A Probabilistic Approach for Human Everyday Activities Recognition using Body Motion from RGB-D Images

Diego R. Faria, Cristiano Premebida and Urbano Nunes

***Abstract*— In this work, we propose an approach that relies on cues from depth perception from RGB-D images, where features related to human body motion (3D skeleton features) are used on multiple learning classifiers in order to recognize human activities on a benchmark dataset. A Dynamic Bayesian Mixture Model (DBMM) is designed to combine multiple classifier likelihoods into a single form, assigning weights (by an uncertainty measure) to counterbalance the likelihoods as a posterior probability. Temporal information is incorporated in the DBMM by means of prior probabilities, taking into consideration previous probabilistic inference to reinforce current-frame classification. The publicly available Cornell Activity Dataset [1] with 12 different human activities was used to evaluate the proposed approach. Reported results on testing dataset show that our approach overcomes state of the art methods in terms of precision, recall and overall accuracy. The developed work allows the use of activities classification for applications where the human behaviour recognition is important, such as human-robot interaction, assisted living for elderly care, among others.**

I. Introduction

Human behaviour is an important issue in indoor environments namely for assistant and service robots applications. By exploring recent advances in human pose detection using an RGB-D sensors, many researches have been focused on activity recognition [2] [3] [4]. Works relying on an RGB-D sensors usually extract the human body silhouette and 3D skeleton from depth images for computing motion features. In [5], maximum entropy Markov model (MEMM) for human activities classification was adopted where features were modelled using a skeleton tracking system combined with Histogram of Oriented Gradient (HOG) [6]. In [7], each activity is modelled into sub-activities, while object affordances and their changes over time were used with a multi-class Support Vector Machine (SVM) for the classification. In [8], a bag of kinematic features was used with a set of SVMs for activity classification.

Other works on recognition of human activities focus their research on how to efficiently model the attributes 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.

Our research aims at developing artificial cognitive skills towards endowing a robot to identify human behaviours in order to cope and interact with humans. In this context, a robot that can recognize human everyday activities will be useful for assisted care, e.g., interacting with elderly people and monitoring them regarding strange or non-usual behaviours. We use a RGB-D sensor in order to perceive the environment. RGB-D data is used to generate a human 3D skeleton model with semantic matching of body parts linked by its joints. Based on this model, we extract periodical joints motion and distances to describe multi-classes of actions. This work thus brings contributions on human everyday activities recognition using a new method called Dynamic Bayesian Mixture Model (DBMM).

We incorporate temporal information in the classification model propagating previous information to reinforce the classification in the next time instant. This model is inspired in the well-known Dynamic Bayesian Network (DBN) modelling. DBMM allows the combination of multiple classifiers into a single form, assigning a weight (confidence level) given by an uncertainty measure (entropy) after analysing the previous behaviour of each single classifier. We will demonstrate that the proposed DBMM counterbalances single classifiers, achieving classification performance superior than benchmark methods.

The structure of the paper is as follows: Section II presents the background of the Bayesian Mixture Models and introduces the proposed approach; Section III presents the proposed feature models using the 3D skeleton and the learning stages; Section IV reports the obtained results; and Section V brings the conclusion and final remarks.

II. Probabilistic Classification Model

In order to increase classification performance on "unseen" human activity recognition, in this work we propose a combination of single classifiers through the DBMM. The concept of Bayesian Mixture Models (BMM) is used and integrated into a dynamic process that incorporates the temporal information (i.e., frame by frame classification). In our approach, a DBMM is learned in order to combine conditional probability outputs (likelihoods) from single classifiers. A weight is assigned to each classifier according to previous knowledge (learning process), using an uncertainty measure as confidence level.

Figure 1 depicts an overview of our approach (learning and classification steps) where single classifiers are joint and used as weighted posterior distributions in a designed general dynamic model to classify everyday activities.

This work was supported by the Portuguese Foundation for Science and Technology (FCT) under Grants PDCS10:PTDC/EEA-AUT/113818/2009 and AMS-HMI12: RECI/EEI-AUT/0181/2012. The authors are with Institute of Systems and Robotics, University of Coimbra, Polo II, 3030-290 Coimbra, Portugal (emails: diego, cpremebida, urbano@isr.uc.pt).

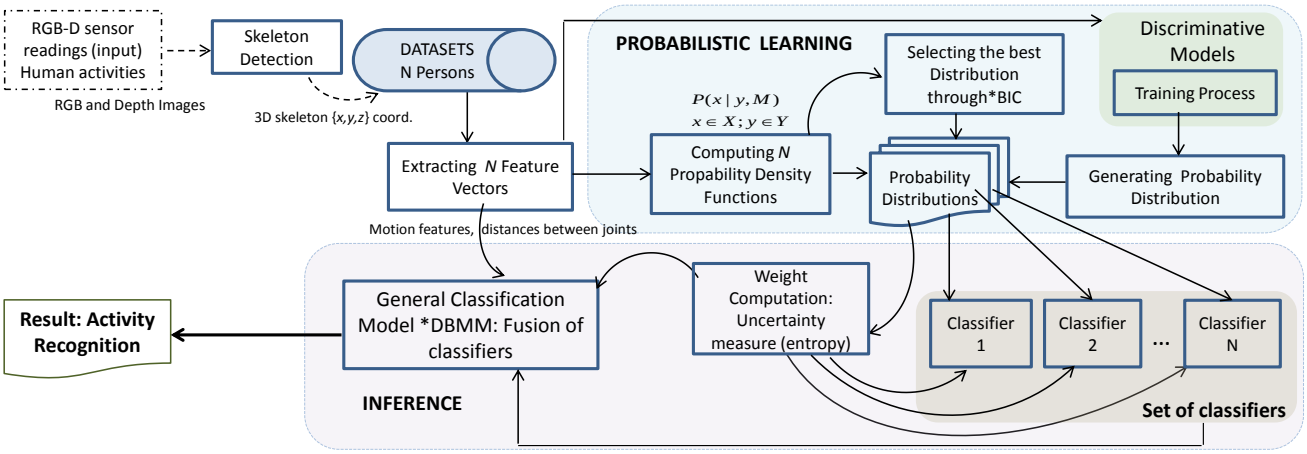


Fig. 1: Overview of our proposed approach for human everyday activities recognition using the proposed DBMM strategy for single classifiers combination.

A. Bayesian Mixture Models

Bayesian modelling has often been used for multimodal fusion [13]. Mixture models are known as distributions of parametric forms with multiple components where the probability distributions are commonly assumed Gaussian. Here, the mixture model allows the combination of heterogeneous classifier models into a single form. This way, a global classifier can be updated as a filter using the weighted mixture of classifiers. The general BMM is given as follows:

$$P(A) = \sum_{i=1}^{N} w_i \times P_i(A), \tag{1}$$

where N is the number of components (here represented by the number of classifiers); w_i is the weight of each Bayesian classifier output $P_i(A)$, and $\sum_{i=1}^{N} w_i = 1$.

The weights computation can be estimated from different ways. In this work, we propose the DBMM using an uncertainty measure as confidence level to weight multiple classifiers as detailed in sequel.

B. Proposed DBMM - Dynamic Bayesian Mixture Model

The DBMM is comprised of a set of models $A = \{A_n^1, A_n^2, ..., A_n^T\}$ where A_n^t is a model with n attributes, i.e., observed variables generated for some dynamic process at each time instant $t = \{1, 2..., T\}$. DBMM can be represented as $\Pi = (\varphi, \theta)$ with a model structure φ composed of n classifiers combined in a mixture model, and with the parameters θ. The DBMM has the following general probability distribution function for each class C:

$$P(C,A) = \prod_{t=1}^{T} P(C^t|C^{t-1}) \times \sum_{i=1}^{n} w_i \times P_i(A|C^t). \tag{2}$$

We assumed the process holds the Markov property (recursion) by taking the posterior of the previous time instant as the prior for the present time instant (i.e., a dynamic update). The formalization of the DBMM for a specific time instant is achieved by rewriting (2), then obtaining the general model for classification as follows:

$$P(C,A) = \beta P(C^t|C^{t-1}) \times \sum_{i=1}^{n} w_i \times P_i(A|C^t),$$
$$\text{with} \begin{cases} P(C^t|C^{t-1}) \equiv \frac{1}{C} \text{ (uniform)}, & t = 1 \\ P(C^t|C^{t-1}) = P(C^{t-1}|A), & t > 1 \end{cases}, \tag{3}$$

where:

- $P(C^t \mid C^{t-1})$ is the class transition probability distribution among class variables over time. A class C at time t (C^t) is conditioned to the class at time $t-1$ (C^{t-1}). This step describes the non-stationary behaviour of the process and is applied recursively, where the previous posterior of each class becomes the current prior, thus it can be seen as a reinforcement from $t-1$ to the current classification at t.
- $P_i(A^t)$ is a-posteriori result of each single classifier model $\psi_i \in \Psi$ at time t. In this work, $i = \{1,2,3\}$ representing three single classifiers.
- The weight w in the model is estimated using an Entropy-based confidence measure. Details are given in the following 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 avoiding the problem of numerical stability once continuous update of belief is done, i.e., product in each frame between the mixture model and prior, making that C escapes from a large decimal number with probability close to zero at the end of many multiplications.

C. Assigning Weights using Entropy

The Shannon entropy H [14] can be used as a measure of the uncertainty associated with a random variable. In the DBMM framework, H is adopted as a confidence level to update the global probabilistic model. The weights are obtained using the entropy value for each single classifier. In a Bayesian framework, each model contributes to the result of the inference in proportion to its probability. The mixture model is presented directly as weighted sums of the distributions, then the combination of different models into one can be obtained.

We can compute the entropy of the posterior probabilities previously observed as follows:

$$H(P_i(A)) = -\sum_i P_i(A)\log(P_i(A)), \tag{4}$$

where $P_i(A) = P(C|A)$ represents the conditional probability given the model of a specific classifier $\Psi = \psi_i, i = \{1,...,3\}$, computed for a class C given a set of features model $A = \{A_1, A_2, ..., A_n\}$. From the learning stage, a likelihood is given by a probability density function (*pdf*) $P(A|C)$.

Knowing H, the weights w for each classifier i is estimated by:

$$w_i = \frac{1 - \left(\frac{H_i}{\sum_{i=1}^{n} H_i}\right)}{\sum_i^n \left(1 - \left(\frac{H_i}{\sum_{i=1}^{n} H_i}\right)\right)}, \; i = \{1,...,n\}, \tag{5}$$

where w_i is the weight result for each one of the n possible classifiers; H_i is the current value of entropy resultant from (4) for each classifier. The denominator in (5) guarantees that $\sum_i w_i = 1$.

Given the confidence for each classifier that can be obtained by analysing the performance of each classifier after a period of time, the general model of classification will then have the knowledge of the most reliable belief, thus each classifier score will be smoothed by continuously multiplying the classification belief by the correspondent weight.

D. Single Classifier Models Integrated in the DBMM

The first classifier used in the DBMM is a naive Bayes (NB). Assuming the features are independent from each other given the class variable, thus different *pdf* (i.e., one for each feature model) is used, obtaining the following expression:

$$P(C_i|A) = \alpha P(C_i)\prod_{j=1}^{N} P(A_j|C_i), \tag{6}$$

where $\alpha = \frac{1}{\sum_i P(A|C_i)P(C_i)}$ is a normalization factor ensuring that the left side of the equation sums up to one over C_i; N is the number of independent feature models.

The second classifier used in the DBMM is a Bayesian classifier without the naive assumption and modelled by a mixture of Gaussian distributions (GMM). The GMM learning process uses the Expectation Maximization (EM) algorithm to estimate the parameters of each individual density function which attempts to find the maximum likelihood estimation of a parameter. A global parameter that needs to be set is the number of clusters k_{max}. An optimal k_{max} can be estimated by Minimum Description Length (MDL) penalty function.

Finally, the third classifier adopted in this work is a multi-class SVM with a linear kernel. To obtain proper probabilistic outputs, the SVM scores are converted into a distribution by using a Sigmoid function as follows:

$$y = \frac{1}{(1 + e^{-f(x)})}, \tag{7}$$

where $f(x)$ is the SVM output, and y is the normalized value between $[0,1]$.

III. 3D Skeleton-based Features and Learning Process

In this work, the features rely on existent relations between body parts to capture motions with meaningful characteristics of a person performing an activity. The features used for activities recognition are extracted only from the 3D skeleton.

Skeleton detection is made given the raw data containing the depth images, then the human skeleton is tracked using the SDK (Software Development Kit) for RGB-D sensor: the OpenNi's [15] skeleton tracker is used for obtaining the locations of the 15 joints of the human body.

A set of features $A = \{A_1, A_2, ..., A_n\}$ are then extracted from such skeleton as shown in Figure 2.

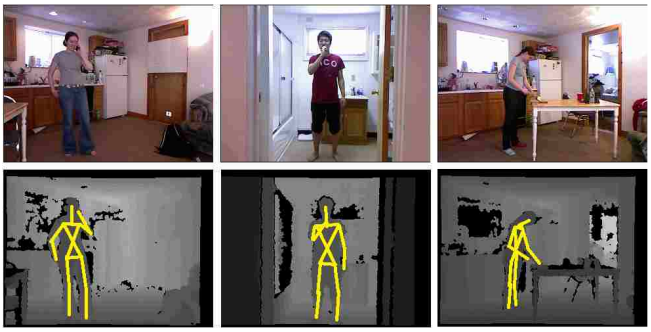


Fig. 2: Example of RGB-D images and the skeleton (Top row: RGB images; bottom: depth images with the skeleton in yellow). Fig. from the Cornell Activity Datasets [1].

We considered the skeleton frame of reference obtaining all joints relative to the torso centroid instead of using the sensor frame of reference. This step is applied for redundancy reduction in the data to better represent the features during an activity. This is done by defining the centroid of the torso as origin and computing the joints distances to the torso centroid.

Figure 3 presents an example of relative and absolute motion during an activity, as well as some types of features that we used in this work, distance between hands, distance between hands and face, distance between shoulder/hip and feet (stand or sit position), and changes in direction of the hands, elbows and head by computing the distances of the initial position of the member to the current position in $\{x, y\}$ directions.

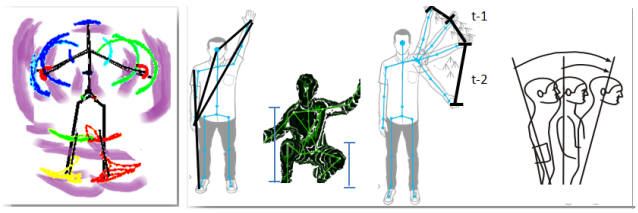


Fig. 3: An example of absolute (purple) and relative motion patterns from the skeleton joints (red, green, blue, yellow) is shown in the left, while examples of features are given in the right image: distances between body parts and torso inclination.

A total of 14 features to characterize 12 activities (as described in Section IV) are used as follows:

- The distances between hands and face, between the left and right hands, shoulders and feet, hip and feet, distance between the initial position of the hands at instant t_0 and the next frames are similarly computed using the Euclidean distance. Let $\{x,y,z\}$ be the 3D coordinates of some body member $b = \{f,h,w,s,f_e,t\}$ meaning the face, hands, waist/hip, shoulders, feet and torso receptively, which are given by the skeleton computation, where the index j denotes a specific joint of the 3D skeleton. All the distances are computed as follows:

$$\delta_{\{j_{b1},j_{b2}\}} = \sqrt{(j^x_{b1} - j^x_{b2})^2 + (j^y_{b1} - j^y_{b2})^2 + (j^z_{b1} - j^z_{b2})^2}. \quad (8)$$

- To find out if the two hands are close to the face at same time, we compute:

$$\delta_{\{j_{hh},j_f\}} = \sqrt{(j^x_{h1} - j^x_f)^2 + (j^y_{h1} - j^y_f)^2 + (j^z_{h1} - j^z_f)^2} + \\ + \sqrt{(j^x_{h2} - j^x_f)^2 + (j^y_{h2} - j^y_f)^2 + (j^z_{h2} - j^z_f)^2}. \quad (9)$$

  The smaller this value, the smaller the distance between the two hands and face.
- To compute the torso inclination, the initial distance between the shoulders to the feet is represented by $\delta_{\{j_s,j_{f_e}\}}$ as in (8), and consequently the difference to the consecutive frames is also computed, then we have:

$$\iota = \delta^{t=0}_{\{j_s,j_{f_e}\}} - \delta^{t}_{\{j_s,j_{f_e}\}}. \quad (10)$$

  Positive values of ι represents the torso inclination and the negative ones are the opposite direction.
- The difference between the initial hand position at time $t = 0$ (for left and right hands) and the consecutive frames, as well as the left and right elbows and the head in x and y coordinates are computed similarly as shown in (10). Thus, 10 feature vectors are acquired for the variations of hands, elbows and head in both, x and y coordinates.

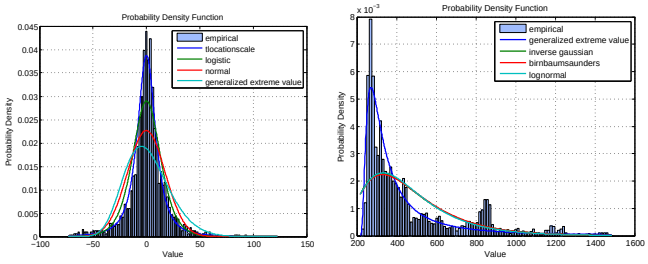
In case of modelling the features from a specific single hand to the face, to avoid misunderstanding of right or left-handed person, we extract the features by selecting from both hands the one that has more distance variations relative to the face position during the motion.

A. Learning by Fitting Probability Density Functions

For the NB Classifier, a set of valid parametric probability distributions were tested using the Bayesian Information Criterion (BIC) [16] for model selection, a score is assigned for each *pdf* describing the best distribution to represent the data. The list of distributions used to fit the data were: *Beta, Birnbaum-Saunders, Exponential, Extreme value, Gamma, Generalized extreme value, Generalized Pareto, Inverse Gaussian, Logistic, Log-logistic, Lognormal, Nakagami, Normal, Rayleigh, Rician, t location-scale, Weibull.*

Among all tested *pdf*, usually the three distributions selected were: *Generalized Extreme Value Distribution*, useful to model the smallest or largest value among a large set of random values allowing a continuous range of possible shapes; *t Location-Scale Distribution*, useful for modelling data distributions with heavier tails than the normal distribution; and *Weibull Distribution*, positive only for positive values of x, zero otherwise.

Figure 4 shows some examples of probability distribution selection given different feature vectors (changes in direction of the right hand and torso inclination). These distributions were acquired for different activities, namely rinsing water and wearing contact lens.



(a) t Location-scale (b) Generalized Extreme Value

Fig. 4: Examples of probability distribution selection for the NB classifier using the modelled features. The left image presents the right hand change in directions and the image at right side shows the torso inclination. The values are in *mm*.

The learning process for the second classifier is obtained by using all features vector as a multidimensional information through the GMM learning.

The entire set of GMM parameters is denoted as $\theta = \{(w_j,\mu_j,\Sigma_j)\}^k_j$, where μ_j represents the mean of a specific cluster j, the covariance matrix is represented by Σ_j, and w_j represents the weight of the cluster, which specifies how likely each Gaussian is selected. The EM algorithm is used to estimate the set of GMM parameters θ (input) and any μ_j,Σ_j, is denoted as Gaussian according to the following expression:

$$\phi(\mathbf{x}|\mu_j,\Sigma_j) \triangleq \\ \triangleq \frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_j)^T\Sigma_j^{-1}(\mathbf{x}-\mu_j)\right). \quad (11)$$

The *pdf* for the combination of the k models to search for the most likely combination θ of models to explain the observed data is obtained by (12). This means a learning of mixture models, so that we are searching for the combination of the proper clusters that better describes the input data, achieving the subset of feature values $\mathbf{x}_i$, representing the proper cluster j, where $j = \{1,...,k\}$.

$$P(\mathbf{x}_i|\theta) = \sum_{j=1}^{k} w_j \phi(\mathbf{x}_i|\mu_j,\Sigma_j), \quad (12)$$

where $w_j > 0$, $\sum_k^j w_j = 1$ and $\theta = \{(w_j,\mu_j,\Sigma_j)\}^k_j$. More details about the theory and use of the EM algorithm and the GMM learning can be found in [17].

Finally, the third classifier is a linear-kernel SVM that has been implemented using the LibSVM package [18]. Normalization was applied to the features set in such a way that the values of minimum and maximum obtained during the training stage were applied on the testing set. The SVMs were 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.

IV. Experimental Results

A. Human Everyday Activities Dataset

Our approach was evaluated using the publicly available Cornell Activity Datasets: CAD-60 [1] [5]. This dataset comprises video sequences of human everyday activities acquired from a RGB-D sensor. There are 12 human everyday activities performed by 4 different subjects (two male and two female, one of them being left-handed) in 5 different environments: office, kitchen, bedroom, bathroom, and living room. The 12 activities are: rinsing mouth, brushing teeth, wearing contact lens, talking on the phone, drinking water, opening pill container, cooking (chopping), cooking (stirring), talking on couch, relaxing on couch, writing on whiteboard, working on computer. Additionally, and for generalization purposes, the CAD-60 dataset has two more activities (random and still) which are used only during classification performance on testing sets.

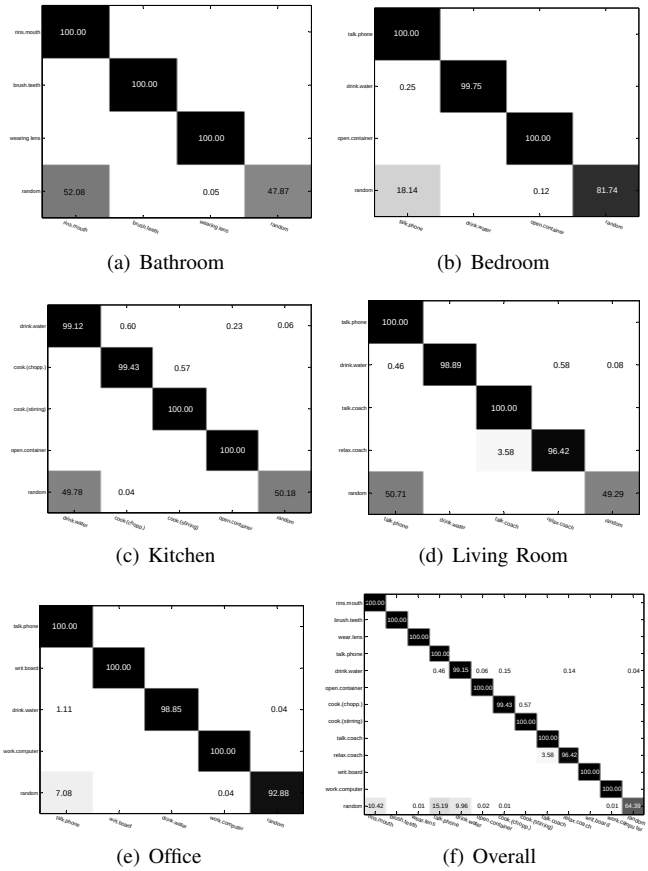
Table I summarizes the dataset information, i.e., number of frames performed by each person for each activity of this dataset in different scenarios.

TABLE I: CAD-60 dataset: number of frames performed by each person for the 12 activities in this dataset divided into five different scenarios.

Location	Activity	Person 1	2	3	4	Total
Bathroom	rinsing mouth	1746	1446	1503	1865	6560
	brushing teeth	1351	1675	1783	1580	6389
	wearing lens	835	1415	822	1100	4172
	random+still	2962	2352	3007	3063	11384
Bedroom	talk. on phone	1525	830	1288	1308	4951
	drinking water	1587	778	1310	1529	5204
	opening container	749	963	621	1012	3345
	random+still	2962	2352	3007	3063	11384
Kitchen	cook. chopping	1565	1664	1754	1910	6893
	cook. stirring	1346	1349	1467	1835	5997
	drinking water	1587	778	1310	1529	5204
	opening container	749	963	621	1012	3345
	random+still	2962	2352	3007	3063	11384
Living room	talk. on phone	1525	830	1288	1308	4951
	drinking water	1587	778	1310	1529	5204
	talk. couch	1681	1539	1712	1812	6744
	relax. couch	1447	1497	1379	1853	6176
	random+still	2962	2352	3007	3063	11384
Office	talk. on phone	1525	830	1288	1308	4951
	writ. board	1792	1637	1597	1792	6818
	drinking water	1587	778	1310	1529	5204
	work. computer	1265	1530	1222	1662	5679
	random+still	2962	2352	3007	3063	11384

B. Classification Results

As long as the CAD-60 dataset brings the activities for each scenario, we are also using the same strategy [5] used by all the approaches reported in [1]. We will present the classification results in terms of Precision (Prec), Recall (Rec) and confusion matrix for each scenario and overall (shown in Fig.5). The assessment criteria was done adopting the leave-one-out cross validation test. The idea is to verify the capacity of generalization of the classifier by using the strategy of "new person", i.e, learning from different persons and testing with an unseen person. The classification is made for each individual frame to account for the accuracy of the frames correctly classified.



(a) Bathroom (b) Bedroom

(c) Kitchen (d) Living Room

(e) Office (f) Overall

Fig. 5: Classification Results: Leave-one-out cross-validation confusion matrix for each scenario and the overall confusion matrix using the DBMM for the "new person" setting.

Figure 5 shows the classification results where the last column of each confusion matrix has the random activity as neutral class, enclosing the activities that were not classified with a high confidence, thus when the classification is smaller than 0.5, even with correct classification, we set as neutral. This is done to show the confidence characteristic of our approach. For each scenario we have added a new activity (random+still movements) as presented in the last row of the confusion matrices. Notice that, the random activity was not trained, but tested against the other trained activities expecting the random+still movements will be set in the last column (random).

The results show that using our designed ensemble, we obtained improvements in the classification compared with

other state of the art methods presented in [1]. Actually, our results overcame all works presented in the ranked table (precision and recall rates) at the CAD-60 website up to the current date. The overall measure obtained in this work (taking the non-trained random activity into consideration to calculate the precision and recall rates) were: global precision 91.26% and recall 89.56%. For comparison purposes, Table II summarizes the results from single classifiers and a simple averaged ensemble compared with the proposed DBMM for the living room (scenario with more misclassification), demonstrating that our approach outperformed the other classifiers. Table III presents the classification rates for the "new person" tested in each scenario. Finally, Table IV shows our approach compared with other state of the art methods that used the CAD-60. This table shows only some selected works (the ones with higher precision) up to date.

TABLE II: Results on the living room scenario of CAD-60 dataset ("new person") using single classifiers, a simple averaged ensemble (AV) and the proposed DBMM.

Activities: 1-talk.on phone; 2-drink.water; 3-talk.couch; 4-relax.couch.

Location	Act.	SVM	Bayes	NB	AV	**DBMM**
Liv.Room	1	96.5%	92.8%	44.9%	78.1%	100%
	2	92%	82.1%	71.4%	81.8%	98.9%
	3	99%	98.3%	100%	99.1%	100%
	4	82.9%	85.6%	75%	81.1%	96.4%
Average:		92.6%	89.7%	72.8%	85%	98.8%

TABLE III: Performance on the CAD-60 testing dataset ("new person"). Results are reported in terms of Precision (Prec) and Recall (Rec).

Location	Activity	**DBMM** Prec	Rec
Bathroom	rinsing mouth	52.53 %	100.00 %
	brushing teeth	99.86 %	100.00 %
	wearing lens	99.95 %	100.00 %
	random+still	100.00 %	47.87 %
	Average	88.10 %	86.97 %
Bedroom	talking on phone	70.44 %	100.00 %
	drinking water	100.00 %	99.75 %
	opening container	99.58 %	100.00 %
	random+still	100.00 %	81.74 %
	Average	92.50 %	95.37 %
Kitchen	cooking chopping	47.65 %	99.12 %
	cooking stirring	99.49 %	99.43 %
	drinking water	99.35 %	100.00 %
	opening container	99.64 %	100.00 %
	random+still	99.95 %	50.18 %
	Average	89.22 %	89.75 %
Living room	talking on phone	46.06 %	100.00 %
	drinking water	100.00 %	98.89 %
	talking on couch	96.83 %	100.00 %
	relaxing on couch	99.50 %	96.42 %
	random+still	99.93 %	49.29 %
	Average	88.46 %	88.92 %
Office	talking on phone	85.14 %	100.00 %
	writing on whiteboard	100.00 %	100.00 %
	drinking water	100.00 %	98.85 %
	working on computer	99.91 %	100.00 %
	random+still	99.98 %	92.88 %
	Average	97.01 %	98.34 %
	Overall Average	**91.06** %	**91.87** %

TABLE IV: Comparison of methods that used the CAD-60.

Method	Prec.	Rec.
Proposed **DBMM**	91.1%	91.9%
Zhang *et al*. [8]	86%	84%
Koppula *et al*. [7]	80.8%	71.4%
Gupta *et al*. [12]	78.1%	75.4%
Ni *et al*. [11]	75.9%	69.5%
Yang *et al*. [9]	71.9%	66.6%
Piyathilaka *et al*. [10]	70%	78%
Sung *et al*. [5]	67.9%	55.5%

V. Conclusion

A probabilistic approach, named DBMM, for activities recognition using 3D skeleton features from RGB-D images was proposed. DBMM combines multiple classifiers into a designed dynamic model given a confidence level for each single classifier. An uncertainty measure to weight each single classifier during a learning phase is computed and afterwards the general classification model compensates the probability outputs into a single distribution form. This weighting strategy demonstrated to be very effective given a set of suitable feature models. The CAD-60 dataset was used to evaluate the performance of our approach. Results show that our approach is suitable for activities recognition. The classification performance overcame other state-of-the-art methods ranked at the CAD-60 website. Future work will address exploitation of our approach in other applications.

References

[1] "Cornell activity datasets CAD-60. Web: http://pr.cs.cornell.edu/humanactivities/data.php, acessed on 02/2014."

[2] 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 *CVPR'11*, 2011.

[3] B. Ni, P. Moulin, and S. Yan., "Order-preserving sparse coding for sequence classification," in *ECCV'12*, 2012.

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

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

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection." in *CVPR'05*, 2005.

[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] F. Colas, J. Diard, and P. Bessire, "Common bayesian models for common cognitive issues," *Acta Biotheoretica*, vol. 58, 2010.

[14] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley & Sons, 1991, iSBN-13: 978-0471062592.

[15] "Openni sdk, Website: http://www.openni.org/."

[16] G. E. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, no. 2, p. 461464, 1978.

[17] M. R. Gupta and Y. Chen, "Theory and use of the EM algorithm," *Found. and Trends in Signal Proc.*, vol. 4, no. 3, pp. 223–296, 2011.

[18] 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.