MULTI-SCALE F-FORMATION DISCOVERY FOR GROUP DETECTION

Francesco Setti[1] *Oswald Lanz*[2] *Roberta Ferrario*[1] *Vittorio Murino*[3,4] *Marco Cristani*[3,4]

[1] ISTC–CNR, via alla Cascata 56/C, I-38123 Povo (Trento), Italy
[2] Fondazione Bruno Kessler (FBK), via Sommarive 18, I-38123 Povo (Trento), Italy
[3] Università degli Studi di Verona, Strada Le Grazie 15, I-37134 Verona, Italy
[4] Istituto Italiano di Tecnologia (IIT), via Morego 30, I-16163 Genova, Italy

ABSTRACT

We present an unsupervised approach for the automatic detection of static interactive groups. The approach builds upon a novel multi-scale Hough voting policy, which incorporates in a flexible way the sociological notion of group as *F-formation*; the goal is to model at the same time small arrangements of close friends and aggregations of many individuals spread over a large area. Our technique is based on a competition of different voting sessions, each one specialized for a particular group cardinality; all the votes are then evaluated using information theoretic criteria, producing the final set of groups. The proposed technique has been applied on public benchmark sequences and a novel cocktail party dataset, evaluating new group detection metrics and obtaining state-of-the-art performances.[1]

Index Terms— Group detection, F-formation

1. INTRODUCTION

After decades of studies on automated modeling of individuals, the videosurveillance community in the last few years has started focusing on the new problem of analyzing *groups*.

In this paper we focus on group detection, which aims at individuating group formations in still images, without exploiting temporal reasoning (proper of the tracking issue). Group detection is desirable for a wide range of applications, including group initialization for tracking [1], semantic tagging of pictures [2], estimation of social relations [3] and many others. In the literature, one of the earlier group detection method utilized Voronoi diagrams with positional features [4]; successively, head orientation has been exploited, considering as a group those individuals that are close and looking at each other [5].

Social signal processing [6], *i.e.*, a research area emerged at the conjunction between social psychology and pattern recognition, introduced the notion of *F-formation* [7]; roughly speaking, F-formations are spatial patterns that characterize groups of two or more people. The most important part of an F-formation is the o-space (see Fig. 1), a convex empty space surrounded by the people involved in a social interaction, in which every participant looks inward, and no external people are allowed. The approach of [8] detects F-formations by individuating maximal cliques in weighted graphs. In [9], a Hough voting approach selects o-space center locations by checking a set of social constraints. All these approaches are supervised, in the sense that the F-formation is a supervised (by sociological theories) model for groups.

Our approach is inspired by the latter work, but goes far beyond, relaxing a strong and constraining limitation: in [9], the radius of an o-space was modeled by a simple Gaussian distribution, and this amounts to individuate people who stay in a restricted range of distances from each other. As Fig. 2 suggests, this is not always the case: inter-personal distances depend on many aspects, and one of the most evident is the cardinality of the groups into play, in the sense that large groups constrain people to lie far enough to allow everyone's participation in the group activity.

This consideration, founded on recent sociological studies [10], has driven us in the design of a novel multi-scale F-formation detection algorithm: the idea is of having different voting modules specifically suited for particular F-formations cardinalities. All these modules collapse their votes in a joint accumulation space, where the final groups are extracted. To analyze the voting space, looking for plausible F-formations, a novel measure has been employed, based on the weighted Boltzmann entropy [11, 12]: the idea is that a group should be voted with a similar intensity by all the different participants, pruning away unbalanced votings.

The approach has been validated on public benchmarks, characterized by people position and head orientation. In particular, we focus on the CoffeeBreak [9] and the novel CocktailParty datasets. Adopting the metrics of [9], our approach sets new state-of-the-art performances. In addition, we define two novel metrics which analyze 1) the ability of modeling groups of different cardinalities, and 2) the capability of individuating all, or part of the group members. Using these metrics, we provide for all the datasets convincing comparative performances, setting the best scores in all the cases.

[1] F. Setti and R. Ferrario are supported by the VisCoSo project grant, financed by the Autonomous Province of Trento through the "Team 2011" funding programme.

The remaining is organized as follows: Sec. 2 recaps the single-scale approach of [9]; our proposal is detailed in Sec. 3 and the experiments are shown in Sec. 4. Finally, in Sec. 5, conclusions are drawn and future perspectives outlined.

2. SINGLE SCALE F-FORMATION DETECTION

In our previous work [9], the *state* of each individual with label $i \in L$ is characterized by its floor position (x_i, y_i) and head orientation θ_i. State uncertainty is injected by sampling from $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu}_i = [x_i, y_i, \theta_i]$, and $\boldsymbol{\Sigma}$ is a diagonal matrix with variances $\sigma_x^2, \sigma_y^2, \sigma_\theta^2$; we thus generate a set of N samples $\{\mathbf{s}_{i,n}\}, n = 1, \ldots, N$ associated to subject i. Each sample has a weight $w_{i,n} \propto \mathcal{N}(\mathbf{s}_{i,n}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma})$. Each sample votes for an o-space center, considering a radius R along its orientation, with an *intensity* equal to its weight. All the votes of the different subjects are stored in two accumulation spaces: an *intensity accumulation space* $\mathcal{A}_I(x, y)$ which collects the sum of the intensities of the votes for the location (x, y); and a *label accumulation space* $\mathcal{A}_L(x, y)$ that records the ID labels $\{i\}$ of the people that voted for (x, y). A final accumulation space $\tilde{\mathcal{A}}$ is built as:

$$\tilde{\mathcal{A}}_I(x, y) = \text{card}(x, y) \cdot \mathcal{A}_I(x, y) \quad \forall (x, y) \in \mathcal{A}_I(x, y) \quad (1)$$

where $\text{card}(x, y)$ counts the *diverse* subjects that voted for x, y, and such information is easily extracted from $\mathcal{A}_L(x, y)$. *Legal* groups are found by evaluating in descending order the $\tilde{\mathcal{A}}_I(x, y)$ values, and checking the o-space *emptiness condition* (no intruders in the o-space), iteratively, pruning away the votes of those people who have been already assigned to another legal group, until $\tilde{\mathcal{A}}_I(x, y)$ becomes empty.

3. MULTI-SCALE APPROACH

As described recently in [10], an F-formation can have various configurations when only 2 people are involved (*L-shape*, *vis-a-vis* or *parallel*), but if three or more individuals are interacting, they usually arrange in a circle. In such a case, people locations can be modeled as the vertices of a regular polygon with side s, whose value belongs to the *personal* range of distances defined in [13]. In particular, we set $s = 95cm$ as it lies in the middle of the *personal* range.

Therefore, we can formally define the radius R_k associated to an F-formation with cardinality k as the circumradius (i.e. the radius of the circumscribed circle) of this polygon (see Fig. 1); in formulae:

$$R_k \simeq \frac{s}{2 \sin \frac{\pi}{k}} \quad (2)$$

This allows us to gather a set of radii, which can be exploited for individuating groups of different *scales*, *i.e.*, cardinalities. As we will see next, the rigid geometry imposed by the radius R_k and the distance s will be relaxed by our voting treatment.

Our algorithm is composed by two steps: 1) *Fixed-cardinality group detection*, in which several scales (*i.e.* group cardinalities) are examined and 2) *Groups merging*, in which all the votes are merged in a multi-scale voting space.

Fig. 1. Example of polygonal arrangement of 5 persons in an F-formation (left) and a real F-formation with the related o-space highlighted in red (right).

3.1. Fixed-cardinality group detection

After having decided the set K of group cardinalities that we want to analyze (usually from 2 to the maximum number of individuals in the scene), for each $k \in K$ we estimate the F-formation radius R_k, using Eq. 2.

Therefore, for each scale, we apply the single scale F-formation detection, modifying the approach of [9] explained in Sec. 2 with respect to three fundamental aspects: first of all, we modify the way the votes are employed to build the accumulator $\tilde{\mathcal{A}}_I(x, y)$ of Eq. 1. In that form, the equation essentially may reward *unbalanced* groups, in which the presumed members vote in an uneven fashion, with only one person voting with many samples; this situation intuitively generates false positives.

Here, we want F-formations as result of an homogeneous choral poll. For this reason, we revise Eq. 1 as follows:

$$\tilde{\mathcal{A}}_I(x, y) = \text{card}(x, y) \cdot \tilde{E}(x, y) \quad \forall (x, y) \in \mathcal{A}_I(x, y) \quad (3)$$

with

$$\tilde{E}(x, y) = -\sum_{i \in L} h_i(x, y) \cdot p_i(x, y) \log_2 p_i(x, y) \quad (4)$$

where $h_i(x, y)$ is a normalized count of the times the subject i voted in position (x, y), and

$$p_i(x, y) = \sum_{\{w_{i,n}\} \text{on}(x,y)} w_{i,n} \quad (5)$$

is the sum of the weights $\{w_{i,n}\}$ assigned by the subject i through his samples $\{s_{i,n}\}$ in location (x, y). The quantity $p_i(x, y)$ is then normalized over i, becoming a distribution. Eq. 4 is essentially a weighted entropy [11], where the entropy is presented in its Boltzmann form (see Eq. 4, where $\tilde{E}$ is the weighted entropy) [12]. As one can note, the entropy has maximum value when each of the presumed members of an F-formation votes in an similar way (with the same number of particles, and the same weights). The term $\text{card}(x, y)$ acts as a normalization, rising up the score for groups with high cardinality, whose entropy is usually low.

	k=2	k=3	k=4	k=5	k=6	Avg.
$R_{2,3}$	0.37	0.52	0.67	0.10	NaN	0.41
$R_{4,5}$	0.42	0.51	0.68	0.33	0.60	0.52
$R_{6,7}$	0.37	0.52	0.60	0.44	0.68	0.55
Multi-Scale	**0.50**	**0.70**	**0.87**	**0.80**	**0.95**	**0.72**

Table 2. F_1 score for *CocktailParty* sequence for each radii interval and for each group cardinality.

The second difference is that, for each cardinality k analyzed, we accept in $\tilde{\mathcal{A}}_I(x,y)$ only those groups with k members (*i.e.*, with card$(x,y)=k$), putting to 0 all the remaining locations: in practice, this operation prunes away group hypotheses which are discordant with the geometric model of Eq. 2. For clarity, let us call the accumulator as $\tilde{\mathcal{A}}_I(x,y)^{(k)}$, to highlight its dependence on the cardinality k. We call this procedure *Fixed-cardinality pruning*.

After building $\tilde{\mathcal{A}}_I(x,y)^{(k)}$, *potential* F-formations are found by looking iteratively for the maximum values, in the same way as the single scale approach of Sec. 2 finds all the legal groups.

The third difference is that, for each cardinality k, the legal groups (*i.e.*, their o-space center location) are here collected in a single-scale accumulation space $\tilde{\mathcal{S}}_I^{(k)}(x,y)$, which stores exclusively the weighted entropy scores of their o-space centers, and will serve for the following step of the multi-scale algorithm.

3.2. Groups merging

In the second main step of the algorithm, all the single-scale accumulators $\{\tilde{\mathcal{S}}_I^{(k)}\}_{k\in K}$ are fused in a multi-scale accumulator $\tilde{\mathcal{M}}_I(x,y)$, that for each location stores the possible scores of legal F-formations. Please note that in a single location only F-formations of different cardinalities may co-exist; the aim of this step is to select, for each location, the maximum value, i.e., the F-formation with highest weighted Boltzmann entropy. Since this quantity is normalized over the different cardinalities, this step ensures an inter-scale fair comparison.

4. EXPERIMENTS

Testing group detection methods is hard, due to the lack of datasets[2] and metrics. Here we try to partially fill these gaps.

As benchmarks, we consider three datasets: two from [9], *i.e.*, the Synthetic (100 frames, no tracking or head orientation errors) and the CoffeeBreak (two sequences, 45 + 75 frames of a crowded coffee break during a summer school); the third one, dubbed CocktailParty, is brand-new; due to the lack of space, we fully detail only the latter. The CocktailParty dataset contains 30 minutes of video recordings of a cocktail party in a $30m^2$ lab environment involving 7 subjects. The party was recorded using four synchronized angled-view cameras (15Hz, $1024\times768px$, jpeg) installed in the corners of the room. The dataset is challenging for video analysis due to frequent and persistent occlusions, in a highly cluttered scene. Subject's positions and horizontal head orientations where logged using a particle filter-based body tracker [16] with head pose estimation as in [17]. Groups in one frame every 5 seconds were annotated manually by an expert, resulting in a total of 320 distant frames for evaluation. Compared to the *CoffeeBreak*, this dataset is more accurate and reliable in terms of people position and head pose estimation.

As metrics, we start by employing the one adopted in [9], where we have a matched group if at least $\lceil(2/3\cdot k)\rceil$ of their components are found in a F-formation, with k the group cardinality. Given this metric, for each sequence we estimate the *precision* and *recall*, averaged over all the total amount of groups in all the frames, and we join them together in the standard $F_1 = 2\cdot\frac{precision\times recall}{precision+recall}$ measure.

As a novel metric, for analyzing the effectiveness of the approach in dealing with groups of particular cardinalities, we extract the precision and recall related to groups of cardinality k, and the related F1 score. The second new metric will be detailed later on.

Considering the set-up of our algorithm, as written above, we keep $s=95cm$; the covariance matrix $\mathbf{\Sigma}$ is the same for all the scales, with $\sigma_x^2=\sigma_y^2=500, \sigma_\theta^2=0.001$. The number of samples N per person is a critical parameter for every voting procedure; in our experiments it has been arbitrarily fixed to 800. In average, each frame required 15 seconds of computation on a non-optimized MATLAB implementation, run on a Intel Xeon 2.83GHz CPU, 8GB RAM.

To ease the visualization of the results, we grouped different cardinalities into three sets $K=\{\{2,3\},\{4,5\},\{6,7\}\}$ and the set of radii we used was $R=\{50cm,75cm,95cm\}$[3].

As comparative approaches, we take into account the Inter-Relation Pattern Matrix method (IRPM) [14], the Dominant Set method [8] and the BMVC2011 approach [9][4]; we also show the results for each of the fixed-cardinality group detection step, named *Fixed-Scale*, in which we removed the fixed-cardinality pruning operation, allowing us to obtain alternative, single-scale group detection approaches. Finally, our proposed approach is dubbed *Multi-scale*.

Table 1 shows the results in terms of average precision, recall and F1 score for each benchmark sequence. Several facts can be appreciated: 1) Our approach defines the best performance in terms of F1 score on all the benchmarks, i.e., generally we collect the least amount of false positive and negative alarms. 2) On the Synthetic and CoffeBreaks sequences, groups tend to be smaller, while in the CocktailParty

[2]We are considering datasets equipped with the positions and head orientations of the individuals.

[3]These are obtained by averaging the value obtained from Eq. 2 and $s=95cm$, over the cardinalities grouped together in K. The Fixed-cardinality pruning step has been revised accordingly, accepting groups of two cardinalities

[4]It is worth noting that in [9] we reported the precision and recall mediated over the frames, and not over the total number of groups, as one should expect, so the precision and recall score in that paper and in the current one do not coincide.

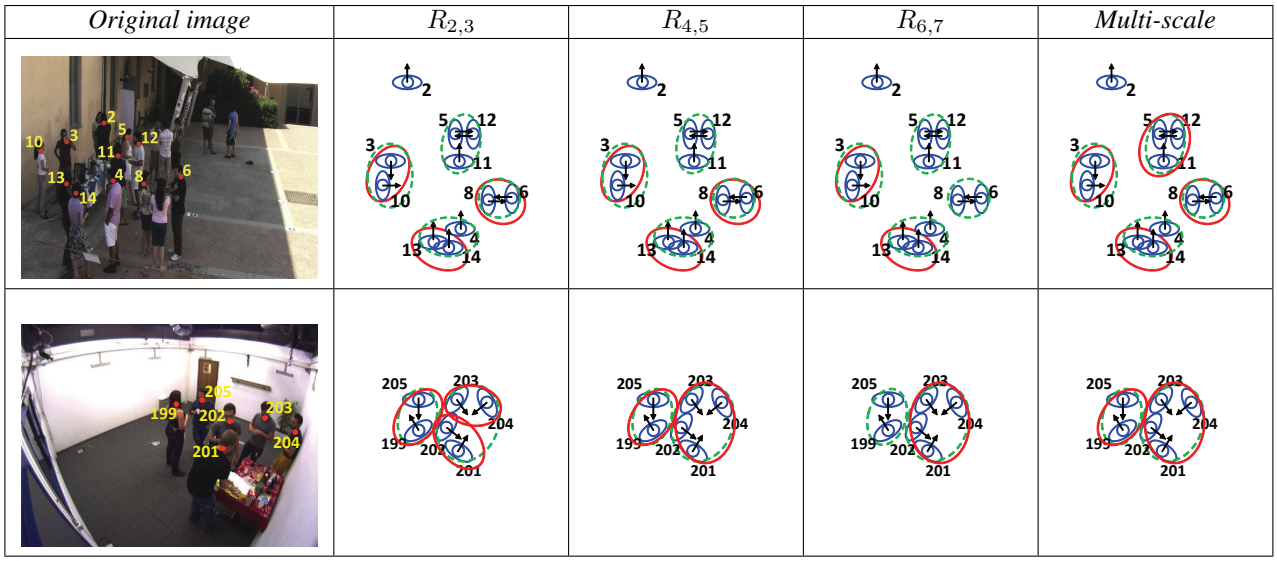

Fig. 2. Qualitative results for one frame of *CoffeeBreak* (first row) and *CocktailParty* (second row) datasets. Green dashed circles represent the ground truth groups, while red continuous circles represent the detected groups. Note how our algorithm does better in detecting group of different sizes w.r.t. the single-scale approaches.

		Synthetic data			**CoffeeBreak *Seq1***			**CoffeeBreak *Seq2***			**CocktailParty**		
Method		Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1
IRPM [14]		0.71	0.54	0.61	0.63	0.54	0.58	0.55	0.19	0.28	0.50	0.46	0.48
Dominant Sets[15]		–	–	–	0.62	0.54	0.58	0.72	0.71	0.72	–	–	–
BMVC2011 [9]		0.73	0.83	0.78	0.42	0.59	0.49	0.58	0.49	0.53	0.59	0.74	0.65
Fixed-Scale	$R_{2,3}$	0.80	0.91	0.85	**0.65**	0.71	0.68	0.92	0.76	0.84	0.36	0.49	0.42
	$R_{4,5}$	0.72	0.79	0.75	0.62	0.63	0.62	0.84	0.62	0.73	0.64	**0.76**	0.69
	$R_{6,7}$	0.59	0.68	0.64	0.59	0.59	0.59	0.69	0.21	0.32	0.65	0.74	0.70
Multi-Scale		**0.86**	**0.94**	**0.90**	0.63	**0.76**	**0.69**	**0.94**	**0.78**	**0.86**	**0.69**	0.74	**0.72**

Table 1. Average precision and recall for all the datasets employed in our experiments.

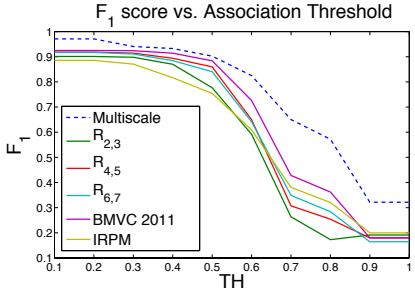

Fig. 3. Partial matching score (best viewed in colors).

benchmark, groups are larger. This is visible by observing the precision and recall scores of the Fixed-scale approaches.

On the CocktailParty dataset, we show the first new metric, *i.e.*, the F1 score for each group cardinality (see Tab. 2): regardless of the group cardinalities, our approach acts better than the fixed-scale voting. Note that for high cardinalities the fixed approaches with longer radii ($R_{5,6}$) get better scores.

Finally, we introduce the second metric, dubbed *partial matching score*, which generalizes the definition of group match $\lceil (2/3 \cdot k) \rceil$, substituting it with $\lceil (TH \cdot k) \rceil$; here, TH is the "acceptance threshold" which may vary from 0 (accept everything) to 1 (accept only groups where you capture all the members). Measuring the average F1 score while varying TH between $]0, 1]$ gives an idea on how well an approach is capable of capturing entirely a group. We estimate this on the CocktailParty dataset, showing the resulting curves in Fig.3: one can note that our approach gives always the highest score, especially when TH becomes higher, *i.e.*, when we accept groups only if most of their members are individuated.

5. CONCLUSIONS

In this paper we proposed a Hough voting approach that detects groups formed by an unconstrained number of people; the method manages the partial analysis of different group detection modules, employing the weighted Boltzmann entropy for scoring each group hypothesis. The results are encouraging, considering also novel metrics of group detection. As future work, we plan to introduce the temporal dimension into play, fusing our method with a group tracking strategy.

6. REFERENCES

[1] L. Bazzani, M. Cristani, and V. Murino, "Decentralized particle filter for joint individual-group tracking," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2012.

[2] M. Marin-Jimenez, A. Zisserman, and V. Ferrari, "Here's looking at you, kid. detecting people looking at each other in videos," in *British Machine Vision Conference*, 2011.

[3] Gloria Zen, Bruno Lepri, Elisa Ricci, and Oswald Lanz, "Space speaks: towards socially and personality aware visual surveillance," in *Proc. ACM international workshop on Multimodal pervasive video analysis*, New York, NY, USA, 2010, MPVA '10, pp. 37–42, ACM.

[4] J. C. S. Jacques, A. Braun, J. Soldera, S. R. Musse, and C. R. Jung, "Understanding people motion in video sequences using voronoi diagrams: Detecting and classifying groups," *Pattern Analysis and Applications*, vol. 10, no. 4, pp. 321–332, 2007.

[5] Neil M. Robertson and Ian D. Reid, "Automatic reasoning about causal events in surveillance video," *EURASIP J. Image and Video Processing*, vol. 2011, 2011.

[6] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social Signal Processing: Survey of an emerging domain," *Image and Vision Computing Journal*, vol. 27, no. 12, pp. 1743–1759, 2009.

[7] A. Kendon, *Studies in the Behavior of Social Interaction*, Lisse: Peter De Ridder Press, 1977.

[8] Hayley Hung and Ben Kröse, "Detecting f-formations as dominant sets," in *Proceedings of the 13th international conference on multimodal interfaces*, New York, NY, USA, 2011, ICMI '11, pp. 231–238, ACM.

[9] Marco Cristani, Loris Bazzani, Giulia Paggetti, Andrea Fossati, Diego Tosato, Alessio Del Bue, Gloria Menegaz, and Vittorio Murino, "Social interaction discovery by statistical analysis of f-formations," in *Proceedings of the British Machine Vision Conference*. 2011, pp. 23.1–23.12, BMVA Press, http://dx.doi.org/10.5244/C.25.23.

[10] Adam Kendon, "Spacing and orientation in co-present interaction," in *Proceedings of the Second international conference on Development of Multimodal Interfaces: active Listening and Synchrony*, Berlin, Heidelberg, 2010, COST'09, pp. 1–15, Springer-Verlag.

[11] Silviu Guiasu, "Weighted entropy," *Reports on Mathematical Physics*, vol. 2, no. 3, pp. 165 – 179, 1971.

[12] E.T. Jaynes, "Gibbs vs Boltzmann Entropies," *American Journal of Physics*, vol. 33, no. 5, pp. 391–398, 1965.

[13] Edward T. Hall, *The Hidden Dimension*, Doubleday, Garden City, NY, USA, Oct. 1966.

[14] L. Bazzani, D. Tosato, M. Cristani, M. Farenzena, G. Pagetti, G. Menegaz, and V. Murino, "Social interactions by visual focus of attention in a three-dimensional environment," *Expert Systems*, 2012, in print.

[15] Francesco Setti, Hayley Hung, and Marco Cristani, "Group detction in still images by f-formation modeling: a comparative study," in *Proceedings of the 14th International Workshop on Image and Audio Analysis for Multimedia Interactive Services (WIAMIS)*. 2013, BMVA Press.

[16] O. Lanz, "Approximate bayesian multibody tracking," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 9, pp. 1436 –1449, sept. 2006.

[17] Oswald Lanz and Roberto Brunelli, "Multimodal technologies for perception of humans," chapter Joint Bayesian Tracking of Head Location and Pose from Low-Resolution Video, pp. 287–296. Springer-Verlag, Berlin, Heidelberg, 2008.