Automatic Detection of Human Interactions from RGB-D Data for Social Activity Classification

Claudio Coppola[1], Serhan Cosar[1], Diego R. Faria[2] and Nicola Bellotto[1]

***Abstract*— We present a system for the temporal detection of social interactions. 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 data. For this reason, it is important to be able to detect temporally the intervals of time in which humans are performing an individual activity or a social one. Recognition of the human activities is a key feature for analysing the human behaviour. In particular, recognition of social activities could be useful to trigger human-robot interactions or to detect situations of potential danger. Based on that, this research has three goals: (1) define a new set of descriptors, which are able to represent the phenomena; (2) develop a computational model, which is able to discern the intervals in which a pair of people are interacting or performing individual activities; (3) provide a public dataset with RGB-D videos where social interactions and individual activities happen in a continuous stream. Results show that using the proposed approach allows to reach a good performance in the temporal segmentation of social activities.**

I. Introduction

Recently the interest in understanding human activities and their automated recognition has been continuously increasing. Indeed the wide spectra of applications such as security, retail or health-care would benefit from such technology. Particular focus has been given to the activities performed in indoor environments due to its potential application for Active and Assisted Living (AAL).

A branch of human activity recognition, for which there has been an increased interest, is the recognition of social behaviour. The latter has acquired the interest from the psychological perspective, to understand how people's behaviours are influenced by the presence of others. Furthermore, it has attracted researchers from the computer vision and robotics fields as well. Indeed it is interesting for them to try to use this knowledge to model and design robots capable of recognising human behaviour and interacting with humans in different contexts, serving as assistants or companions. Therefore, it is important to define methodologies to detect and recognise social interactions and build datasets for social activity recognition in the real world. This paper focuses in particular on the automatic detection and temporal segmentation of social interactions from continuous video streams of RGB-D data, as illustrated in Fig. 1. For this reason, it is important to be able to detect social interactions in continuous video streams, to then properly classify the activities with models like the one we proposed in [1]. The description of the possible settings of a social interaction using body language has been deeply investigated in social science. In these research works, social interactions have been analyzed using the proxemics [2], the field of view of the involved people [3], and through the formation of specific topologies between the interacting agents [4]. In our work, we refer to social interaction as a mutual physical or visual engagement to obtain a certain goal. For example, according to this definition, people, who are talking back to back are not socially interacting, unless there is a physical contact.

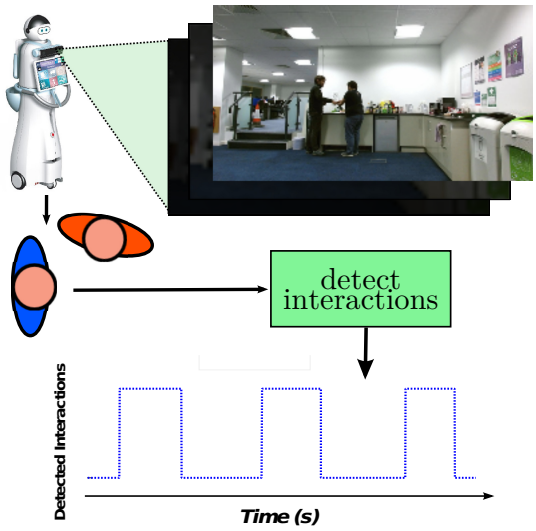


Fig. 1: Social interaction detection in time

In this paper, we perform a temporal detection of the social interactions, identifying the instants when the latter start and end, and distinguishing them from other individual activities. Such task is performed via exploitation of the upper body joints of the skeleton to compute features inspired by the social science studies previously mentioned. These features feed a classifier to estimate the likelihood of the two cases, which will be used alongside a Hidden Markov Model (HMM) to find the most probable sequence of cases. The contributions of the paper are threefold:

1) A new set of features based on the social science studies of social interactions;
2) A novel framework for detection of social interactions in a complex scenario;
3) A new public dataset for the detection of social activities in a realistic indoor setting.

This work has been supported by by the European project: ENRICHME, EC H2020 Grant Agreement No. 643691. Claudio Coppola, Serhan Cosar and Nicola Bellotto are with [1]L-CAS, School of Computer Science, University of Lincoln (UK); Diego Faria is with [2]System Analytics Research Institute, School of Engineering and Applied Science, Aston University (UK); (emails: {ccoppola, scosar, nbellotto}@lincoln.ac.uk; d.faria@aston.ac.uk).

The paper is organized as follows: in section II, we present the state of the art for activity recognition and social interaction detection. In section III, we introduce the features designed for detecting social interactions from RGBD data. In section IV, we describe the model that can temporally detect the social interactions. In section V, we describe the experiments performed and the dataset created to test our approach, and comment on our results. In section VI, we conclude discussing our approach and results, and by presenting possible directions for future research in this area.

II. Related Work

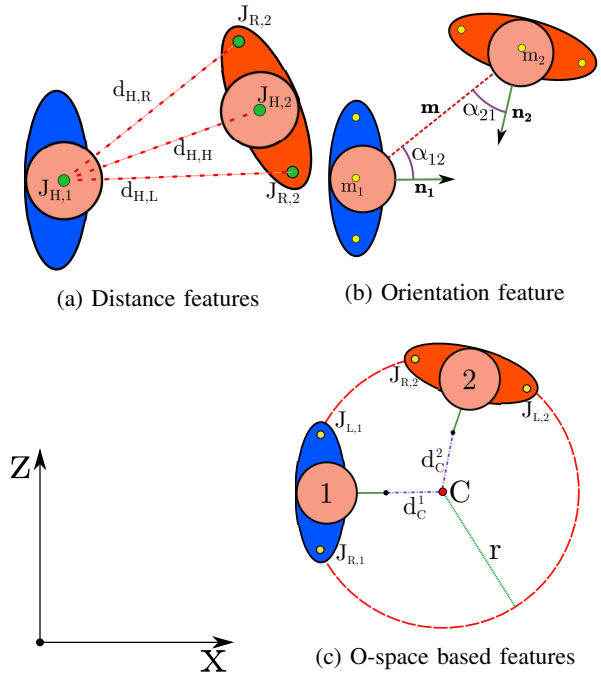
The interest for activity recognition with RGB-D cameras has been increasing due to its potential applications in robotics and in assisted living. Obtaining a stable solution capable of working in real environments is a challenging problem, but many solutions have been presented to be able to recognise activities in datasets. In [5], hierarchical self-organizing neural networks is presented to recognise human actions using depth and auditive information. In [6], the authors have applied a 3D extension of the Qualitative Trajectory Calculus (QTC) to model movements of the body joints, which have been analysed with HMMs. Faria *et al.* [7], [8] have introduced the Dynamic Bayesian Mixture Model (DBMM). It is a probability based ensemble which combines a set of classifiers through their temporal entropy. The approach presented in [9] uses HMMs implemented as a Dynamic Bayesian Network with Gaussian Mixture Models (GMM) to handle the multimodality of the data over time. A Long Short Term Memory (LSTM) based approach [10] was presented to perform classification. In [11], Jalal *et al.* present a system for activity recognition and temporal segmentation based on skeletal and silhouette features. Start and end of the activity time intervals are found comparing fitness between a non-activity model and the models of each activity built using HMMs. The intervals are then classified with an accumulative HMM. The authors in [12] present a system of human activity recognition for autonomous robots on RGB images. Convolutional networks are trained on the computed human silouhettes, while the spatial context is exploited as a prior. Like the individual activity recognition, also social activity recognition is capturing the attention of many researchers. Our previous work [1] presented an extension of DBMM for multiple mixtures, which is used to develop a system for the recognition of social activities by merging the models for two separate individuals and their social characteristics. These approaches were able to identify human activities with a good level of performance. However, they are only applicable to clipped videos of human activities or require a large amount of data. Differently from our previous work, in this paper we address the detection of human interactions to shift the recognition of social activities to more realistic scenarios.

To be able to recognise social behaviours in a real environment, it is important to be able to discern whether we are observing two separate activities from two or more individuals or a social interaction. Social sciences researchers have put a lot of effort to be able to detect social interactions through non-verbal language. In social science social interactions have been analysed through the reciprocal distance [2], the mutual presence in the field of view of the participants [3] and through the formation of specific topologies between the interacting agents [4]. These theories have been already exploited for detection of conversational groups on still images. The authors in [13], [14] present an algorithm to detect visually social interactions on RGB images using the concept of F-Formations [4]. The oriented position of the people is exploited to extract a circle (O-Space) through voting of the centre. The latter is then validated by checking the absence of external objects inside the circle. In [15], instead, the authors detect F-Formations building a graph of the people weighted on an affinity measure and detecting the dominant groups. Furthermore, a classifier is fed with social involvement features to improve the detection with association priors. In [16], the authors present a system for recognising conversational groups exploiting the orientation of the lower part of the body. In [17], social interactions are detected using the field of view of the subjects using the head orientation. The analysis is performed with the inter-relation pattern matrix, which records the cases in which an eye contact occurred. The goal of our research is indeed similar to the aim of the above works, yet different. The main intention of this work is to detect the time intervals in which a group of two people in the same area are performing a social interaction rather than two individual activities, so to be able to recognise social interactions more efficiently.

RGB-D sensors have been used widely for the recognition of human activities as well providing not only the RGB-D data, but also the tracked skeletons and, in some cases, also the objects used in the activities. In [18], the authors have introduced video clips for 16 different daily activities. In [19], the authors proposed video clips of realistic individual activities and sub-activities including information about the used objects. In [1] we introduced a dataset for recognition of social activities. In [10], a dataset containing many video clips of 60 action classes, from 3 different views, including individual and social activities is presented. Those datasets are characterised by a set of clipped videos concerning human activities. However, it is important for activity recognition to be able to perform the recognition in videos coming from continuous streams of data. Therefore, in our work, we created a dataset including long videos in which continuous streams of social interactions happen, alternating individual activities and social ones.

III. Interaction Feature-Set

Inspired by studies in social science, a new set of features have been defined to detect the intervals of time in which a social interaction occurs between two agents. The developed features are computed from the skeleton obtained by the tracking software provided with the Kinect 2 SDK and they are mainly based on geometrical properties and statistical behaviour in time of the upper bodies' position/orientation. It is important to note that our features are computed using



(a) Distance features (b) Orientation feature

(c) O-space based features

Fig. 2: Examples of the Social Feautres

only the upper body joints (e.g. shoulder, head and torso joints) on the top-view 2D plane, as can be seen in Fig 2. The features are the following:

Upper joint distances: According to the proxemics theory [2], humans make wide usage of space in their social activities creating actual sectors (*intimate, personal, social, public*) of distance depending to the intimacy of the interacting agents and several cultural factors. Therefore it is straightforward to exploit the distance between two agents. Such distance is computed over the upper body joints on two dimensions (X and Z in Kinect 2 optical frame), to be invariant to the height of the actors, according to the following formula:

$$d_{k;p} = \delta(J_{k;1}; J_{p;2}) = \sqrt{(J^X_{k;1} \quad J^X_{p;2})^2 + (J^Z_{k;1} \quad J^Z_{p;2})^2}: \quad (1)$$

where $J_{k;1}$ and $J_{p;2}$ are the positions of the upper body joints of the two skeletons with $k; p \quad 2 \quad fhead(H); shoulder\_left(L); shoulder\_right(R); shoulder\_mid(T)g$ while the $J^X_{k;i}$ $J^Z_{p;i}$ are the components on a plane parallel to the floor. In other words $d_{H;R}$ is the distance between the *head* joint of subject 1 and the *right_shoulder* joint of subject 2 (Fig. 2a).

Body mutual orientation:

People's field of view plays an important role in social interactions [3] , therefore considering the relative body orientation between persons would help to detect false positives and false negatives which the distance could not find alone. Letting $\mathbf{n_1}$ and $\mathbf{n_2}$ be the normals to the lines between the shoulder joints of the people, M_1 and M_2 their middle points and $\mathbf{m} = M_1M_2$, we consider as features the following two angles (Fig. 2b):

$$\alpha_{12} = \angle(\mathbf{n_1}; \mathbf{m}) \quad \alpha_{21} = \angle(\mathbf{n_2}; \quad \mathbf{m}) \quad (2)$$

Temporal similarity of the orientations: speakers and listeners tend to show some synchrony of their movements [20]. Therefore, to estimate the temporal similarity (or dissimilarity) of those movements, we compute the matrix logarithm of windowed moving covariance matrix of the aforementioned orientations.

$$\begin{aligned} L^t_w &= log(1 + Cov^t_w(\alpha_{12}; \alpha_{21})) \\ Cov^t_w(\alpha_{12}; \alpha_{21}) &= Cov(\alpha^t_{12\ w;:::;t}; \alpha^t_{21\ w;:::;t}) \end{aligned} \quad (3)$$

where w is the window of reference (in our case we used a window of 1 second), $\alpha^t_{\ w;:::;t}$ indicates the set of samples of α collected in the time interval $[t \quad w; t]$.

O-space radius and oriented distance: According to the theory of the F-Formations [4], a social interaction occurs when there is an overlap between the transactional segments of the actors. In this case, the interacting people form an internal circle (O-space) with their bodies to which centre they are directed. For this reason, we fit a circle using the position of the shoulder joints of the agents.

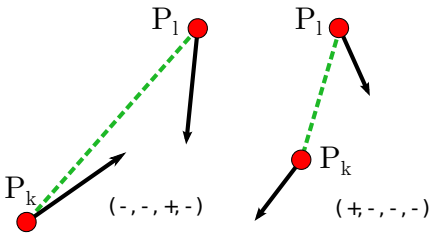
$$\begin{aligned} (C; r) &= CircleFit(J_{L;1}; J_{R;1}; J_{L;2}; J_{R;2}) \\ d^1_C &= \delta(C; M_1 + \mathbf{n_1}) \\ d^2_C &= \delta(C; M_2 + \mathbf{n_2}) \end{aligned} \quad (4)$$

The feature set is composed by $fr; d^1_C; d^2_C g$, where, r is the radius of the fitted circle, C its centre and d^i_C is the distance between the middle point of the shoulders translated in direction of the normal vector n_i and the centre of the circle (Fig.2c). It is clear that if the radius of the circle is too large, it means that the common point of focus of the two people is too far away to be considered. Furthermore, if d^i_C is higher than the radius, the people are oriented outside of the circle.

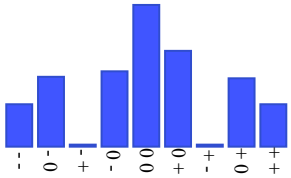
***QTC_C* relation:** The Qualitative Trajectory Calculus (QTC) is a mathematical formalism used to describe qualitative information about moving point trajectories [21]. There are different variants of this calculus, mostly dealing with points in 2D. In this work we consider the QTC_C version, where the qualitative relations between two moving points P_k and P_l are expressed by the symbols $q_i \ 2 \ f \quad ; +; 0g$ 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 $P_k\overrightarrow{P_l}$;
0: P_k is moving along $P_k\overrightarrow{P_l}$;
+: P_k is moving to the right side of $P_k\overrightarrow{P_l}$;
q_4) same as q_3, but swapping P_k and P_l.

A string of QTC symbols $fq_1; q_2; q_3; q_4g$ is a compact representation of the relative motion in 2D between P_k and P_l. For example, $f \quad ; \quad ; 0; 0g$ could be read as "P_k and P_l are moving straight towards each other". In this work, we consider the 2D trajectories of the torso joints of the



(a) Cases of QTC_C string.



(b) Example of QTC_C histogram

Fig. 3: Examples of QTC_C based features.

subjects. Further examples of QTC_C can be observed in Fig. 3a.

Temporal Histogram of QTC_C relations: Due to its compactness, QTC_C can be used to summarise the history of relative 2D trajectories between the actors using histograms. Therefore we build two windowed moving histograms of 9 bins each in time splitting components of QTC_C in two sets. The first histogram is based on the values of the first two components $(q_1; q_2)$, the other on those of the last two $(q_3; q_4)$ as shown in Fig. 3b. Separating the two histograms depends on the different nature of the two parts of the string and makes sense also because it reduces the number of bins. Indeed, having a single histogram would require $3^4 = 81$ bins and would have very sparse values, while having two separate histograms results only $2 \cdot 3^2 = 18$ bins.

Working on features space, a min-max normalisation step was applied so that the values of minimum and maximum obtained during the training stage for each type of feature were used to normalise the training and test set. Furthermore, to reduce noise on the measurement a median filter of 20 temporal samples is applied.

IV. Social Interaction Segmentation

The objective of this work is to be able to distinguish the temporal intervals of social activities, which are performed by pairs of people, from the individual ones. This would help to classify social activities in a complex scenario, rather than doing it on a dataset of clipped videos. Therefore, we need to properly classify input frames, which present a structural dependence on time. This is performed by the combination of two standard models, as shown in Fig. 4.

1) *Hidden Markov Model:* The Hidden Markov Models (HMM) are a tool to represent probability distributions over sequences of observations and therefore the classic solution for labeling sequential data. Thus, they can be used to identify the time intervals in which a social interaction

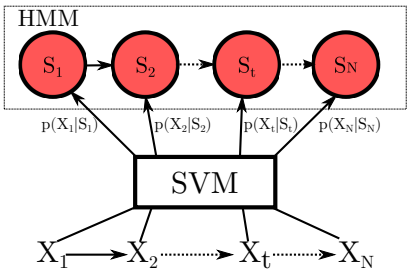


Fig. 4: Our classification approach: X_i and S_i correspond to the input at time i

is happening. HMMs model probability distribution of the world states S sequences and the input observation X through a transition probability $p(S_t|S_{t-1})$ usually modeled as a square matrix $|S| \times |S|$; an initial state probability $p(S_0)$; and an observation model $p(X_t|S_t)$ which models the probabilistic relation, usually represented as an $|S| \times |X|$ matrix or other kinds of distributions (e.g. Gaussian Gaussian Mixtures etc.), between states and observations.

2) *Support Vector Machine:* The Support Vector Machine (SVM) is an algorithm for binary classification. The SVM has been shown to be efficient even in cases of non-linearly separable data. To do so, it exploits kernels to classify data in a space of higher dimensionality. Given a set of training data consisting of N input vectors $x_1; x_2; \ldots; x_N$, where $x_i \in \mathbb{R}^n$, alongside with the labels $\hat{y}_n \in \{-1; +1\}$ of the class they belong to, the hyper-plane which separates the space in two classes is given by

$$y(x) = w^T x + b \tag{5}$$

where w is the weight vector, x is the feature vector and b is the bias. If the training data is linearly separable, then the sign of the function determines the target class assigned to the data points, i.e., $\hat{y}_n y(x_n) > 0$ holds true for all correctly classified instances. SVMs allow to compute a confidence value to their decisions, which is mainly based on the distance between x and the hyper-plane of equation 5.

In our work, we do define an HMM with two states correspondent to the two classes (individual, social), where the transition probability distribution $p(S_t|S_{t-1})$ is learnt counting the transitions in time on the labels of the training set. The observation probability is instead learnt training an SVM classifier. In such way, the output confidence of the classifier is used as a likelihood. During the testing phase, the sequence of labels is estimated computing the most probable path with the Viterbi algorithm using the output confidence of the SVM classifier as $p(X_t|S_t)$.

V. Experimental Results

A. Social-Interaction-Dataset

A new dataset (UoL-3D Social Interaction Dataset) of social interaction between two subjects has been developed to evaluate the performance of our approach and made publicly

available to the research community[1]. This dataset consists of RGB, Depth and tracked skeletons of the participants (i.e. joints' 3D coordinates and quaternions), collected with a Kinect 2 sensor. It differs from the dataset presented in [1] for its realistic scenario and setting of the performed activities. Indeed, the activities are not performed as truncated (or repeated) clips, but as a continuous stream of activities, performed over several minutes.

The dataset is organised into two parts: the first part is divided into 10 sessions composed of 2 clips performed by a specific set of participants, for a total of 20 video clips performed by 20 subjects. In each video, the participants were asked to alternate individual activities (e.g. making coffee, washing dishes etc.) and social interactions (greeting, handshaking, talking etc.) multiple times in every video. This data is meant to evaluate the performance of the temporal detection of the social activities. The second part of the dataset is composed of a single video in which the set of social activities occurs, extracted from the *3D Social Activity Dataset* [1] (handshake, hug, help walk, help stand-up, fight, push, conversation, call attention). Those are performed as a continuous stream of data in the same modality of the first part. This second part is used to evaluate the combined performance of the temporal segmentation and the social activity classification shown in [1]. Some snapshots of the new dataset are shown in Fig. 5

B. Social-Interaction-Segmentation

To test the validity of our approach, we have designed a set of experiments which include the benchmarking of the proposed set of features and the evaluation of the complete segmentation system with all the introduced features. In both cases, the approach will be evaluated using a leave-one-out cross-validation. Thus, in each iteration, 19 clips are used for training and the last one is used to test the performance of the algorithm. To compare the features, accuracy precision and recall have been calculated in a run of the cross-validation in which only one of the features was used. In Table I, we can observe that the upper joint distances outperforms the other features. The orientation similarity, the mutual orientation and, in particular, the O-space based features are also able to provide good performances. The QTC_C based features instead obtained less satisfactory results, probably because of the impossibility of the classifier to define a pattern from such kind of representation of the trajectories, but also because our interactions involved little motion of the human torsos. In any case, the distances alone are not sufficient to reach the results obtained with the concatenation of the whole feature set. The combination shows also an improvement of over 5% of the accuracy, which is significant, given that the total amount of classified frames is more than 40K.

[1] Dataset available at:
https://lcas.lincoln.ac.uk/wp/research/data-sets-software/uol-3d-social-interaction-dataset

Features	Accuracy	Precision	Recall
Upper Joint Distances	80.66	80.79	79.29
Body relative orientations	63.96	62.78	61.82
Temporal Orientation similarity	65.51	74.80	59.60
O-space based features	74.26	73.99	72.59
QTC_C relation	57.83	78.91	50.05
QTC_C histogram	59.40	57.79	54.00
Complete feature set	**85.56**	**85.55**	**84.71**

TABLE I: accuracy, precision and recall of the cross-validation when the features are used singularly and combined.

C. Social-Activity-Classification

As a proof of concept we wanted to observe the combined behaviour of the full classification system, consisting of the interaction segmentation system proposed above with the classifier presented in [1] to perform classification of the segmented intervals. In this way, we can evaluate the potential improvement in the classification performance coming from our temporal segmentation approach. To train the classifier we exploit the dataset presented in [1] (*handshake*,*hug*, *help walk*, *help stand-up*, *fight*, *push*, *conversation*, *draw attention*), while for the testing phase we use the extra video provided in the dataset presented in Section V-A. As the dataset in [1] is recorded using the OpenNI2 skeleton tracker, there is a mismatch between the number of skeleton joints in the two datasets. We have matched the number of joints by reducing the number of joints of the current dataset (25) to those of the OpenNI2 skeleton of the other one (15). To evaluate the performance, we consider the rate of samples in which the correct class is within the best three estimated by the classifier. This will show the performance of the classification system in choosing a reasonably good ranking of the classes (pruning out more than half of the possible ones) even without having the correct one as the highest. We have compared three cases: absence of segmentation (*no segmentation*), segmentation obtained with ground truth annotations (*ideal segmentation*) and the one obtained with our approach (*real segmentation*). The correct class was in the top 3 probable activities for 1 frame out of 4 in the case of *no segmentation*. In the case of *ideal segmentation*, 2 out of 3 are in the top 3, while in the case of *real segmentation* more than half of them were correctly classified. This is definitely due to the higher complexity of the setting of the activities and the way those activities were performed in [1], but also from the difference of the skeletons of the two different datasets. Despite achieving compatibility betweeen the data structures of the joints, they remain different because the quality in the tracking of the skeleton trackers used in the two datasets is different. It is still hard to compare this with other of the state-of-the-art approach since social activity recognition is still relatively new and there are no other datasets available with similar characteristics. Most works focus on the interaction detection from single RGB images, while in this work we want to recognise the time interval in which they occur from RGB-D streams.