Modeling the dynamics of individual behaviors for group detection in crowds using low-level features

Omar Adair Islas Ramírez Giovanna Varni Mihai Andries Mohamed Chetouani
Raja Chatila

Abstract

This paper introduces two novel algorithms for detecting groups of people standing or freely moving in a crowded environment. The proposed algorithms exploit low-level features extracted from videos. The first algorithm, the *Link Method*, uses a learning and forgetting strategy for modeling dynamics of proxemics between individuals. Two versions of this algorithm are proposed: they differ in the analysis of proxemics. The second one, called *Interpersonal Synchrony Method*, explicitly adopts interpersonal synchrony to refine clusters of persons detected by combining together proxemics and 2D field of view of individuals. The algorithms are evaluated on both simulated and real-world video sequences from state-of-the-art databases. Clustering metrics such as the Adjusted Mutual Information shows that our models outperform the approach based on F-formations. This work developed algorithms that can be readily applied in robotics, to allow robots to automatically detect groups in crowded environments.

Keywords. Group Detection, Synchrony, Proxemics

1 INTRODUCTION

Individuals perform a large set of activities within groups of different nature (e.g. private, public). Spontaneous and complex behaviors regulated by explicit and implicit social rules allow individuals to join, interact and leave a group. Frameworks building upon social sciences research have been proposed to describe *proxemics* in terms of private, personal and social spaces [15]. In those frameworks, a *group* is considered as a social unit comprising several members who stand in relationships with one another [11]. Groups are characterized by some durable membership and organization [13]. Furthermore, Goffman states that groups or gatherings in public places consist of any set of two or more individuals in mutual presence at a given moment who are having some form of social interaction [14].

These definitions give insights for the development of automated detection and analysis of human social behavior. In particular, in computer vision, *a group* is an entity whose members are close to each other, with a similar speed and with a similar direction of motion [12, 4]. The increasing number of applications requiring the deployment of mobile robots with possible interactions with humans has opened new challenges. For example, a social robot should be able to safely navigate in an environment populated by humans without hurting their comfort or being the source of dangerous situations. This navigation is often performed without a complete perception of the environment; perception sensing abilities is constrained by the type, the quality and the number of the sensors embedded.

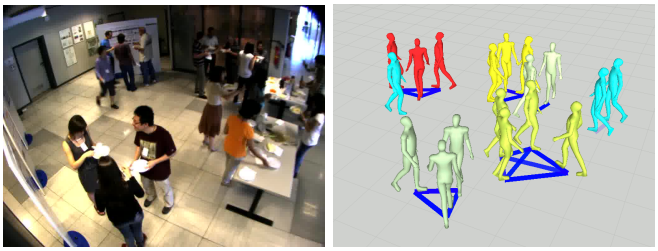

Figure 1: A snapshot from the SALSA database. Left panel shows a frame from the Cocktail party scenario. Right panel shows the groups detected by our algorithm: blue lines stand for the links between two people.

This work focuses on the development of algorithms that allow a robot to identify and track in real-time groups of persons acting in a crowded environment. This is performed by exploiting low-level features such as position, orientation and motion of individuals. The algorithms have been validated on video sequences extracted from state-of-the-art databases (Figure 1 illustrates our algorithm detecting groups of the SALSA database [2]) and they will be further exploited by a mobile robot for engagement or guiding scenarios in airports. This scenario is addressed by the EU-FP7-ICT SPENCER Project (Social situation-aware perception and action for cognitive robots) aimed at deploying a fully autonomous mobile robot to approach passengers in a socially acceptable way and to assist them [31].

Through this work, we present two main contributions:

- We go beyond some traditional approaches [8, 29, 18] that focus on frame based algorithms and whose evaluation is performed in still images by adding tracking capabilities.

- We propose robust real-time algorithms taking into account the current perception sensing abilities of a robot [31].

Sorbonne Universités, UPMC Univ Paris 06, CNRS, UMR 7222, Institut des Systèmes Intelligents et de Robotique, F-75005, Paris, France.
{islas,varni,andries,chetouani,chatila}@isir.upmc.fr

The remainder of this paper is organized as follows: section 2 summarizes the state of the art, section 3 introduces and details the methods, and section 4 presents the results. Finally, in section 5 conclusions are drawn and future research is sketched.

2 Related Work

Social gatherings have been gaining attention from the computer vision and computing communities. We present the research efforts on group detection, classifying them in two categories: identification of static groups, and identification and tracking of dynamic groups.

2.1 Identification of static groups

Tackling the scenario of free-standing conversational groups, Haritaoglu and Flickner [16] proposed a monocular real-time computer vision system for identifying shopping groups. Groups are identified by analyzing distances between the persons waiting in a checkout line or service counter.

Another approach for detecting groups employs the notion of *F-Formations*, as defined by Kendon [22]. F-formation is defined as a spatial organization of people around a shared physical space, to which they have equal, direct, and exclusive access.

A pioneering work developed by Cristani et al. [8] adopts a statistical inference over positions and orientation of standing people. However, this approach is computationally heavy and it is not able to run in real-time. Setti et al. [28] presented an unsupervised approach for group detection, that was based on a multi-scale Hough voting policy, containing voting sessions specialized for particular group cardinalities. Nonetheless, the voting approach is similar to Cristani's approach and therefore there is no improvement in computation time.

Hung and Kröse [20] used an affinity matrix to estimate the relationships among persons. They proposed a socially motivated estimate of focus of orientation based on proxemics to identify when a person is prone to be included in a group. Nonetheless, this approach is susceptible to false positives because only relative position is used to estimate the group membership. However, all these approaches focus on a frame based algorithm, so their evaluation is performed in still images. Furthermore, they are not suitable in real-time processes, because of irregular events (e.g. shaking the head) that generate noise in the outputs.

2.2 Identification and tracking of dynamic groups

Bazzani et al. [5] introduce a visual focus of attention (VFOA) in 3D of each person and then create an Inter-Relation Pattern Matrix suggesting possible social interactions in a window of time. In contrast to this work, we intend to reduce the number of tuning parameters with proxemics. Vázquez et al. [32] used the tracking of lower body pose as an input for these algorithms. They present a distribution for every subject in a scene, mixing the functions and using the Hessian of these functions to localize the centers of groups. They use the strides of a person to calculate the mentioned functions in order to find an *o-space*. Thus, the computation of the Hessian can become expensive and the use of value of fixed strides may be inconvenient with different subjects in a scenario.

Lau et al. [24] and Luber and Arras [25] cluster people using object tracking algorithms that handle fragmentation and grouping, bypassing all proxemics theory.

3 Proposed algorithms for group detection

In this work, we use the definition of *gathering in public places* provided by Goffman [14]: *a gathering consists of any set of two or more individuals in mutual presence at a given moment who are having some form of social interaction.* We argue that this definition is particularly suitable when a robot has to perform group detection tasks, considering that a robot with on-board cameras and laser is able to perceive and recognize people based on state-of-the-art computer vision techniques.

Two algorithms were conceived and developed. The first one, the *Link Method*, relies on evaluating at each instant of time the graph of possible connections between the pairs of people on the scene. Time parameters are inspired by the Ebbinghaus's forgetting curve [9]. The novelty of this approach is to merge dynamic and static analysis for group detection. The second algorithm, the *Interpersonal Synchrony Method*, grounds on the hypothesis by Fiske [10] and Lakens [23]. This hypothesis ascertains that interpersonal synchrony is as antecedent of entitativity, that is the degree to which a collection of people are perceived as a group (Campbell [6]).

The following subsections detail the methods we propose.

3.1 Link Method

This method is performed in three steps: *1) Static Analysis*, subdivided into *Link Method Simple* and *Link Method Gauss*, is inspired by proxemics; *2) Dynamic Analysis* is inspired by Ebbinghaus's forgetting curve; and *3) Forming Groups from Pairs* that allows to cluster people in groups.

Static Analysis

In this step we compute the relationships between all pairs of people acting on a scene at time t. We conceived an approach suitable when the angle between people is available by using (1, 2), and another one when this information is unavailable (3). Further, relationships between moving people are taken into account in (4,5). Then the combination of this information is used as explained subsequently.

Let us consider persons p_i and p_j described by their position and orientation (i.e. $p_i = [x_i, y_i, \theta_i]$). A Gaussian-like function f_g is projected in the space in front of person p_i at a distance $r = 0.6$ (half of personal space as in Cristani [8]).

Within this region a projection of p_j at the same distance r is evaluated inside this function as follows.

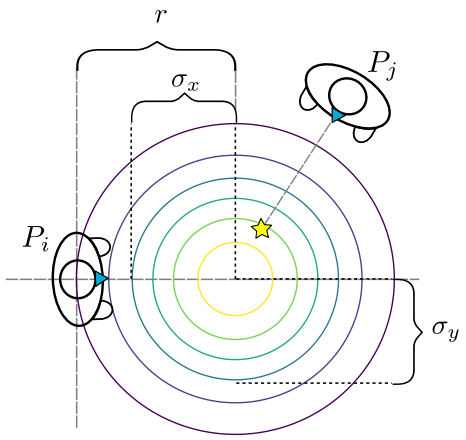

Figure 2: Gaussian-like function f_g. p_i and p_j represent the person i and j respectively. The concentric circles represent the contours of f_g in (2) with variances equal to σ_x and σ_y. The yellow star represents $[\widetilde{x_j}, \widetilde{y_j}]^T$ computed in (1).

First, we transform the projected distance of p_j to the p_i's coordinate system (1):

$$\begin{bmatrix}\widetilde{x_j}\\ \widetilde{y_j}\end{bmatrix} = Rot(-\theta_i)\begin{bmatrix}x_j + rcos(\theta_j) - x_i\\ y_j + rsin(\theta_j) - y_i\end{bmatrix} \tag{1}$$

where $Rot(-\theta_i)$ is the rotation matrix in the direction of $-\theta_i$ and $[\widetilde{x_j}, \widetilde{y_j}]$ is the projection of p_j , this value is represented by the star in Figure 2. Then this position is evaluated as follows:

$$f_g(\widetilde{x_j}, \widetilde{y_j}) = \exp\left(-\left(\frac{(\widetilde{x_j} - r)^2}{2\sigma_x^2} + \frac{\widetilde{y_j}^2}{2\sigma_y^2}\right)\right) \tag{2}$$

However, the correct orientation of people may be impossible to extract, due to constraints of the perception system such as the position of the camera inside the scene, or the type of sensor employed. In these cases, f_g can be replaced by the next equation:

$$f_d(p_i, p_j) = \frac{1}{a\left\|p_j - p_i\right\|^n + 1} \tag{3}$$

where $\|p_j - p_i\|$ is the euclidean distance of the values of position x, y for both p_i and p_j, and $a = 0.6$ and $n = 3$ are parameters empirically tuned. The difference of f_d with respect to f_g, is that a detection system with orientation will create connections with all nearby pairs of people, regardless if they are looking at different places or to the same focus of attention.

To take into account people motion, we define a further function f_v based on relative velocities between pairs of people.

$$\begin{bmatrix}\dot{\widetilde{x_j}}\\ \dot{\widetilde{y_j}}\end{bmatrix} = Rot(-\arctan2(\dot{y}_i, \dot{x}_i))\begin{bmatrix}\dot{x}_j - \dot{x}_i\\ \dot{y}_j - \dot{y}_i\end{bmatrix} \tag{4}$$

$$f_v(\dot{\widetilde{x_j}}, \dot{\widetilde{y_j}}) = \exp\left(-\left(\frac{\dot{\widetilde{x_j}}^2}{2\sigma_{\dot{x}}^2} + \frac{\dot{\widetilde{y_j}}^2}{2\sigma_{\dot{y}}^2}\right)\right) \tag{5}$$

where $\dot{\widetilde{x_j}}$ and $\dot{\widetilde{y_j}}$ are the relative linear velocities between person i and j. For $\sigma_{\dot{x}}$ and $\sigma_{\dot{y}}$, the value of both variances is $(0.2m/s)$, therefore relationships are created with pairs of people having similar velocities.

Dynamic Analysis

When a group is perceived by a person, the person retains the members of the group in mind. This *remembrance* suggests that a person, member of a group, even when he/she leaves the group, will be related to the members of the group for a certain period of time.

This step allows to keep track of pairs for a certain period of time. Thus, for each pair, using the Ebbinghaus forgetting curve [9] as inspiration:

$$g_{ij}(t+T) = \begin{cases} g_{ij}(t)\tau_f^T(\alpha_{ij}) & \text{if } \alpha_{ij} < \alpha_{th} \\ g_{ij}(t) + (1 - g_{ij}(t))\tau_l(\alpha_{ij})T & \text{otherwise} \end{cases} \tag{6}$$

$$\begin{aligned}\tau_f(\alpha_{ij}) &= 1 - \tau_{forget}\left(1 - \frac{\alpha_{ij}}{\alpha_{th}}\right)\\ \tau_l(\alpha_{ij}) &= \tau_{learn}\left(1 - \frac{\alpha_{ij} - \alpha_{th}}{1 - \alpha_{th}}\right)\end{aligned} \tag{7}$$

where t is current time, T is the period of a sampling time, τ_l and τ_f are the learning and forgetting parameters, and g_{ij} is the relationship in time between a pair. Then, $\alpha_{ij} = f_g(p_i, p_j)f_v(\dot{p}_i, \dot{p}_j)$, or without orientation of the person $\alpha_{ij} = f_d(p_i, p_j)f_v(\dot{p}_i, \dot{p}_j)$. These equations will be referred as *Link Method Gauss* and *Link Method Simple* respectively in the results section. Finally α_{th} is the threshold parameter, that means, whenever the value α_{ij} is bigger than the threshold, the "remembrance" between person i and j will increase (learn), or decrease (forget) otherwise. Figure 3 illustrates how these parameters act with respect to α_{ij} value. τ_{learn}, τ_{forget} and α_{th} are tuned parameters with values 0.3, 3 and 0.7 respectively.

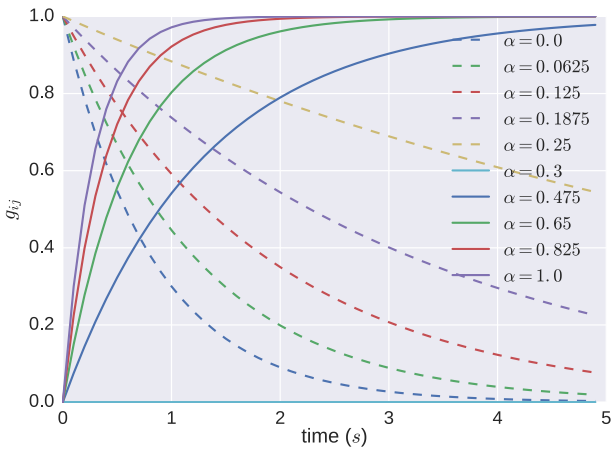

Figure 3: Remembrance curves from eq. (6) and (7) with different values of $\alpha_{ij} \in [0, 1]$. The dotted lines are inspired from Ebbinghaus Forgetting curve. The continuous lines represent the learning strategy.

Forming Groups from Pairs

In this step we define a couple of functions aimed to *1)* cluster people in groups taking computed pairs as input; and *2)* track groups in time through a similarity function Γ.

For this step, we consider all the persons in the scene as nodes of an undirected graph $\mathcal{P}$ and the pair calculation of previous steps as the edges g_{ij} of this graph.

The pseudo code implementing this step is depicted in 1.

Data: $\mathcal{P}^t$: Graph of relationships at time t.
$\mathbb{G}^{t-1}$: all groups at time $t-1$.
Result: Groups at time t ($\mathbb{G}^t$)
Algorithm `computeGroups(`$\mathcal{P}^t$,$\mathbb{G}^{t-1}$`)`
  Initialize idused vector.
  Initialize $\mathbb{G}^t$ as empty.
  `/* Loop computes groups at time` t `*/`
  for *all nodes* k *of graph* $\mathcal{P}$ **do**
    if k *is in* idused **then**
      continue
    Start empty list G
    `computeGroup(`k`)`
    if G *contains more than 1 node* **then**
      Add G to $\mathbb{G}^t$
  `/* Loop tracks groups */`
  for $G_i^t \in \mathbb{G}^t$ **do**
    for $G_j^{t-1} \in \mathbb{G}^{t-1}$ **do**
      if $\Gamma_g(G_i^t, G_i^{t-1}) > \Gamma_{th}$ *Equation (8)* **then**
        Same group, assign identical group ID
        break
  Groups without tracking ID, assign unused ID
  return $\mathbb{G}^t$
Procedure `computeGroup(`k`)`
  Add k to G
  Add k to idused
  for *all edges* g_{ki} *of node* $\mathcal{P}(k)$ **do**
    if $g_{ki} >$ *group threshold* **then**
      `computeGroup(`i`)`

Algorithm 1: Compute groups given the persons' Graph $\mathcal{P}^t$ and groups at time $t-1$ with recursive function computeGroup(k) to find the group related to person k.

The similarity function Γ is defined as follows:

$$\Gamma(G_a, G_b) = \frac{2}{N_{G_a} + N_{G_b}} \sum_{i}^{N_{G_a}} \sum_{j}^{N_{G_b}} \delta_{ij} \tag{8}$$

where δ_{ij} is the Kronecker delta. G_a and G_b are the groups to compare, each variable contains the ids of the people inside the group. N_{G_a} and N_{G_b} are the number of people that contained in each group respectively. The value of the similarity $\Gamma \in [0, 1]$ where 1 is complete similarity, therefore all the members of group in G_a are exactly the same as in G_b and 0 when none of the members of G_a is in G_b. Finally, we empirically chose $\Gamma_{th} = 0.66$ as the similarity threshold used in Algorithm 1.

3.2 Interpersonal Synchrony Method

This algorithm is performed in three steps: *(1) Pairing People from Possible Interactions*; *(2) Forming groups from Pairs*; and *(3) Thresholding of candidate groups* through intra-group synchrony. Unlike the Link Method, in which there is a step for static analysis and another one for the dynamic analysis, the Interpersonal Synchrony Method runs over sliding time-windows of a prefixed length.

Pairing People from Possible Interactions
This step is devoted to detect the relationships between all the pairs of persons acting on a scene. We conceived a strategy that combines together the inter-body distance between a couple of persons and the potential space of their interaction here defined as the area resulting from the geometrical intersection of their 2D FoV (Field of View). At each time t in a time-window of size N, for each person p_i a search of neighbors in his/her personal space of radius R is performed. When a neighbor p_j is detected, the *instantaneous intersection* of the $p_i p_j$ 2D Field of View (FoV) is checked to determine if it is empty (0) or not (1). FoV of each person is approximated with a 6-vertices polygon. The overall intersection of the FoV of p_i and p_j in the time-window N is referred as $\Psi_{i,j}$. It is computed as the summation of the instantaneous FoVs' intersections as follows:

$$\Psi_{i,j} = \frac{1}{N} \sum_{t=0}^{N-1} \psi_{i,j}^t \tag{9}$$

where N is the length of the observational window (2 s) and $\psi_{i,j}$ is the FoV intersection at the time t that can assume the value of 0 (empty intersection) or 1 (not empty intersection). $\Psi_{i,j}$ is estimated *not empty* when it is greater than 0.7, that is when p_i and p_j share their FoVs for more than 1.4 s. Then $\Psi_{i,j}$ is used as g_{ij} in Algorithm 1.

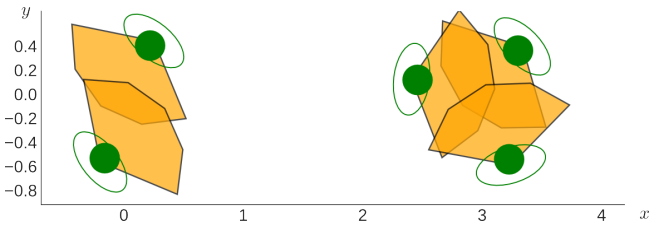

Figure 4: Reconstruction of FoVs (orange polygons) for two groups of participants (in green) acting in a synthetic scene.

Forming groups from Pairs
This second step is similar to the third step of the Link Method (see Subsection 3.1).

Thresholding of candidate groups through intra-group synchrony
This step allows to finalize the groups' detection by computing an intra-cluster synchrony index among the speed of each person supposed to belong to the same group. Starting from these speeds, the *S-estimator* synchrony index is computed and a threshold on its value is applied to identify the final groups as explained below. This index was conceived by [7] and provides the amount of synchrony relying on the eigenspectrum of the correlation matrix of a set of signals.
Let us consider a group candidate G_i composed of K persons: the speed v_i of each person is a vector that can be arranged in a matrix N x K, where N is the length of the observational window (2 s). The corresponding correlation

matrix is:

$$C = \frac{1}{N} \sum_{n=0}^{N-1} v_n v_n^T \tag{10}$$

having the following associated Λ-spectrum:

$$\Lambda = \{\lambda_1^{'}, .., \lambda_K^{'}\} \qquad \text{where} \qquad \lambda_i^{'} = \frac{\lambda_i}{\sum_{j=1}^{K} \lambda_j} \tag{11}$$

are the the normalized eigenvalues. Thus, the S-estimator is defined as:

$$S = 1 + \frac{\sum_{i=1}^{K} \lambda_i^{'} \log(\lambda_i^{'})}{\log(K)} \tag{12}$$

and has a range between 0 (for completely independent signals) and 1 (for fully synchronized signals). In our algorithm, the S-estimator is computed at each time t for each of the candidate groups, and its value is compared with the threshold value $S_{th} = 0.4$ to decide if retain or not the persons as a group (this value is defined by rule of thumb). We expect that persons having similar speeds (e.g. people traveling together) will reach a synchrony value close to 1, whereas people acting in a disjointed way (e.g. a person stands watching a notice-board and another one passes by) will have a low value of synchrony.

4 Experimental Evaluation

This section includes a description of the data sets from which we extracted video sequences used as benchmarks for our models and the evaluated results.

4.1 Data sets

Our algorithms were tested on synthetic and real video sequences. The adoption of synthetic data set is devoted to demonstrate the effectiveness of our models in ideal experimental settings, that is in scenarios where a priori occlusions, bad tracking and so on are missing.

The synthetic data set employed in this study includes simulations performed using a Robot Operating System (ROS) implementation[1] of PedSim. This simulator is based on the social force model [17, 27].

The other two data sets are the *Friends Meet* [3] and the SALSA [1] real-world data corpus. Both data sets contain annotated video sequences with humans standing or walking.

The *Friends Meet* dataset contains 15 annotated video sequences at 30fps, with lengths ranging between 20 s and 90 s with people standing and walking in outdoor area where usually they meet to have coffee breaks. The data set provides the following information: *id*, *position* (x, y) and *velocity of people* $(\dot{x}, \dot{y})$. We have inferred the people orientation θ by computing the arc tangent of the ratio of the two velocity components. This angle assumption is going to affect the algorithms performance when people are quasi-static because the orientation vector will become noisy.

The *SALSA* data set includes two 30 minutes long video sequences recorded by four synchronized static RGB cameras (1024 x 768, 15 fps). These sequences were recorded in an indoor space where 18 participants were involved in a poster session and a cocktail party, respectively. SALSA data set includes multimodal data as position, head and body orientation for each person in the scene and data from microphones, accelerometers, bluetooth and infrared sensors. This work focuses only on group detection from position data of the cocktail party scenario. However, the ground-truth annotations provided by this data set were performed only every 3 seconds; for this reason, in order to reach a finer resolution, we re-annotated both position of the people and groups. Further, groups are re-annotated following the focused and unfocused gatherings taxonomy proposed by Kendon [21]. The SALSA data set is, at the present, the most challenging data set for groups detection in ecological scenario: a large number of people interact really close to each other in an indoor environment, there are not scripted behaviors, furniture accessories influence the geometry of groups, illuminations settings changes during recordings.

Images from both data sets are shown in Figure 5.

(a) (b)

Figure 5: Images from the real data sets. **a)** Image from *SALSA* [1]. **b)** Image from *Friends Meet* [3]

4.2 Results

To evaluate the performance of our group detection models on the several data sets, two external cluster validation indexes are computed at each frame and then they are averaged over the whole length of the video sequences. These indexes measure the extent to which cluster labels match an externally supplied ground truth. Here, we adopted these measures to determine how well the groups detected by our algorithms match the ground-truth annotations. The following mutual information-based scores are chosen.

The first one, the Normalized Mutual Information (NMI) [30], is commonly used in the literature. It ranges from 0 (all the persons in a detected groups are assigned to different groups in the annotations) to 1 (all the persons in a detected groups are assigned exactly as in the annotations), but it does not have a constant baseline. To tackle this problem, we have also computed a second score, the Adjusted Mutual Information (AMI) [33]. This score is a normalized against chance variation of NMI guaranteeing a

[1]ROS implementation of PedSim `https://github.com/srl-freiburg/pedsim_ros`

constant baseline around 0 for random group assignment. In this way, we filter out the possible agreement between grouping solely due to chance. This score is upper bounded at 1 indicating a perfect agreement with the annotations. AMI is independent of the absolute values of the labels, so a permutation of the class or cluster label values will not change the score. This is more suitable when comparing people labeled as a group in ground truth and deduced as another group, but having the same members of people as in ground truth. In the state of the art other metrics are used, for example Cristani et al. [8] provide an accuracy measure based on the cardinality of a group. They assume that a group G_i with more than two participants is correctly estimated when at least $(\frac{2}{3} * |G_i|)$ of its component are found, where $|G_i|$ is the cardinality of G_i. For groups having cardinality equal to 2, all participants have to be found. The indexes we chose are less tolerant than this cardinality-based approach that assume that there is a perfect group matching when at least 67% of persons are correctly detected in a group. In other studies (e.g., [19, 29]) F_1 score is used. However, this score, defined as a combination of precision and recall is suitable for classification problems and it is not applicable when the number of detected groups is different from the number of of ground-truth groups. Two alternative F1-scores: the pairwise F1-score [26] and the cluster F1-score [18] are proposed as more specific measures to evaluate the quality of clustering.

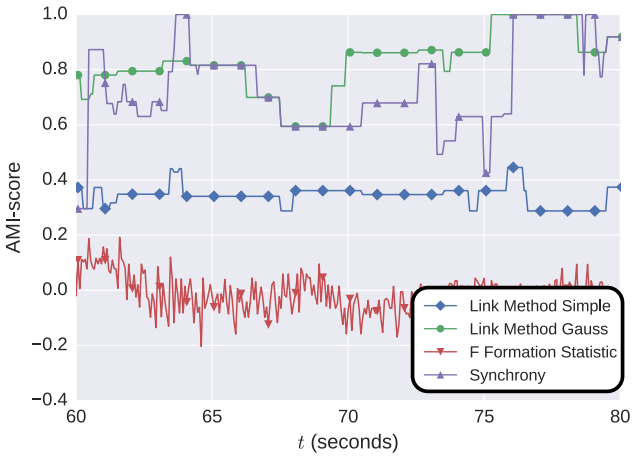

Figure 6: AMI shape during a time segment of 20 s extracted from the video sequence SALSA S1.

Evaluation was performed on video sequences extracted from the three data sets mentioned above. We tested our algorithms on two video sequences (S1 and S2) for each data set. The two sequences from SALSA are chopped at the beginning and at the end of the video to have considerable changes in groups. Table 1 shows the results in terms of average NMI and AMI for each benchmark data set. The time course of AMI is illustrated for 2 sequences in Figure 6.

Table 1 reports an R value. For *Link Method Gauss* it represents the following: $\sigma_x = \sigma_y = R$ applied on (2) and for *Interpersonal Synchrony Method* R is defined as the radius of the FoV. The indexes show that our algorithms globally work well over all the sequences extracted from the three data sets. We compared the performance of our algorithms with the most widely known F-Formation approach [8] using a grid with resolution of 10 divisions per meter and local maxima footprint of a 20x20 divisions to find centers of groups. In the future, we intend to compare our results with other state of the art dynamic approaches.

Our algorithms outperform F-formations approach over all the sequences. F-Formations implementation of Cristani have the expected performance on synthetic data. However the performance of this method is low on the real data sets having their worst performance on the SALSA sequences where the value of the indexes is very low due to random assignment according to the AMI metric. In Table 1, R* represents the value used in the Interpersonal Synchrony Method. When R* is not present, its value is equal to R. All R values are those of interpersonal space defined by Hall [15].

Link Method Simple generally exhibits a very good behavior. However its performance on the sequence SALSA S2 is not convincing at all because the people during that scene are really close to each other, and it includes all the people (even when they are not facing) due to the lack of orientation. This, however, it is expected and it was developed to be applied within systems incapable to provide orientation of people.

Link Method Gauss and *Interpersonal Synchrony Method* proved to be the most robust against both inter data sets and intra data set variations. For example, sequences S1 and S2 of SALSA differ in how cluttered people are gathered in one specific area.

The algorithms *Link Method Simple* and *Link Method Gauss* run in at around 2.5 ms and *Interpersonal Synchrony Method* in around 10 ms with 35 persons in a scene on a 2.2GHz Intel Core i7-4702MQ. Videos of the results can be seen in: `http://chronos.isir.upmc.fr/~islas/group_analysis/`

5 Conclusion

This paper presented two algorithms to detect and track groups of people in crowded environments. The first algorithm is inspired by learning and forgetting curves combined with proxemics. The second one exploits interpersonal synchrony to refine clusters of people obtained mixing proxemics and the intersections of the 2D fields-of-view of people. The algorithms are evaluated both on synthetic and real data sets through standard external cluster validation indexes and the results are encouraging. However, they revealed some limitations of our methods. For example, the *Link Simple Method* in SALSA sequences performs poorly due to the lack of orientation and its counterpart *Link Gauss Method* performs well (AMI=0.74) given the cluttered scenario where it is applied. Furthermore, dependency from some parameters and from the scenarios does not allow, at the present, a complete generalization and it will be investigated. These limitations will be addressed through a more extensive test on sequences from other synthetic and real data sets. We aim to reduce the number of

	Method	PedSim S1 (R=1.2m)	PedSim S2 (R=1.2m)	FRiends Meet S1 (R=1.2m)	FRiends Meet S2 (R=1.2m)	SALSA S1 (R=0.6m, R*=0.45m)	SALSA S2 (R=0.6m, R*=0.45m)
NMI	**F-formations**	0.93 (SD=0.02)	**0.94 (SD=0.01)**	0.45 (SD=0.16)	0.52 (SD=0.09)	0.51 (SD=0.06)	0.47 (SD=0.07)
	Link Simple	**0.98 (SD=0.01)**	**0.98 (SD=0.01)**	0.98 (SD=0.07)	0.87 (SD=0.26)	0.71 (SD=0.08)	0.05 (SD=0.11)
	Link Gauss	0.96 (SD=0.01)	**0.98 (SD=0.01)**	0.96 (SD=0.15)	0.83 (SD=0.34)	0.90 (SD=0.06)	0.91 (SD=0.91)
	Int. Synchrony	0.96 (SD=0.02)	**0.97 (SD=0.02)**	0.91 (SD=0.09)	0.88 (SD=0.08)	0.87 (SD=0.08)	0.91 (SD=0.04)
	Method	**PedSim S1 (R=1.2m)**	**PedSim S2 (R=1.2m)**	**Friends Meet S1 (R=1.2m)**	**Friends Meet S2 (R=1.2m)**	**SALSA S1 (R=0.6m, R*=0.45m)**	**SALSA S2 (R=0.6m, R*=0.45m)**
AMI	**F-formations**	**0.52 (SD=0.07)**	0.51 (SD=0.01)	-0.05 (SD=0.18)	0.17 (SD=0.09)	-0.02 (SD=0.07)	-0.06 (SD=0.07)
	Link Simple	0.85 (SD=0.09)	0.88 (SD=0.09)	**0.96 (SD=0.15)**	0.79 (SD=0.34)	0.74 (SD=0.15)	0.74 (SD=0.07)
	Link Gauss	0.80 (SD=0.07)	0.84(SD=0.08)	**0.96 (SD=0.15)**	0.79 (SD=0.34)	0.74 (SD=0.15)	0.74 (SD=0.07)
	Int. Synchrony	0.76 (SD=0.09)	**0.79 (SD=0.12)**	0.71 (SD=0.28)	0.72 (SD=0.16)	0.66 (SD=0.17)	0.75 (SD=0.08)

Table 1: Average NMI and AMI for the video sequences on which we evaluated our algorithms. In bold the best performance reached by each algorithm.

parameters as well as to find optimal parameters based on methods such as Monte-Carlo simulation.

In Human Aware Robotics, fast algorithms as ours can be advantageous for detecting groups. These methods can provide a level of membership that a robot has with respect to a group of people, i.e. at what level the robot itself is a member of a group. We aim to use them in the future to interact with groups of people. These algorithms could be used to enable social aware navigation where the robot is able to understand groups of people in order to interact with them, e.g. approaching people in shopping malls to advertise products, guiding people in an airport in order to find their boarding gate or guide people during emergencies.

Acknowledgements

This research has been supported by the European Commission under contract number FP7-ICT-600877 (SPENCER) and by Laboratory of Excellence SMART (ANR-11-LABX-65) supported by French State funds managed by the ANR within the Investissements d'Avenir programme (ANR-11-IDEX-0004-02). The authors thank Oswald Lanz for his support with the SALSA database.

References

[1] X. Alameda-Pineda, J. Staiano, R. Subramanian, L. Batrinca, E. Ricci, B. Lepri, O. Lanz, and N. Sebe. Salsa: A novel dataset for multimodal group behaviour analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PP(99):1–1, 2015.

[2] X. Alameda-Pineda, J. Staiano, R. Subramanian, L. Batrinca, E. Ricci, B. Lepri, O. Lanz, and N. Sebe. SALSA: A Novel Dataset for Multimodal Group Behaviour Analysis. 2015.

[3] L. Bazzani, M. Cristani, and V. Murino. Decentralized particle filter for joint individual-group tracking. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1886–1893, 2012.

[4] L. Bazzani, M. Cristani, and V. Murino. Decentralized particle filter for joint individual-group tracking. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1886–1893. IEEE, 2012.

[5] L. Bazzani, M. Cristani, D. Tosato, M. Farenzena, G. Paggetti, G. Menegaz, and V. Murino. Social interactions by visual focus of attention in a three-dimensional environment. *Expert Systems*, 30(2):115–127, May 2013.

[6] D. T. Campbell. Common fate, similarity, and other indices of the status of aggregates of persons as social entities. *Behavioral science*, 3(1):14–25, 1958.

[7] C. "Carmeli, M. G. Knyazeva, G. M. Innocenti, and O." De Feo. Assessment of {EEG} synchronization based on state-space analysis. *NeuroImage*, 25(2):339 – 354, 2005.

[8] M. Cristani, L. Bazzani, G. Paggetti, A. Fossati, D. Tosato, A. Del Bue, G. Menegaz, and V. Murino. Social interaction discovery by statistical analysis of F-formations. In *BMVC*, pages 1–12, 2011.

[9] H. Ebbinghaus. *Memory: A contribution to experimental psychology*. Number 3. University Microfilms, 1913.

[10] A. P. Fiske. Four modes of constituting relationships: Consubstantial assimilation; space, magnitude, time, and force; concrete procedures; abstract symbolism. *Relational models theory: A contemporary overview*, pages 61–146, 2004.

[11] D. R. Forsyth. *Group Dynamics*. Wadsworth/Cengage Learning, 2010.

[12] W. Ge, R. T. Collins, and R. B. Ruback. Vision-based analysis of small groups in pedestrian crowds. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(5):1003–1016, 2012.

[13] E. Goffman. Encounters: Two studies in the sociology of interaction. 1961.

[14] E. Goffman. *Behavior in Public Places*. 1966.

[15] E. T. Hall. *The hidden dimension*, volume 1990. Anchor Books New York, 1969.

[16] I. Haritaoglu and M. Flickner. Detection and tracking of shopping groups in stores. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001. CVPR 2001*, pages I–431–I–438 vol.1, 2001.

[17] D. Helbing and P. Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995.

[18] Jian Huang, Seyda Ertekin, and C. Lee Giles. Efficient name disambiguation for large-scale databases. In *Proceedings of the 10th European Conference on Principle and Practice of Knowledge Discovery in Databases*, PKDD'06, pages 536–544, Berlin, Heidelberg, 2006. Springer-Verlag.

[19] H. Hung and D. Gatica-Perez. Estimating Cohesion in Small Groups Using Audio-Visual Nonverbal Behavior. *IEEE Transactions on Multimedia*, 12(6):563–575, 2010.

[20] H. Hung and B. Kröse. Detecting F-formations as dominant sets. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 231–238. ACM, 2011.

[21] A. Kendon. How gestures can become like words. *Crosscultural Perspectives in Nonverbal Communication*, 1988.

[22] A. Kendon. *Conducting interaction: Patterns of behavior in focused encounters*, volume 7. CUP Archive, 1990.

[23] D. Lakens. Movement synchrony and perceived entitativity. *Journal of Experimental Social Psychology*, 46 (5):701–708, 2010.

[24] B. Lau, K. O. Arras, and W. Burgard. Multi-model hypothesis group tracking and group size estimation. *International Journal of Social Robotics*, 2(1):19–30, 2010.

[25] M. Luber and K. O. Arras. Multi-Hypothesis Social Grouping and Tracking for Mobile Robots. In *Robotics: Science and Systems*, 2013.

[26] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, 2008.

[27] M. Moussaïd, N. Perozo, S. Garnier, D. Helbing, and G. Theraulaz. The walking behaviour of pedestrian social groups and its impact on crowd dynamics. *PloS one*, 5(4):e10047, 2010.

[28] F. Setti, O. Lanz, R. Ferrario, V. Murino, and M. Cristani. Multi-scale F-formation discovery for group detection. In *Image Processing (ICIP), 2013 20th IEEE International Conference on*, pages 3547–3551.

[29] F. Setti, H. Hung, and M. Cristani. Group detection in still images by F-formation modeling: A comparative study. In *Image Analysis for Multimedia Interactive Services (WIAMIS), 2013 14th International Workshop on*, pages 1–4. IEEE, 2013.

[30] Alexander Strehl and Joydeep Ghosh. Cluster ensembles-a knowledge reuse framework for combining partitionings. In *AAAI/IAAI*, pages 93–99, 2002.

[31] R. Triebel, K. O. Arras, R. Alami, L. Beyer, S. Breuers, R. Chatila, M. Chetouani, D. Cremers, V. Evers, M. Fiore, and others. SPENCER: A socially aware service robot for passenger guidance and help in busy airports. 2015.

[32] 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 *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pages 3010–3017. IEEE, 2015.

[33] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research*, 11:2837–2854, 2010.