Journal of Intelligent & Robotic Systems (2019) 93:745–761
https://doi.org/10.1007/s10846-018-0815-7

Predicting Human Actions Taking into Account Object Affordances

Vibekananda Dutta[1] · **Teresa Zielinska**[1]

Received: 18 November 2017 / Accepted: 14 March 2018 / Published online: 4 April 2018

Abstract

Anticipating human intentional actions is essential for many applications involving service robots and social robots. Nowadays assisting robots must do reasoning beyond the present with predicting future actions. It is difficult due to its non-Markovian property and the rich contextual information. This task requires the subtle details inherent in human movements that may imply a future action. This paper presents a probabilistic method for action prediction in human-object interactions. The key idea of our approach is the description of the so-called object affordance, the concept which allows us to deliver a trajectory visualizing a possible future action. Extensive experiments were conducted to show the effectiveness of our method in action prediction. For evaluation we applied a new RGB-D activity video dataset recorded by the Sez3D depth sensors. The dataset contains several human activities composed out of different actions.

Keywords Intention recognition · Human-object relation · Object affordance · Action prediction · Feature extraction · Probability distribution

1 Introduction

In everyday life a human performs various actions. Being able to detect and anticipate which action is going to be performed in a complex environment is important for assistive robots, social robots and healthcare assistants. Such ability requires reasoning tools and methods.

With such capability [20], a robot is able to plan ahead with reactive responses together with avoiding potential accidents. When a partial observation is available, we should be able to predict what is going to happen next (e.g., a person is about to open the door as shown in the Fig. 1).

Predictive models are also useful in detecting abnormal actions in surveillance videos with alerting emergency responders [38]. It is necessary that a reliable prediction is done at the early stage of an action, e.g., when only 60% of a whole action was observed.

Recent research focuses on actions recognition problem [16, 24, 32]. Although few recent works addressed the problem of ongoing activity recognition with partial information avilable [31, 36], they do not answer how to perform activity prediction. A reliable action prediction relies on selecting and processing the crucial information, e.g., scene context, object properties (affordance, object texture) and relative human-object posture. The action prediction has two features:

- anticipating human actions requires identifying the subtle details inherent in human movements that would lead to a future action,
- the action prediction problem must be carried out with the focusing on temporal human interactions with the environment (e.g., interaction with the objects or with the other people).

In this work, we discuss the problem of action prediction in natural scenarios using collection of examples of human actions in the real world sampled by video records (WUT-ZTMiR[1] dataset, CAD-60[2] dataset). We investigate how the user behaviors evolves dynamically in a short time. Our goal

The preliminary version of the paper presented during "Intentional workshop on Robot Motion Control (RoMoCo)", 2017, Poland.

✉ Vibekananda Dutta
vibek@meil.pw.edu.pl

Teresa Zielinska
teresaz@meil.pw.edu.pl

[1] Institute of Aeronautics and Applied Mechanics, Warsaw University of Technology, ul. Nowowiejska 24, 00-665 Warsaw, Poland

[1] Warsaw University of Technology, Division of Theory of Machines and Robots. https://ztmir.meil.pw.edu.pl/web/eng/Pracownicy/Vibekananda-Dutta-M.Sc

[2] Cornell Acivity Dataset. http://pr.cs.cornell.edu/humanactivities/data.php

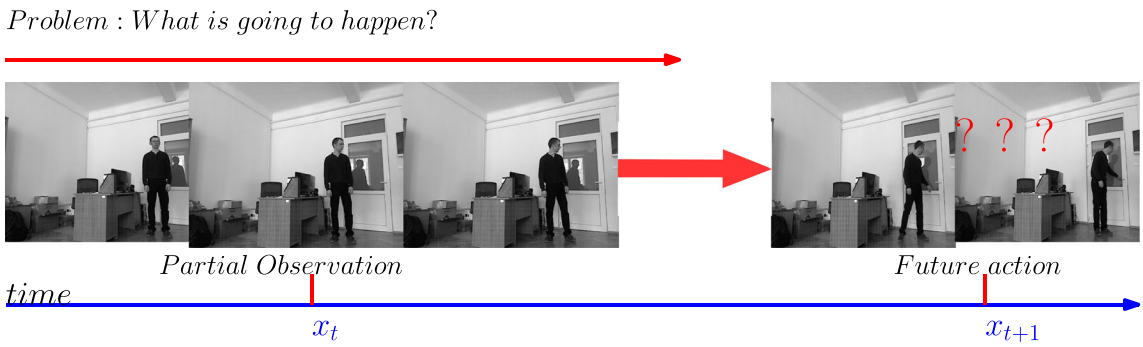


Fig. 1 Selected pictures illustrating the action prediction: **a** available observation, **b** the final action (post recording, this should be prognosed)

is to infer the action that a person is going to execute in the nearest future.

This paper is an extension of [10]. Comparing to previous material which was a short description, the method presented in this paper is an enhanced version with all relevant details. Moreover an improved method for temporal segmentation and feature extraction is introduced. The applied probability functions are here justified and the so-called limiting condition (Section 7) is summarized. Additionally, besides of previously described online experiments, the testing using offline data is discussed. We evaluated the action prediction problem in both: real-time and offline settings across the two datasets covering a wide range of actions. The contribution of this paper is four-fold:

- an improved method for action prediction is proposed,
- the concept of object affordances and scene context for human action prediction is formally described,
- a rapid training and testing method for action prediction is summarized,
- the proposed method is tested together with its efficiency evaluation.

The remaining part of the paper discusses this contribution in details. In Section 2 we had reviewed the related works. Section 3 describes the physical setup of the experiment. Video pre-processing is discussed in Section 4. The probability functions are summarized in Section 6. The description of motion trajectories is presented in Section 6. While Sections 7 and 8 present the implementation method and experimental results. The paper is ended with conclusions.

2 Related Works

Action Recognition Human action recognition becomes an extremely important research topic. Earlier research addressed mostly recognizing simple human actions, such as running, walking and standing in constrained settings [32]. However, recent research has gradually moved towards understanding complex actions in real-time records and in still images collected in various conditions [21, 22, 44]. These data typically involves occlusions, noisy background, changing viewpoints, etc and requires significant efforts on action recognition. Most of the action recognition approaches based on the still images, treat the problem as a pure image classification problem using i.e., mutual context model [43]. The mutual context model consider the bounding boxes of objects and human body parts, which is difficult to obtain especially with a large number of images. Another works consider the human "skeleton" features collected using the Kinect sensor together with object position. Recent contributions rely on the scene modeling [17] and human pose description [34]. The work presented in [6] presents transition of a "skeleton" pose through a Riemannian manifold. Riemannian manifolds have been confirmed useful for dealing with features and models that do not lie in Euclidean spaces. Those manifolds are used to analyse the human action similarity graphs that are mapped to a new space. A similar approach was adopted by Slama et al. [35], who classified activities using a Linear Support Vector Machines (LSVM) taking into account trajectory of human position using a learned Grassmann manifolds, which are special cases of Riemannian manifolds. The work described in [28] utilizes a multivariable Gaussian distribution to model the intermediate poses. The temporal deviations of activities were considered. Papadopoulos et al. [26] proposed a real-time "skeleton" tracking-based method for human action recognition which uses as an input a sequence of depth maps captured by a single Kinect sensor. The approach applies a motion energy-based concept, the spherical angles between the selected joints are evaluated with their respective angular velocities, for handling the execution differences among the individuals for the same actions.

Action Prediction Recent research has attempted to expand the concept of human action recognition to future actions. Some recent contributions on predicting actions are aiming at recognizing of the unfinished actions. The method

described in [15] uses the so-called max-margins for discriminating the action classes. Lan et al. [20] proposed a hierarchical representation of future possible actions. Li and Fu [23] explored the prediction problem for long duration actions. However, their work concentrated on identifying motion fragments by finding associated velocity peaks, it is not applicable to the unconstrained set of movements. The work presented in the article [39] describes how to consider object affordances for predicting in a static scenario what action will happen. An activity forecasting, which aims at reasoning of a human preferred motion path for a given goal has been explored in [14].

Other works capture human actions by representing the possible motion trajectories taking into account the detected point of interest [30, 40], the so-called key-frames were used for this purpose in [29]. Most of the previous contributions on action recognition methods were designed for recognizing complete actions, assuming that the action in each test will be fully executed. This makes these methods not appropriate for predicting actions in partial trials.

We concentrate on a probabilistic approach to model an action prediction (Fig. 2) taking into account object affordances (the affordance will be explained later in the text). We represent the types of human actions using a dynamic representation of human-object relation. Our method incorporates an action as a sequence of human postures and relates it to the information about object affordances, which is a new approach comparing to [20, 29, 40]. The proposed method allows also to predict long-term activities and allows to visualize how a human is going to perform an action, using trajectories prediction. Appropriate training method together with affordance function reduces the computational time comparing to the other methods [14, 18]. The experiments indicated that the proposed method is promising and be advantageous comparing to the state-of-the-art results.

Action Prediction Methods A general overview for action prediction is summarized in Fig. 2. As it can be seen that methods consist of 3 phases, namely: (a) preprocessing, (b) feature extraction, and (c) model formulation. Preprocessing is the term for low-level operations using videos. Generally, the aim of preprocessing is to divide the input video into multi-temporal segments, each segment represents an action or a sub-activity. The feature extraction means the finding of most compact and informative set of parameters (features) which should be selected in such a way that they provide sufficient information and assure the processing efficiency. The next step is to model formulation which means building a model representing an action and to recognize an activity.

The feature extraction methods are in general categorized into three categories: (a) low-level features, (b) mid-level features, and (c) high-level features.

Low-level methods (HoG) [5] focuses on a static appearance and shape within an image frame that can be described by the distribution of intensity gradients or edge directions. To this group belong also the methods in which the motion trajectories are obtained by tracking densely sampled points using dense optical flow fields [41].

The mid-level category considers mainly the semantic meaning of a scene and usually is build using low-level features. In actionlet [23] method belonging to this category, the first step is to temporally divide the activity into actions (reaching, drinking). These segments are called actionlet which represents the atomic actions. Poselet [2] feature extraction method describes a particular part of a human pose under specific viewpoint. Poslets are not necessarily semantic. Onset [31] feature extraction approach captures activity information from the sequence of actions which are components of an activity. The onset concept summarizes pre-activity observation in addition to ongoing observation.

High-level feature concept [12] uses input videos for extracting together with spatial and temporal features.

The created models can be of three types: (a) discriminative model, (b) generative model, (c) deep network respectively. Discriminative model generally use conditional probability distribution. A Support Vector Machine (SVM) is the most popular approach used for data classification [42]. Conditional random field (CRF) models are used for describing the predictions [18].

The generative model concentrate also on modeling conditional probability distribution but they require more detailed descriptions parameterized by time. The most common tool used for action recognition and action prediction is Hidden Markov Model (HMM) [3, 27].

Example of deep network models are the models using "memory" [25] created in the form of recurrent neural network. Such models are able to capture the useful information using previous observation and holding the long-range context. In [38] a regression network was used to anticipate human actions. Such networks was trained to predict the visual representation in the future.

3 Physical Setup

The setup for the data recording consists of two fixed viewpoint cameras placed on tripods with adjustable height 1.5 - 1.7m. Applied by us Senz3D [4] RGB-D cameras consist of an RGB and of a depth sensor. We used down-sampled images (640 × 480 pixels) since real-time decoding and display of multiple streams of a high-resolution video is a bottleneck problem. The recording rate was 60 frames per second. The camera system has the ability to register the 3D locations such as a human pose and object

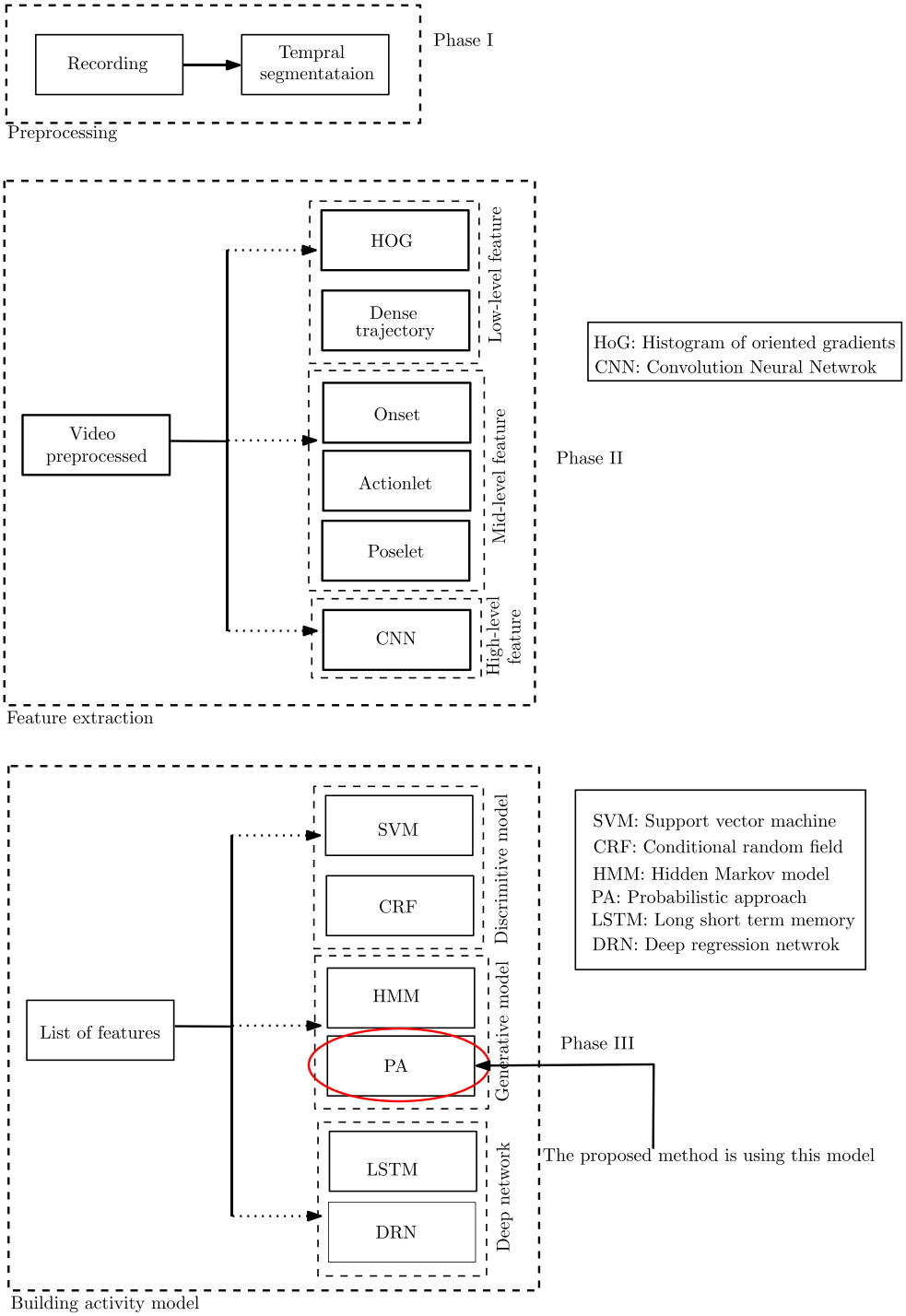


Fig. 2 General framework for human action prediction

position etc. During the experiments, the orientation of the cameras is fixed. Two cameras are delivering one planar picture in which (after proper preprocessing) the x, y, z coordinates of the points of interest are provided.

We used several objects on which the manipulations were performed. The cameras range for human observation is 1 - 3m, as it is shown in the Fig. 3b. The experiments were recorded both - in a day and - in an evening time, thus the lighting conditions had varied from daylight to an artificial one, with involving both plain and textured background.

Customized programming tools were developed for data extraction from the raw images. The software was developed using the C++ language and robot operating system (ROS). We applied a single coordinate system

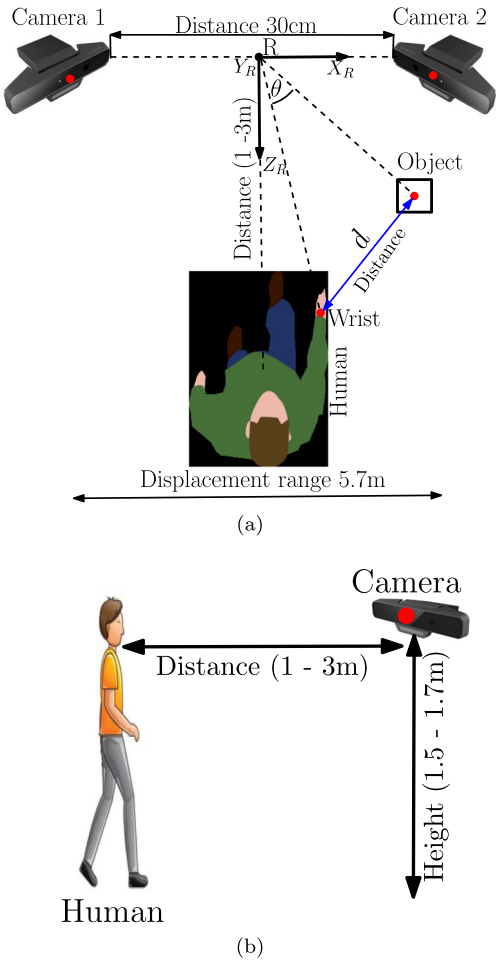


Fig. 3 Camera settings for video recording: **a** top view, **b** side view

with the origin placed in the mid point between the two cameras (see Fig. 3a). The transformation of the positions registered by each of the cameras was done respectively [13].

3.1 Basic Assumptions

A human activity is a something that a person is doing or is going to do, activity is a state of doing. We assume that an activity consists of a sequence of elementary actions. The aim of our work is to predict the actions that can help next in anticipating the human performance in the fragments of the whole activity what finally, after more research, can end with a whole activity prediction.

The first stage of our work consists of data collection, next the data are preprocessed and used for establishing the probabilistic models describing the possible motion trajectories. Finally, those models are used for actions prediction. On the end, we had done the set of tests evaluating the correctness of predictions.

3.2 Recording

We first recorded activities performed by 4 different persons: 3 males and 1 female respectively. Let us denote by A an activity. For each activity we had done M recording. Let's F_m^A denotes m-th record of an A-th activity (in our case $M = max(m) \geq 50$). Each record F_m^A consists of f frames, where f can vary from case to case. The m-th record of A-th activity consisting of f frames is denoted by $F_m^A(f)$. In our work, we considered temporal segmentations by partitioning the activity into group of actions.

It means that $F_m^A(f)$ is divided into smaller parts by the human expert (see Fig. 4). Each part represents an action (actions are the parts of activity). Therefore, the a-th part of $F_m^A(f)$ is denoted by $F_m^A(f)_a$, where $max(a) \geq 1$.

The segmentation is made having in mind that the final goal is to predict an action. The segmentation must be made in such way that the groups of frames in each segment are representing atomic movements of the human and/or of an object subjected to an action. It must be noted that a mistake in segmentation affects following-up prediction procedure and all the predictions can perform poorly. We carefully followed the segmentation approach described in [15].

4 Video Preprocessing

4.1 Temporal Segmentation

The part representing a_i-th action is again divided into segments. It not requires the expertise and the division method is fair enough. Such segmentation is needed for creating the probability functions. As an action can start when the human hand is in different distance towards the object of interest and can be made with different speed for gathering the relevant data is needed to selected and analyse some fixed amount of video segments covering the period till end of an action.

Let us consider a complete a_k action segment containing f_k frames. We divide into K uniform segments (in this work K is fixed to 10). In general, each segment contains $\frac{f_k}{K}$ frames. For different video lengths, f_k will vary. In certain cases, it requires an appropriate unification. For example, if an action video a_k is containing 233 frames, it is not possible to segment them uniformly into K segments. In such cases, we do additional pre-processing. Let us take an example, if $[f_k - int(\frac{f_k}{K}) \cdot K] = 0$, the frames are uniformly segmented into K segments, otherwise, we select the first image frame $f(1)_k$ from the action a_k and multiply it $[(f_k - int(\frac{f_k}{K}) \cdot K)]$ times (see Fig. 5). With such modification is possible to make the further automatic processing.

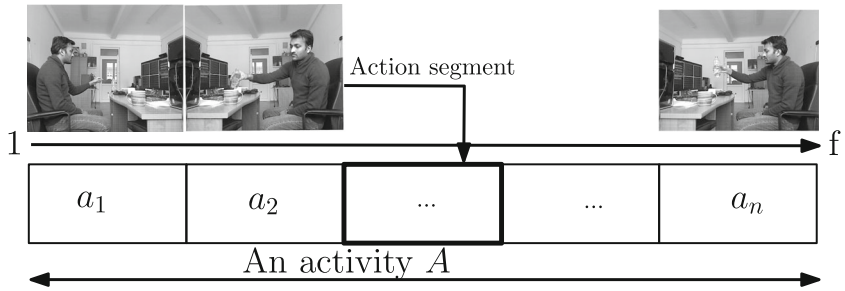


Fig. 4 Graphical illustration of an activity segmentation into actions. A human expert produced the groups of action (a_i - *i-th* action)

4.2 Features Extraction

We can recognize human action by looking at his/her current pose and interaction with the object/objects over a time, this is captured by a set of the so-called features. Features are quantities which are relevant for establishing the probabilistic models of actions. Finding a good feature extraction is very target oriented. Used by us features are rather simple.

In this work, we extracted three important features: (a) hand and torso position H, (b) object position O, (c) spatio-temporal features which are in our case – the distance d and angle θ as it is shown in Fig. 3a.

Hand Position The feature H describing the hand wrist joint (for both hands) position and torso position is obtained using the library Skeltrack API[3] which automatically delivers the positions in the camera fame (next the simple transformation converts it to the global reference frame). The information about the hand is very important. In particular, we want to capture information such as "hand is near to the object" or "hand is near the mouth". To do this, we evaluated the distance of the hand to the object and to the camera respectively. In general, the Skeltrack library provides the tool for the stick diagram visualization of the human body is described by the length of the links and the joints (Fig. 6). More specifically, in this work, the tracking algorithm detects the position of the following set of points in the 3D space, denoted by $J = \{$*Torso (T), Left Wrist (LW), Right Wrist (RW)*$\}$. The features are the (x, y, z) coordinates of each above point. Therefore the features matrix H is defined by,

$$H = \begin{bmatrix} x_T, x_{RW}, x_{LW} \\ y_T, y_{RW}, y_{LW} \\ z_T, z_{RW}, z_{LW} \end{bmatrix}, \quad (1)$$

where x, y are the positions of the points as they are seen in video frame. These positions, expressed first in terms of pixel coordinates, are converted to the metric coordinates in the global reference frame. The z coordinate comes from the cameras distance sensors and is expressed in the same units as the first two coordinates (see Fig. 3).

Object Features O represents an vector containing the x, y, z coordinate of the object center. For full description we use object identificator and position information (identificator can be represented by QR code of an object). In our work, we did both: the object detection, and tracking respectively (Fig. 7).

We consider two types of objects: (a) larger objects (i.e. door, table, box, whiteboard, etc.), (b) smaller objects (e.g. marker, bottle, cup, etc.). Larger objects are labelled by QR codes which can be properly recognized from different points of view using label-based object detection method [8].

For the smaller objects we use the "Lucas-Kanade" descriptor (KLD) field test[4] which in simplicity means the search of an object which picture is stored in the data base. Moreover we evaluate the distance passed by the objects when they are manipulated by the human being.

Spatio Temporal Features The spatio-temporal features, namely the distance d and angle θ (Fig. 3a) describe the relation between human hands and the objects feature.

We evaluate the distance d from some moment of time till the end of an action (in our case we consider the frame which marks 40% of frames from the end of the action segment). For each action, we collect such data repeating the recording M times. For each record we store the distance between the hand wrist position in the mentioned above moment of time and the object of interest (object to be manipulated). The collected data are used for evaluating the mean value μ_d, μ_θ and variance σ_d^2, σ_θ^2 which are applied later as the probability function parameters. Those functions are used for concluding about the destination of performed motion.

For each objects which can be manipulated (each action) in the human vicinity the video recordings are made as it was described above and the values of μ_d, μ_θ, σ_d^2, σ_θ^2 are gathered for each of those objects. Next

[3] https://people.igalia.com/jrocha/skeltrack/doc/latest/

[4] 3D tracking with descriptor fields https://cvlab.epfl.ch/page-107683-en.html

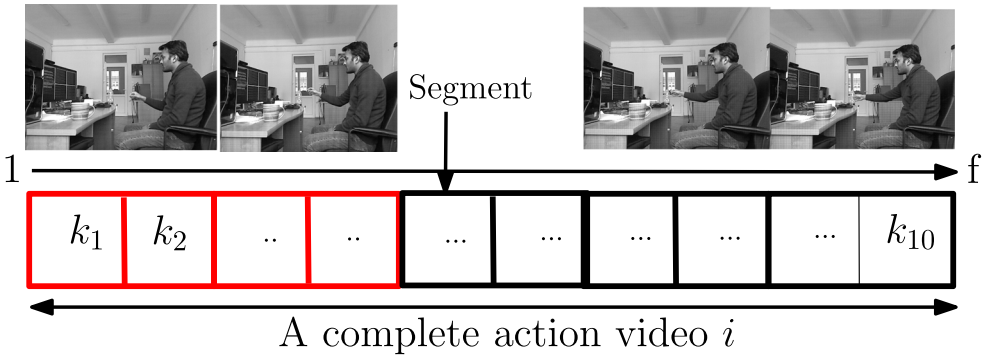


Fig. 5 Example of a temporal segmentataion of an action

for each object the probability function representing the chance of being manipulated is produced. The function consists of two terms – the term which is the so called distance preference (DP) and the term which is the angular preference (AP). The distance preference probability $P(DP_i)$ (for action a_i) represented by the normal Gaussian distribution and the angular preference probability $P(AP_i)$ is described by modified Wrapped Normal distribution. The justification of such functions selection will be discussed in Section 5.

For simplicity we can say that during the human hand motion as the most possible object to be manipulated (this is associated with the action) will be indicated such object to which the current distance (and the angle) is closest to μ_d (μ_θ). More precisely, applied probability functions will be delivering the probability of reaching each of the objects of interest providing for each of them probability created on the basis of current value of d, θ and the set of μ_d, μ_θ, σ_d^2, σ_θ^2. This is an action selection. Such action a_k is selected among all possible actions a_i $(i = 1, 2, ..., N)$, therefore:

$$P(a_k) = \max_{i=1,...,N} P(a_i) = \max_{i=1,...,N} (P(DP_i) \cdot P(AP_i)) \quad (2)$$

5 Probability Functions

The probability of an action is naturally related to the object of interest and the " easiness" of reaching/manipulating it. Therefore we call it the **object affordance**. The object affordance in our case results from the **angular** and **distance** preferences which are expressed as a product of two probability functions justified by experiments.

Distance Preference The distance preference is described by normal Gaussian distribution parameterized by mean μ_d and variance ${\sigma_d}^2$.

$$P(DP) = \frac{1}{\sqrt{2\pi {\sigma_d}^2}} \exp^{-\frac{1}{2}\left(\frac{d-\mu_d}{{\sigma_d}^2}\right)}. \quad (3)$$

The standard statistical test was applied to check whether the data are consistent with the selected distribution. A common test in such case is a *Shapiro Wilk* normality test, it has good performance for the smaller amount of samples as it was in our case.

The normal distribution plot given in Fig. 8 proofs that the distance features are following a normal distribution. Figure 8 also visualizes the probability distribution when

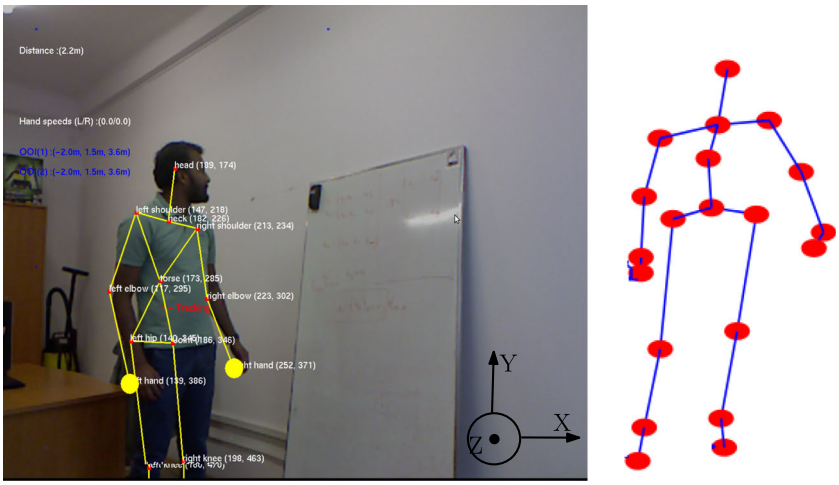


Fig. 6 Pictorial representation of human pose. The left image illustartes the RGB image (ground truth) with "skeleton" detected and the right image shows the extracted sketch diagram representing the human body

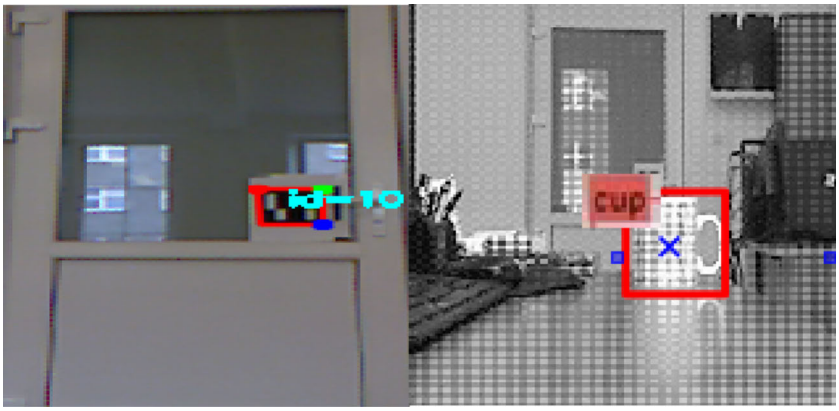


Fig. 7 Object detection results. The left figure shows the marker-based object detection for larger object and recognized object label which was defined for the object by the human expert. The picture on the right illustrates that the object of interest (a cup) was successfully detected

reaching an object (it is an alternative to $P(DP)$ expressed by an Eq. 3).

Angular Preference Angular positions of the human towards an object is very relevant in certain actions. For example, reaching action covers a wider range of angles than the drinking action. It is a "circular" statistics [33], where the data are expressed in an angular scale, typically around the circle. Here we applied the Wrapped Normal (WN) distribution introduced in the article [19]. The WN distribution is one that it is expressing the probability density function of a linear random variable to the circumference of a (unit) circle. It can be added that the von-Mises and the WN distribution are very similar. They both are the circular analogs of the normal distribution. However, the Wrapped Normal Distribution is more convenient for reasoning and is well explored in various activity recognition approaches [1, 7, 11]. Therefore, in our work, the probability function of angular preference is expressed as,

$$P(AP) = \frac{1}{2\pi}\left(1 + 2\sum_{j=1}^{J}\left(\exp^{-\frac{{\sigma_\theta}^2}{2}}\right)^{j^2}\cos(j(\theta - \mu_\theta))\right). \tag{4}$$

At the very beginning, we considered von-Mises distribution to capture the angular data [9]. However, due to its poor performance for larger values of σ, our choice moved to such distribution that possesses the normality feature for the larger values of σ. In such case, the modified version of the WN distribution, which is expressed in terms of Jacobi theta function is an appropriate choice.

The corresponding angular probability distribution function integrates to the unit in $[0, 2\pi]$. The justification of

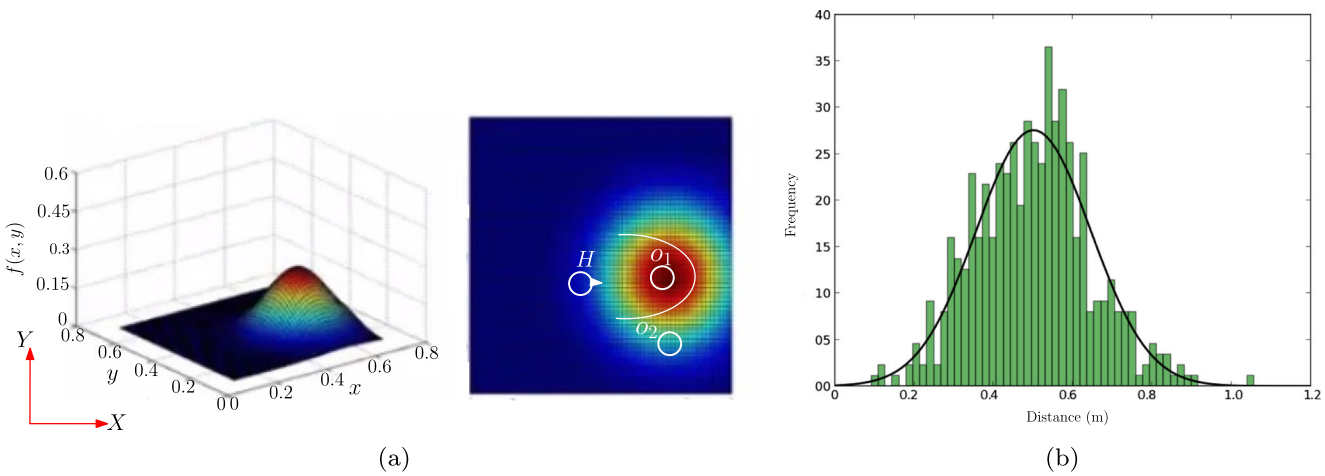


Fig. 8 From left to right: 2D Gaussian distribution (both side view and top view) for reaching an object, x and y represents the coordinate of the points (as described in the text) marking the hand position (the figure is taken from [10]). On the right the figure shows the histogram plot of the data justifying the normal distribution

the angular preference probability function was made using goodness-of-fit tests based on *Watson's* U^2 [37] statistic. A goodness-of-fit test enables to determine whether or not more complex models need to be considered. The advantage of *Watson's* U^2 statistic is that it is location invariant and thus does not depend on how the starting direction is assigned to the circle. A circular plot (Fig. 9) of the chosen statistical test for different values of the parameter σ shows that the data successfully follow the considered normal distribution (i.e. $U^2 < U^2_{critical}$).

6 Motion Trajectories

Prediction An object can be approached by various types of motion trajectories depending on the action that is going to be performed [10]. Once a location is estimated basis on probability function given in Eq. 2, we generate a nominal future trajectory form the current location (i.e., depends on the situation, can be hand, object and a considered joint) to the predicted target location.

A nominal future trajectory of the human hand is produced using the parameterized cubic equation of Bezier curve (see Fig. 10). The advantage of this equation is that it will not generate a fragment that lies outside the outline of the so-called the control points (commonly called the "hull" for the curve). In fact, we can control how the relevant points

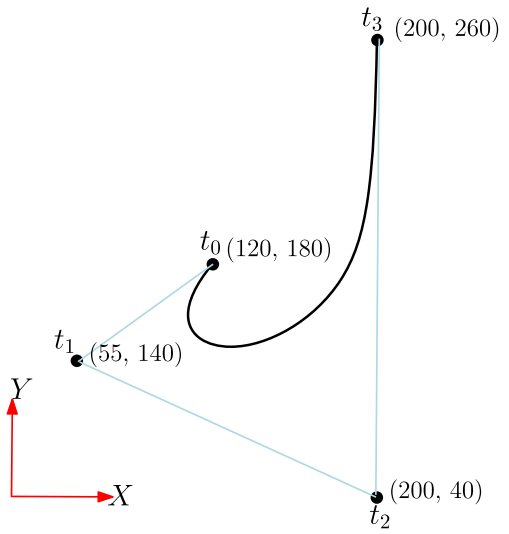


Fig. 10 Graphical representation of cubic Bezier curve

contribute to the value generated by the function, so we can influence how the points are important to the curve. The Bezier curve is a polynomial of p, with p Bezier interval being in the range $< 0, 1 >$:

$$t = (1-p)^3 t_0 + 3(1-p)^2 p t_1 + 3(1-p)p^2 t_2 + p^3 t_3, \quad (5)$$

where $t_i = \{x_i, y_i, z_i\}$ and $i = 0, 1, 2, 3$. Such a cubic Bezier curve is parameterized by a set of four points: the start and

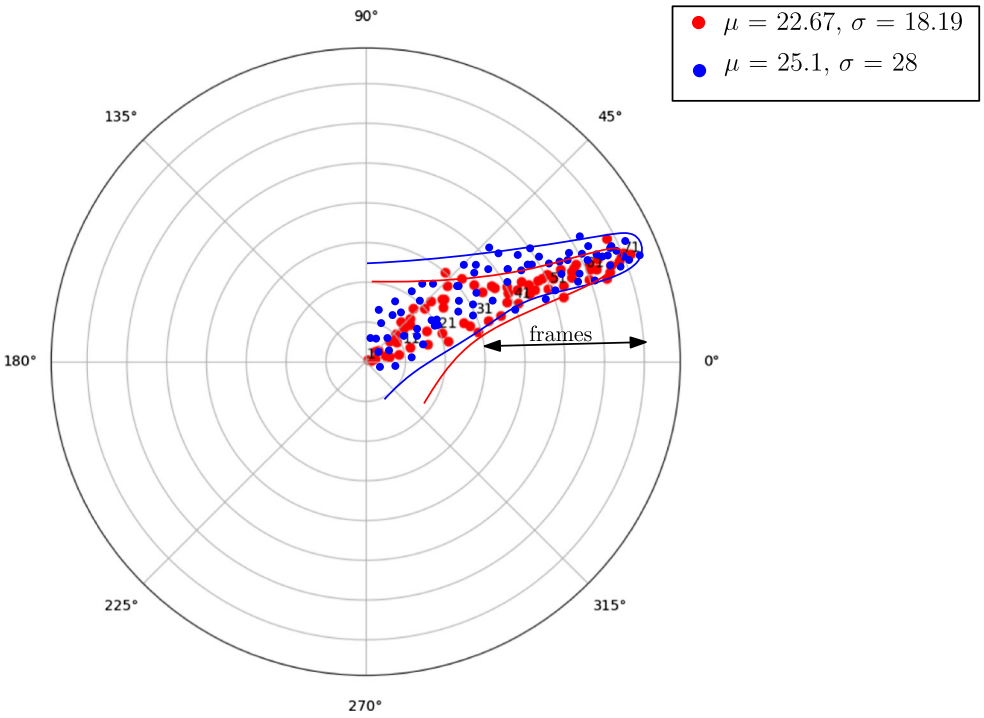


Fig. 9 Circular plot of proposed angular distribution for recahing an object. The figure is best viewed in color

(a) Pouring (b) Placing (c) Reaching

Fig. 11 Example images of pouring, placing, reaching actions from WUT-ZTMiR dataset

end point of the trajectory (t_0 and t_3), and two control points (t_1 and t_2) which define the shape of the curve. In our case, t_0 is the current position of the hand. The point t_3 is the end point of the action indicated by the probability function. The control points t_1 and t_2 are produced using the training data. Point t_1 and t_2 are the points taken from the previously recorded trajectory which has its beginning closet to the t_0.

Heat Map Around Trajectory We defined a potential function that allowed us to visualize the possible motion area. The heat-map visualization was implemented using exponential Gaussian kernel function. The map visualizes the active region around the trajectories and the target location, when the corresponding affordance is active. We implemented the heat-map visualization model in a software module using exponential Gaussian kernel function $f(h_m)$.

$$f(h_m) = \exp^{-\left(\frac{||d_T - \mu_{d_T}||}{2\sigma^2}\right)}, \tag{6}$$

d_T represents the point in question (e.g., the current position of a human hand) and μ_{d_T} is the point of the anticipated trajectory or the target location. Parameter σ denotes the radius of Gaussian kernel (the value is adjustable in this work). With the above estimate by Gaussian kernel, we can visually represent the expected regions with marking the greater heat by "warmer" colors. Accordingly, the "red color" denotes the maximum likelihood region.

7 Implementation

During implementation, the collected sets of d and θ were grouped with respect to the objects of interest. When the person starts the motion and in the environment are several objects of interest (objects which can be used when performing the activity) it is not clear to which object the person will focus. To make the reasoning process simpler we introduced a limiting condition $f(R)$ for selecting which set of objects (that means also which sets of actions) should be considered. The condition is as follows:

$$f(R) = \begin{cases} d \leq T^{o_t}_{near} & \text{near,} \\ T^{o_t}_{far} > d > T^{o_t}_{near} & \text{medium,} \\ d \geq T^{o_t}_{far} & \text{far.} \end{cases} \tag{7}$$

Where $T^{o_t}_{near}$ and $T^{o_t}_{far}$ represent near and far distance limits to the object. Using this condition only the objects which

(a) Drinking (b) Placing (c) Reaching

Fig. 12 Example images of drinking, placing, reaching actions from CAD-60 dataset

are relatively nearer to human hand are considered. For example, in real-time experiments, the following objects of interest were identified (a) a glass, (b) a bottle, and (c) the door. The distance between human torso and the door is relatively smaller than the distance between human torso and the glass, or the bottle. In our software, system we also defined the associations between the objects and the actions type which can be performed on them [15, 18]. For example, if a hand is near to the glass, the possible action will be grasping. But if we consider the same situation and the object is a computer monitor, the possible action would be turning on/off instead of grasping. It does not requires any sophisticated algorithms.

8 Testing

In this section, we describe the evaluation of the presented method for both: (a) offline data and (b) real-time settings. We first give the details of the dataset in Section 8.1. We then present the experimental results in Section 8.2, and the performance analysis of the proposed approach is discussed in Section 8.3.

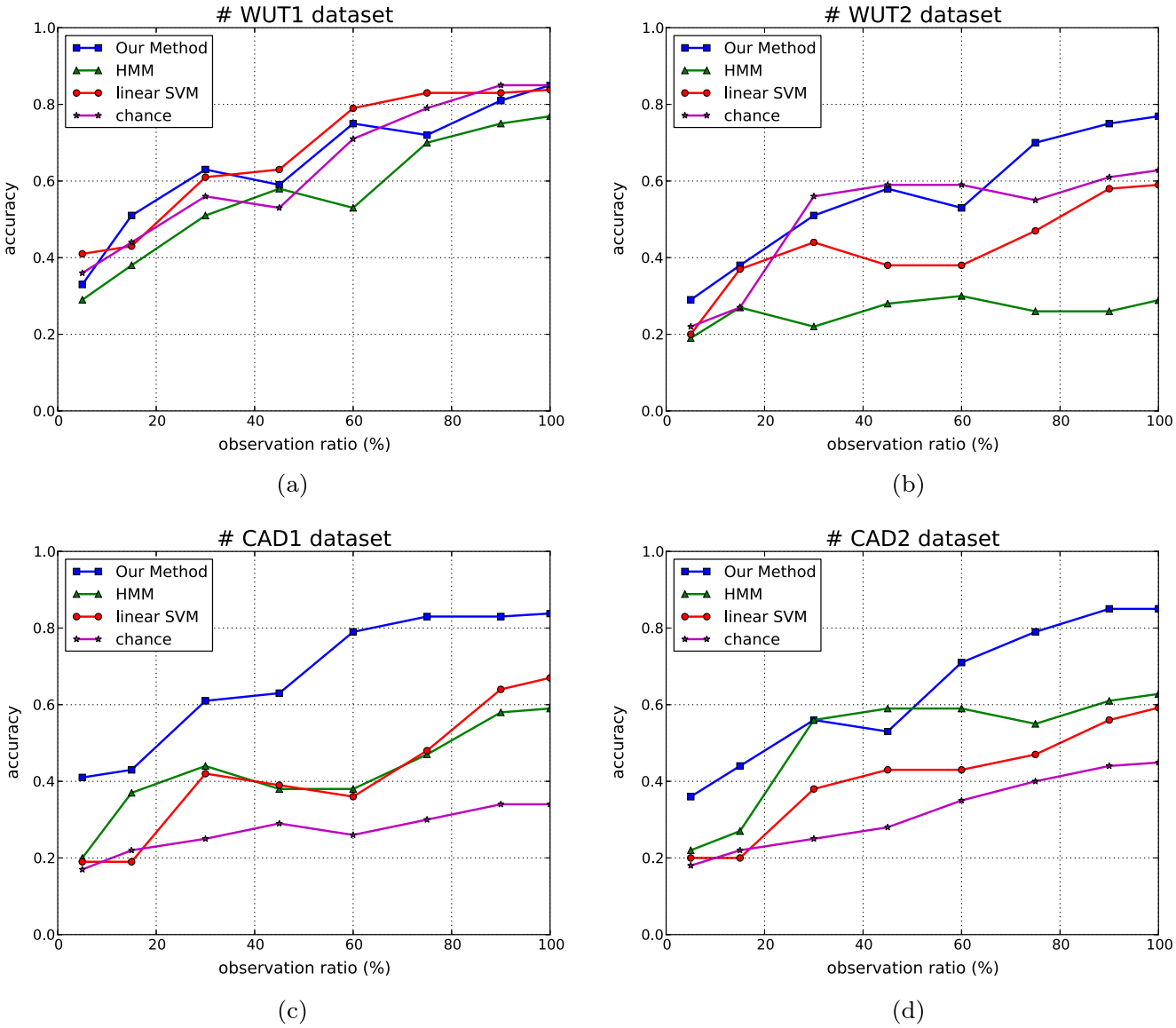


Fig. 13 Action prediction accuracy results. The comparisons of the proposed method against other methods on both WUT and CAD datasets (the figures are taken from [10]). The figure is best viewed in color

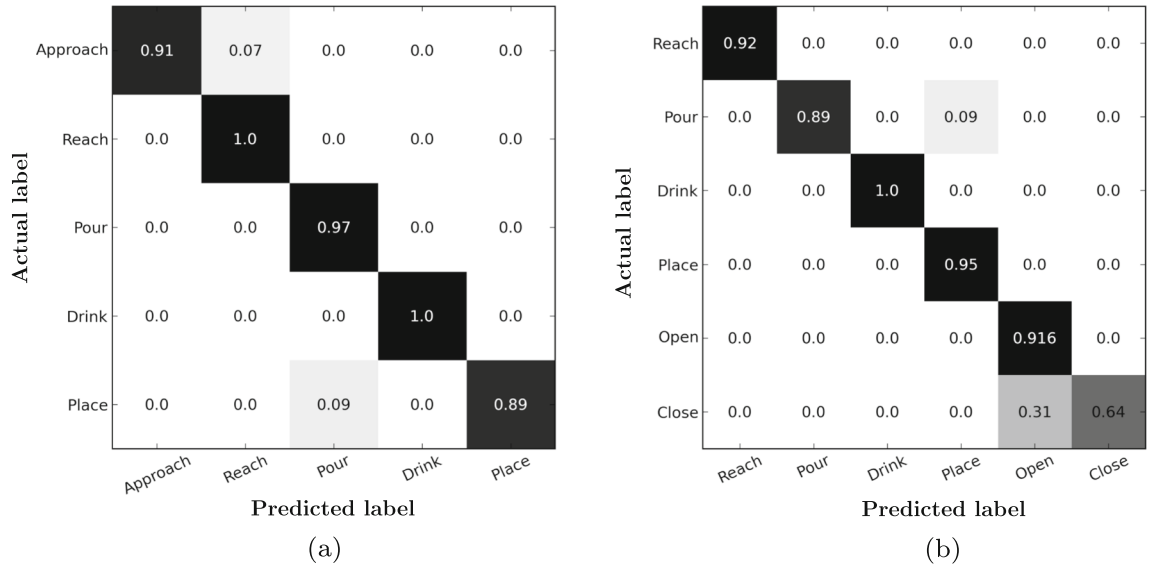


Fig. 14 Error matrices of action prediction on the test video records of both WUT and CAD dataset. Figure 14a shows the confusion matrix of prediction accuracy for WUT dataset. The confusion matrix of prediction accuracy for CAD test dataset is shown in Fig. 14b

8.1 Datasets

We tested the proposed method using two datasets: the WUT-ZTMiR dataset (WUT) set 1 (WUT # 1) and set 2 (WUT # 2), the publicly available CAD-60 dataset set 1 (CAD # 1) and set 2 (CAD # 2).

We created a publicly available dataset (named as WUT-ZTMiR) of human activities recorded in the office environment under RGB-D settings, i.e. color plus depth as shown in Fig. 11. The following activities are the part of the dataset: *drinking water, opening a door, object placing, etc.*.

The Cornell Activity Dataset (CAD-60) is composed of 12 different activities (see Fig. 12), performed in 5 different environments: (a) office, (b) kitchen, (c) bedroom, (d) bathroom, and (e) living room. The activities are performed by 4 people: 2 males and 2 females. The dataset is a

Table 1 Confirming the correctness of trajectory prediction on WUT and CAD dataset, showing average precision, reall and F-score for the actions

Action	Number of users	Number of objects	WUT-ZTMiR dataset			CAD-60 dataset		
			Pr	Re	Fc	Pr	Re	Fc
Reaching	1	2	0.63	0.68	0.65	0.618	0.63	0.642
passing	1	1	0.65	0.43	0.52	–	–	–
Reaching	1	1	0.71±0.08	0.53±0.31	0.57±0.24	0.69	0.59±0.48	0.59
Approaching	1	2	**0.76±0.28**	0.79	0.77	–	–	–
Pouring	1	2	0.56	**0.83**	0.67	0.58	0.79	0.79
Drinking	1	2	0.68±0.01	0.80±0.07	**0.77±0.02**	**0.76**	**0.85±0.18**	**0.81±0.42**
Placing	1	1	0.66±0.28	0.59±0.7	0.52±0.02	0.67	0.62±0.7	0.65
Opening	1	1	–	–	–	0.56	0.59±0.7	0.53±0.09
Closing	1	1	–	–	–	0.42±0.28	0.49±0.7	0.46±0.02
Approaching	1	1	0.69±0.28	0.74±0.7	0.70	–	–	–

Bold font indicates the best results

Table 2 Early and late predicted actions in both WUT and CAD dataset

Early prediction	Lately prediction
Reaching	Pouring
Approaching	Opening
Drinking	Closing
Placing	–

Actions are sorted according to the testing video records fallen in the category of EP and LP

collection of RGB images, depth images, and skeleton data with 15 joints. The activities are: *rinsing mouth, brushing teeth, wearing contact lens, talking on the phone, drinking water, etc.*.

8.2 Experimental Results

We conducted an experimental evaluation comparing our method to other methods using: (a) the so-called "chance model" which randomly selects the time moments and makes the prediction for that time, we use its published code and followed the settings given in [17], (b) the method using Hidden Markov Model (HMM) in which the hidden state sequences corresponding to the observation is considered, (c) Linear Support Vector Machine (LSVM) method where the transitions between the actions are focused [42]. All methods requires the ground truth progress to be known in the testing phase.

We were following the settings given in [42] and tuning parameters according to our needs. We actually achieved comparable performance to those reported in [17, 42]. The proposed method was evaluated using testing video records. The observation ratio is defined as the proportion between the frames considered as observed towards the total amount of frames. Figure 13 gives the comparison of our method with the other baseline using the two datasets described in Eq. 8.1. We applied a test video with different combinations of action. The accuracy (prediction rate) is defined by Eq. 8 and the results are shown in Fig. 13.

$$a_c = \frac{Number\ of\ correctly\ predicted\ actions}{Number\ of\ total\ actions}. \quad (8)$$

In order to evaluate the interpretable aspect of our method, we demonstrated its ability with using error matrix, also known as a confusion matrix (see Fig. 14). Note that in Fig. 14b a diagonal indicates few errors, such as *closing* sometimes was predicted as an *opening*, the reason is that both movements range is small. Moreover the *placing* action was predicted as a *pouring* due to the problem with light sensitive object recognition.

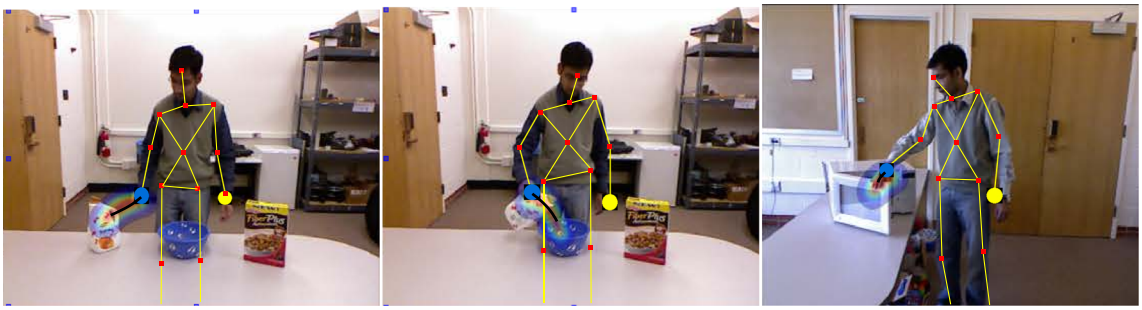
In the experiments, we found that the proposed method is generally capable of improving the high-level detection using joint reasoning. For example, a "closing microwave" video has an input action prediction accuracy of 48.9%. After joint reasoning, the output action prediction accuracy raised to 64.0%.

8.3 Performance Analysis

The evaluation of the proposed method was made using binary scores $t_{p_i}, t_{n_i}, f_{p_i}, f_{n_i} \in [0,1]$.

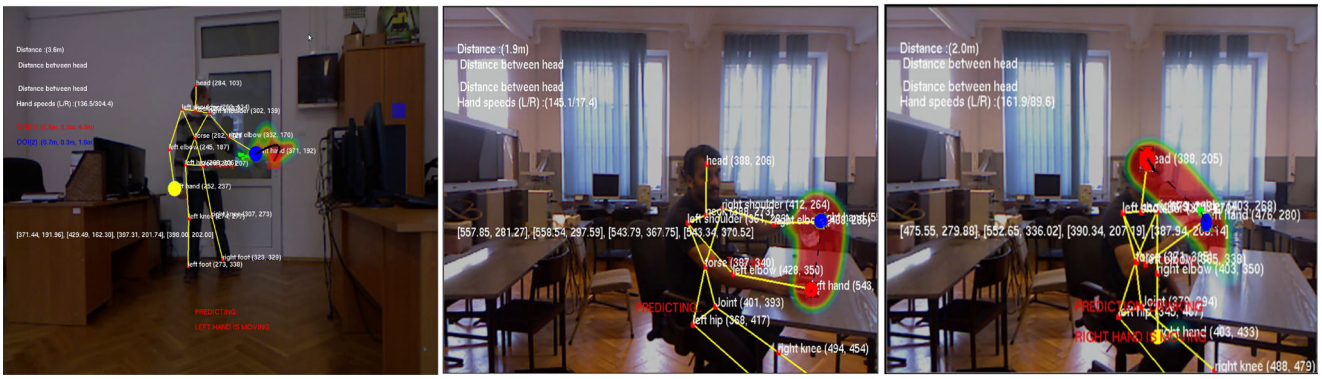
- $t_{p_i} = 1$ if the condition (c_1) is true,
- $t_{n_i} = 1$ if the condition (c_2) is true,
- $f_{p_i} = 1$ if the condition (c_3) is true,
- $f_{n_i} = 1$ if the condition (c_4) is true.
- $t_{p_i}, t_{n_i}, f_{p_i}, f_{n_i} = 0$, otherwise.

c_1 is true - when the system successfully identified an action that match a real situation (ground truth), c_2 is true - when the system rejected an action but in reality there it is the actual action, i.e. ground truth. c_3 is true - the system identified an action which does not match the real scenario,



(a) Reaching (b) Pouring (c) Closing

Fig. 15 Qualitative results of action prediction on CAD-60 dataset.The figures show the predicted right hand trajectories with heat maps. The following actions are: **a** reaching, **b** pouring, and **c** closing. The figure is best viewed in color



(a) Reaching (b) Pouring (c) Drinking

Fig. 16 Experimental results. The figures show the visualization of predicting the actions in the real-time scenario. The following actions are: **a** reaching, **b** pouring, and **c** drinking. Figure 16b and c are taken from [10]. The figure is best viewed in color

and c_4 is true - when the system successfully rejected an action which matches the real situation.

In Eq. 9, the global scores t_p, t_n, f_p, and f_n are evaluated as follows.

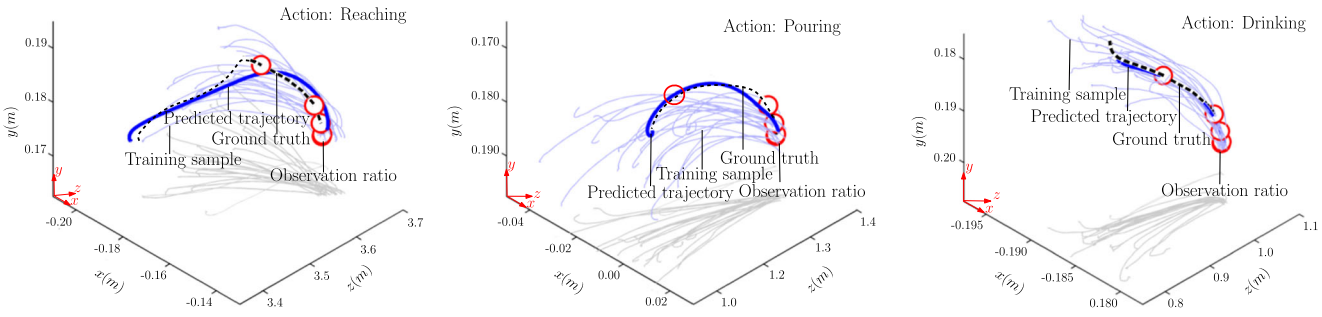
$$\begin{aligned} t_p &= \sum_i t_{p_i}, \quad t_n = \sum_i t_{n_i}, \\ f_p &= \sum_i f_{p_i}, \quad f_n = \sum_i f_{n_i}. \end{aligned} \tag{9}$$

Where t_p, t_n, f_p, and f_n are known as *true positive, true negative, false positive and false negative* respectively. The above binary scores were used to evaluated the recognition accuracy of an unfinished action and its limitations. Following [10], the precision $(Pr) = \frac{t_p}{t_p+f_p}$, recall $(Re) = \frac{t_p}{t_p+f_n}$, and F-score $(Fc) = 2\frac{Pr \cdot Re}{Pr+Re}$ were calculated, results are summarized in Table 1.

We analyzed the prediction of all the selected actions in CAD-60 and WUT-ZTMiR dataset and observed at what stage an action was predicted. In general, we defined two categories of action prediction according to the prediction stage: early prediction (EP) and lately prediction (LP). Early prediction means that the action was predicted if no more than 30% of the video was observed. However, the LP means that an action was predicted if more than 30% but less than 60% of the video was observed. Results are shown in Table 2.

Basis on the results is can be concluded that the proposed method performs well with partial observation (up to 60%), and is capable to make the real-time prediction with our equipment. It was collected 60 frames per second using 3.7 GHz Intel core 7 computer with 16 GB of RAM, with 64-bit Linux operating system. The average prediction time vary from 0.18s to 0.32s, what is acceptable for real-time applications.

Figures 16 and 15 show the visual output of the human action prediction for both: offline and online datasets. The blue circle indicates the moving hand and yellow circle indicates that the hand is stationary. The red curve with green and yellow outline along around black trajectory describes the possible future action. Figure 17 shows the predicted trajectories in 3D space (blue color) with respect to the ground truth trajectories (black color) of a particular action as well as the training sample trajectories defined by violet color.



(a) Reaching (b) Pouring (c) Drinking

Fig. 17 The 3D graphs of both ground truth and predicted trajectories of an action while performing a task. The actions are following: **a** reaching, **b** pouring, and **c** drinking. The figure is best viewed in color

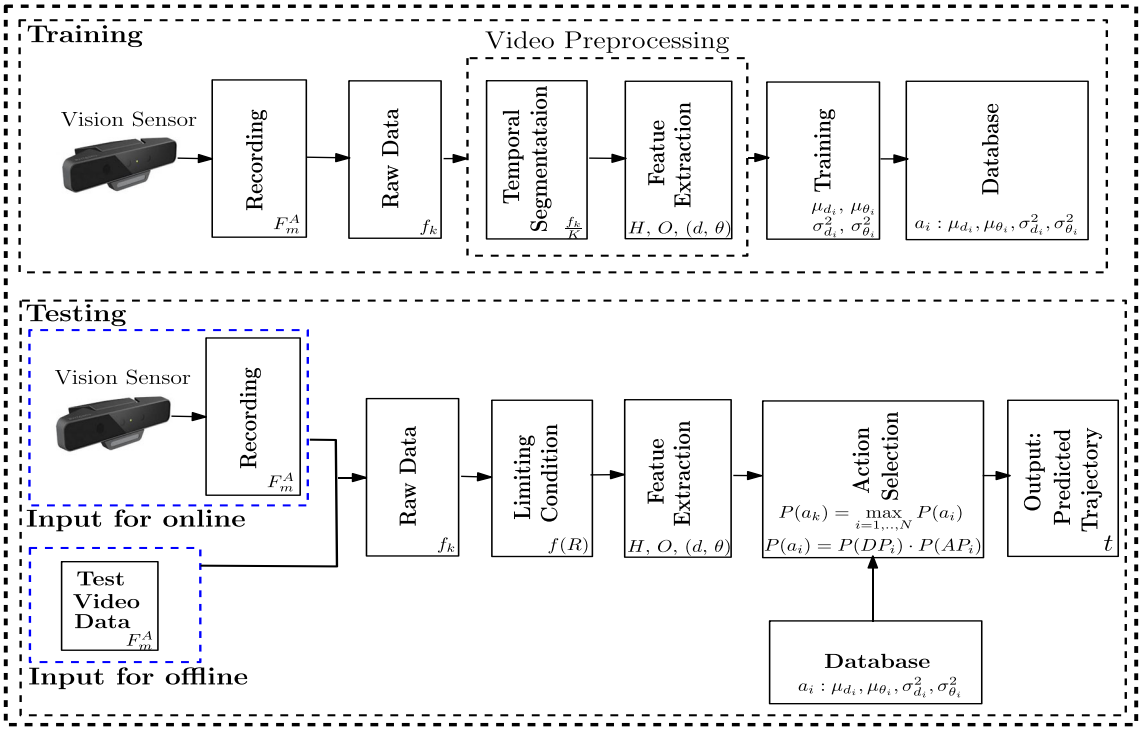


Fig. 18 Block diagram of the proposed method

9 Conclusions

In this paper, we presented the problem of human action prediction. The scheme illustrating the main components of the method is given in Fig. 18. The object affordance concept for predicting future actions was applied. The most possible motion trajectory was used as the "kernel" for producing the heat-maps representation of expected trajectory disparity area. The method was tested using on both: offline and online data. Obtained results were quantified and the method was validated as satisfactory. For selecting the possible actions we considered the probability functions which is based on the normal distributions. The choice of such function was justified, however, it would be interesting in the future to investigate the other possible distributions. We also showed that it is important to model the different properties (object affordances, temporal interactions, appropriate segmentation, etc.) in order to achieve good performance. In future, our intention is to study a wider range of actions with different environments and to expand the prediction process from the selecting one action among several alternatives to the chain prediction of actions aiming an activity.

Acknowledgements The initial stage of the work was supported by "HERITAGE" EU program (Grant Agreement 2012-2648/001-001 EM Action 2 Partnership) and in the later stages, the work was supported by the Preludium 11 (Grant No. 2016/21/ N/ST7/ 01614) funded by Polish National Science Center (NCN).

Publisher's Note

References

1. Baltieri, D., Vezzani, R., Cucchiara, R.: People orientation recognition by mixtures of wrapped distributions on random trees. In: Computer Vision–ECCV, pp. 270–283 (2012)
2. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3d human pose annotations. In: 12th Int. Conf. on Computer Vision, pp. 1365–1372. IEEE (2009)
3. Brand, M., Oliver, N., Pentland, A.: Coupled hidden markov models for complex action recognition. In: Proc., Computer Society Conf. on Computer Vision and Pattern Recognition, pp. 994–999. IEEE (1997)
4. Creative: Senz3D. https://us.creative.com/p/web-cameras/creative-senz3d. Accessed (2017)
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Society Conf. on Computer Vision and Pattern Recognition, CVPR, vol. 1, pp. 886–893. IEEE (2005)

6. Devanne, M., Wannous, H., Berretti, S., Pala, P., Daoudi, M., Del Bimbo, A.: Space-time pose representation for 3d human action recognition. In: Int. Conf. Image Analysis and Processing, pp. 456–464. Springer (2013)
7. Diethe, T., Twomey, N., Flach, P.: Bayesian modelling of the temporal aspects of smart home activity with circular statistics. In: Joint European Conf. Machine Learning and Knowledge Discovery in Databases, pp. 279–294. Springer (2015)
8. Dutta, V.: Mobile robot applied to qr landmark localization based on the keystone effect. In: Mechatronics and Robotics Engineering for Advanced and Intelligent Manufacturing, pp. 45–60. Springer (2017)
9. Dutta, V., Zielinska, T.: Predicting the intention of human activities for real-time human-robot interaction (hri). In: Int. Conf. Social Robotics, pp. 723–734. Springer (2016)
10. Dutta, V., Zielinska, T.: Action prediction based on physically grounded object affordances in human-object interactions. In: 11th Int. Workshop on Robot Motion and Control (RoMoCo), pp. 47–52. IEEE (2017)
11. Fablet, R., Black, M.J.: Automatic detection and tracking of human motion with a view-based representation. In: European Conf. Computer Vision, pp. 476–491. Springer (2002)
12. Ke, Q., Bennamoun, M., An, S., Boussaid, F., Sohel, F.: Human interaction prediction using deep temporal features. In: European Conf. Computer Vision, pp. 403–414. Springer (2016)
13. Kim, Y., Baek, S., Bae, B.C.: Motion capture of the human body using multiple depth sensors. In: ETRI Journal, vol. 39, pp. 181–190. Electronics and Telecommunications Research Institute (2017)
14. Kitani, K., Ziebart, B., Bagnell, J., Hebert, M.: Activity forecasting. Computer Vision–ECCV, 201–214 (2012)
15. Kong, Y., Fu, Y.: Max-margin action prediction machine. In: IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 38, pp. 1844–1858. IEEE (2016)
16. Kong, Y., Jia, Y., Fu, Y.: Interactive phrases: Semantic descriptionsfor human interaction recognition. In: IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 36, pp. 1775–1788. IEEE (2014)
17. Koppula, H.S., Gupta, R., Saxena, A.: Learning human activities and object affordances from rgb-d videos. In: The Int. Journal of Robotics Research, vol. 32, pp. 951–970. SAGE Publications (2013)
18. Koppula, H.S., Saxena, A.: Anticipating human activities using object affordances for reactive robotic response. In: IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 38, pp. 14–29. IEEE (2016)
19. Kurz, G., Gilitschenski, I., Hanebeck, U.D.: Efficient evaluation of the probability density function of a wrapped normal distribution. In: Sensor Data Fusion: Trends, Solutions, Applications (SDF), pp. 1–5. IEEE (2014)
20. Lan, T., Chen, T.C., Savarese, S.: A hierarchical representation for future action prediction. In: European Conf. Computer Vision, pp. 689–704. Springer (2014)
21. Lan, T., Sigal, L., Mori, G.: Social roles in hierarchical models for human activity recognition. In: Computer Vision and Pattern Recognition (CVPR), Conf., pp. 1354–1361. IEEE (2012)
22. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Computer Vision and Pattern Recognition, CVPR. Conf., pp. 1–8. IEEE (2008)
23. Li, K., Fu, Y.: Prediction of human activity by discovering temporal sequence patterns. In: IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 36, pp. 1644–1657. IEEE (2014)
24. Liu, J., Kuipers, B., Savarese, S.: Recognizing human actions by attributes. In: Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 3337–3344. IEEE (2011)
25. Ma, S., Sigal, L., Sclaroff, S.: Learning activity progression in lstms for activity detection and early detection. In: Proc. of the IEEE Conf. Computer Vision and Pattern Recognition, pp. 1942–1950 (2016)
26. Papadopoulos, G.T., Axenopoulos, A., Daras, P.: Real-time skeleton-tracking-based human action recognition using kinect data. In: MMM (1), pp. 473–483 (2014)
27. Pelc, L., Kwolek, B.: Activity recognition using probabilistic timed automata. In: Pattern Recognition Techniques, Technology and Applications. InTech (2008)
28. Pérez-D'Arpino, C., Shah, J.A.: Fast target prediction of human reaching motion for cooperative human-robot manipulation tasks using time series classification. In: Robotics and Automation (ICRA), Int. Conf., pp. 6175–6182. IEEE (2015)
29. Raptis, M., Sigal, L.: Poselet key-framing: A model for human activity recognition. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 2650–2657 (2013)
30. Raptis, M., Soatto, S.: Tracklet descriptors for action modeling and video analysis. In: European Conf. Computer Vision, pp. 577–590. Springer (2010)
31. Ryoo, M.S.: Human activity prediction: Early recognition of ongoing activities from streaming videos. In: IEEE Int. Conf. on Computer Vision (ICCV), pp. 1036–1043 (2011)
32. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local svm approach. In: ICPR. Proc. 17th Int. Conf. on Pattern Recognition, vol. 3, pp. 32–36. IEEE (2004)
33. Sengupta, A., SenGupta, A.: Topics in Circular Statistics. Series on Multivariate Analysis. World Scientific (2001)
34. Shotton, J., Girshick, R., Fitzgibbon, A., Sharp, T., Cook, M., Finocchio, M., Moore, R., Kohli, P., Criminisi, A., Kipman, A., et al.: Efficient human pose estimation from single depth images. In: IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 35, pp. 2821–2840. IEEE (2013)
35. Slama, R., Wannous, H., Daoudi, M., Srivastava, A.: Accurate 3d action recognition using learning on the grassmann manifold. In: Pattern Recognition, vol. 48, pp. 556–567. Elsevier (2015)
36. Sminchisescu, C., Kanaujia, A., Metaxas, D.: Conditional models for contextual human motion recognition. In: Computer Vision and Image Understanding, vol. 104, pp. 210–220. Elsevier (2006)
37. Thode, H.C.: Testing for normality. In: STATISTICS: Textbooks and monographs, vol. 164. CRC Press (2002)
38. Vondrick, C., Pirsiavash, H., Torralba, A.: Anticipating visual representations from unlabeled video. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 98–106 (2016)
39. Vu, T.H., Olsson, C., Laptev, I., Oliva, A., Sivic, J.: Predicting actions from static scenes. In: European Conf. Computer Vision, pp. 421–436. Springer (2014)
40. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 3169–3176 (2011)
41. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. In: Int. Journal of Computer Vision, vol. 103, pp. 60–79. Springer (2013)
42. Wu, H., Pan, W., Xiong, X., Xu, S.: Human activity recognition based on the combined svm&hmm. In: IEEE Int. Conf. on Information and Automation (ICIA), pp. 219–224 (2014)
43. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 17–24. IEEE (2010)
44. Yao, B., Jiang, X., Khosla, A., Lin, A.L., Guibas, L., Fei-Fei, L.: Human action recognition by learning bases of action attributes and parts. In: Int. Conf. on Computer Vision (ICCV), pp. 331–338. IEEE (2011)

Vibekananda Dutta received the B.Sc degree in computer science from Cotton College, Assam, India, in 2010 and the M.Sc degree in computer science with specialization in artificial intelligence from Central University of Rajasthan, Rajasthan, India, in 2012. Currently, he is working toward the Ph.D degree at the Faculty of Power and Aeronautical Engineering, Warsaw University of Technology. His research focuses on the mobile robots, computer vision and human-robot interaction.

Teresa Zielinska Ph.D, D.Sc, M.Sc in Eng. focuses her research interest on novel robotic systems, walking machines, humanoids, mobile robots, she also works on real-time control systems and interfacing methods for non-conventional robotic applications. She is the senior member of IEEE and Secretary General of IFToMM. She is the member of editorial board of several international journals devoted to the robotics and mechanics. She authored or co-authored of over 250 research publications.