2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Daejeon Convention Center
October 9-14, 2016, Daejeon, Korea

Social Activity Recognition based on Probabilistic Merging of Skeleton Features with Proximity Priors from RGB-D Data

Claudio Coppola[1], Diego R. Faria[2,3], Urbano Nunes[2] and Nicola Bellotto[1]

Abstract— **Social activity based on body motion is a key feature for non-verbal and physical behavior defined as function for communicative signal and social interaction between individuals. Social activity recognition is important to study human-human communication and also human-robot interaction. Based on that, this research has threefold goals: (1) recognition of social behavior (e.g. human-human interaction) using a probabilistic approach that merges spatio-temporal features from individual bodies and social features from the relationship between two individuals; (2) learn priors based on physical proximity between individuals during an interaction using proxemics theory to feed a probabilistic ensemble of activity classifiers; and (3) provide a public dataset with RGB-D data of social daily activities including risk situations useful to test approaches for assisted living, since this type of dataset is still missing. Results show that using the proposed approach designed to merge features with different semantics and proximity priors improves the classification performance in terms of precision, recall and accuracy when compared with other approaches that employ alternative strategies.**

I. Introduction

In recent years, there has been a growing interest in recognizing social behavior. The main focus from the psychology side is on understanding how people's thoughts, feelings, and behaviors are influenced by the actual, imagined, or implied presence of others and also the way humans are influenced by ethics, attitudes, culture, etc. From the robotics side, roboticists try to use this knowledge to model and design robots with capabilities not only to recognize human behavior, but also to interact with humans in different contexts to serve as assistants. Active & Assisted Living (AAL) is becoming a central focus for robotics research since there is a drastic increase of aging population. Robots could be used to improve the quality of life for those people by assisting them in their daily life or detecting anomalous situations. In this context, human activity recognition plays a central role in identifying potential problems to apply corrective strategies as soon as possible. In particular, a robot that is able to analyze the daily social interaction between humans, can also detect dangerous situations such as identification of social problems, aggression, etc. Due to the aforementioned reasons, big effort has been made for creation of datasets with RGB-D data [1], [2], [3] and development of approaches for recognition of Activities of Daily Living (ADL) [4], [5]. In [6], a simple way to apply a qualitative trajectory calculus to model 3D movements of the tracked human body using hidden Markov models (HMMs) is presented. Faria *et al*. [7], [8] have proposed a probabilistic ensemble of classifiers called Dynamic Bayesian Mixture Model (DBMM) to combine different posterior probabilities from a set of classifiers with interesting performance on datasets. A biologically inspired approach adopting artificial neural network to combine pose and motion features for action perception is proposed in [9]. The approach presented in [10] uses HMMs combined with Gaussian Mixture Models (GMM) to model the multimodality of continuous joint positions over time for activity recognition. All the aforementioned works have in common the fact that they attempt to recognize daily activities from one individual performing an activity or interacting with some object during the activity. Nowadays, publicly available RGB-D datasets for ADL present only one subject performing the activities. In this work, we are going further, focusing on social interaction between two subjects, since this topic is still challenging in robotics and when it comes to RGB-D data, it is still little explored.

Approaches based on other types of sensors (one or a network of monocular cameras or IMUs) can be found in the literature for social interaction analysis. However when IMUs are used in social interaction, most of datasets analyze only one individual using wearable technology. In [11], the authors show a model based on the orientation of the lower part of the body to recognize conversational groups. In [12], the authors resort to Laban Movement Analysis (LMA) to recognize the social role of a human in a social interaction. In [13], proxemics theory is adopted to define qualitative features for social behavior. In this work, we also take support from proxemics theory, which was introduced by Edward T. Hall [14] to associate proximity features with social space surrounding a person as a key-feature to study the effect of distance on communication and how the effect varies between cultures and other environmental factors. The space is divided into intimate, personal, social and public spaces. In robotics, this is a topic that was carried out by [15], [16], [17], yet in a simpler way, using only the concept of defined distances based on thresholds observed from social science. Differently from others, our approach extracts proximity-based features learned from social interaction as prior for the recognition module.

There are three main contributions in this paper: (i) A

This work has been supported by: Santander in the form of an International Travel Grant, by the European project: ENRICHME[1], EC H2020 Grant Agreement No. 643691, and by the FCT project AMS-HMI12: RECI/EEI-AUT/0181/2012, COMPETE[2]. Claudio Coppola and Nicola Bellotto are with [1]L-CAS, School of Computer Science, University of Lincoln (UK); Diego Faria is with [3]System Analytics Research Institute, School of Engineering and Applied Science, Aston University (UK); and Urbano Nunes is with [2]Institute of Systems and Robotics, Department of Electrical and Computer Engineering, University of Coimbra, Portugal. (emails: {ccoppola, nbellotto}@lincoln.ac.uk; d.faria@aston.ac.uk; urbano@isr.uc.pt).

probabilistic approach that merges spatio-temporal features from individual bodies and social features from the relationship between two individuals for social activity recognition; (ii) Learned priors using physical proximity between individuals during an interaction based on proxemics theory to feed the probabilistic classification model; (iii) Public social activity dataset with RGB-D data that will be useful to test approaches for assisted living scenarios.

The remainder of this paper is organized as follows. Section II introduces the models that we are based on. Section III and IV introduces our approach, detailing how we extended previous works to be adapted for social activity recognition. Section V presents the performance of our approach, and finally, Section VI presents the conclusion and future work.

II. Preliminaries: Classification Background

This section brings a brief review of the Dynamic Bayesian Mixture Model (DBMM), which was first proposed by Faria *et al* [7] for individual activity recognition, also employed in other classification contexts [8], [18], [19], [20], [21]. This background section aims to facilitate the understanding of the next section that will introduce our re-design of the classification model as an extension in order to allow the fusion of multiple set of features with different semantics as multiple mixture and incorporating the learned prior from proximity features.

A. Dynamic Bayesian Mixture Model

The DBMM [7] is a probabilistic ensemble of classifiers that was designed based on a dynamic Bayesian network (DBN) with the concept of mixture model to fuse different classifier outputs, also adding temporal information through time slices. The DBMM as a DBN representation with different time slices can be obtained by employing the Markov property for a finite set of priors and mixture models. The random variable A (e.g. feature model for a specific classifier) is considered to be independent on previous A-nodes: $P(A^t|A^{t-1},C^t,C^{t-1}) = P(A^t|C^t,C^{t-1})$, where C represent a set of possible classes (e.g. activities). The nodes are not conditionally dependent of future nodes e.g., $P(A^{t-2}|C^t,C^{t-1},C^{t-2}) = P(A^{t-2}|C^{t-2})$. As a consequence, the transition probabilities between classes reduces to the probability of the current-time class $P(C^t) = P(C^t|C^{t-1})$. Knowing that $P(A^t|C^t)$ is a mixture of probabilities, then the explicit expression for the DBMM with T time slices assumes the following form as shown in [20]:

$$P(C^t|C^{t-1:t-T},A^{t:t-T}) = \frac{\prod_{k=t}^{t-T}\left(\sum_{i=1}^{n} w_i^k \times P_i(A^k|C^k)\right)\times P(C^k)}{\sum_{j=1}^{nc}\left[\prod_{k=t}^{t-T}\left(\sum_{i=1}^{n} w_i^k \times P_{i,j}(A^k|C^k)\right)\times P_j(C^k)\right]}, \quad (1)$$

where n is number of classifiers; nc is the number of classes; w is the weight for each base classifier learned from the training set. In this work, (1) can be simplified for a single time slice $T = 1$ as follows:

$$P(C^t|A^t) = \frac{P(C^t)\times\sum_{i=1}^{n} w_i^t \times P_i(A^t|C^t)}{\sum_{j=1}^{nc} P_j(C^t|C^{t-1})\times\left(\sum_{i=1}^{n} w_i^t \times P_{i,j}(A^t|C^t)\right)}, \quad (2)$$

where $P(C^t|A^t)$ is the posterior probability; the prior assumes the form $\forall t > 1, P(C^t) = P(C^t|C^{t-1})$, otherwise, $t = 1$, $P(C^t) = 1/nc$ (uniform); $P_i(A^t|C^t)$ is the likelihood model in the DBMM as the posterior probability of a base classifier; and the mixture model is obtained by $mix = w_i^t \times P_i(A^t|C^t)$. Each weight w_i, $i = \{1,2,...,n\}$ is learned using entropy-based confidence for each base classifier based on their performance in the training set as shown in [7].

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. For the linear-kernel multiclass SVM implementation, we adopted the LibSVM package [22], 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 feedforward network, where the hidden layer transfer function is a hyperbolic tangent sigmoid and a normalized exponential (*softmax*) for the output as posterior probability estimates.

III. Social Activity Recognition

In this section we describe a proposed strategy to combine multiple set of features as individual entities - one for each mixture in the model - to be merged into the DBMM. Figure 1 depicts a flowchart of the proposed modified structure in the DBMM. Basically, for each set of features with different semantics (i.e. one representing an individual with active or non-active role during the social interaction, another representing social features), we employ multiple base classifiers conditioned to this specific set of features representing a mixture that will feed the final fusion. Each mixture adopts entropy-based weighting as shown in [7] and afterwards the resulting posterior will have a new weight assigned to it based on the normalization of the outputs from each mixture. Once the most probable class for each mixture model is known based on the higher posterior probability, then in order to quantify the uncertainty of each mixture, a simple normalization is employed using these posteriors to obtain the weights $w_{mix_y}^t$ for the final fusion as follows:

$$w_{mix_y}^t = \frac{P_y(C^t|A^t)}{\sum_{y=1}^{N} P_y(C^t|A^t)}, \quad (3)$$

where $P_y(C^t|A^t)$ is the posterior probability of the y^{th} mixture model (herein, with a total of 3 mixtures, one for each set of features: individual 1, individual 2 and social features); t is an index for each time instant.

Given a set of mixture models and their computed weights (3), we reformulate the DBMM presented in (2) as the new classification strategy for social activity recognition assigning weights and prioritizing the features set with higher confidence as follows:

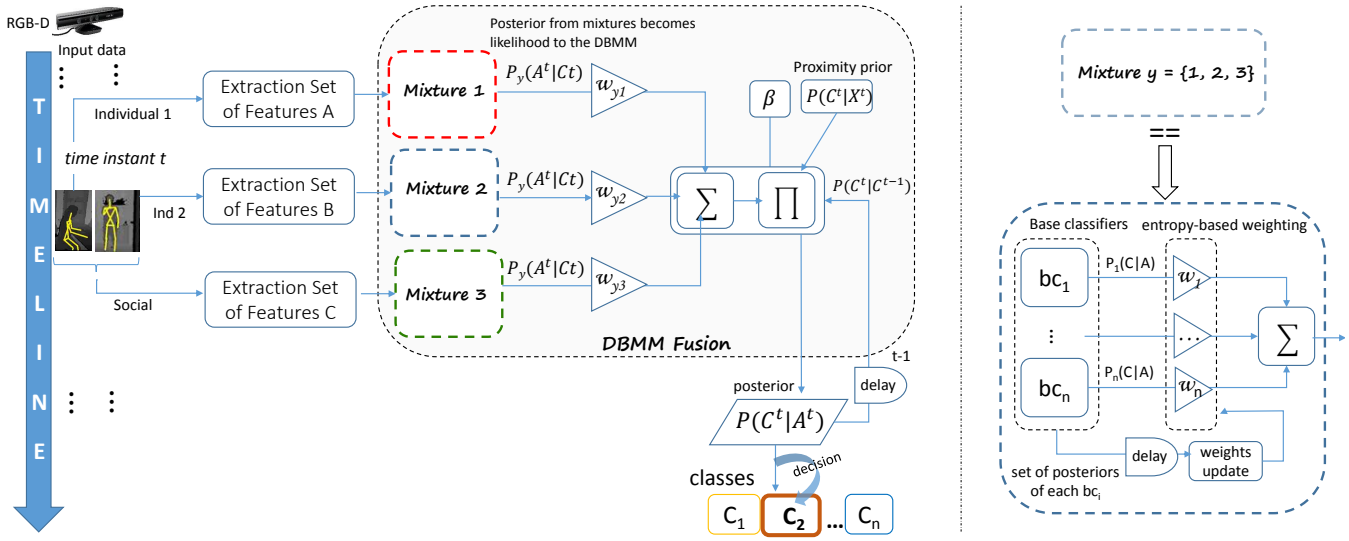


Fig. 1: Proposed multiple mixtures to feed the DBMM fusion based on different set of features for social activity recognition.

$$P(C^t|A^t,X^t) = \beta \times \underbrace{P(C^t|C^{t-1})}_{\text{dynamic transitions}} \times \underbrace{P(C^t|X^t)}_{\text{proxemity prior}} \times \underbrace{\sum_{y=1}^{m}\left[w^t_{mix_y} \times \left(\sum_{i=1}^{n} w^t_{i,y} \times P_{i,y}(A|C^t)\right)\right]}_{\text{fusion of multiple mixtures}}, \tag{4}$$

where $w^t_{mix_y}$ is the weight for fusion of each y^{th} mixture model; $w^t_{i,y}$ is the weight for each i^{th} base classifier in a mixture model y; $P(C^t|X^t)$ is the learned priors from the dataset based on proxemics, with X representing the proximity between skeletons; and the normalization factor $\beta = \frac{1}{\sum_{j=1}^{n}\left(P(C^t_j|C^{t-1}_j)\times P(C^t_j|X^t_j)\times\sum_{y=1}^{m}\left[w^t_{mix_y}\times\left(\sum_{i=1}^{n} w^t_{i,y}\times P_{i,y}(A_j|C^t_j)\right)\right]\right)}$.

IV. Features Extraction and Proximity Priors for Social Activity Recognition

This section describes the steps to obtain a set of features extracted from skeleton data, e.g. both subjects individually, and also a set of features from the relationship between both skeletons interacting. Before features extraction, a moving average filter with five neighbors was applied on the raw skeleton data to smooth some noise.

A. Spatio-Temporal Features from Individual Skeleton Data

In order to characterize the labeled social activity dataset using body posture and movements, we have exploited skeleton spatio-temporal features from both individuals performing a social interaction individually. To do so, we follow the features extraction step for single daily activity developed in [7], [8], since these features have been successful used in human daily activity recognition. Thus, 51 spatio-temporal features were used, such as: Euclidean distances between joints; angles formed between joints (e.g. shoulders, elbows and hands, and hips knees and feet from left and right side); torso inclination; joints velocities; energy of upper body joints velocities; log-energy entropy over the skeleton joints; auto-correlation between skeleton poses; some statistics such as mean and standard deviation over joints distances. More details can be found in [8]. However, we aim to gain in performance, herein we are also using a new set of features based on log-covariance of the joints distances of a body pose. We first built a matrix D with distances computed between all joints of a skeleton S with indexes $i, j = \{1,...,15\}$. Since we have fifteen joints represented by 3D coordinates $\{x,y,z\}$, we computed the 3D Euclidean distance among all joints, resulting a 15×15 matrix. We removed the null diagonal, due to the distances between the same joints are zero, obtaining a 15×14 matrix. Subsequently, we applied the log-covariance and we kept the upper triangular elements as features as follows:

$$\mathbf{M}_{lc} = \mathbf{U}(\log(\text{cov}(\mathbf{M}))), \tag{5}$$

where the covariance for each element in $\mathbf{M}$ is given by $cov_{ij} = \text{cov}(\mathbf{M}) = \frac{1}{N}\sum_{k=1}^{N}(\mathbf{M}^{ik} - \mu_i)(\mathbf{M}^{kj} - \mu_j)$; $\log(\cdot)$ is the matrix logarithm function (logm) and $\text{U}(\cdot)$ returns the upper triangular matrix elements composed of 120 features. The rationale behind the log-covariance is the mapping of the convex cone of a covariance matrix to the vector space by using the matrix logarithm. A covariance matrix forms a convex cone, so that it does not lie in Euclidean space, e.g., the covariance matrix space is not closed under multiplication with negative scalars. A total of 171 features per frame for each individual skeleton was obtained.

B. Social Features: Skeletons Proximity over Time

We define here social features as the ones that describe the relationship between two skeletons based on physical proximity, i.e. inter-bodies distance during the interaction. This set of features encompasses different subsets of features as follows:

- As first subset of features, the 3D Euclidean distance among all joints of an individual skeleton to the other individual skeleton corresponding joints were computed:

$$\delta(S1_{1,\ldots,15}, S2_{1,\ldots,15}) = \\ = \sqrt{(S1_i^x - S2_i^x)^2 + (S1_i^y - S2_i^y)^2 + (S1_i^z - S2_i^z)^2}. \quad (6)$$

  With the resulting 15×15 matrix, we have computed the log-covariance and kept the upper triangular elements as shown in (5), obtaining 120 features from this step.
- The second subset of features consists of two features that were computed by considering the minimum 3D Euclidean distance among all joints from individual one to the torso of individual two, and vice-versa: $d1_{min} = \min(\delta(S1_i, S2_{torso}))$ and $d2_{min} = \min(\delta(S2_i, S1_{torso}))$.
- The third subset of features consists of three features that helps to figure out the most active person (i.e., the one is approaching the other individual space). The first one is the computed distance from torso to torso $\delta(S1_{torso}, S2_{torso})$. The other two features were obtained from the energy over the 3D euclidean distances from all joints of skeleton one to the torso of skeleton two and vice-versa as follows:

$$ed_{1_{\{S1,S2\}}} = \sum \mathbf{v}_1^2 \quad \text{and} \quad ed_{2_{\{S2,S1\}}} = \sum \mathbf{v}_2^2, \\ \text{with} \quad \mathbf{v} = \delta(Si_{\{1,\ldots,15\}}, Si_{torso}), \quad i = \{1,2\}. \quad (7)$$

- The fourth subset has 120 features that were computed similarly to the first subset, however, in a temporal way. The same steps are computed for time instant t and $t-1$ regarding the Euclidean distances among the joints of skeleton one to the corresponding joints of skeleton two, $\delta(S1_{\{1,\ldots,15\}}, S2_{\{1,\ldots,15\}})^t$ and $\delta(S1_{\{1,\ldots,15\}}, S2_{\{1,\ldots,15\}})^{t-n}$, where n is a temporal window, herein defined as 10 frames. Following this step, we computed the difference between them, $\mathbf{r} = \delta(S1_{\{1,\ldots,15\}} S2_{\{1,\ldots,15\}})^t - \delta(S1_{\{1,\ldots,15\}} S2_{\{1,\ldots,15\}})^{t-n}$ as input for the log-covariance, getting the upper triangular elements from that.

From all subsets, we have acquired 245 social features per frame given both skeletons interacting.

C. Features Normalization

Normalization, standardization or filtering may be a requirement for many machine learning estimators, as they can behave badly if these steps are not applied to the features set. Working on features space, a normalization step was applied in such a way that the values of minimum and maximum obtained during the training stage for each type of feature were used to normalize the training and test set as follows:

$$\mathbf{F}_{tr}^{set_k} = \frac{\mathbf{F}_{tr}^{set_k} - \min(\mathbf{F}_{tr}^{set_k})}{\max(\mathbf{F}_{tr}^{set_k}) - \min(\mathbf{F}_{tr}^{set_k})}, \quad (8)$$

$$\mathbf{F}_{te}^{k} = \frac{\mathbf{F}_{te}^{set_k} - \min(\mathbf{F}_{tr}^{set_k})}{\max(\mathbf{F}_{tr}^{set_k}) - \min(\mathbf{F}_{tr}^{set_k})}, \quad (9)$$

where $\mathbf{F}_{tr}$ and $\mathbf{F}_{te}$ are the training and test sets respectively; set_k represents a feature type (i.e., column in the training and test set matrices).

D. Learning Priors based on Proximity Features

In this work we use the proxemics theory assuming that certain types of social interactions happens in specific social space based on distances. Although these social spaces are not unique for every person and are not easy to define, we try to learn this information by extracting distance features between two individuals during the interaction. To do so, given the skeletons data with 3D coordinates of body joints, we compute 3D Euclidean distances between the two skeletons for each social interaction. The features set encloses seven distances: torso to torso distance, $\delta(S1_{torso}, S2_{torso})$; the minimum joint distance of individual one to the torso of the individual two, $d1_{min} = \min(\delta(S1_{\{1,\ldots,15\}}, S2_{torso}))$ and vice-versa, $d2_{min} = \min(\delta(S2_{\{1,\ldots,15\}}, S1_{torso}))$; similarly as the latter, however, with maximum distance, $d1_{max} = \max(\delta(S1_{\{1,\ldots,15\}}, S2_{torso}))$ and $d2_{max} = \max(\delta(S2_{\{1,\ldots,15\}}, S1_{torso}))$; and the minimum and maximum joint to joint distances between both individuals, $d12_{min} = \min(\delta(S1_{\{1,\ldots,15\}}, S2_{\{1,\ldots,15\}}))$ and $d12_{max} = \max(\delta(S1_{\{1,\ldots,15\}}, S2_{\{1,\ldots,15\}}))$.

In order to learn the priors given the set of features for each activity using the training set, we resorted to a multivariate Gaussian distribution. The estimation of the parameters is based on mean and covariance matrix. Once we computed the distribution for each activity, then, during the test set, we have extracted these proximity features followed by a fitting step to the learned distribution. This fitting probability is used as likelihood to a recursive Bayesian model to predict from the test set the most probable activity. This posterior probability will be the prior to the DBMM classification. The rationale is that, given the test set, based on the observed distance between the individuals during the interaction, we have an estimation/guess about the activity based on the social space between them, as a prior knowledge to our proposed approach.

The Bayesian update of the priors is an estimation performed at every 30 frames (1 second), in order to get more variance during the interaction. It is given by $P(a^t|x^t) = \frac{P(x^t|a^t)P(a^t)}{\sum_j P(x_j^t|a_j^t)P(a_j^t)}$, where $P(a^t|x^t)$ is the posterior of each frame for an activity a, given the proximity features set x. The initial prior of each class is defined as uniform, $1/nc$, where nc is the number of classes, and afterwards the last posterior of the Bayesian update is used as prior to the next frame classification, $P(a^t) = P(a^t|a^{t-1})$. The fitting process given the proxemics features from the test set is obtained for each activity, which is used as likelihood to the Bayesian update:

$$P(x^t|a^t) = \phi(x_i|\mu_j, \Sigma_j)_j^k \triangleq \\ \triangleq \frac{1}{(2\pi)^{d/2}|\Sigma|} \exp\left(-\frac{1}{2}(x_i - \mu)^T \Sigma^{-1}(x_i - \mu)\right). \quad (10)$$

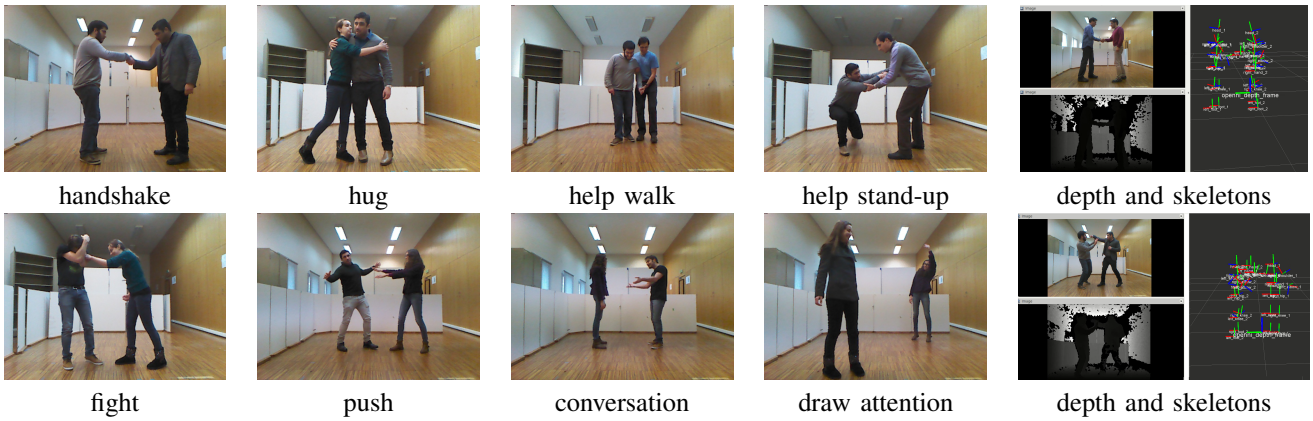


Fig. 2: Samples of the 3D Social Activity Dataset.

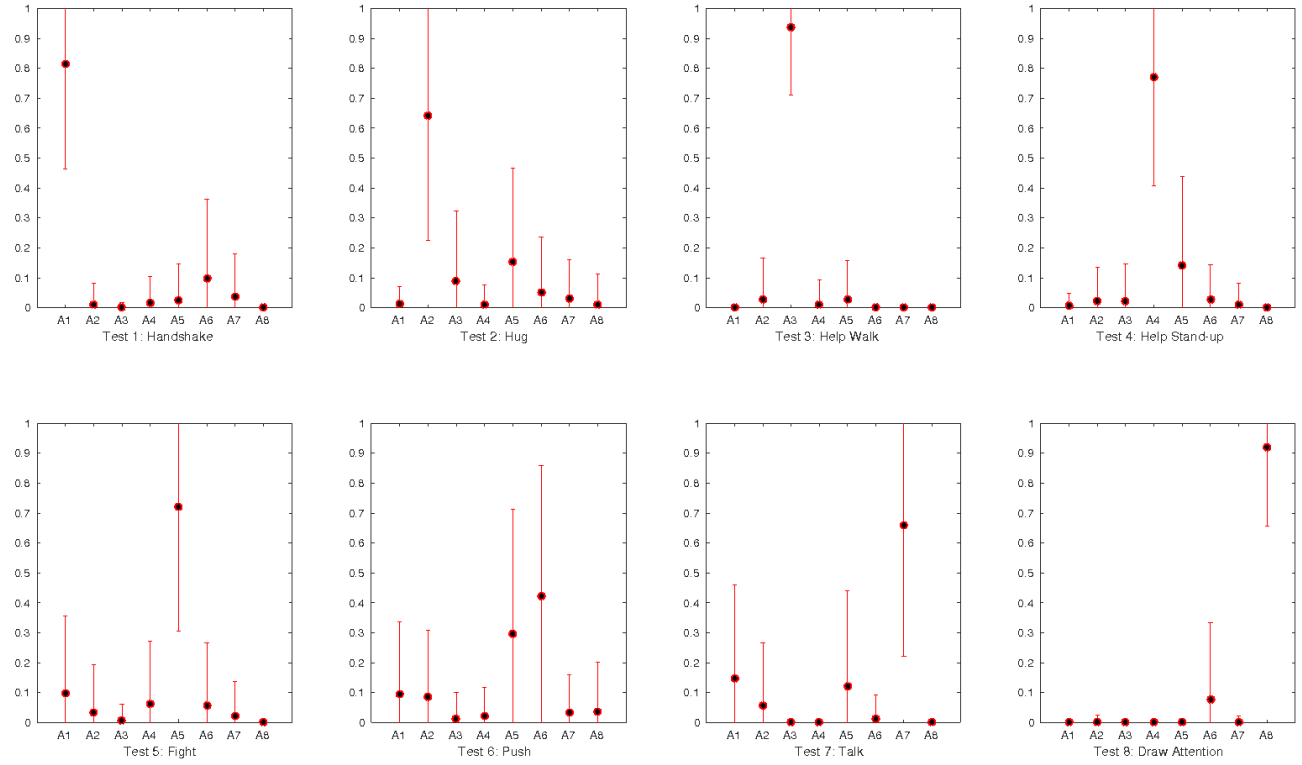


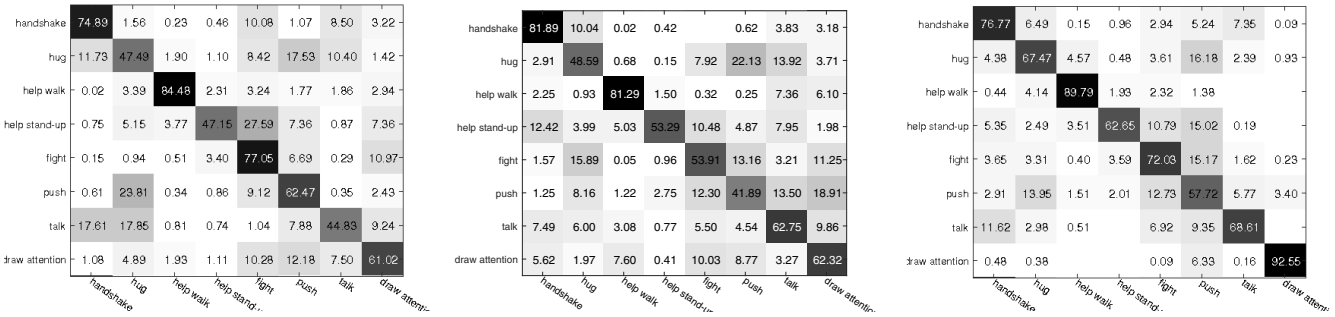
Fig. 3: Proximity priors performance: Average and standard deviation for each social activity over different tests.

V. Experimental Results

A. Social Activity Dataset

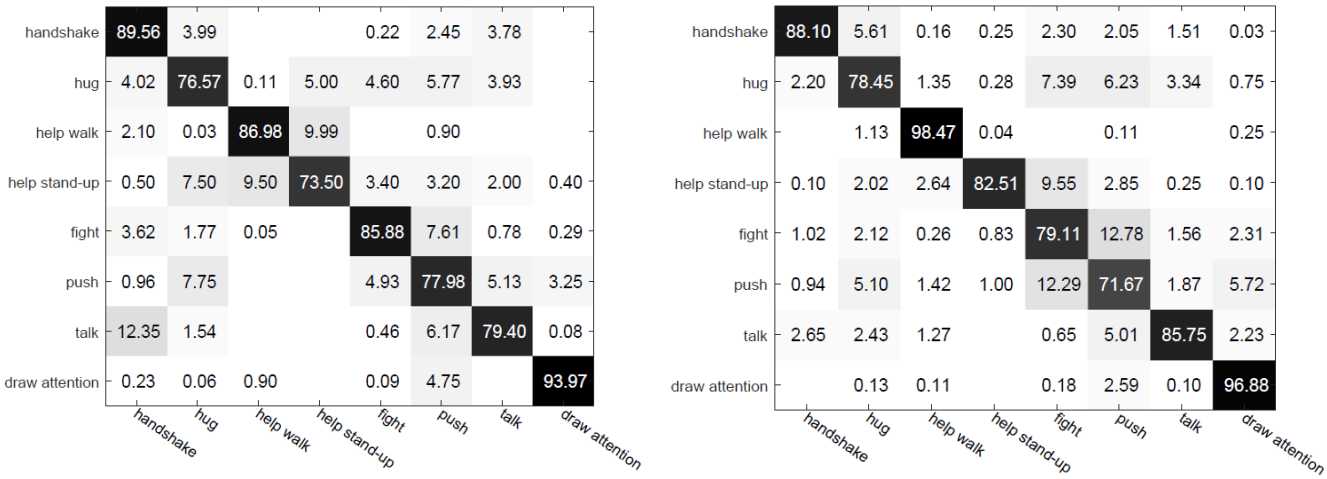
A new dataset of social interaction (ISR-UoL 3D Social Activity Dataset) between two subjects was built and it is publicly available for the community[1]. This dataset consists of RGB and depth images, and tracked skeleton data (i.e. joints 3D coordinates and rotations) acquired by an RGB-D sensor. It includes 8 social activities: handshake, greeting hug, help walk, help stand-up, fight, push, conversation, call attention. Each activity was recorded in a period around 40 to 60 seconds of repetitions within the same session at a frame rate of 30 frames per second. The only exceptions are help walking (at a short distance) and help stand-up, which were recorded 4 times as the same session, regardless of the time spent on it. The activities were selected to address the assisted living scenario (e.g. happening in a health care environment: help walking, help stand-up and call attention), with potential harm situations, such as aggression (e.g. fighting, pushing), and casual activities of social interactions (e.g. handshake, greeting hug and conversation). The activities were performed by 6 persons, 4 males and 2 females with an average age of 29.7 ± 4.2, from different nationalities

[1]Dataset available at: `https://lcas.lincoln.ac.uk/wp/isr-uol-3d-social-activity-dataset`



(a) Prec: 64.45%; Rec: 62.42%; Acc: 62.42% (b) Prec: 62.10%; Rec: 60.74%; Acc: 60.74% (c) Prec: 75.18%; Rec: 73.44%; Acc: 73.44%.

Fig. 4: Confusion matrices (leave-one-out cross validation) for (a) learning individual features from person 1 and 2 and test on person 1; (b) learning individual features from person 1 and 2 and test on person 2; (c) learning and test using only social features.



(a) Overall prec: 83.22%; rec: 82.98%; and acc: 82.98% (b) Overall prec: 85.60%; recall: 85.12%; acc: 85.12%

Fig. 5: Left image: results for test case (d), combining all features as one mixture (DBMM) and without proximity priors. Right image: results for test case (e), proposed approach using all features distributed in 3 mixtures and learned proximity priors.

(Italian, Brazilian and Portuguese), which can influence the proxemics parameters. A total of 10 different combinations of individuals (or sessions) were performed, with variation of the roles (active or passive person) between the subjects. Each subject has participated at least with 3 combinations, acting each role at least once. Half of the recorded sessions have been performed by a pair of persons who never met before the interaction, This was done in order to increase the generalization of the study regarding individual behavior.

B. Classification Results

In order to emphasize the advantage of using priors based on promexics theory (i.e. using proximity features), Fig. 3 shows the performance of the priors on a test set for each activity. Priors based on proximity features alone are not enough for frame to frame classification, however, when combined with the classification approach, it helps to obtain a faster convergence, since the initial guess tend to reduce the chances of less probable classes.

Regarding the classification, the protocol for the tests was leave-one-out cross validation strategy given 10 sessions and 8 activities for: (a) learning with individual features from subject one and two - excluding the social features - and test on subject one (i.e. individual with passive role during interactions); (b) the same as the latter, but testing on subject 2 (i.e. individual with active role during interactions, e.g. aggressor, call attention); (c) learning and test with social features only; (d) learning and test using all features as only one mixture in the classification model; and (e) full approach as depicted in Fig. 1: learning and test with all features as different mixtures and using the proximity (i.e. proxemics-based) priors. Fig. 4, presents the test cases (a)-(c). Looking at these test cases, we can notice that our approach has interesting performance in recognizing social activities even observing only one role, i.e., one subject performing the social interaction (as active or passive role).

The results obtained are attractive, since this dataset is very challenging, with many variations in subject's role, different persons with different behavior/response, showing that our approach has big potential for generalization in social activity recognition. The results obtained from tests (a) and (b) are consistent, since they are very similar for both individuals. The results obtained from test (c) were computed using one single mixture in the model. The proposed social features can distinguish the classes of activities, however improvements can be obtained as shown next. Fig. 5 (left image) presents the result attained for the test case (d). We can observe that encompassing all features as one entity (i.e. one mixture), the classification performance is improved when compared with the previous tests (a)-(c), where the features sets are tested individually. Finally, the test case (e) is presented in Fig. 5, showing that the proposed approach outperforms the previous test cases in terms of overall precision, recall and accuracy, reaching the performance above 85%. We can state that using the proximity priors and three different mixtures - one for each feature model - we obtained an improvement on the overall result around 3% in terms of precision, recall and accuracy when compared with test case (d), which is a significant improvement for this dataset due to the amount of frames for classification.

VI. Conclusion and Future Work

Recognition of social behavior and interaction is an important and challenging topic in ambient assisted living and robotics. This paper presented a new challenging RGB-D dataset for social activity recognition that is publicly available [23], introducing a new set of social features based on two subjects interacting. An adaptation on the dynamic Bayesian mixture model, first proposed in [7], was presented to deal with social activity recognition, showing that this proposed design, when combined with priors learnt from proximity features, improves the classification performance, reducing the likelihood of the less probable activities. Results show that the proposed approach can recognize different social activities and has the potential to be exploited in robotics for assisted living. Future work will address the design of other social features (e.g. bodies orientation), and the integration of this approach for monitoring tasks performed by mobile robots in contexts of assisted living.

References

[1] G. Yu, Z. Liu, and J. Yuan, "Discriminative orderlet mining for real-time recognition of human-object interaction," in *Computer Vision–ACCV 2014*. Springer, 2014, pp. 50–65.

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

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

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

[5] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Learning actionlet ensemble for 3D human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, pp. 914–927, 2014.

[6] C. Coppola, O. Martinez Mozos, N. Bellotto *et al.*, "Applying a 3d qualitative trajectory calculus to human action recognition using depth cameras," *IEEE/RSJ IROS Workshop on Assistance and Service Robotics in a Human Environment*, 2015.

[7] 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*, 2014.

[8] D. R. Faria, M. Vieira, C. Premebida, and U. Nunes, "Probabilistic human daily activity recognition towards robot-assisted living," in *IEEE RO-MAN'15: IEEE Int. Symposium on Robot and Human Interactive Communication. Kobe, Japan.*, 2015.

[9] G. Parisi, C. Weber, and S. Wermter, "Self-organizing neural integration of pose-motion features for human action recognition," *Name: Frontiers in Neurorobotics*, vol. 9, no. 3, 2015.

[10] L. Piyathilaka and S. Kodagoda, "Human activity recognition for domestic robots," in *Field and Service Robotics*. Springer, 2015.

[11] M. Vázquez, A. Steinfeld, and S. E. Hudson, "Parallel detection of conversational groups of free-standing people and tracking of their lower-body orientation," in *IEEE IROS'15, Germany*, 2015.

[12] K. Khoshhal Roudposhti, U. Nunes, and J. Dias, "Probabilistic social behavior analysis by exploring body motion-based patterns," *IEEE Trans. PAMI*, 2015.

[13] R. Mead, A. Atrash, and M. J. Mataric, "Automated proxemic feature extraction and behavior recognition: Applications in human-robot interaction," in *Int. Journal of Social Robotics*. Springer, 2013.

[14] E. T. Hall, "A system for the notation of proxemic behavior," *American Anthropologist*, 1963.

[15] I. Chakraborty, H. Cheng, and O. Javed, "3d visual proxemics: Recognizing human interactions in 3d from a single image," in *IEEE CVPR*, June 2013, pp. 3406–3413.

[16] L. Takayama and C. Pantofaru, "Influences on proxemic behaviors in human-robot interaction," in *IEEE IROS'09*, 2009.

[17] Y. Kim and B. Mutlu, "How social distance shapes human–robot interaction," *Int. Journal of Human-Computer Studies*, 2014.

[18] M. Vieira, D. R. Faria, and U. Nunes, "Real-time application for monitoring human daily activities and risk situations in robot-assisted living," in *Robot'15: 2nd Iberian Robotics Conf.*, 2015.

[19] C. Premebida, D. R. Faria, F. A. Souza, and U. Nunes, "Applying probabilistic mixture models to semantic place classification in mobile robotics," in *IEEE IROS'15, Germany*, 2015.

[20] C. Premebida, D. R. Faria, and U. Nunes, "Dynamic bayesian network for semantic place classification in mobile robotics," *AURO Springer: Autonomous Robots*, 2016.

[21] J. Vital, D. R. Faria, G. Dias, M. Couceiro, F. Coutinho, and N. Ferreira, "Combining discriminative spatio-temporal features for daily life activity recognition using wearable motion sensing suit," *PAA Springer: Pattern Analysis and Applications*, 2016.

[22] 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.

[23] "ISR-UoL 3D Social Activity Dataset. Web: https://lcas.lincoln.ac.uk/wp/isr-uol-3d-social-activity-dataset."