International Journal of Social Robotics
https://doi.org/10.1007/s12369-019-00541-y

Social Activity Recognition on Continuous RGB-D Video Sequences

Claudio Coppola[1] · **Serhan Cosar**[2] · **Diego R. Faria**[3] · **Nicola Bellotto**[2]

Accepted: 8 March 2019

Abstract
Modern service robots are provided with one or more sensors, often including RGB-D cameras, to perceive objects and humans in the environment. This paper proposes a new system for the recognition of human social activities from a continuous stream of RGB-D data. Many of the works until now have succeeded in recognising activities from clipped videos in datasets, but for robotic applications it is important to be able to move to more realistic scenarios in which such activities are not manually selected. For this reason, it is useful to detect the time intervals when humans are performing social activities, the recognition of which can contribute to trigger human-robot interactions or to detect situations of potential danger. The main contributions of this research work include a novel system for the recognition of social activities from continuous RGB-D data, combining temporal segmentation and classification, as well as a model for learning the proximity-based priors of the social activities. A new public dataset with RGB-D videos of social and individual activities is also provided and used for evaluating the proposed solutions. The results show the good performance of the system in recognising social activities from continuous RGB-D data.

Keywords Social activity recognition · Activity recognition · Activity temporal segmentation · Machine learning

1 Introduction

In many applications of service and domestic robots, for example to help customers in a shopping centre or assist elderly people at home, it is important to be able to identify and recognise human activities. Particular attention has been given to indoor activities for potential application in security, retail and Active & Assisted Living (AAL) scenarios. In the latter case, for example, human activity recognition with a domestic robot can be useful to identify potential problems and apply corrective strategies. Many researchers therefore have developed methodologies and techniques for human activity recognition exploiting smart-home or mobile robot sensors, such as RGB-D cameras, to collect and analyse large datasets of indoor activities.

Besides individual activities, the detection and recognition of social activities and recognition of social activities is also important to understand social behaviours, and therefore increasingly of interest to the scientific community. In psychology for example, social activity recognition can help to understand how people's behaviours are influenced by the presence of others [13,16,31]. Furthermore, the subject attracts the attention of many researchers in computer vision and robotics, since it enables them to build robots capable of interacting with humans in different social contexts, and to provide tailored robot services for assistance and companionship. A robot that can detect and recognise human social activities, could also be used to identify dangerous situations, antisocial behaviours, aggressions, etc.

Similarly to the case of individual activity recognition, the challenges in social activity recognition are the high intra-class and low inter-class variability of the data, due to the different ways in which the same activity can be performed and to the similarities between different activities. In addition, social activity recognition has to deal with the extra degrees of freedom introduced by the presence of multiple

✉ Claudio Coppola
c.ccoppola@qmul.ac.uk

Serhan Cosar
scosar@lincoln.co.uk

Diego R. Faria
d.faria@aston.ac.uk

Nicola Bellotto
nbellotto@lincoln.co.uk

[1] Queen Mary University of London, Mile End Road, London, England E1 4NS, UK

[2] University of Lincoln, Brayford Pool, Lincoln, England LN6 7TS, UK

[3] Aston University, Aston Express Way, Birmingham, England B4 7ET, UK

Published online: 29 April 2019

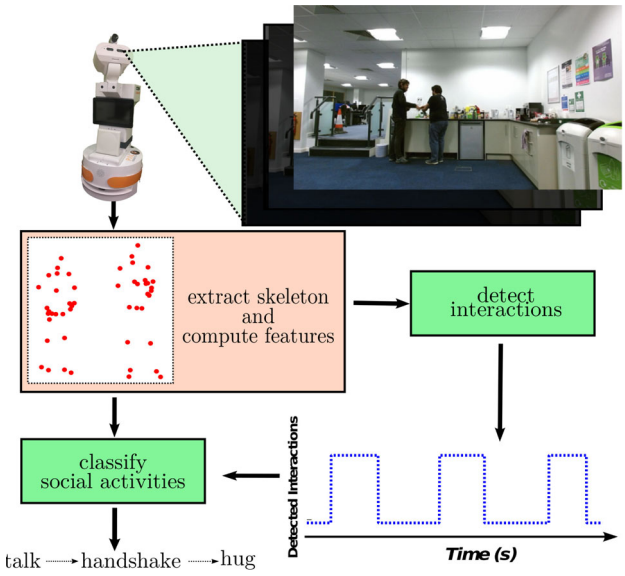


Fig. 1 Overview of the social activity recognition system segmenting and classifying interactions from continuous RGB-D skeleton data

actors. Social activities are also affected by cultural differences (e.g. interaction distance and social space), which complicate the classification problem.

In order to recognise social activities in realistic scenarios, we propose a system that deals with continuous streams of RGB-D data, rather than cropped videos of activities as in many previous datasets. The system detects when two subjects engage in an interaction and classifies the underlying social activity (see Fig. 1). In our work, a social activity is defined as a mutual physical or visual engagement between two persons in order to obtain a certain goal. In our previous work on social activity recognition [6], a set of DBMM Classifiers using different sets of features is presented. These features model the relational information between the two peoples movements (i.e. how ones movement affects the other) and the individual's movement information. Furthermore, in [5], a SVM-HMM model is used to segment the intervals of time in which social interactions occur. Since the performance of these two models has been only evaluated individually, their combined performance needs to be assessed to consider using them in robotic applications. Compared to those works, the new contributions of this paper are fourfold:

1. a novel framework and full pipeline implementation for recognising social activities in realistic scenario from continuous RGB-D data;
2. an improved method to learn proximity-based priors, based on Gaussian Mixture Models, which are used in the probabilistic classification of social activities;
3. a new public dataset with continuous RGB-D sequences of individual and fully labelled social activities for the evaluation and future comparison of our method;
4. An extensive experimental analysis, including a comparative study of our social activity classification;

The paper is organized as follows: Sect. 2 summarizes the state of the art for activity recognition and detection of interactions; Sect. 3 provides a high level overview of the system and its components; Sect. 4 introduces the features designed for the detection of interactions and recognition of social activities from RGB-D data; Sect. 5 describes our model for temporal detection and segmentation of interactions; Sect. 6 explains the approach used for the classification of social activities, including the improved proximity-based priors, and shows how the final estimation on continuous activity sequences is computed; Sect. 7 illustrates the dataset and the experiments performed to evaluate our system, including a detailed analysis of its key components; Finally, Sect. 8 concludes the paper discussing our approach and results, as well as presenting possible directions for future research in this area.

2 Related Work

2.1 Classification of Human Activities

Automatic recognition of human activities has become increasingly important in the computer vision and robotics research communities, in particular after the release of affordable RGB-D cameras and software for human tracking and pose estimation. For example, in [7], a 3D extension of the Qualitative Trajectory Calculus (QTC) was applied to model movements of the body joints on RGB-D skeletal data. In [9,10], the Dynamic Bayesian Mixture Model (DBMM) combining a set of classifiers based their temporal entropy is introduced. The approach presented in [23] uses HMMs implemented as a Dynamic Bayesian Network with Gaussian Mixture Models (GMM). In [39], a Multiple Instance Learning-based approach for social activity recognition is proposed. In [20], a social activity recognition system based on the detection of posture clusters and used to train a set classifiers, is presented. In [11], relation history images are introduced. This descriptor is able to characterise individual, social and ego-centric activities. The approach presented in [30], performs classification using a pool of Long Short Term Memory (LSTM) cells with common output gate. In [22], instead, used hierarchical self-organizing neural networks to recognise human actions from depth and audio information. To obtain a semi-supervised behaviour the previously presented growing network [21] has been extended, adding a layer to associate human words with the activities. A social activity recognition system which merged multiple DBMMs to represent two separate individuals and their social characteristics was introduced in [6]. Finally, [17] used

a qualitative representation of human motion based on Laban Movement Analysis (LMA) for modelling and estimating social behaviours using Dynamic Bayesian Networks.

All the approaches considered so far were able to recognise human activities, but they were only applied on manually clipped videos. In case of continuous data streams, it is necessary to determine the actual beginning and end of each activity [1]. Described an approach suitable for continuous RGB videos, in which the temporal segmentation of the activities is performed by opportune active learning-based methods [14]. Presented a system for activity recognition and temporal segmentation based on skeletal and silhouette features from RGB-D videos. The beginning and the end of the activity were found comparing the fitness value coming from a non-activity model or a HMM for each activity. The time intervals were then classified with a cumulative HMM [26]. Proposed an activity recognition system for autonomous robots based on RGB images. Convolutional networks were trained using pre-computed human silhouettes to recognise human body motions [19]. Describes an approach to recognise sequences of simultaneous individual human actions that compose complex activities using a hierarchical approach. This approach recognise human poses from skeleton descriptors, atomic actions from a sequence of poses and finally activities from a sequence of actions. All these approaches extracted and recognised individual activities from continuous video streams. However, they did not consider the social activity case, which is addressed instead by the current paper.

2.2 Detection of Social Interactions

Social scientists have since long being studying social interactions and non-verbal communication. Previous work include theories on the reciprocal distance by [13], mutual presence in the participants' field of view by [31] and topology formation of interacting agents by [16].

These theories have already been exploited for detecting conversational groups on still images. For example, [4] estimated 3D proxemics parameters to identify social interactions in internet images [8,28,29]. Detected social interactions on RGB images using the concept of F-Formations by [16], where the centre of a circular space (O-Space) is induced by people's orientation [40]. Detected F-Formations by building a graph of people locations. A classifier is fed with social involvement features to perform the detection. A system for recognising conversational groups was presented by [34], who exploited the orientation of the lower body part [2]. Detected social interactions using the subjects field of view modelled as subjective view frustum, which is characterised by the head orientation.

These works informed our choice and definition of spatial features for the detection and temporal segmentation of social interactions, and used by our system to improve the classification of the underlying social activities.

2.3 Activity Recognition Datasets

In order to train and evaluate systems for human activity recognition, several datasets have been created using RGB-D sensors. These datasets usually provide also body pose and possibly objects used in the activities [37,38]. Provided video clips of 16 different daily activities [18], instead, collected video clips of realistic individual activities and sub-activities, including information about the objects used. Another dataset for the recognition of social activities in video clips was presented by [30,39]. Built a dataset containing video clips of 60 action classes from 3 different points of view, including individual and social activities. A dataset with 60 videos of individual activities occurring in 5 different locations was finally proposed by [32].

All these RGB-D datasets of human activities are characterised by short clipped videos. However, an activity recognition system for real-world and robot-assisted scenarios should be able to work on continuous video streams of RGB-D data. Therefore, our work includes a new public dataset in which long, continuous sequences of individual and social activities are included for training and evaluation purposes.

3 System Overview

Our approach for social activity recognition focuses on continuous streams of skeleton data whenever two individuals are in the RGB-D camera's field of view. The systems consists of three main parts (Fig. 2):

- **Temporal segmentation of interactions:** This component is responsible for finding the temporal intervals in which the social activities occur. It uses features based on social science theories, measured on the upper bodies. In practice, this behaves like a switch, which decides when the following components need to be activated and when not.
- **Classification of the social activities:** This component performs the classification of the detected social activities. It consists of three classifiers, which use three different sets of features based on individual poses, movements, and spatial relations. The output likelihoods are then merged to obtain a final likelihood vector of the activities.
- **Estimation of the proximity-based priors:** This component is responsible for estimating the probability priors from learnt distributions of the proximity between two subjects. These priors are then merged with the likelihood

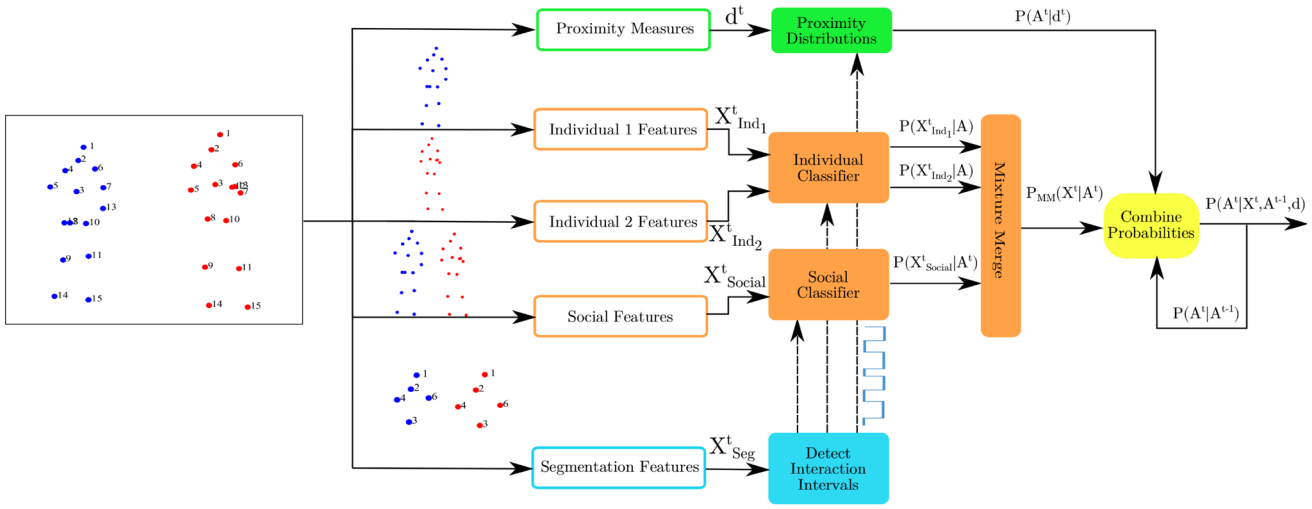


Fig. 2 The proposed approach for continuous social activity recognition: temporal segmentation modules (blue); classification modules (orange); priors estimation modules (green). (Color figure online)

from the classifiers to obtain the posterior probability of the activities.

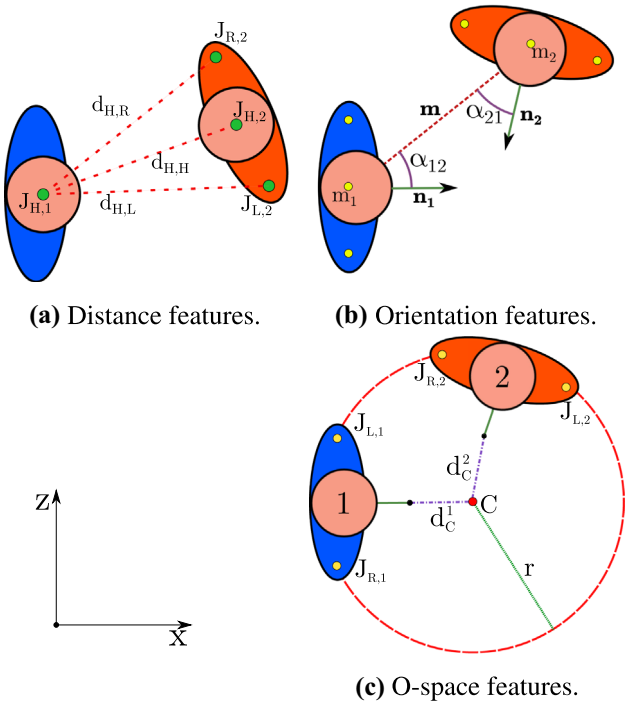
4 Feature-Sets

Our system exploits the estimated 3D body joints from a skeleton tracker provided by Microsoft Kinect SDK2. The software is very stable and it is able to detect and track human skeletons in challenging situations, although its application is limited to Kinect 2 sensors only. Using skeletal data, we define two sets of features:

- **Segmentation features:** used to detect the temporal intervals of the social interactions (X_{Seg}) , based on the upper bodies of the two actors and originally proposed by [5]. These features are computed on two dimensions only (x and z of the Kinect 2 optical frame, see Fig. 3).
- **Classification features:** consisting of individual and social features. The first ones serve the two individual mixtures (X_{Ind_1},X_{Ind_2}) of the classification model. They are based on single skeletons and used for individual activity classification, as suggested by [9,10]. The second ones are for the social mixture of the classification model (X_{Social}). They are based on both skeletons and are used for social activity classification, as proposed by [6].

4.1 Segmentation Features

This set of features is inspired by studies in social science and refer only to the upper body joints of the skeletons (head, left shoulder, right shoulder, torso). They are computed on a planar view, as illustrated in Fig. 3, so that they are invariant



(a) Distance features. **(b)** Orientation features.

(c) O-space features.

Fig. 3 Examples of the segmentation features. Distances d are computed between different joints J of the two subjects, including head (H), left shoulder (L), right shoulder (R) and torso (T)

to human height. This set of features is based on geometrical properties and statistics of the upper body position, orientation and motion. The features are the following:

- **Upper joint distances:** According to the proxemic theory of [13], humans create spacial sectors around them, the size of which depends on the personal intimacy and

cultural background of the subjects. Extracting these sectors from the distance between two persons' skeletal joints is relatively straightforward. As shown in Fig. 3a, the 2D distance $d_{i,j}$, on the (x, z) plane of the camera's frame, is computed between the upper body joints $J_{i,1}$ and $J_{j,2}$ of the two persons, where $i, j \in \{H, L, R, T\}$ – i.e. head, left shoulder, right shoulder and torso, respectively resulting in 16 different distances. For example, $d_{H,R}$ is the distance between the head of subject 1 and the right shoulder of subject 2.

- **Body orientation angle to the reference line:** According to [31], being in each other's field of view plays an important role in the social interaction between two persons. The relative body orientation between them is therefore an important clue to discriminate between interactions and non-interactions, where distance alone would not be sufficient. As shown in Fig. 3b, we consider the following two angles:

$$\alpha_{12} = \angle(\mathbf{n_1}, \mathbf{m}) \quad \alpha_{21} = \angle(\mathbf{n_2}, -\mathbf{m}) \tag{1}$$

where $\mathbf{n_1}$ and $\mathbf{n_2}$ are the orientation vectors of the subjects (normal to the torso) and $\mathbf{m}$ is the vector between their torsos.

- **Temporal similarity of the orientations:** [15] demonstrated that speakers and listeners often synchronise their movements. Based on this, we compute the logarithm L of windowed moving covariance matrices (4 features) to estimate the temporal similarity between relative changes of the subject orientations during the time interval $[t - w, t]$:

$$L = \log(1 + \mathrm{cov}(\alpha_{12}^{t-w,\ldots,t}, \alpha_{21}^{t-w,\ldots,t})) \tag{2}$$

where w is the window of reference (in our case $w = 1$s).

- **O-space radius and oriented distance:** According to the F-Formations theory by [16], social interactions occur when the transactional segments of the two subjects are overlapping, Interacting people stand on the border of a circular area (O-space), with their bodies oriented towards the centre. As shown in Fig. 3c, the O-space can be defined by (approximately) fitting a circle on the shoulders of the subjects and checking whether the normal vectors $\mathbf{n_1}$ and $\mathbf{n_2}$, from their torsos, lie inside or outside this space. The situation is fully captured by a set of features $[r, d_1^C, d_2^C]$, where r is the radius of the circle, and d_k^C (with $k = 1, 2$) is the distances between the extremity of the normal $\mathbf{n_k}$ and the centre C. If $d_k^C > r$, it means subject k is oriented towards the outside of the circle. Also, if $r > r_{max}$, the two people are considered too far to be interacting. Note that, in this system, $\mathbf{n_k}$ is a unit vector (1m).

- **QTC$_C$ relation:** The Qualitative Trajectory Calculus (QTC) is a mathematical formalism introduced by [33] to describe spatial relations between two moving points. We use a particular version of the calculus, called QTC$_C$, where the qualitative relations between two points P_k and P_l are expressed by the symbols $q_i \in \{-, +, 0\}$ as follows:

  (q_1) $-$: P_k is moving towards P_l
  0: P_k is stable with respect to P_l
  $+$: P_k is moving away from P_l
  (q_2) same as q_1, but swapping P_k and P_l
  (q_3) $-$: P_k is moving to the left side of $\overrightarrow{P_k P_l}$
  0: P_k is moving along $\overrightarrow{P_k P_l}$
  $+$: P_k is moving to the right side of $\overrightarrow{P_k P_l}$
  (q_4) same as q_3, but swapping P_k and P_l.

  A string of QTC symbols $\{q_1, q_2, q_3, q_4\}$ is therefore a compact representation of the 2D relative motion between P_k and P_l. For example, $\{-, -, 0, 0\}$ means "P_k and P_l are moving straight towards each other". Other examples can be observed in Fig. 4a. The 2D trajectories considered in our work are those of the people's torsos.

- **Temporal Histogram of QTC$_C$ relations:** QTC$_C$ can be used to analyse sequences of torso trajectories using temporal histograms. In particular, we build two windowed moving histograms, with 9 time bins each, splitting the QTC$_C$ components in two sets: the first one considers the distance relations (q_1, q_2), while the second captures the side relations (q_3, q_4). This separation has also the advantage of reducing the total number of bins ($2 \cdot 3^2$ rather than 3^4). An example of QTC$_C$ histogram is shown in Fig. 4b.

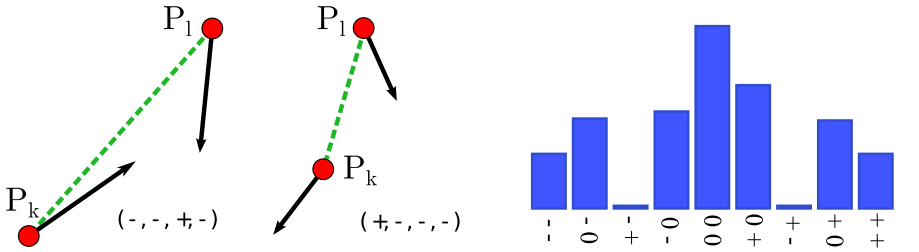
4.2 Classification Features

This set of features is used to classify social activities considering both individual and social properties of the subjects.

Individual features characterise poses and movements of each single person involved in a social activity. They have been designed and successfully applied for individual activity recognition by [9,10]. In total, there are 171 of these spatio-temporal features, computed from the joints of each subjects, and broadly categorised in geometrical, energy-based and statistical features.

Social features, instead, describe the relation between the joints of both skeletons. There are in total 245 social features per frame, details of which are as follows:

- **Covariance of inter-body joint distances:** Similar to the upper joint distances of Sect. 4.1, but extended to 3D and computed on the full set of joints to deal with the more complex task of activity classification. All the 3D



(a) Cases of QTC_C string. **(b)** Example of QTC_C histogram.

Fig. 4 Examples of QTC_C based features

Euclidean distances between the 15 joints of an individual skeleton are used to fill a 15 matrix **D**. The upper 120 triangular elements of its log-covariance matrix constitutes then the actual features, which basically represent the relative variation in the position and body posture of the subjects. The matrix logarithm makes the covariance based features more robust by mapping the covariance space into a euclidean space [12].

- **Temporal covariance of inter-body joint distances:** The temporal variation of the previous features is also considered by computing $\mathbf{D}^t$ and $\mathbf{D}^{t-n}$ at time t and $t-n$, respectively, and their difference $\mathbf{R}^t = \mathbf{D}^t - \mathbf{D}^{t-n}$. The upper triangular elements of the log-covariance of $\mathbf{R}^t$ are the final features in this case. Like the previous set, this is also composed by 120 features.
- **Minimum distance to torso:** Two more social features are derived by calculating all the 3D distances between the joints of subject 1 and the torso of subject 2, then taking the minimum, and vice-versa (subject 2 to subject 1).
- **Accumulated energy of the torsos:** These features allow to discriminate the most active person (e.g. who is approaching the individual space of the other). They include the distance from torso to torso, plus the energy E depending on the distance variations of all the joints of a subject to the torso of the other:

$$E = \sum_i v_i^2 \quad \text{and} \quad v_i = d_{i,T}^t - d_{i,T}^{t-n} \tag{3}$$

where $d_{i,T}^t$ is the distance, at time t, of the ith joint of a subject to the torso T of the other, and $[t-n, t]$ is the considered time interval. Two energy features, one for each subject, are computed.

5 Interaction Segmentation

To recognise social activities from continuous data, we need to detect the time intervals in which some interaction between two or more people occurs.

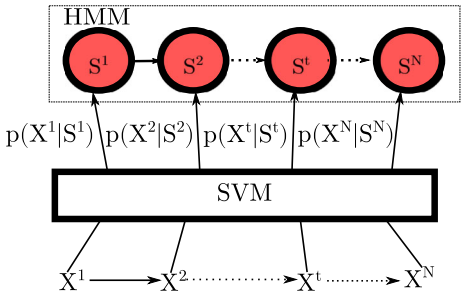


Fig. 5 Interaction segmentation module: X_i and S_i are, respectively, the observed features and the activity state (individual, social) at time i

In order to perform this temporal segmentation, we combine two standard techniques for frame classification and sequential state estimation:

1. *Support Vector Machine* (SVM), which is an algorithm for binary classification, shown to be efficient even in cases of non-linearly separable data.
2. *Hidden Markov Model* (HMM), which is a tool to represent probability distributions over sequences of observations, suitable for labelling sequential data.

In our work, we implemented a HMM with two activity states (individual, social), where the transition probability distribution $p(S^t|S^{t-1})$ is learnt from the number of state changes in a training set. The observation probability, instead, is defined by an SVM classifier trained on the same data, using its output confidence as a likelihood $p(X^t|S^t)$ for the HMM. The SVM is implemented with a linear kernel and with cost $c = 1$. In the testing phase, the activities are labelled by estimating the most probable state paths using a standard Viterbi algorithm. A graphical representation of the temporal segmentation process can be seen in Fig. 5.

The role of the HMM is to avoid potential errors in the estimated likelihood, which cause a 'flickering' effect on the estimated segmentation. In Fig. 6, where a threshold-based approach is compared to the HMM output for three consecutive interactions. It can be seen that a simplistic threshold of the likelihood would have caused a flickering in the segmen-

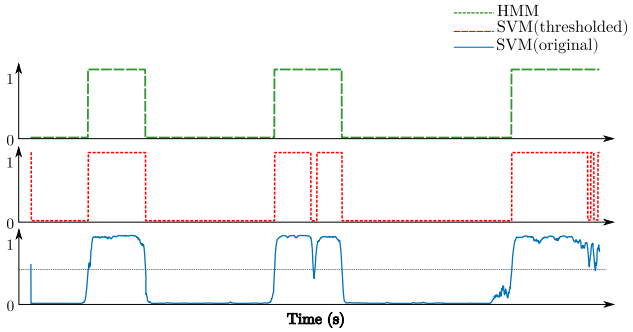


Fig. 6 Example of segmentation of the social interaction. In green the the estimated segmentation output of the HMM; in blue the likelihood output of the SVM ($p(X^t|S^t)$); in red the segmentation obtained via thresholding of the likelihood in blue. (Color figure online)

tation, while exploiting temporal information with the HMM corrects such problem.

6 Social Activity Classification

In this section we first introduce the Dynamic Bayesian Mixture Model (DBMM) originally proposed by [9] for individual activity recognition, which was also used for other classification problems by [10,24,25,35] and [36]. We present then our approach to fuse semantically-different sets of features as a multiple mixture of DBMMs, incorporating also additional priors learnt from proximity features.

6.1 Dynamic Bayesian Mixture Model

A DBMM is a probabilistic ensemble of classifiers using a Dynamic Bayesian network (DBN) and a mixture model to fuse the outputs of different classifiers, exploiting also temporal information from previous time slices. The method was originally proposed in [9] and is here summarised with details of our current implementation.

Let X^t be an observation at time t, assumed independent from previous observations, and $A^t \in \mathscr{A}$ the activity at time t belonging to the set $\mathscr{A}$ of all possible activities. Assuming A^t is conditionally independent from future activities, we can formulate a DBMM with n time slices as follows:

$$P_h(X_h^t|A^t) = \sum_{i=1}^{N} w_{i,h}^t \times P_{i,h}(X_h^t|A^t) \tag{4}$$

where N is the number of classifiers and the weight $w_{i,h}$ of each base classifier is learnt from the samples training set using the feature set X_h and the likelihood $P_{i,h}(X_h^t|A^t)$ is the output of the ith classifier at time t.

Our DBMM implementation includes the following base classifiers: a Naive Bayes Classifier (NBC), a Support Vector Machine (SVM) with linear kernel, and an Artificial Neural Network (ANN) with 70 hidden neurons and a softmax output.

6.2 Multi-Merge DBMM

The Multi-Merge DBMM (MM-DBMM) is an ensemble, defined in [6], that combines multiple DBMMs classifiers, each one processing a specific set of features. The orange part in Fig. 2 shows the structure of this extended DBMM scheme. The three sets of features (i.e. one for each individual component of the activity, plus one for the social information of the activity) are given as input to two independent classifiers, namely the Individual Classifier and the Social Classifier. Each one of these classifiers outputs the likelihood that a certain activity occurs. The likelihoods are then weighted and fused by the Mixture Merge block.

The previous Eq. (4) can be rewritten as follows:

$$\begin{aligned} P(A^t|X^t, A^{t-1}) &= \beta \times P(A^t|A^{t-1}) \times P_{MM}(X^t|A^t) \\ P_{MM}(X^t|A^t) &= \sum\nolimits_{h\in\mathscr{H}} w_h^t \times P_h(X_h^t|A^t) \end{aligned} \tag{5}$$

where $P_{MM}(X^t|A^t)$ is the merged likelihood of all the available DBMMs in $\mathscr{H} = \{Ind_1, Ind_2, Social\}$. $P_h(X_h^t|A^t)$ is the likelihood obtained from the hth DBMM with the feature set X_h^t. The quantities w_h^t and $w_{i,h}^t$ are weights for the hth DBMM and its ith base classifier, respectively. Finally, β is just a normalisation factor. As already mentioned, each DBMM is a weighted combination of base classifiers. In our MM-DBMM though, a new set of normalised weights w_h^t are used for the merged likelihood P_{MM}, based on the normalised outputs of the DBMMs:

$$w_h^t = \frac{P_h(X_h^t|A^t)}{\sum_{g\in\mathscr{H}} P_g(X_g^t|A^t)} \tag{6}$$

Decomposing the classification in individual and social mixtures allows to break the complexity of the social activities into components dependant on each person pose and movement and a component dependant on their mutual relation. In this way, our system can cope with the challenging high intraclass and low inter-class variability of the data.

6.3 Proximity-Based Priors

Similarly to our previous work in [6], To boost the classification results, we generate prior probabilities of social activities based on proxemics, assuming that certain interactions occur within social spaces defined by the distance between the subjects. These social spaces are not unique and therefore not easy to define deterministically due to personal and cultural differences, and therefore better described in the form of probability distributions. The aim of our probability priors

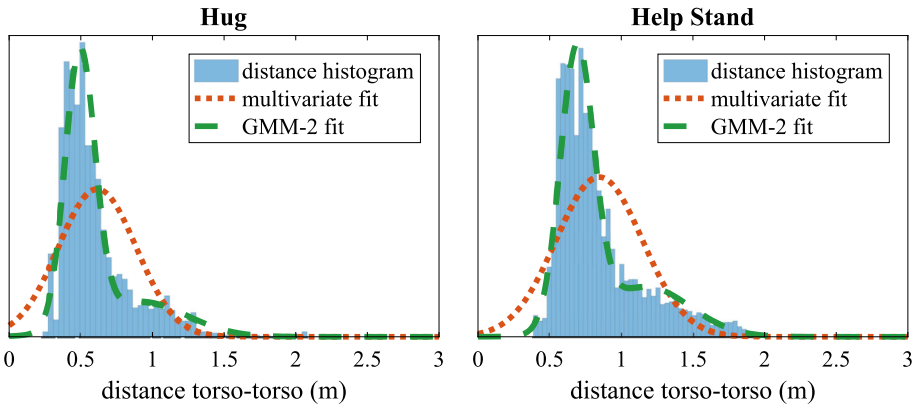


Fig. 7 Examples of histograms of torso-torso distances, in two different activities, fitting a multivariate Gaussian model and a Gaussian Mixture Model

is to improve the classification performance by filtering out unlikely social activities, based on the distance between the actors.

Let d^t be the proximity measure. We can compute the posterior probability of an activity A^t given an observation X^t using the Bayesian rule:

$$P(A^t|X^t, d^t) = \beta \times P(X^t|A^t) \times P(A^t, d^t) \tag{7}$$

where $P(A^t|X^t, d^t)$ is the merged posterior probability of the system, $P(X^t|A^t)$ is the likelihood of a classifier (assuming X^t and d^t are conditionally independent given A^t) and $P(A^t, d^t)$ is the probability prior. In our specific case, the likelihood $P(X^t|A^t)$ corresponds to $P_{MM}(X^t|A^t)$ Note that $P(A^t, d^t) \propto P(A^t|d^t)$, since $P(d^t)$ is assumed uniform and therefore incorporated in the normalisation factor β.

For this model we consider the following seven distances:

(a) Torso to torso distance;
(b) The minimum distance between any joint of one person and the torso of the other (two values);
(c) As in (b), but in this case maximum distance (two values);
d) The minimum/maximum distance between any two joints, one per each subject (two values).

The latter measures in particular provide information about the closest and farthest joints of the two skeletons.

Unlike the model proposed by [6], which was based on a multivariate Gaussian, with mean μ and covariance matrix Σ, fitted on the distances, in this new model for priors we use a Gaussian Mixture Model (GMM) to represent the proximity priors:

$$P(A^t|d^t) = \sum_j \alpha_j \mathcal{N}(\mu_j, \Sigma_j) \tag{8}$$

where α_j,μ_j and Σ_j are the mixture weights, the mean and the variance of the jth component, respectively. The advantage of using GMMs can be seen in Fig. 7, where the distance

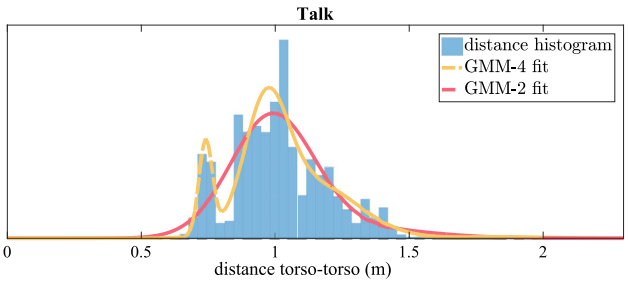


Fig. 8 Histograms of the torso-torso distance during the *talk* activity, comparing Gaussian Mixture Model fits with two and four mixtures

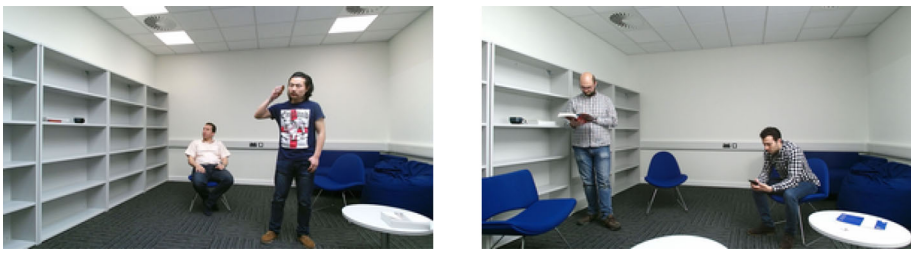
distributions are non-Gaussian (and sometimes multimodal). The non-Gaussianity of the distributions depends on the variability of the social activities, which could occur at different distance sectors. The GMM parameters are estimated by the Expectation Maximisation (EM) algorithm initialised with random samples, uniform mixing proportion and diagonal covariance matrix.

The risk with GMMs, however, is to over-fit the data using an excessive number of mixtures (see for example Fig. 8). Thus, it is important to decide how many components to use for each activity without including noise into the model. For each activity, we choose the number of GMM components through minimisation of the Bayesian Information Criterion (BIC):

$$\text{BIC} = \ln(n)k - 2\ln(\hat{L}) \tag{9}$$

where n is the number of samples, k is the number of the estimated parameters (i.e. each parameter of the GMM components), and $\hat{L}$ is the maximised likelihood obtained from the estimated model. This formula limits the number of components, during the model estimation phase, thanks to the logarithmic penalty term $\ln(n)k$. In our case we consider a maximum of 4 GMM components. The BIC penalises the models with higher number of parameters more strongly than the Akaike Information Criterion (AIC), therefore it is more suitable to avoid overfitting.

(a) Examples of social activities *handshake*, *help walking*, *fight* and *talk*.



(b) Examples of individual activities *phonecall* and *read*.

Fig. 9 RGB snapshots of the new social activity dataset

6.4 Combined Model

Given the transition probability $P(A^t|A^{t-1})$, the proximity prior $P(A^t|d^t)$, and the output likelihood $P_{MM}(X^t|A^t)$ of the MM-DBMM, we can compute the final posterior as follows:

$$P(A^t|X^t, A^{t-1}, d^t) = \beta \times P(A^t|A^{t-1}) \times P(A^t|d^t) \times P_{MM}(X^t|A^t) \quad (10)$$

The last equation merges the transition probability and the likelihood coming from the full MM-DBMM model in Eq. (5) with the proximity priors according to the approach shown in Eq. (7).

The final system integrates the MM-DBMM classifier with the new proximity-based priors and the interaction segmentation presented in Sec. 5 to implement a full software pipeline to recognise social activities on continuous RGB-D data streams.

7 Experiments

In this section we first introduce our new dataset for social activity recognition, and then present the performance of the overall system. We finally analyse more in detail the behaviour of each module—segmentation, classification, proximity priors—to better understand how their role in the social activity recognition task.

7.1 Social Activity Dataset

We created a new dataset ("3D Continuous Social Activity Dataset") for social activity recognition to validate the performance of our system on continuous stream RGB-D data. The dataset is publicly available[1] for the research community. It consists of RGB and depth images, plus skeleton data of the participants (i.e. 3D coordinates and orientation of the joints), collected indoor with a Kinect 2 sensor. The dataset includes 20 videos, containing individual and social activities with 11 different subjects. The approximate length of each video is 90 s, recorded at 30 fps (more than 50 K samples in total). In particular, the social activities in the videos are *handshake*, *hug*, *help walking*, *help standing-up*, *fight*, *push*, *talk*, *draw attention*. Some snapshots from the dataset are shown in Fig. 9. Differently from a previous "3D Social

[1] Dataset available at: https://lcas.lincoln.ac.uk/wp/research/data-sets-software/continuous-social-activity-dataset

Table 1 Statistics of the final social activity recognition

	No Segm. [6]	Man. Segm.	Aut. Segm.
MM-DBMM with no priors			
% Accuracy	92.13	97.65	97.02
% Precision	45.39	76.96	68.46
% Recall	83.79	76.08	68.20
MM-DBMM with multivariate priors			
% Accuracy	91.33	97.86	97.08
% Precision	59.72	79.42	70.62
% Recall	79.3	81.06	71.23
MM-DBMM with GMM priors			
% Accuracy	92.11	98.69	97.86
% Precision	67.65	88.01	78.52
% Recall	84.19	86.56	76.13

Activity Dataset" by [6], the social activities in this new dataset appear in uninterrupted sequences, within the same video, alternating 2 or 3 social activities with individual ones such as *read*, *phonecall*, *drink* or *sit*. Furthermore, unlike the dataset introduced in [5], which was focused exclusively on the segmentation, the occurrence of all social activities is consistent in every video and the number of activities is higher, allowing to perform experiments for the performance evaluation of the classifier. The activities of this dataset, therefore, are not manually selected and cropped in short video clips, as in previous cases.

The dataset is used to train both the temporal segmentation and the classification modules, and to evaluate the performance of the whole recognition system.

7.2 Overall System Performance

To evaluate the performance of the whole recognition system and verify the impact of the segmentation and the proximity-based priors, we calculate accuracy, precision and recall from the results of a leave-one-out cross-validation. Table 1 shows the results of our MM-DBMM classification alone and in combination with proximity-based priors generated by the simple multivariate or the GMM approximations. Three more cases are also compared: without interaction segmentation, with manual segmentation (i.e. ground truth by human expert) and with automatic segmentation. From the results, we can observe that the segmentation greatly improves the accuracy and, in particular, the precision. Indeed, the latter is affected by the number of individual activities (about half of total in the dataset) successfully excluded by the segmentation process. When using pure MM-DBMM, the recall seems the highest in absence of segmentation. This occurs because of the internal filtering of the DBMM, which tends to improve itself in longer sequences. Although, the recall

Table 2 Performance of the interaction segmentation only

	% Accuracy	% Precision	% Recall
Segmentation	92.26	92.26	92.26

Table 3 Performance of the segmentation in relation to the time interval of the HMM

HMM—interval	1s	2s	3s	Full-Seq.
Segm. accuracy (%)	92.15	92.31	92.20	92.26

in the case of Automatic Segmentation it gets lower than the other cases in all the configurations. The drop in performance is mainly due to the non-perfect segmentation, as can be see in Table 2, and it is further discussed in the next section. As expected, the results in case of automatic segmentation are not as good as with manual segmentation, although still considerably high.

Finally, Table 1 shows that integrating the proximity-based prior in the classification process improves the overall recognition performance. In particular, the GMM approximation leads to better accuracy, precision and recall than the previous multivariate Gaussian case.

The current implementation of The combined-system with non-optimised code can classify RGB-D video streams at 16 fps on average. This can further be improved by executing the different modules of the MM-DBMM and priors in parallel, since they are independent until the final merge. The component that introduces the greatest limitation in time is the segmentation module. Indeed, the HMM requires the full input sequence to perform its elaboration. In order to reduce its impact on the processing speed we have reduced the time interval processed by the HMM. In Table 3, we can observe how much the accuracy of the segmentation module decreases by decreasing the interval on which the HMM is applied.

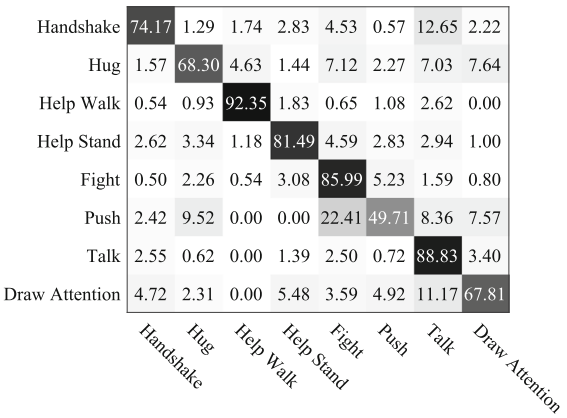
7.3 Analysis of Interaction Segmentation

To examine the performance of the segmentation model in Sect. 5, we evaluate accuracy, precision and recall with a leave-one-out experiment on our dataset (Table 2). In addition, to measure the impact of the segmentation errors on the different social activities, in Table 4 we report the percentages of false positives and negatives in segmenting each one of them.

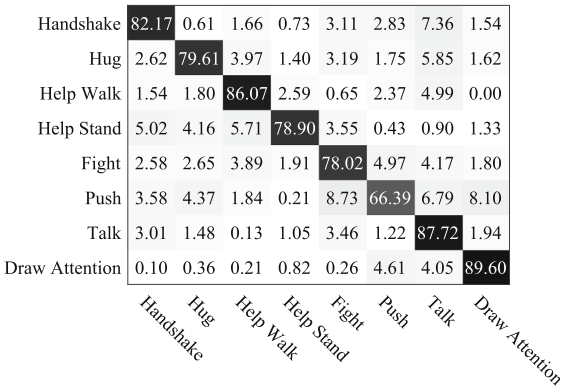
What these two tables show is that, in general, our segmentation module works very well. Although, in the last table we can notice that the segmentation errors are not equally distributed among the activity classes. The *draw attention* activity, in particular, generates more false negatives and pos-

Table 4 Percentage of the errors of the segmentation over the different classes

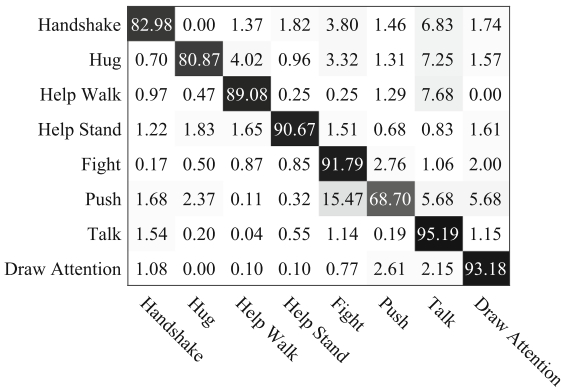
	Handshake	Hug	Help walk	Help stand	Fight	Push	Talk	Draw attention
% False negatives	2.06	1.64	1.45	7.56	16.66	7.98	3.89	58.75
% False positives	1.25	1.00	0.52	18.04	14.45	4.85	24.41	35.48



(a) Absence of Priors. Accuracy: 97.65% Precision: 76.96% Recall: 76.08%



(b) Normal based Priors. Accuracy: 97.86% Precision: 79.42% Recall: 81.06%



(c) GMM based Priors. Accuracy: 98.69% Precision: 88.01% Recall: 86.56%

Fig. 10 Confusion matrix of the MM-DBMM Classifier with manually segmented social activities

itives because often it starts before the actual interaction takes place, and it is therefore harder to detect.

It should be noticed, however, that even for a human expert it is difficult to detect precisely when an activity starts or ends, simply because an exact moment in time does not really exists. These results should therefore be taken with a 'pinch of salt' and considered only an approximate measure of the segmentation performance. As shown in the previous section, however, the segmentation module affects significantly the final results of the social activity recognition, and it is therefore a crucial component of our system.

7.4 Analysis of Social Activity Classification

A further analysis of the social activity classification, with a leave-one-out cross-validation experiment, was carried out by manually segmenting the actual interactions. This allows us to evaluate the performance of our MM-DBMM independently of the other components. From the confusion matrices in Fig. 10a, we can see that the classification of social activities is in general very good. The less accurate cases are those where the activity is very short (e.g. *push*, *draw attention*), since they provide the least number of samples. It can be observed that some activities, where the two subjects right in

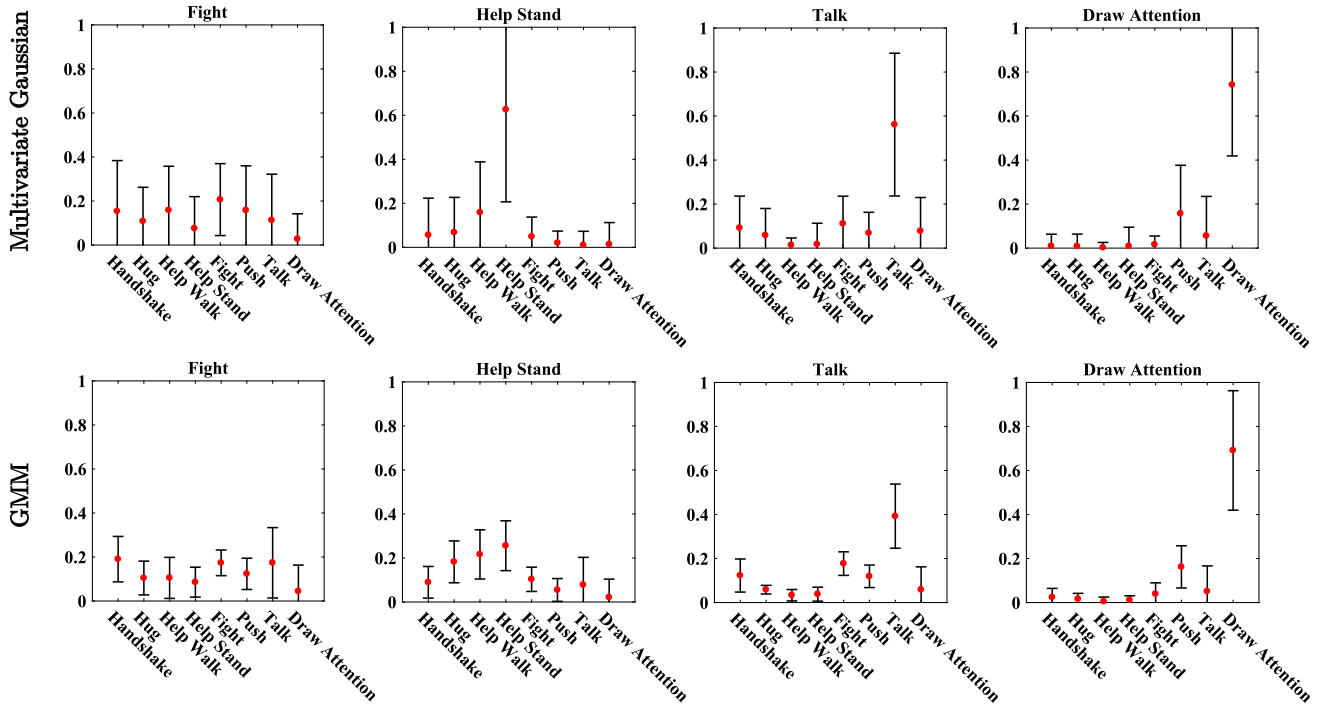


Fig. 11 Mean and standard deviation of the multivariate Gaussian (first row) and GMM (second row) priors of each activity when a the the fight, help stand, Talk and Draw Attention activities (on the top of each graph) are occurring

front of each other (e.g. *handshake*, *push*), are often confused with the *talk* case. As shown in the next section, this problem is mitigated by the introduction of our proximity-based priors.

7.5 Analysis of Proximity-Based Priors

To analyse the reliability of our proximity-based priors, we consider a specific activity and compute the means and the standard deviation of the all the remaining ones, assuming perfectly segmented videos. Even in this case we do a leave-one-out cross-validation. What we expect is that the probability of the actual activity is higher than all the other ones. Comparing the priors obtained from a simple multivariate Gaussian and a GMM approximation (Fig. 11) for some social activities, we can see that in the multivariate case the mean probability of the actual activity is higher than in GMM case, but the variance of the latter is much smaller and therefore more reliable.

The effect of these two different priors on the activity classification is shown by the confusion matrices in Fig. 10b, c. In both cases, it is clear that the proximity-based priors improve the classification of social activities. However, we can also see that the improvement is higher when GMM priors are used.

7.6 Comparative Study

To compare our classification performance with other works we tested our social activity classification model also on the SBU Kinect Interaction dataset 2.0 [39]. The latter also includes 8 dyadic social activities (*approaching*, *departing*, *pushing*, *kicking*, *punching*, *exchanging objects*, *hugging*, *shaking hands*), but in a cropped video scenario. To be more precise, the dataset includes 2 different types of segmented social activity clips (*clean*,*noisy*). In the *clean* case the clip starts and stops tightly around the activity, while in the *noisy* includes the same videos but more loosely segmented, including other random movements. For these reasons we can only compare our classification model enriched with the social priors discussed respectively in Sect. 6.

In [39], the authors evaluate the performance of their MIL-Boost classifier using the two parts of the dataset. The first evaluates the classification done on each frame of the video, while the second evaluates the performance on the classification of the full video clip. The method proposed in [20], is evaluated on full sequences on the *noisy* part of the dataset.

We compare our classification approach to the above ones, providing the accuracy achieved in all the four scenarios of the SBU Dataset, as can be seen in Table 5. Since our approach is meant for frame by frame classification, to classify the full sequence we select the most frequent label assigned in that videoclip. In our experiments, we have observed that the most frequent label occurs at least twice

Table 5 Accuracy on the SBU dataset

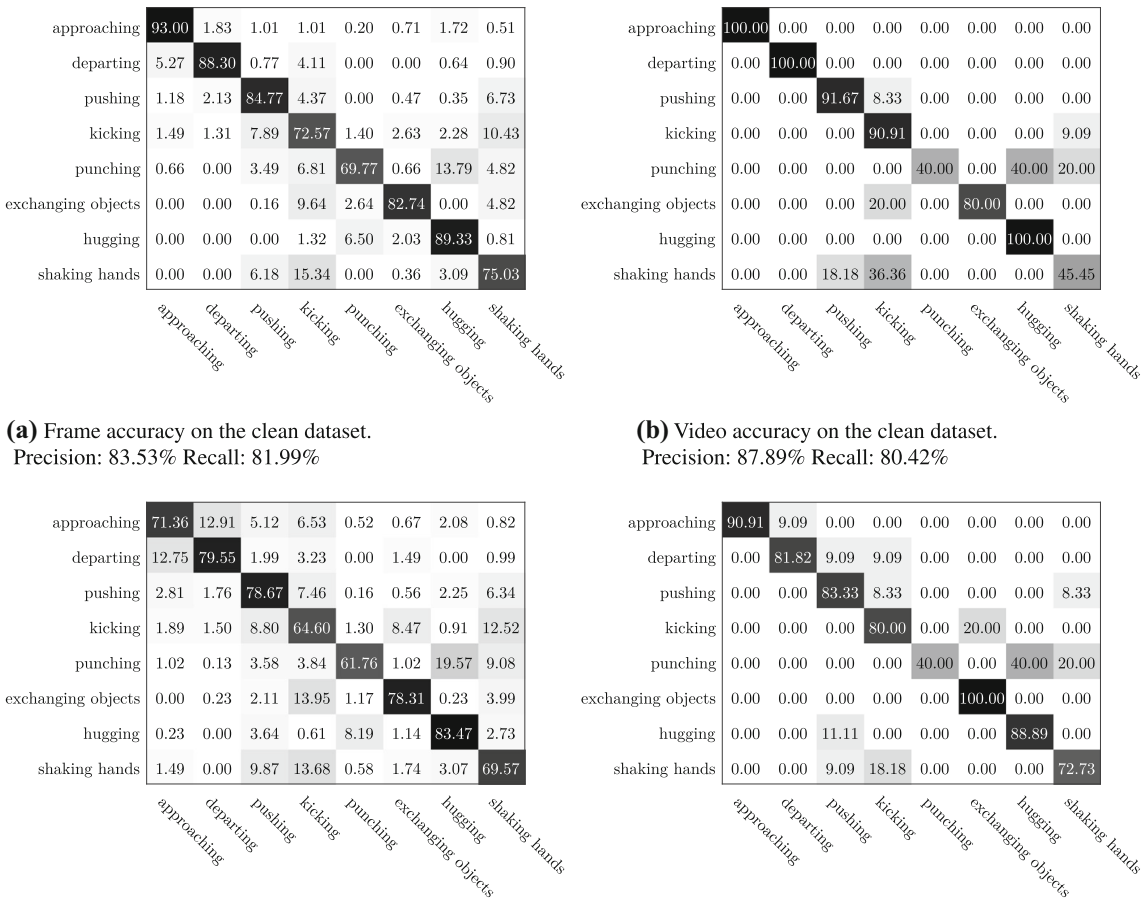
%	Reference [39]		Reference [20]		Our approach	
	Frame	Video	Frame	Video	Frame	Video
Clean	80.30	92.11	–	–	95.55	96.09
Noisy	–	87.30	–	88.00	93.40	95.14

as often as the second most frequent one. Thus, we have not seen an influence of this approach in the results. The results show how our approach outperforms the others in terms of accuracy on this dataset. More detailed information about the our classification performance is provided by the confusion matrices in Fig. 12 including precision and recall, which were provided only by [20].

8 Conclusion

Recognising social activities from a continuous stream of data is a challenging and important problem for robots to understand people's behaviour in real-world scenarios. This paper presented a novel approach for social activity recognition from continuous RGB-D skeleton data, which integrates detection and segmentation of interactions, social activity classification, and estimation of probability priors from people's proximity. Furthermore, it introduced a new dataset including individual and social activities in challenging situations. Experiments demonstrated the good performance of both the segmentation and the classification of various social activities, and that modelling the proximity distributions as a mixture of Gaussians improves the recognition even further.

An obvious limitation of the current system is the reliability on robust RGB-D skeleton trackers and (almost) full visibility of the human subjects. Such limitation could be overcome by using the most recent human pose estimation algorithms, such as [3,27]. The identification of social activities from videos, like many other problems in machine learning, are still limited by the number of cases considered in the training sets. This can reduce the applicability of the system to the real world and its relatively infinite possibilities. Future research should explore alternative ways to learn



(a) Frame accuracy on the clean dataset. Precision: 83.53% Recall: 81.99%

(b) Video accuracy on the clean dataset. Precision: 87.89% Recall: 80.42%

(c) Frame accuracy classification on the noisy dataset. Precision: 75.58% Recall: 73.49%

(d) Video accuracy on the noisy dataset. Precision: 83.13% Recall: 79.58%

Fig. 12 Confusion matrices computed in the four experiments on the SBU Dataset

from and adapt to the actual human environment where the robot operates. Extensions of this work should also consider social activities of groups with more than two persons. This could be achieved splitting groups of people considering all pairs composing it and introducing additional mixtures to the MM-DBMM model using features regarding the full groups. Further extensions should also look at new solutions, perhaps supported by the integration of alternative sensing modalities, for dealing with partial occlusions of one or both subjects.

Acknowledgements This work has been partially supported by the European project: ENRICHME, EC H2020 Grant Agreement No. 643691.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Bandla S, Grauman K (2013) Active learning of an action detector from untrimmed videos. In: Proceedings of the IEEE international conference on computer vision, pp 1833–1840
2. Bazzani L, Cristani M, Tosato D, Farenzena M, Paggetti G, Menegaz G, Murino V (2013) Social interactions by visual focus of attention in a three-dimensional environment. Expert Syst 30(2):115–127
3. Cao Z, Simon T, Wei SE, Sheikh Y (2017) Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR
4. Chakraborty I, Cheng H, Javed O (2013) 3d visual proxemics: recognizing human interactions in 3d from a single image. In: IEEE CVPR
5. Coppola C, Cosar S, Faria D, Bellotto N (2017) Automatic detection of human interactions from rgb-d data for social activity classification. In: IEEE international symposium on robot and human interactive communication
6. Coppola C, Faria DR, Nunes U, Bellotto N (2016) Social activity recognition based on probabilistic merging of skeleton features with proximity priors from rgb-d data. In: 2016 IEEE/RSJ international conference on intelligent robots and systems (IROS), IEEE. pp 5055–5061
7. Coppola C, Mozos OM, Bellotto N (2015) Applying a 3D qualitative trajectory calculus to human action recognition using depth cameras. In: IEEE/RSJ IROS workshop on assistance and service robotics in a human environment
8. Cristani M, Bazzani L, Paggetti G, Fossati A, Tosato D, Del Bue A, Menegaz G, Murino V (2011) Social interaction discovery by statistical analysis of f-formations. In: BMVC, vol 2, p 4
9. Faria DR, Premebida C, Nunes U (2014) A probalistic approach for human everyday activities recognition using body motion from RGB-D images. In: IEEE RO-MAN'14
10. Faria DR, Vieira M, Premebida C, Nunes U (2015) Probabilistic human daily activity recognition towards robot-assisted living. In: IEEE RO-MAN'15: IEEE international symposium on robot and human interactive communication. Kobe, Japan
11. Gori I, Aggarwal JK, Matthies L, Ryoo MS (2016) Multitype activity recognition in robot-centric scenarios. IEEE Robot Autom Lett 1(1):593–600. https://doi.org/10.1109/LRA.2016.2525002
12. Guo K (2012) Action recognition using log-covariance matrices of silhouette and optical-flow features. Boston University, Boston
13. Hall ET (1963) A system for the notation of proxemic behavior. American Anthropologist, Arlington
14. Jalal A, Kim YH, Kim YJ, Kamal S, Kim D (2017) Robust human activity recognition from depth video using spatiotemporal multi-fused features. Pattern Recognit 61:295–308
15. Kendon A (1970) Movement coordination in social interaction: some examples described. Acta Psychol 32:101–125
16. Kendon A (1990) Conducting interaction: patterns of behavior in focused encounters, vol 7. CUP Archive
17. Khoshhal Roudposhti K, Nunes U, Dias J (2015) Probabilistic social behavior analysis by exploring body motion-based patterns. In: IEEE PAMI
18. Koppula HS, Gupta R, Saxena A (2012) Learning human activities and object affordances from RGB-D videos. In: IJRR journal
19. Lillo I, Niebles JC, Soto A (2017) Sparse composition of body poses and atomic actions for human activity recognition in rgb-d videos. Image Vis Comput 59:63–75
20. Manzi A, Fiorini L, Limosani R, Dario P, Cavallo F (2017) Two-person activity recognition using skeleton data. IET Comput Vis 12:27–35
21. Parisi G, Weber C, Wermter S (2015) Self-organizing neural integration of pose-motion features for human action recognition. Name Front Neurorobot 9:3
22. Parisi GI, Tani J, Weber C, Wermter S (2016) Emergence of multimodal action representations from neural network self-organization. Cognit Syst Res 43:208–221
23. Piyathilaka L, Kodagoda S (2015) Human activity recognition for domestic robots. In: Field and service robotics. Springer, pp 395–408
24. Premebida C, Faria DR, Nunes U (2016) Dynamic bayesian network for semantic place classification in mobile robotics. Auton Robots 41:1161–1172
25. Premebida C, Faria DR, Souza FA, Nunes U (2015) Applying probabilistic mixture models to semantic place classification in mobile robotics. In: IEEE IROS'15, Germany
26. Rezazadegan F, Shirazi S, Upcroft B, Milford M (2017) Action recognition: from static datasets to moving robots. In: International conference on robotics and automation (ICRA)
27. Alp Güler R, Neverova N, Kokkinos I (2018) Densepose: dense human pose estimation in the wild
28. Setti F, Hung H, Cristani M (2013) Group detection in still images by f-formation modeling: a comparative study. In: 2013 14th International workshop on image analysis for multimedia interactive services (WIAMIS), IEEE. pp 1–4
29. Setti F, Russell C, Bassetti C, Cristani M (2015) F-formation detection: individuating free-standing conversational groups in images. PLoS ONE 10(5):e0123783
30. Shahroudy A, Liu J, Ng TT, Wang G (2016) Ntu rgb+d: A large scale dataset for 3d human activity analysis. In: The IEEE conference on computer vision and pattern recognition (CVPR)
31. Sommer R (1959) Studies in personal space. Sociometry 22(3):247–260
32. Sung J, Ponce C, Selman B, Saxena A (2012) Unstructured human activity detection from RGBD images. In: ICRA'12
33. Van de Weghe N (2004) Representing and reasoning about moving objects: a qualitative approach. Ph.D. thesis, Ghent University

34. Vázquez M, Steinfeld A, Hudson SE (2015) Parallel detection of conversational groups of free-standing people and tracking of their lower-body orientation. In: IEEE IROS'15, Germany
35. Vieira M, Faria DR, Nunes U (2015) Real-time application for monitoring human daily activities and risk situations in robot-assisted living. In: Robot'15: 2nd Iberian robotics conference
36. Vital J, Faria DR, Dias G, Couceiro M, Coutinho F, Ferreira N (2016) Combining discriminative spatio-temporal features for daily life activity recognition using wearable motion sensing suit. Pattern Anal Appl 20:1179–1194
37. Wang J, Liu Z, Wu Y, Yuan J (2012) Mining actionlet ensemble for action recognition with depth cameras. In: 2012 IEEE conference on computer vision and pattern recognition (CVPR), IEEE. pp 1290–1297
38. Wang J, Liu Z, Wu Y, Yuan J (2014) Learning actionlet ensemble for 3d human action recognition. IEEE Trans Pattern Anal Mach Intell 36(5):914–927. https://doi.org/10.1109/TPAMI.2013.198
39. Yun K, Honorio J, Chattopadhyay D, Berg TL, Samaras D (2012) Two-person interaction detection using body-pose features and multiple instance learning. In: 2012 IEEE computer society conference on computer vision and pattern recognition workshops (CVPRW), IEEE
40. Zhang L, Hung H (2016) Beyond f-formations: determining social involvement in free standing conversing groups from static images. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1086–1095

Claudio Coppola is a postdoctoral researcher at Queen Mary University of London and member of the Advanced Robotics Queen Mary (ARQ). He completed his PhD at the Lincoln Centre for Autonomous Systems Research (L-CAS), University of Lincoln, United Kingdom. He received his M.Sc. and B.Sc. degrees in Computer Engineering and Computer Science from the University Federico II of Napoli, Italy in 2011 and 2013, respectively. He is the author of 6 international conference papers (IROS, ECAI, ROMAN) and he has participated in a number of European and National projects. His research interest includes robot perception, activity recognition, biometrics, and Human-Robot Interaction.

Serhan Cosar is a postdoctoral research fellow at Lincoln Centre for Autonomous Systems Research (L-CAS), University of Lincoln, United Kingdom. He received his M.Sc. and Ph.D. degrees of Electronics Engineering and Computer Science from Sabanci University, Istanbul, Turkey in 2008 and 2013, respectively. He is author of 3 book chapters, 7 international journal papers (CVIU, IMAVIS) and 15 international conference papers (ACPR, ICCV, AVSS) and he has participated in a number of European and National projects. His research interest includes robot perception, human tracking, activity recognition, sparse representation, and distributed estimation.

Diego R. Faria is a Lecturer (Assistant Professor) in Computer Science, School of Engineering and Applied Science, Aston University, Birmingham, UK, since July 2016. He received his Ph.D. degree in Electrical and Computer Engineering from the University of Coimbra, Portugal, in 2014. During 2014 to 2016, he carried his research as postdoctoral fellow at the Institute of Systems and Robotics, University of Coimbra within the Automation and Robotics for Human Life group. Currently, Dr Faria is the project co-ordinator of a prestigious EU CHIST-ERA project (2019–2022): InDex - Robot In-hand Dexterous manipulation by extracting data from human manipulation of objects to improve robotic autonomy and dexterity. He has also been principal investigator of multiple seed-corn projects within the context of assistive robotics and applied machine learning. In the past, he collaborated on two large scale integrated EU projects, and multiple projects funded by the Portuguese foundation for science and technology, within a variety of topics such a as cognitive robotics, assisted living, autonomous vehicles, artificial perception, and dexterous manipulation. His research interests are: Social Robotics, Machine Perception and Applied Machine Learning.

Nicola Bellotto is a Reader in the School of Computer Science, University of Lincoln, UK, and a member of the Lincoln Centre for Autonomous Systems. His main research interests are in machine perception, especially for human detection, tracking, identification and activity recognition with autonomous mobile robots. He has a Master in Electronic Engineering from the University of Padua, Italy, and a PhD in Computer Science from the University of Essex, UK. Before joining the University of Lincoln, he was a researcher in the Active Vision Lab at the University of Oxford. Dr Bellotto is the recipient of a Google Faculty Research Award and a PI/Co-I in several EU and UK projects on autonomous mobile robots.