasets	Without Object				With Object			
	Precision	Accuracy	Recall	TNR	Precision	Accuracy	Recall	TNR
TUM	0.571	0.867	0.533	0.927	0.825	0.961	0.891	0.972
GTEA Gaze+	0.5384	0.838	0.583	0.892	0.7272	0.909	0.727	0.945



(a) TUM DATASET WITHOUT OBJECT

(b) TUM DATASET WITH OBJECT

(c) GTEA GAZE+DATASET WITHOUT OBJECT

(d) GTEA GAZE+DATASET WITH OBJECT

(e)COLORBAR

FIG. 9. CONFUSION MATRICES FOR ACTIVITY PREDICTION: FIG. (A-B) SHOW THE TUM DATASET WITH AND WITHOUT CONSIDERING OBJECT INFORMATION. FIG. (C-D) SHOW THE SAME FOR GTEA GAZE+DATASET

5. CONCLUSION

In this paper, we used the context information for a complete activity prediction system. This includes identifying objects in FPV streams, generating an activity recognition model by training an ANN on hand-motion information from TPV images and combining the object information with activity information for improved prediction precision and accuracy. We used standard cooking datasets of TUM and GTEA Gaze+ to identify activities of reaching, picking, sprinkling, spreading, opening/ closing and cutting etc. for a variety of objects like bread, knives, pepper, cheese, cereal, milk etc. and a number of scenarios like sandwich making, omelette, eating cereal etc.

For activity prediction, in the first round, we extracted the current activities and trained our network. Next, we included the object information received from FPV images and re-trained the network. The precision (and accuracy) of predicted activities increased from 55.21% (respectively 85.25%) to 77.61% (respectively 93.5%). Hence, we showed that including the focus of attention of the actor in the form of object seen in FPV can help in predicting activities better. As a further stretch, we were able to predict the next activity using only the objects observed and achieved an accuracy of 71.4%. This is pretty decent given the fact that it avoids all the setup costs and efforts of external cameras.

The applications of this work are numerous. The findings of this work can be used to improve the human activity recognition in a number of areas like robotic navigation in homes and industries, smart homes, health-care systems etc. It can also be used for image-based content retrieval and indexing of huge image datasets.

ACKNOWLEDGEMENT

Part of this research was funded by a grant from the DAAD under German-Pakistani Research Cooperation Program.

REFERENCES

[1] Shaharyar, K., Ahmad, J., and Kim, D., "Depth Images-Based Human Detection, Tracking and Activity Recognition Using Spatiotemporal Features and Modified HMM", Journal of Electrical Engineering & Technology, Volume 11, No. 3, pp. 1921-1926, 2016.

[2] Ni, Q., Hernando, A.B.G., and Cruz, I.P., "A Context-Aware System Infrastructure for Monitoring Activities of Daily Living in Smart Home", Journal of Sensors, 2016.

[3] Segundo, R.S., Montero, J.M., Pimentel, J.M., and Pardo, J.M., "HMM Adaptation for Improving a Human Activity Recognition System", Algorithms, Volume 9, No. 3, 2016.

[4] Kim, E., Helal, S., and Cook, D., "Human Activity Recognition and Pattern Discovery", IEEE Pervasive Computing, Volume 9, No. 1, pp. 48-53, January, 2010.

[5] Jatoba, L.C., Grossmann, U., Kunze, C., Ottenbacher, J., and Stork, W., "Context-Aware Mobile Health Monitoring: Evaluation of Different Pattern Recognition Methods for Classification of Physical Activity", 30th IEEE Annual International Conference on Engineering in Medicine and Biology Society, 2008.

[6] Anguita, D., Ghio, A., Oneto, L., Parra, X., and Reyes-Ortiz, J.L., "Energy Efficient Smartphone-Based Activity Recognition Using Fixed-Point Arithmetic", Journal of University Computer Science, 2013.

[7] Maurer, U., Smailagic, A., Siewiorek, D., and Deisher, M., "Activity Recognition and Monitoring Using Multiple Sensors on Different Body Positions", Proceedings of International Workshop on Wearable and Implantable Body Sensor Networks, 2006.

[8] http://easynn.com/ ((Last Visit: 27 June 2017)

[9] Yang, J., "Toward Physical Activity Diary: Motion Recognition Using Simple Acceleration Features with Mobile Phones", Proceedings of 1st ACM International Workshop on Interactive Multimedia for Consumer Electronic, 2009.

[10] Kwapisz, J.R., Weiss, G.M., and Moore, S.A., "Activity Recognition Using Cell Phone Accelerometers", SIGKDD Explore News Letters, Volume 12, No. 2, pp. 74-82, March, 2011 (Last Visit: 15 June 2017). [Online]. Available: https://en: wikipedia.orgwiki/Deeplearning

[11] Noor, S., and Uddin, V., "Using ANN for Multi-View Activity Recognition in Indoor Environment", International Conference on Frontiers of Information Technology, pp. 258-263, December, 2016.

[12] DamaeviIius, R., Vasiljevas, M., AlkeviIius, J., and Wofniak, M., "Human Activity Recognition in AAL Environments Using Random Projections", Computational and Mathematical Methods in Medicine, pp. 17, 2016.

[13] Zhang, M., and Sawchuk, A.A., "USC-HAD: A Daily Activity Dataset for Ubiquitous Activity Recognition Using Wearable Sensors", Proceedings of ACM Conference on Ubiquitous Computing, pp. 1036-1043, New York, USA, 2012.

[14] Yang, X., and Tian, Y., "Super Normal Vector for Human Activity Recognition with Depth Cameras", IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 39, No. 5, pp. 1028-1039, May, 2017.

[15] Cippitelli, E., Gasparrini, S., Gambi, E., and Spinsante, S., "A Human Activity Recognition System Using Skeleton Data from RGBD Sensors", Computational Intelligence and Neuroscience, pp. 14, 2016.

[16] Kumar, A., Singh, S.K., and Kala, R., "Human Activity Recognition in Real-Times Environments Using Skeleton Joints", International Journal of Interactive Multimedia and Artificial Intelligence, Volume 3, No. 7, pp. 61-69, 2016.

[17] Mo, L., Li, F., Zhu, Y., and Huang, A., "Human Physical Activity Recognition Based on Computer Vision with Deep Learning Model", Proceedings of IEEE International Conference on Instrumentation and Measurement Technology, pp. 1-6, May, 2016.

[18] Ronao, C.A., and Cho, S.B., "Human Activity Recognition with Smart-Phone Sensors Using Deep Learning Neural Networks", Expert Systems Applications, Volume 59, pp. 235-244, 2016.

[19] Dengel, A., Sefen, B., Baumbach, S., and Abdennadher, S., "Human Activity Recognition Using Sensor Data of Smartphones and Smartwatches", Proceedings of 8thInternational Conference on Agents and Artificial Intelligence, pp. 488-493, February 26-28, 2016.

[20] Wang, L., "Recognition of Human Activities Using Continuous Autoencoders with Wearable Sensors", Sensors, Volume 16, No. 2, 2016.

[21] Ponce, H., Villaseor, M., and Pechun, L.M., "A Novel Wearable Sensor-Based Human Activity Recognition Approach Using Artificial Hydrocarbon Networks", Sensors, Volume 16, pp. 1033, 2016.

[22] Yordanova, K., Krger, F., and Kirste, T., "Context Aware Approach for Activity Recognition Based on Precondition-Effect Rules", Proceedings of IEEE International Conference on Pervasive Computing and Communications Workshops, pp. 602-607, March, 2012.

[23] Zhu, Y., Nayak, N.M., and Chowdhury, A.K.R., "Context-Aware Modeling and Recognition of Activities in Video", Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2491-2498, June, 2013.

[24] Hasan, M., and Chowdhury, A.K.R., "Context Aware Active Learning of Activity Recognition Models", Proceedings of IEEE International Conference on Computer Vision, pp. 4543-4551, 2015.

[25] Amaro, K.R., Minhas, H.N., Zehetleitner, M., Beetz, M., and Cheng, G., "Added Value of Gaze-Exploiting Semantic Representation to Allow Robots Inferring Human Behaviors", ACM Transactions on Interactive Intelligent Systems, 2017.

[26] Das, D., Rashed, M.G., Kobayashi, Y., and Kuno, Y., "Supporting Human Robot Interaction Based on the Level of Visual Focus of Attention", IEEE Transactions on Human-Machine Systems, Volume 45, No. 6, pp. 664-675, 2015.

[27] Lowe, D.G., "Object Recognition from Local Scale-Invariant Features", Proceedings of International Conference on Computer Vision, Volume 2, pp. 1150, 1999.

[28] Lowe, D.G., "Distinctive Image Features from Scale Invariant Keypoints", International Journal of Computer Vision, Volume 60, No. 2, pp. 91-110, 2004.

[29] Hinton, G.E., "Learning Multiple Layers of Representation", Trends in Cognitive Sciences, Volume 11, No. 10, pp. 428-434, 2007.

[30] Krizhevsky, A., Ilya, S., and Hinton, G.E., "Imagenet Classification with Deep Convolutional Neural Networks", Advances in Neural Information Processing Systems, Volume 25, Pereira, F., Burges, C.J.C., Bottou, L., and Weinberger, K.Q., (Editors), pp. 1097-1105, Curran Associates, Inc., 2012.

[31] Taylor, G.W., Fergus, R., LeCun, Y., and Bregler, C., “Convolutional Learning of Spatio-Temporal Features”, Proceedings of 11th European Conference on Computer Vision, Part-VI, Heraklion, Crete, Greece, September 5-11, 2010.

[32] Laptev, I., “On Space-Time Interest Points”, International Journal of Computer Vision, Volume 64, No. 2, pp. 107-123, September, 2005.

[33] Yang, J., Yu, K., Gong, Y., and Huang, T., “Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification”, IEEE Conference on Computer Vision and Pattern Recognition, pp. 1794-1801, June 20-25, 2009.

[34] Harris, C., and Stephens, M., “A Combined Corner and Edge Detector”, Proceedings of 4th Conference on Alvey Vision, pp. 147-151, 1988.

[35] Dalal, N., and Triggs, B., “Histograms of Oriented Gradients for Human Detection”, IEEE Conference on Computer Vision and Pattern Recognition, Computer Society, Volume 1, pp. 886-893, June, 2005.

[36] Mahmudul, H., Chowdhury, R., and Amit, K., “Continuous Learning of Human Activity Models Using Deep Nets”, Proceedings of 13th European Conference on Computer Vision, Part-III, Zurich, Switzerland, September 6-12, 2014.

[37] “Psychology Dictionary”, (Last Visit: 2 June 2017), [Online]. Available: http://psychologydictionary.org/goal-directed-behavior/

[38] Sutton, C., and McCallum, A., “An Introduction to Conditional Random Fields for Relational Learning”, Getoor, L., and Taskar, B., Editors, Introduction to Statistical Relational Learning, MIT Press, 2006.

[39] https: //www.smivision.com/eye-tracking/product/eye-tracking-glasses/ (Last Visit: 2 May 2017).