Probabilistic Human Daily Activity Recognition towards Robot-assisted Living

Diego R. Faria, Mario Vieira, Cristiano Premebida and Urbano Nunes

Abstract— **In this work, we present a human-centered robot application in the scope of daily activity recognition towards robot-assisted living. Our approach consists of a probabilistic ensemble of classifiers as a dynamic mixture model considering the Bayesian probability, where each base classifier contributes to the inference in proportion to its posterior belief. The classification model relies on the confidence obtained from an uncertainty measure that assigns a weight for each base classifier to counterbalance the joint posterior probability. Spatio-temporal 3D skeleton-based features extracted from RGB-D sensor data are modeled in order to characterize daily activities, including risk situations (e.g.: falling down, running or jumping in a room). To assess our proposed approach, challenging public datasets such as MSR-Action3D and MSR-Activity3D [1] [2] were used to compare the results with other recent methods. Reported results show that our proposed approach outperforms state-of-the-art methods in terms of overall accuracy. Moreover, we implemented our approach using Robot Operating System (ROS) environment to validate the DBMM running on-the-fly in a mobile robot with an RGB-D sensor onboard to identify daily activities for a robot-assisted living application.**

I. Introduction

Nowadays with the advances of technology and the broad research worldwide, a cognitive robot can act as human assistant in the context of robot-assisted living, and also having the potential to offer social and entertaining interaction experiences through human-robot interaction. For that, in order to enable this natural human-robot interaction, the robot needs to infer the human intentions, their daily routine and potential risk situations by observing them. In this work, we focus our attention in the domain of human daily activity recognition. In this context, a robot that can recognize daily activities will be useful for assisted care: human-robot or child-robot interaction (e.g. in coping tasks); and also monitoring elderly and disabled people regarding their activities at home. In our previous work [3], we proposed a Dynamic Bayesian Mixture Model (DBMM) that was applied as a probabilistic loop, where the model recursively uses the prior information to reinforce current classification as a first-order Markov process. Herein, we are extending this model by using the memory of the system for dynamic update of the weighted ensemble, adjusting the weights based on previous behaviors of the base classifiers to improve the performance of classification. We validated the DBMM performance using different datasets and also using a mobile robot in an on-the-fly application for monitoring tasks. In the scope of human daily activity recognition, experimental results show that our proposed probabilistic ensemble is robust and with better performance than single classifiers and state-of-the-art approaches as well. Notice that, our framework relies only on 3D skeleton-based features, which is enough to characterize different classes of activities. The main impact of this work are the following:

- Employing a local update of weights on the DBMM using the memory of the system (i.e. previous base classifier behaviors) to obtain better classification performance.
- Modeling meaningful spatio-temporal features relying on skeleton distances, energy model and autocorrelation of joint translational movements, which can successfully characterize different activities.
- Assessment and validation: (i) comparing with single classifiers and state-of-the-art activity recognition approaches; and (ii) on-the-fly tests using a mobile robot for robot-assisted living.

The remainder of this paper is organized as follows. Section II covers selected related works. Section III introduces our approach, detailing the extended model with dynamic update of weights. The proposed skeleton-based features is presented in section IV. Section V presents the performance of the DBMM using state-of-the-art datasets and using a mobile robot for assisted living. Finally, Section VI brings the conclusion and future work.

II. Related Work

By looking to recent advances of works that use RGB-D sensors, several works focus on human-pose detection for human activity recognition [4] [5]. In [6], a maximum entropy Markov model (MEMM) for human activities classification was adopted, where features were modeled using the Histogram of Oriented Gradient (HOG). In [7], each activity is modeled into sub-activities, while object affordances and their changes over time were used with a multi-class Support Vector Machine (SVM) classifier. In [8], a bag of kinematic features was used with a set of SVMs, for activity classification. Other works on the recognition of human activities focus their research on how to model the attributes efficiently, to successfully obtain reliable classification [9] [10] [11]. In [12], a descriptor which couples depth and spatial information to describe humans body-pose was proposed. This approach is based on segmenting masks from depth images to recognize an activity. Sparse coding and temporal pyramid

This work has been supported by the Portuguese Foundation for Science and Technology, COMPETE and QREN programs under Grant AMS-HMI12 RECI/EEI-AUT/0181/2012. The authors are with Institute of Systems and Robotics, Dept. of Electrical and Computer Engineering, University of Coimbra, Polo II, 3030-290 Coimbra, Portugal (emails: diego, mario, cpremebida, urbano@isr.uc.pt).

matching is proposed in [13] for human action recognition. They use depth data for a learning algorithm that employs a discriminative class-specific dictionary. In [14], a feature descriptor for action recognition is presented. Depth motion maps are built given projection views in order to capture motion cues. Later on, a compact feature representation is obtained by using local binary patterns. Regarding our proposed framework, it allows the combination of different classifier models, which is advantageous to increase the classification performance. The DBMM dynamically reinforces the classification as a probabilistic loop, updating the initial learned weights given a confidence level to generate a distribution conditioned to the previous posteriors. Moreover, the DBMM approach has success in obtaining better results compared with benchmarked methods for activity recognition.

III. Probabilistic Classification Model: DBMM

DBMM is an ensemble of classifiers designed to combine a set of single classifiers (also referred as base classifiers) towards obtaining more accurate results than any of its individual members. For that, a probabilistic approach was adopted, using the concept of mixture models in a dynamic form in order to combine conditional probabilities. A weight is assigned to each base classifier, according to previous knowledge (learning process), using an uncertainty measure as a confidence level, and can be updated locally during the online classification. In our solution, the local weight update assigns priority to the base classifier with more confidence along the temporal classification, since they can vary along the different frame classifications. Figure 1 depicts an example of DBMM classification, where base classifiers are integrated as weighted posterior distributions, and previous posteriors and weights are used to update the model. The DBMM uses a set of models $A = \{A_m^1, A_m^2, ..., A_m^T\}$ where A_m^t is a model with m attributes; i.e., observed variables generated for some dynamic process at $t = \{1, 2..., T\}$. The DBMM probability distribution function for each class $P(C,A) = \prod_{t=1}^{T} P(C^t|C^{t-1}) \times \sum_{i=1}^{n} w_i \times P_i(A|C^t)$ can be rewritten holding the Markov property by taking the posterior of previous time instant as the new prior as follows:

$$P(C|A) \;=\; \beta \;\times\; \underbrace{P(C^t|C^{t-1})}_{\text{dynamic transitions}} \;\times\; \underbrace{\sum_{i=1}^{n} w_i^t \times P_i(A|C^t)}_{\text{mixture model with dynamic w}},$$

$$\text{with} \begin{cases} P(C^t|C^{t-1}) \equiv \frac{1}{C} \;\text{(uniform)}, & t = 1 \\ P(C^t|C^{t-1}) = P(C^{t-1}|A), & t > 1 \end{cases}, \tag{1}$$

where:

- $P(C^t \mid C^{t-1})$ is the transition probability distribution among class variables over time. A class C^t is conditioned to C^{t-1}. This means a non-stationary behavior applied recursively, then reinforcing the classification at time t.
- $P_i(A|C^t)$ is the posterior result of each i^{th} base classifier at time t, becoming the likelihood in this model.
- The weight in the model for each base classifier w_i^t is initially estimated using an entropy-based confidence on the training set (offline) as shown in our previous work [3], and afterwards ($t > 5$) it is updated as explained in the next subsection.
- $\beta = \frac{1}{\sum_j \left(P(C_j^t|C_j^{t-1}) \times \sum_{i=1}^{n} w_i \times P_i(A|C_j^t) \right)}$ is a normalization factor, ensuring numerical stability once continuous update of belief is done.

A. Dynamic Update of Weights using the System's Memory

During a classification task, base classifiers can change the performance over time. Thus, the local update of the weights during the on-line classification will benefit from the fact that the adjusted weights will produce a higher belief when priority is assigned to a base classifier with more confidence on previous classifications. We update the ensemble model using the temporal information on the test set as the memory of the system (set with previous posteriors for each base classifier $\Omega_i^s = \{P(C|A)^{t-1} ... P(C|A)^{t-s}\}$ together with the weights at the previous time instant w_i^{t-1}. Thus, in order to apply an update on the current weights, we compute:

$$w_i^t = \frac{w_i^{t-1} \times P(w_i|H_i(\Omega^s))}{\sum_{i=1}^{n} w_i^{t-1} \times P(w_i|H_i(\Omega^s))}, \tag{2}$$

where w_i^t is the estimated weight for each base classifier (updated); w_i^{t-1} is the previous weight at $t-1$. In order to obtain $H_i(\Omega^s)$, we use the memory of the system during the classification by keeping the previous posteriors (up to 5^{th} order), and consequently, we acquire the the entropy on each set of posteriors $H_i(\Omega^s)$ as follows:

$$H_i(\Omega^s) = -\sum_{j}^{s} H_i(\Omega^j) \log(H_i(\Omega^j)). \tag{3}$$

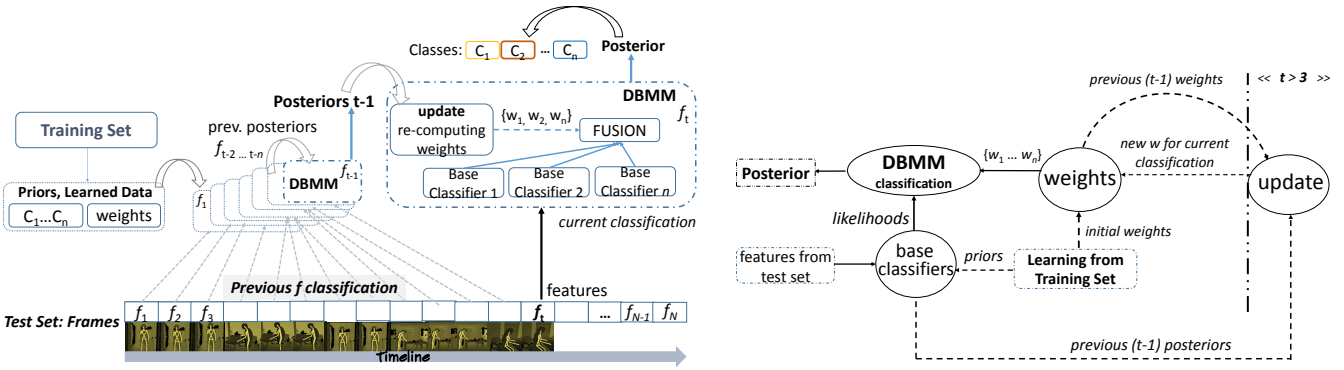
Knowing $H_i(\Omega^s)$ for each base classifier, the weights $P(w_i|H_i(\Omega^s))$ are estimated inversely proportional to the entropy:

$$P(w_i|H_i(\Omega^s)) = \frac{\left[1 - \left(\frac{H_i(\Omega^s)}{\sum_{i=1}^{n} H_i(\Omega^s)}\right)\right]}{\sum_{i}^{n} \left[1 - \left(\frac{H_i(\Omega^s)}{\sum_{i=1}^{n} H_i(\Omega^s)}\right)\right]}, \; i = \{1, ..., n\}, \tag{4}$$

where w_i is the result for each base classifier, and H_i is the current value of entropy given by (3). The denominator in (4) ensures that $\sum_i w_i = 1$.

B. Base Classifiers for DBMM Fusion

In this work, we have used the Naive Bayes Classifier (NBC), Support Vector Machines (SVM) and an Artificial Neural Network (ANN) as base classifiers for the DBMM. The NBC assumes the features are independent from each other given a class, $P(C_i|A) = \alpha P(C_i) \prod_{j=1}^{m} P(A_j|C_i)$. For the linear-kernel multiclass SVM implementation, we adopted the LibSVM package [15], trained according to the 'one-against-one' strategy, with *soft margin* (or Cost) parameter set to 1.0, and classification outputs were given in terms of probability estimates. The ANN adopted is a multilayer



(a) DBMM classification approach

(b) DBMM update cycle

Fig. 1: Example of DBMM during frame to frame classification in activity recognition. The left image shows that during the dynamic classification, initially the weights are learned from the training set, and later on, during the test they are updated.

feedforward network (with 40 neurons in the hidden layer), where the hidden layer transfer function is a hyperbolic tangent sigmoid and a normalized exponential (*softmax*) is used for the output of the transfer function as posterior probability estimates, conditional on the input, i.e., $\sum_{i=1}^{n} P(C_i|x) = 1$.

IV. Spatio-Temporal Skeleton-based Features

It is of utmost importance to find discriminative features of daily activity relying on existing relations between skeleton body parts to model their motion by correlating different time instants. The skeleton detection and tracking is made using depth images, adopting the OpenNi's software development kit for RGB-D sensor to obtain the joint locations of the human body.

We defined a set $\mathbf{F}$ with 51 features per frame to discriminate daily activities. Features based on skeleton joint distances, velocities and difference of skeleton poses along different frames are used in this work. Three types of spatio-temporal features are substantiated in the energy concept: 1) energy-based features using the joint velocities, 2) log-energy entropy-based features using skeleton poses, and 3) sample autocorrelation-based features using the distances of skeleton poses in different time instants. The velocities energy of the upper joints of the skeleton (i.e. seven joints: head; left and right shoulders, hands and elbows) are computed as follows:

$$E_{uv} = \sum_{j=1}^{N}(V_{jx})^2 + \sum_{j=1}^{N}(V_{jy})^2 + \sum_{j=1}^{N}(V_{jz})^2, \quad \text{with} \quad V_{jd} = \frac{\mathbf{S}^t_{jd} - \mathbf{S}^{t-s}_{jd}}{\Delta T}, \quad d = \{x,y,z\}, \tag{5}$$

where for each dimension $\{x,y,z\}$, $\mathbf{S}_j$ is a vector of dimension 7×1, whose elements are the skeleton joints; for the computation of V_{jd}, the numerator corresponds to the skeleton joints distances from t to t_s preceding frames (herein, $s = 10$), and the denominator corresponds to the elapsed time $\Delta T = f_{rate} \times \varpi$ (a frame rate $f_{rate} = 1/30$ and a temporal slide window $\varpi = 10$ were used).

The second feature is based on the sum of log-energy entropy $logE_s$ using the global skeleton joints in each dimension as follows:

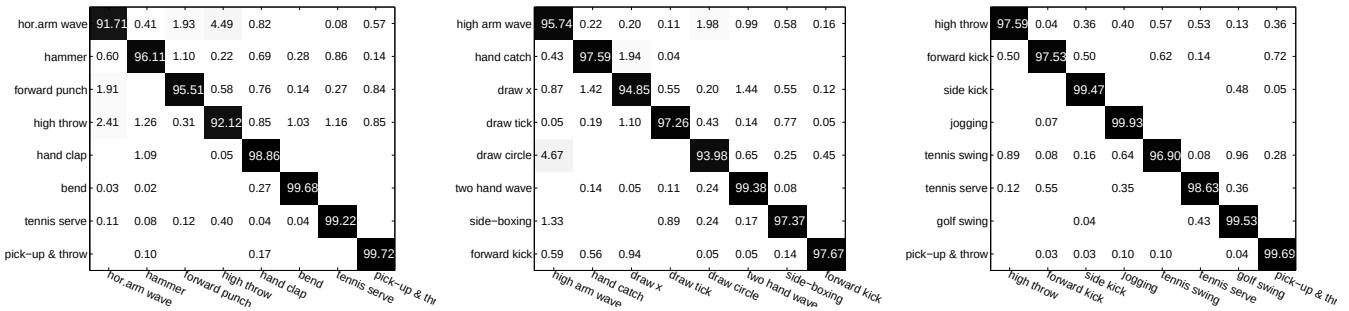
$$logE_s = \sum_j \log(\mathbf{S}^2_{jx}) + \sum_j \log(\mathbf{S}^2_{jy}) + \sum_j \log(\mathbf{S}^2_{jz}). \tag{6}$$

The two aforementioned features enclose key poses of movements, i.e., when the skeleton joints alternately show acceleration and deceleration in repeated movements that leads to changes in the energy model representation. This information helps the characterization of drastic changes in direction and velocities of the skeleton. The energy model (5) is applied to the upper body part and the log-entropy (6) is applied to all body joints.

The third feature is based on the autocorrelation function employed on the difference of skeleton poses at time t and $t-1$. The first step before computing the autocorrelation is to obtain the translation of each skeleton joint S_j from a time instant $t-1$ to the current time instant t by employing the Euclidean distance $\delta_{\{S^t_{jd}, S^{t-1}_{jd}\}} = \sqrt{(S^t_{jd} - S^{t-1}_{jd})^2}$, $d = \{x,y,z\}$, obtaining a matrix of $N \times d$ (i.e., number of joints N and d-dimensional space). Subsequently, the sample autocorrelation is computed by:

$$r(\tau) = \frac{\frac{1}{T-1}\sum_{t=1}^{T-\tau}\left(\delta^t_{\{S^t,S^{t-1}\}} - \mu^t_\delta\right)\left(\delta^{t+\tau}_{\{S^t,S^{t-1}\}} - \mu^{t+\tau}_\delta\right)}{\sigma^2} \tag{7}$$

where $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\delta_{\{S^t,S^{t-1}\}} - \mu_\delta\right)^2$ is the sample variance and μ_δ is the sample mean value; and τ is the lag variable of a process at different times. Since we are working with 3D skeleton arranged in a matrix $\delta_{\{S^t,S^{t-1}\}}$ of 20×3 (joints by 3 dimensions), then in order to facilitate the autocorrelation computation, we applied a self-convolution, whereas the autocorrelation is alike to a convolution, apart from it does not need to flip an input about the origin. Thus, 2D convolution in spatial form for finite intervals is achieved by $f * g = c(i,j) = \sum_k^p \sum_l^q f(k,l) \times g(i-k, j-l)$, where $f = \delta_{\{S^t,S^{t-1}\}}$, and g which commonly has the role of the filter in convolution, herein it is in charge of the shift of f with respect to itself (rotates about the origin) in the plane $p \times q$. A resulting matrix that is given by $f * g$ has a



(a) AS1: Prec:96.59%; Rec:96.61%; Acc:96.61% (b) AS2: Prec:96.76%; Rec:96.73%; Acc:96.73% (c) AS3: Prec:98.66%; Rec:98.65%; Acc:98.65%

Fig. 2: Classification results: cross-validation confusion matrix for each action set using the DBMM for the "new person" setting (training five persons and testing on other "unseen" five). Global Prec.: 97.34%, Rec.: 97.33%, Acc.: 97.33%.

dimension of $m \times n$ $(2 \times \text{size}(\delta_{\{S^t,S^{t-1}\}}) - 1) = 39 \times 5$, was then reshaped as a feature vector $\mathbf{r}$ of $(m \times n) \times 1$ elements to compute the autocorrelation energy $E_r = \sum_i \mathbf{r}_i^2$.

Additionally, a set of features based on Euclidean distances of the skeleton joints $\delta_{\{S^d_{j1},S^d_{j2}\}}$ was used, as similarly presented in [3]: 1) the minimum distance from hand (left or right) to the head, e.g. $\min(\delta_{\{S^d_{j1},S^d_{j2}\}}, \delta_{\{S^d_{j1},S^d_{j3}\}})$; 2) the minimum distance from elbow (left or right) to the head; 3) the minimum distance from hand (left or right) to the center of the skeleton; 4) distance from the left hand to the right hand; 5) distance from the head to the center of the hip; 6) distance from the central knee (mean coordinate taking into account the left and right knees) to the center of the hip; 7) the minimum distance from foot (left or right) to the head; 8) the hand with higher changes in directions (i.e., using the difference of the current position to a previous one); 9) six angles obtained from triangles formed by: shoulder, hand and elbow; hip, shoulder and knee; hip, knee and foot, all considering left and right sides. The angle computation is given by $\theta_i = \arccos(\delta^2_{j_{12}} + \delta^2_{j_{23}} - \delta^2_{j_{13}} \;/\; 2 \times \delta_{j_{12}} \times \delta_{j_{23}})$, where $\delta_{j_{12}}$ is the Euclidean distance between two joints. These angles are useful to discriminate stand and seated positions or torso inclination.

Then, a stage consisting of derivatives and accumulative values was employed on the aforementioned set of extracted features $\mathbf{F}$. We first applied a discrete derivative $y = \frac{\mathbf{F}^t - \mathbf{F}^{t-s}}{\Delta T}$ on each feature, where s represents a temporal slide window of ten frames. Subsequently, we accumulated each feature value over the frames: $y^t_{cum} = \sum_{k=1}^{t} \mathbf{F}_k$. Thus, with these two steps we obtained more 34 features, and $\mathbf{F}$ sums up to a total of 51 features. To ensure a higher classification performance, an essential step is employed; the extracted set of features are normalized in such a way that, values of minimum and maximum obtained during the training were applied on the normalization of the test set.

V. Assessment of the Proposed Framework on Datasets and Robotic Application

Experimental tests using a mobile robot and two datasets were performed to assess our framework. Looking at the performance attained, we can state that our framework has good potential for activity recognition in robot-assisted living.

A. Performance on MSR-Action3D Dataset

The MSR-Action3D dataset [1] contains skeleton data from depth images captured by an RGB-D sensor at 15Hz. MSR-Action3D comprises twenty actions, and each action was performed by ten subjects for three times. The actions cover various movement of arms, legs, torso and their combinations. For performance evaluation purposes, and concerning this dataset, we followed the same methodology as described in [1] [2], where the dataset is split into 3 action sets with eight actions each one as shown in Fig. 2. As stated in [1], AS1 and AS2 group actions with similar movements, while AS3 groups actions that are more complex. We follow the cross-validation test as defined by [2] and [16]. The tests were performed by training five subjects out of ten, and testing on the other five subjects (testing on "unseen persons"), e.g., training persons {1,3,5,7,9} and testing on persons {2,4,6,8,10}; afterwards the opposite (even, odd); then, training on persons {1...5} and testing on persons {6...10}, and so on. Taking into consideration 5×5 splits, there are 252 possible splits in total. The overall accuracy (average) was computed to compare our proposed framework with other state-of-the-art methods. Results show that our proposed framework outperforms other state-of-the-art benchmarked methods using this dataset up to the current date. The overall accuracy obtained with the DBMM was 97.33%, taking the average of all attained performances. Figure 2 presents the overall confusion matrix for the cross-subject classification for each action set. Table I summarizes the results attained by the DBMM in comparison with each single classifier and an averaged ensemble for AS1, AS2 and AS3, showing that our approach outperforms the other classifiers (all using our skeleton features). Finally, Table II presents the results of our DBMM approach in comparison with other state-of-the-art methods evaluated using the MSR-Action3D dataset. This table references some selected works, the ones with higher overall accuracy up to date.

Our approach using only 3D skeleton features outperforms other approaches that use features from skeleton, from depth

TABLE I: Accuracy on action sets using single classifiers, a simple averaged ensemble (AV) and the proposed DBMM.

Action Set	SVM	Bayes	ANN	AV	**DBMM**
AS1	92.8%	89.3%	90.8%	90.9%	96.6%
AS2	91.7%	88.4%	90.4%	90.1%	96.7%
AS3	94.6%	89.9%	92.7%	92.4%	98.6%
Average	93.0%	89.2%	91.3%	91.1%	97.3%

TABLE II: Comparison of approaches that use the MSR-Action3D in terms of overall accuracy. Columns 3 an 4 point out the feature types used by the approaches.

Method	Acc	SK joints	Depth
Proposed framework (DBMM)	**97.33**%	✗	
* Luo *et al.* [13]	97.26%	✗	✗
Chen *et al.* [14]	94.90%	✗	
Ohn-Bar and Trivedi [17]	94.84%	✗	✗
Yang, Zhang and Tian [18]	91.63%	✗	
Chaudhry *et al.* [19]	90.00%	✗	
Evangelidis *et al.* [20]	89.86%	✗	
Oreifej and Liu [16]	88.89%		✗
Wang *et al.* [2]	88.20%	✗	✗

*The approach in [13] obtained 96.7% when using only skeleton features

and even approaches that combine both.

B. Performance on MSR-DailyActivity3D Dataset

The MSR-DailyActivity3D [2] is another dataset with depth images and 3D skeleton data from an RGB-D sensor that was used herein to evaluate our approach. It contains 16 activities: *1-drink, 2-eat, 3-read book, 4-call cellphone, 5-write on a paper, 6-use laptop, 7-use vacuum cleaner, 8-cheer up, 9-sit still, 10-toss paper, 11-play game, 12-lie down on sofa, 13-walk, 14-play guitar, 15-stand up, 16-sit down* performed by 10 subjects twice, where one trial is in standing position, and the second in sitting position on a sofa. We followed the state-of-the-art methodology [2] for evaluation of our framework. This dataset has all 16 activities in a single scenario, i.e., a multi-class cross-subject classification. The tests were performed in the same way of the MSR-Action3D by training five subjects out of ten, and testing on the other five subjects ("unseen persons"). The results attained are shown by means of a confusion matrix in Fig. 3. To the best of our knowledge, our results outperforms other state-of-the-art methods applied on MSR-DailyActivity3D dataset up to the current date. The overall performance obtained with the DBMM approach are: precision of 97.39%; recall of 96.83%; and accuracy of 96.83%. Table III shows the overall accuracy of our approach compared with some selected works of the state-of-the-art, i.e. the ones with higher accuracy for this dataset up to the current date.

C. Performance using a Mobile Robot

In order to evaluate our approach using a mobile robot, we built a dataset (e.g. Fig. 4) with RGB-D image sequences and skeleton data to learn human daily activities, such as *1-walking, 2-stand/still, 3-talking on the phone, 4-working on a computer and 5-sitting*; and for suspicious or risk situations: *6-jumping, 7-falling down, 8-running*. We recorded 4 persons performing 3 times each activity during 30 up to 45 seconds.

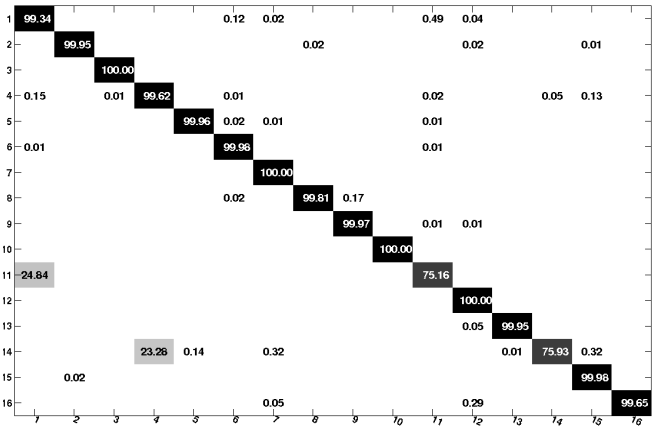


Fig. 3: Confusion Matrix obtained from the DBMM classification applied on the MSR-DailyActivity3D dataset.

TABLE III: Comparison of approaches that use the MSR-DailyActivity3D in terms of overall accuracy. Columns 3 an 4 point out the feature types used by the approaches.

Method	Acc	SK joints	Depth
Proposed framework (DBMM)	**96.83**%	✗	
Luo *et al.* [13]	95.00%	✗	✗
Xia and Aggarwal [21]	88.20%	✗	✗
Wang *et al.* [2]	85.75%	✗	✗

Robot Operating System (ROS) packages in *hydro* version were used to program the mobile robot to navigate in an indoor environment. For that, the robot has different sensors onboard, such as laser for mapping and localization, avoiding obstacle collision, and an RGB-D sensor for human body detection for skeleton tracking and human activity recognition. Reminding that, in this work, the focus of our attention is on the evaluation of our probabilistic approach for activity recognition on-the-fly, thus, herein we do not detail other robot functionalities (e.g., navigation and robot (re)actions). Once the skeleton is detected in a range of two up to five meters to the RGB-D sensor, the robot starts the activity recognition. In this experiment, a robot response is assigned for each activity that is recognized (e.g. during a monitoring task, when a usual activity is classified, the robot will just re-position itself to keep monitoring). For each risk situation detected, the robot is supposed to assist somehow, by sending warnings or calling relatives to report the current situation. Figure 5 shows the cognitive system for activity recognition in robot-assisted living (monitoring task) using ROS environment[1].

The strategy to test an on-the-fly application using a mobile robot is a little different than the evaluation on datasets. In this case, the DBMM classification is made in 3 up to 5 seconds to guarantee a confidence for a final decision, i.e., after recognizing the activity, the robot will respond with an action. Figure 6 shows few snapshots of the experiments of daily activities including a risk situation that

[1]A video demonstrating our approach for robot-assisted living can be seen at `https://youtu.be/FAfLj28_iSM`

Fig. 4: Few examples of the dataset (RGB and depth images) which was built to learn some daily and risk situations.

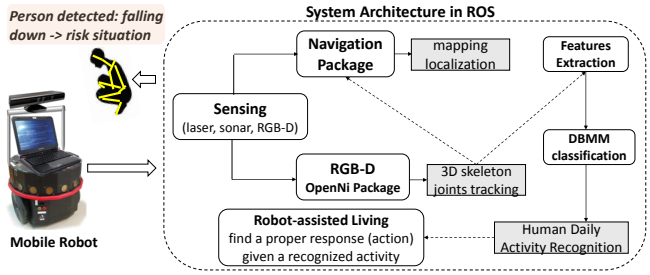


Fig. 5: Architecture in ROS of our artificial cognitive system for robot-assisted living.

our mobile robot correctly recognized. During the on-the-fly experiments using a mobile robot, all activities performed twice by two "unseen" persons were correctly classified. The overall confidence of classification in the context of robot-assisted living is presented in a confusion matrix as shown in Fig. 7, with overall accuracy of 90.46%. We noticed that the activities can be correctly classified with a high certainty within 3 up to 6 seconds of frames by frame classification. The activities *walking* and *running* were the ones with more misclassification due to their strong similarities.

VI. Conclusion and Future Work

A dynamic probabilistic ensemble of classifiers (DBMM) using a local update of weights was designed for activity recognition. The local weighting strategy to update the model has shown through experimental results to be very effective given a set of suitable features. Two well-known state-of-the-art datasets of human daily activities, Microsoft Research [1] [2], were used to evaluate the performance of our approach. The classification performance in terms of overall accuracy has shown that our proposed framework outperforms other methods in the scope of human daily activity recognition. In addition, we performed experimental tests of our approach running on-the-fly in a mobile robot for monitoring daily activities and risk situations, showing that it has potential to successfully be used in robot-assisted living applications. Future work will exploit and extend our framework for robot-assisted living and natural human-robot interaction scenarios.

References

[1] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points,," in *IEEE CVPRW: Human Comm. Behav. Analysis*, 2010.

[2] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *IEEE CVPR*, 2012.

[3] D. R. Faria, C. Premebida, and U. Nunes, "A probalistic approach for human everyday activities recognition using body motion from RGB-D images," in *IEEE RO-MAN'14, * Kazuo Tanie Award Finalist*, 2014.

Fig. 6: Few snapshots of daily activities recognition experiments ('unseen" person) using a mobile robot.

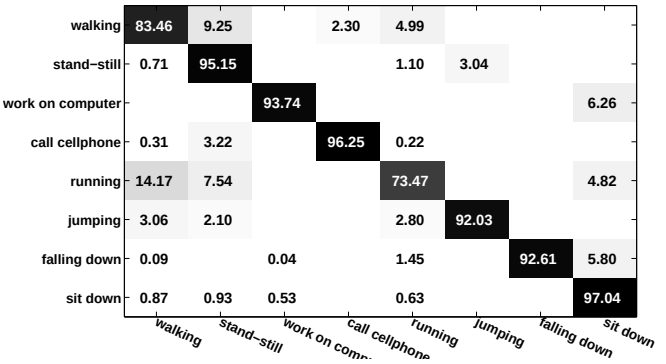


Fig. 7: DBMM on-the-fly classification confidence (average).

[4] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from a single depth image," in *IEEE CVPR*, 2011.

[5] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Learning actionlet ensemble for 3d human action recognition," in *IEEE Transactions on PAMI*, 2013.

[6] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Unstructured human activity detection from RGBD images," in *ICRA'12*, 2012.

[7] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from RGB-D videos," in *IJRR journal*, 2012.

[8] C. Zhang and Y. Tian, "RGB-D camera-based daily living activity recognition," in *J. of Comp. Vision and Image Proc.*, 2012.

[9] X. Yang and Y. Tian, "Effective 3d action recognition using eigenjoints," *J. of Visual Comm. and Image Repr.*, vol. 25, pp. 2–11, 2013.

[10] L. Piyathilaka and S. Kodagoda, "Gaussian mixture based HMM for human daily activity recognition using 3d skeleton features," in *IEEE 8th Conf. on Ind. Electronics and App.*, 2013.

[11] B. Ni, Y. Pei, P. Moulin, and S. Yan, "Multilevel depth and image fusion for human activity detection," *IEEE Trans. on Cybern.*, 2013.

[12] R. Gupta, A. Y.-S. Chia, and D. Rajan, "Human activities recognition using depth images," in *21st ACM Int. Conf. on Multimedia*, 2013.

[13] J. Luo, W. Wang, and H. Qi, "Group sparsity and geometry constrained dictionary learning for action recognition from depth maps," in *ICCV'13*.

[14] C. Chen, R. Jafari, and N. Kehtarnavaz, "Action recognition from depth sequences using depth motion maps-based local binary patterns," in *IEEE Winter Conf. on App. of Computer Vision (WACV)*, 2015.

[15] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM TIST*, 2011, http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[16] O. Oreifej and Z. Liu, "HON4D: Histogram of oriented 4d normals for activity recognition from depth sequences," in *IEEE CVPR*, 2013.

[17] E. Ohn-Bar and M. M. Trivedi, "Joint angles similiarities and HOG2 for action recognition." in *CVPRW*, 2013.

[18] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in *ACM International Conf. on Multimedia*, 2012.

[19] R. Chaudhry, F. Ofli, G. Kurillo, R. Bajcsy, and R. Vidal, "Bio-inspired dynamic 3d discriminative skeletal features for human action recognition," in *Comp. Vision and Pattern Rec. WS. (CVPRW)*, 2013.

[20] G. Evangelidis, G. Singh, and R. Horaud, "Skeletal quads: Human action recognition using joint quadruples," in *ICPR*, 2014.

[21] L. Xia and J. Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera," in *CVPR*, 2013.