# Robust detection of conversational groups using a voting scheme and a memory process

Victor Fortier, Isabelle Bloch, Catherine Pelachaud

# Robust detection of conversational groups using a voting scheme and a memory process *

Victor Fortier[1], Isabelle Bloch[1][0000−0002−6984−1532], and Catherine Pélachaud[2][0000−0003−1008−0799]

[1] Sorbonne Université, CNRS, LIP6, Paris, France
victor.fortier@etu.sorbonne-universite.fr,
isabelle.bloch@sorbonne-universite.fr
[2] CNRS, ISIR, Sorbonne Université, Paris, France
catherine.pelachaud@sorbonne-universite.fr

**Abstract.** Studies in human-human interaction have introduced the concept of F-formation to describe the spatial organization of participants during social interaction. This paper aims at detecting such F-formations in images of video sequences. The proposed approach combines a voting scheme in the visual field of each participant and a memory process to make the detection in each frame robust to small, irrelevant changes of participant's behavior. Results on the MatchNMingle data set demonstrate the good performances of this approach.

**Keywords:** F-formation · Clustering · Temporal regularity.

## 1 Introduction

Participants during social interaction place themselves in certain spatial formations, so as to see each other and respect social and cultural distancing [4]. They can face each other, be side by side... Their position and behavior such as body orientation and gaze behavior can indicate a great quantity of information; they can reveal information about their level of engagement, their focus of interest but also the quality of their relationship, their degree of intimacy, to name a few [1]. Participants' position and behavior evolve continuously to accommodate others' behaviors and to obey to some socio-cultural norms. A group can be defined as an entity where individuals are spatially close, and each member is able to see and know the other members. Group members perform a common, shared activity by interacting socially. People can be simply gathered spatially (e.g. people in a queue), doing an action together but do not interact together (e.g. watching a film at the cinema) or discussing together on a given topic. Studies in human-human interaction have introduced the concept of F-formation [8] that defines three zones: O-space, P-space and R-space. The O-space corresponds to the convex space between the participants of a group; the P-space corresponds

* This work was partly supported by the chair of I. Bloch in Artificial Intelligence (Sorbonne Université and SCAI).

to the belt where the participants are; and the R-space is the space outside the participants.

Lately computational models have been designed to detect if individuals form a group and what is its formation based on proxemics and behaviors [2]. Further analysis can be pursued to characterize the dynamics of the social interaction between participants. Such models can then be used to drive the behaviors of robots when interacting with humans.

The aim of our study is to detect, analyze and understand social interactions from images and videos, in order to build computational models of social interactions. We focus on free-standing conversational groups with limited size (typically 2-6 persons) that are discussing with each other [9].

To this aim, we rely on the existing database MatchNMingle [2]. This database contains videos of group interaction that have been annotated at different levels (activity, speaking, laughing, non-verbal behavior). As a first step, we detect group formations using still images and consider only two visual cues, namely distance and gaze direction. The proposed approach is based on a voting procedure to find the O-spaces and the groups. It does not require any heavy learning method. Moreover, a new feature of the approach is that the detection in still images is made more robust by exploiting the temporal information in the video sequence.

Related work is briefly summarized in Section 2, the proposed approach is described in Section 3, and results on the MatchNMingle data set [2] are provided in Section 4.

## 2 Related work

One of the pioneering methods to detect F-formations from images is called the "Hough Voting for F-formations", which constructs a Hough accumulator and where groups are extracted from it by searching for local maxima [3]. The method reduces the detection of F-formations to that of O-spaces. An O-space corresponds to the intersection of the visual fields of its participants. Thus, the method models each individual's field of attention by drawing many samples from a 2D Gaussian distribution centered at some distance from its position and respecting its orientation. Each sample corresponds to a vote that remains to be aggregated in the Hough accumulator. Finally, the local maxima correspond to the positions that received votes from most of the individuals in the scene. That is, they correspond to the positions of the O-space centers and thus to the searched F-formations. Further studies applied the paradigm of the voting process in a Hough space [12, 13]. Later on, the same authors proposed an approach based on graph-cut to optimize an objective function defined from the probability of the assignment of participants to groups, under minimum description length constraint to limit the number of clusters [14]. These authors also addressed detection in videos using a game-theoretic approach [16]. With respect to the static approach, this approach includes an additional fusion step over a sequence. However groups are fixed in time. Other fusion approaches have been

proposed, with the same drawback. To account for evolution over time, tracking methods have been used, but this is out of the scope of this paper.

Hung and Kröse [7] proposed to model the interactions between individuals as a graph. The analysis of the graph gives information on group formations. The authors defined the "Dominants Sets for F-formations" method that sees F-formations as entities where the affinity (probability of interaction) between all members of a group is higher than that between a member of the group and an individual outside the group. This definition is closed to the dominant set problem which is known to be an NP-complete problem. The dominant sets of a graph are subsets of vertices of a graph for which any of its vertices is either a leaf, or is connected by an edge to an element [5]. A group can be viewed as an undirected graph where each vertex corresponds to an individual and edges are weighted. A weight on an edge corresponds to the affinity between individuals linked by the edge. It is estimated from their body positions and orientations. The dominant sets can be detected by optimization methods that provide an approximate solution to determine the groups in a scene.

Lately, Thompson and colleagues [15] applied to dominant sets a message-passing Graph Neural Network (GNN) to predict how individuals are grouped together. This approach requires a lot of data and annotations, so as to train properly the network. Other methods based on neural network and deep learning have similar strong requirements, and are out of the scope of this paper.

The approach we propose is lighter as only a very limited learning step is required to set parameters once for all, and the whole method has a low complexity. Moreover, we exploit the temporal information to improve the detection in each frame, providing a better regularity over time, including the short changes in position or gaze direction, while keeping the meaningful changes (i.e. not assuming that the groups are fixed over time).

## 3 Proposed approach: Multiple votes and exploiting temporal information

### 3.1 Voting in each frame

The main idea of the proposed approach is to detect F-formations by identifying the O-spaces of each group. To this end, we use the position, orientation and field of view of each person to model "votes" for a O-space center, drawn from uniform distributions in the field of view.

More precisely, let $(x_i, y_i)$ denote the position of person (or participant) $i$ in the image, and $\theta_i$ her orientation. Her field of view is defined as a cone of aperture $\alpha \in [0, 2\pi]$, truncated at a minimal radius $\gamma d_{\max}$ and a maximal radius $\beta d_{\max}$ where $\gamma$ and $\beta$ are parameters in $[0, 1]$, with $\gamma \leq \beta$, and $d_{\max}$ the maximal distance on the scene (in practice the length of the image diagonal). Each vote provides a potential O-space center $(x_i^k, y_i^k)$, $k \leq n_s$, where $n_s$ denotes the number of samples, defined as:

$$x_i^k = x_i + d^k \cos \theta_i^k \tag{1}$$

$$y_i^k = y_i + d^k \sin \theta_i^k \tag{2}$$

where $d^k \sim U([\gamma d_{\max}, \beta d_{\max}])$ and $\theta_i^k \sim U([\theta_i - \frac{\alpha}{2}, \theta_i + \frac{\alpha}{2}])$.

If the votes of two persons are close to each other, this means that the persons are close to each other and in a spatial configuration that allows for interaction. Otherwise, they are either too far from each other or not looking at each other, or not looking at a common region of space. Relevant O-spaces (and thus, F-formations) can therefore be identified by clustering the votes according to the Euclidean distance. Since the number of clusters is not known, a simple idea consists in applying any clustering method (e.g. K-means), with different numbers of clusters, and choosing the best one. To this end, we propose to define a score of a clustering based on the silhouette, measuring the similarity of a vote with the cluster it belongs to (mean distance to the other votes in the cluster), and its dissimilarity with the other clusters (smallest mean distance to the votes in the other clusters) [11]. The scores for each vote are then averaged for a cluster, and then for all the clusters, thus providing a global score for a given clustering. The clustering with the highest global score is finally chosen. Each person $i$ is then assigned to the cluster containing the majority of her votes. All persons assigned to the same cluster belong to the same F-formation.

The principle of the proposed approach is illustrated in Figure 1.

To fasten the computation, a simplified approach consists in setting $\alpha = 0$, and $\gamma = \beta$, which means that the field of view of each person is reduced to one point, corresponding to a unique vote for this person. However, this is a strong limitation with respect to the human perception, which impacts the results, as will be shown in Section 4.

### 3.2 Increasing robustness by taking into account temporal memory

Instead of applying the previous approach in a static way, where each frame is processed independently, we now propose to exploit the dynamics of videos. The underlying hypothesis is that F-formations are usually quite stable over time, with a few changes from time to time, where some persons can leave a group to join another one. Moreover, small changes for instance in a person's body orientation or gaze direction, during a very short time, do not imply that this person has left the group. To model this behavior, we propose a method inspired by the work in [10]. We define a memory process, where two persons $i$ and $j$ in a group increase progressively their interaction (learning process) while forgetting progressively their interaction when they are no more in the same group (forgetting process). In particular, if a person briefly looks in another direction (or towards another F-formation) and rapidly returns to looking at her partners of her initial F-formation, this change of behavior will not be considered as changing group, thus increasing the stability of the detected F-formations. More formally, the memory process at time $t$ and during a time interval $\Delta t$ is

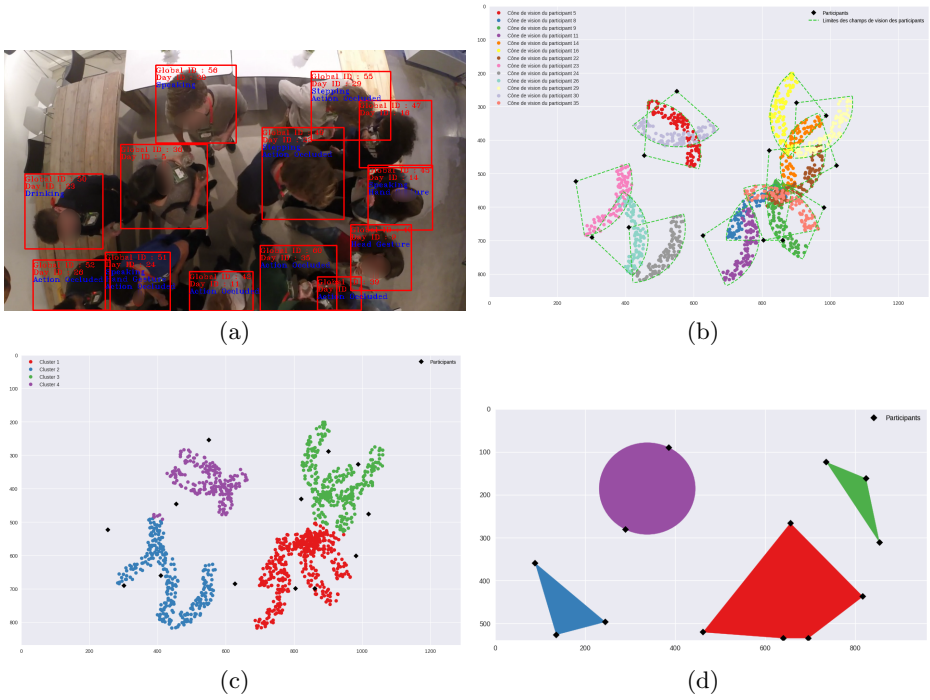



**Fig. 1.** Illustration of the identification of F-formations: (a) original image, (b) votes within the field of view, (c) clustering of the votes, and (d) obtained F-formations.

modeled as:

$$M_{i,j}(t+\Delta t) = \begin{cases} (1-\frac{\Delta t}{\tau_l})M_{i,j}(t) + \frac{\Delta t}{\tau_l} & \text{if } i \text{ and } j \text{ belong to the same group} \\ (1-\frac{\Delta t}{\tau_f})M_{i,j}(t) & \text{otherwise} \end{cases} \tag{3}$$

where $\tau_l$ is the learning rate and $\tau_f$ the forgetting rate. They are not necessarily equal, and if the learning process is considered faster than the forgetting process, then we will choose $\tau_l < \tau_f$. The process is initialized by setting $M_{i,j} = 0$ for all $i$ and $j$, $i \neq j$. The time interval $\Delta t$ is typically defined from the step $s$ between frames to be analyzed and the video frequency $f$ (e.g. $\Delta t = s/f$). For instance, to analyze every 20th frame in a video (such as in the Mingle base) with 20 frames per second, then we can set $\Delta t = 1$. Figure 2 illustrates two different behaviors of a person's memory with respect to a second person.

Two persons $i$ and $j$ are considered to interact at time $t$ if $M_{i,j}(t)$ is above some threshold value (0.5 in our experiments). A graph is then defined, where vertices are the persons present in the scene and there is an edge between $i$ and $j$ if the corresponding persons do interact. Groups are then obtained by selecting

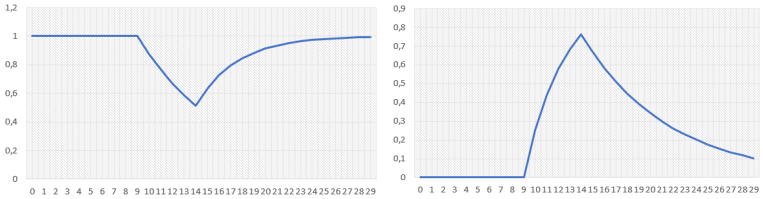


**Fig. 2.** Examples of the evolution of the memory in time. Left: a person knows a second person, then moves to another group and progressively forgets the second person, then joins the group again, and so "re-learns" the second person. Right: no interaction between two persons, then progressive learning, and then forgetting again.

the maximal cliques in the graph. An example is illustrated in Figure 3. Note that a person can belong to several groups.

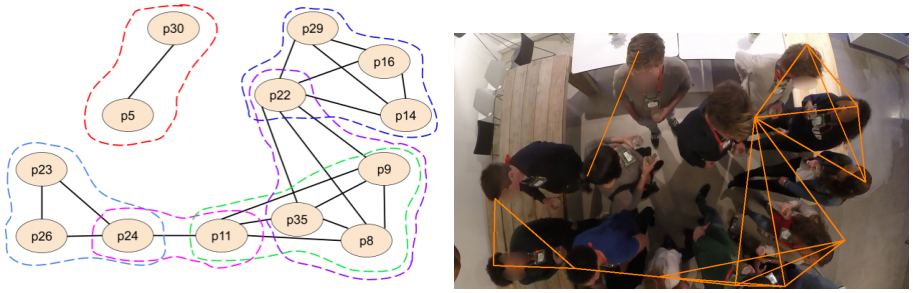



**Fig. 3.** Left: graph where vertices are persons and edges represent their interactions, and maximal cliques. Right: edges of the graph superimposed on the image.

Now, to restrict the groups to F-formations where one person can belong to only one of them, as is the case in the ground-truth of the Mingle data base, a last assignment step is required. The membership of person $i$ to a group $C_k$ at time $t$ is defined as:

$$\mu(i,k,t) = \frac{1}{|C_k|-1} \sum_{j \in C_k, j \neq i} M_{i,j}(t) \tag{4}$$

and the assignment is then $\arg\max_k \mu(i,k,t)$. A result is illustrated in Figure 4. The comparison with the static approach, using only the information from this particular frame, without using the memory from frame to frame, shows a better consistency in the result when using the memory process, and better matches with the ground truth. In particular persons unassigned in the static mode are now correctly assigned to a group, and wrong assignments are also corrected.

**Fig. 4.** Left: final result, after maximal cliques detection and assignment of each person to one group. Right: result on the same frame, but without using the memory process.

## 4 Experiments and results

### 4.1 MatchNMingle data set

Experiments were carried out on the MatchNMingle data base [2]. This data base consists of "speed-dating" videos (the Match data base), and of cocktail videos (the Mingle data base). This second part was used here. Several videos show participants moving freely in space, from one group to another one, with potentially complex interactions. These data are therefore relevant to demonstrate the usefulness of the proposed approach. Annotations are available and include:

- the list of participants;
- the spatial coordinates $(x_i, y_i)$ of each participant $i$ in each frame;
- the head orientation $\theta_i^{head}$ and body orientation $\theta_i^{body}$ of each participant in each frame;
- the coordinates $(x_1, y_1), (x_2, y_2)$ of the diagonal points of the bounding box of each participant in each frame;
- the F-formations, where each F-formation is defined as the set of its participants, its starting frame and its ending frame.

In our experiments, we used the position of the participants and the head orientation ($\theta_i = \theta_i^{head}$) as input data (see Section 3), and infer the F-formations.

### 4.2 Evaluation criteria

The obtained F-formations can be compared to the ground-truth (provided as annotations with the data set). This amounts to compare two clusterings of the same data. A common measure to this end is the adjusted rand index (ARI) [6], which allows comparing two partitions of the same data, with potentially different cardinalities. This index takes values in $[-1, 1]$, and two partitions are considered as approximately similar if the index is higher than 0.5.

Let $F_t$ denotes the partition (i.e. the set of F-formations) obtained in frame $t$ of a video, and $G_t$ the corresponding ground-truth. We denote by $ARI(F_t, G_t)$

the adjusted rand index comparing $F_t$ and $G_t$. Averaging over all processed frames ($t = 1...T$) leads to a global score $EC = \frac{1}{T}\sum_{t=1}^{T} ARI(F_t, G_t)$, which should be as close to 1 as possible.

### 4.3 Parameter setting

In our experiments, we set the number of samples in the voting procedure to $n_s = 100$. The field of view is parametrized by the aperture $\alpha$, and the coefficients $\beta$ and $\gamma$ defining the truncation in terms of distance. According to the human perception, the maximum angle of vision is $\frac{2\pi}{3}$, and we performed experiments with $\alpha \in \{0, \frac{\pi}{6}, \frac{\pi}{3}, \frac{\pi}{2}, \frac{2\pi}{3}\}$. The parameter $\beta$ defines the maximal distance $\beta d_{\max}$ at which votes can occur. In our experiments, $d_{\max}$ was set to the length of the image diagonal, and we limited $\beta$ to 0.25 so as not to allow for votes that are too far away. Experiments were performed with $\beta \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$, and $\gamma$ is defined as $\gamma = (1 - \gamma')\beta$, where $\gamma' \in \{0, 0.25, 0.75, 1\}$, with the constraint $\gamma \leq \beta$. The choice of particular values of these three parameters was done using a small part of the data base (every 20th frames in a sequence of 5000 frames in one of the videos of the Mingle data base). All the values were tested, and the ones providing the best $EC$ values (computed over these frames only) were chosen. An example of the $EC$ values obtained for $\alpha = \pi/2$ is illustrated in Figure 5. Similar tests were also performed for other values of $\alpha$. Although this is a greedy approach, it is performed only once, and the number of parameters combinations to test remains limited. Then the obtained optimal parameters were fixed for all other frames and all videos. Finally, all experiments have been performed with parameters $\alpha = \pi/2, \beta = 0.15, \gamma = 0.11$. An example of the corresponding visual field is illustrated in Figure 6. To illustrate the usefulness of the visual field and the votes, results are also compared with $\alpha = 0, \beta = \gamma = 0.1$, i.e. each participant votes for only one point, the center of the O-space.

In the clustering procedure, the number of clusters varies from 2 to $n_p$, where $n_p$ denotes the number of persons in the scene. The parameters of the memory process are set as follows in our experiments: $\Delta t = 1$, $\tau_l = 3$ and $\tau_f = 8$.

### 4.4 Results

Qualitative results have been illustrated in Section 3, and demonstrate the improvement brought by the memory process over a purely static, frame by frame, approach. In this section, we propose a quantitative evaluation.

Figure 7 illustrates, on a sequence of one of the videos, the ARI values computed using:

1. a simple method, where only the distance between two persons is used to decide whether there should be grouped into a F-formation. This is performed by determining the connected components of a graph where vertices represent participants, and two vertices are linked by an edge if the distance between the corresponding participants is less than $\beta d_{\max}$. This method is obviously too simple, as shown by the low ARI values, below 0.5 most of the time, and is no longer considered in the remaining of our evaluation;

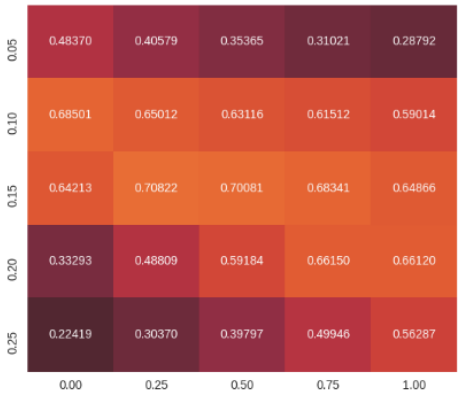



**Fig. 5.** $EC$ values obtained on the frames used for training, for $\alpha = \pi/2$, and different values of $\beta$ (ordinate) and $\gamma' = 1 - \gamma/\beta$ (abscissa). The highest score is obtained for $\beta = 0.15, \gamma' = 0.25$ (i.e. $\gamma = 0.11$).

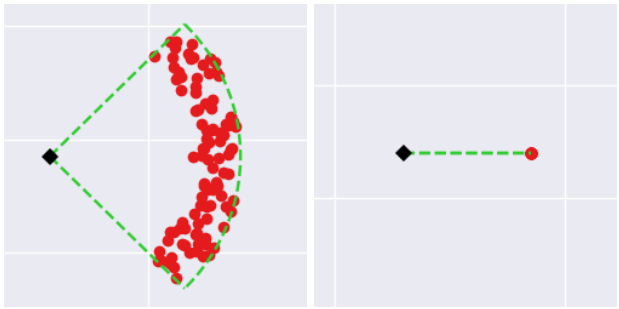
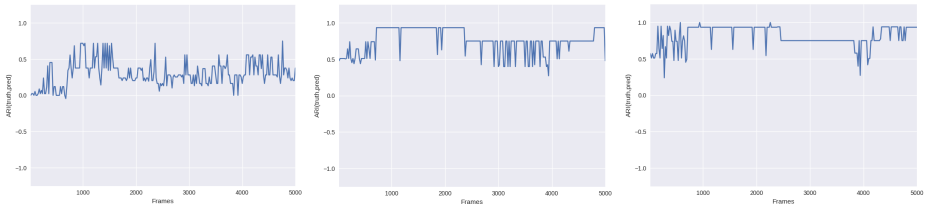


**Fig. 6.** Left: example of a visual field, in which votes are drawn, for the final parameter setting $\alpha = \pi/2, \beta = 0.15, \gamma = 0.11$. Right: vote for only one point, the center of the O-space, corresponding to $\alpha = 0, \beta = \gamma = 0.1$.

2. method 2: our proposed method, using distance and angle, but simplified by using parameters $\alpha = 0, \beta = \gamma$, i.e. each participant votes for only one point (center of the O-space). This shows a clear improvement over the previous naive approach;
3. method 3: our proposed method where several votes are drawn in the visual field. The results are even better, both in terms of ARI values and regularity of the detections. The method is also less sensitive to orientation due to the use of the visual field.

One can notice a few bad detections, e.g. frame 3940, where the ARI values are 0.27 for methods 2 and 3. This is improved by the memory process, as demonstrated next. The overall scores over the considered frames for the three methods are $EC_1 = 0.3, EC_2 = 0.76, EC_3 = 0.81$, respectively.



(a) (b) (c)

**Fig. 7.** Comparison of the ARI values on a sequence of frames of one of the Mingle videos. (a) Only distance is used (global score $EC_1 = 0.3$). (b) Distance and angle are used, with only one vote per participant ($EC_2 = 0.76$). (c) Proposed voting method, where each participant votes several times in her visual field ($EC_3 = 0.81$).

Figure 8 compares the results for methods 2 and 3, without and with the memory process. Without the memory process, the global scores are $EC_2 = 0.76, EC_3 = 0.81$, while with the memory process they are $EC_2 = 0.78, EC_3 = 0.82$. This shows a slight improvement of the detection of F-formations, and a high improvement of the temporal regularity, which better matches what is intuitively expected in such scenarios. The results also confirm the superiority of the proposed method as described in Section 3, enhanced with the memory process to increase robustness to brief, non significant changes of head orientation.

Without any specific code optimization, the computation time is 1.5 second in average on a standard computer, and most of the time is spent on the clustering part. The memory process is only 1% of the total time. Note that the time is reduced to 0.23 second in average if any individual votes only once.

## 5 Conclusion

We proposed in this paper a simple yet robust method to detect F-formations in images. In particular, the obtained groups are relevant and the influence of brief changes of position or gaze direction, which do not mean changes of group, is reduced thanks to a memory process. The results demonstrated the good performance of this approach, both in each frame using the ARI index as an evaluation measure, and over a video sequence using the global score $EC$.

While we assumed here that the position and orientation of the participants were known or obtained in a preliminary step, the proposed approach could be directly combined with one the numerous existing methods for this preliminary step.

The obtained values of ARI values over time could also be a hint for change detection, using graphs such as the one in Figure 8 (d). Lower ARI values indicate changes in the F-formations. As an example, frames 720, 740 and 760, illustrated in Figure 9, show such changes. The stability of the F-formations along time can be evaluated by $SC = \frac{1}{t_2 - t_1} \sum_{t=t_1}^{t_2-1} ARI(F_t, F_{t+1})$, which will be close to 1 if the

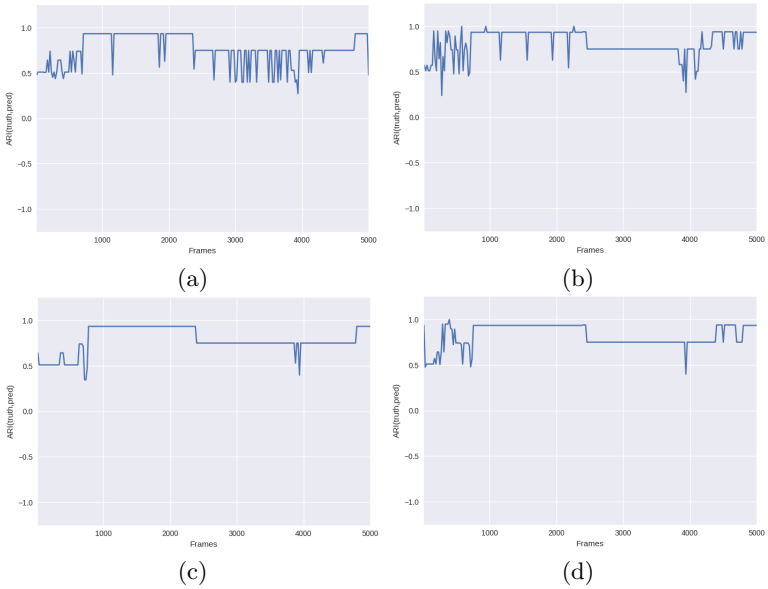



**Fig. 8.** (a) Method 2 (only one vote per participant) without memory, $EC_2 = 0.76$. (b) Method 3 (several votes in the visual field) without memory, $EC_3 = 0.81$. (c) Method 2 with memory process, $EC_2 = 0.78$. (d) Method 3 with memory process, $EC_3 = 0.82$.

F-formations do not evolve much from $t_1$ to $t_2$ (i.e. always the same groups of discussion). This can help detecting in which time interval significant changes occur. A deeper analysis is left for future work.

**Fig. 9.** Excerpt of frames 720, 740, 760 in a video from the Mingle data set, exhibiting changes in the F-formations, correctly detected by the analysis of ARI values over time.

## Acknowledgments

The experiments in this paper used the MatchNMingle Dataset made available by the Delft University of Technology, Delft, The Netherlands [2].

## References


1. Beebe, S.A., Masterson, J.T.: Communicating in small groups. Pearson Education, Boston, MA (2003)
2. Cabrera-Quiros, L., Demetriou, A., Gedik, E., van der Meij, L., Hung, H.: The MatchNMingle dataset: a novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates. IEEE Transactions on Affective Computing **12**(1), 113–130 (2018)
3. Cristani, M., Bazzani, L., Paggetti, G., Fossati, A., Tosato, D., Del Bue, A., Menegaz, G., Murino, V.: Social interaction discovery by statistical analysis of F-formations. In: British Machine Vision Conference. vol. 2, p. 4 (2011)
4. Hall, E.T.: The hidden dimension, reprint. Anchor books New York, NY (1990)
5. Hu, S., Liu, H., Wu, X., Li, R., Zhou, J., Wang, J.: A hybrid framework combining genetic algorithm with iterated local search for the dominating tree problem. Mathematics **7**(4), 359 (2019)
6. Hubert, L., Arabie, P.: Comparing partitions. Journal of Classification **2**, 193–218 (1985)
7. Hung, H., Kröse, B.: Detecting F-formations as dominant sets. In: 13th International Conference on Multimodal Interfaces. pp. 231–238 (2011)
8. Kendon, A.: Conducting interaction: Patterns of behavior in focused encounters, vol. 7. Cambridge University Press (1990)
9. Mohammadi, S., Setti, F., Perina, A., Cristani, M., Murino, V.: Groups and crowds: Behaviour analysis of people aggregations. In: International Joint Conference on Computer Vision, Imaging and Computer Graphics. pp. 3–32 (2017)
10. Ramírez, O.A.I., Varni, G., Andries, M., Chetouani, M., Chatila, R.: Modeling the dynamics of individual behaviors for group detection in crowds using low-level features. In: 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). pp. 1104–1111 (2016)
11. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics **20**, 53–65 (1987)
12. Setti, F., Hung, H., Cristani, M.: Group detection in still images by f-formation modeling: A comparative study. In: 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS). pp. 1–4. IEEE (2013)
13. Setti, F., Lanz, O., Ferrario, R., Murino, V., Cristani, M.: Multi-scale f-formation discovery for group detection. In: 2013 IEEE International Conference on Image Processing. pp. 3547–3551 (2013)
14. Setti, F., Russell, C., Bassetti, C., Cristani, M.: F-formation detection: Individuating free-standing conversational groups in images. PloS One **10**(5), e0123783 (2015)
15. Thompson, S., Gupta, A., Gupta, A.W., Chen, A., Vázquez, M.: Conversational group detection with graph neural networks. In: International Conference on Multimodal Interaction. pp. 248–252 (2021)
16. Vascon, S., Mequanint, E.Z., Cristani, M., Hung, H., Pelillo, M., Murino, V.: Detecting conversational groups in images and sequences: A robust game-theoretic approach. Computer Vision and Image Understanding **143**, 11–24 (2016)