The GRODE Metrics: Exploring the Performance of Group Detection Approaches

Francesco Setti
ISTC - CNR
via alla Cascata 56/C, I-38121 Trento
francesco.setti@loa.istc.cnr.it

Marco Cristani
University of Verona
Strada Le Grazie 2, I-37134 Verona
marco.cristani@univr.it

Abstract

The detection of groups of people is attracting the attention of many researchers in diverse fields, with a growing number of approaches published each year; despite this, the evaluation metrics are not consolidated, with some measures inherited from the people detection fields, other ones designed specifically for a particular approach, generating a set of not comparable figure of merits. Moreover, existent methods of analysis are scarcely expressive, for example ignoring the fact that groups have different cardinalities, and that obviously larger groups are harder to find. This paper fills this gap presenting GRODE, a suite of measures of accuracy which defines precision and recall on the groups, including the group cardinality as variable. This gives the possibility to investigate aspects never considered so far, such as the tendency of a method of over- or under-segmenting groups, or of better dealing with specific group cardinalities. The metrics have been applied to all the publicly available approaches of group detection, discovering interesting strength and pitfalls of the current literature, and promoting further research in the field.

1. Introduction

The detection of groups of people has become a relevant task in many computer vision systems, in order to understand social activities and performing high-level profiling, thus impacting on videosurveillance, social robotics, social signal processing, to quote a few.

On the one side, some approaches build over object tracking frameworks, enriching the usual per-person estimation with labels of group membership: these are added after the tracking in a post-processing step [4, 8, 14, 24], or directly embedded in the state-space of the filtering mechanism [2, 6, 11]. On the other side, many more approaches discover groups without temporal reasoning, directly on still images [1, 5, 7, 10, 13, 15, 16, 17, 18].

In most of the cases, the individuation of a group lies on intuitive observations of proxemics, that is, mutual proximity and common dynamics (especially in the tracking approaches), while in some other cases the sociological notion of F-formation [7, 10] has been embedded into different minimization problems solved by Hough voting strategies [1, 7, 16, 17] or graph theory methods [10, 18, 21, 22], or directly faced by classification approaches [5, 13, 15].

In all the cases, the quantitative evaluation of a group detection method is not lying on a widely accepted protocol, generating a considerable set of metrics which are specific for the scenario-at-hand, making hard a fair comparison among different researches.

The most straightforward measures are inherited from the tracking literature, that is, the CLEAR MOT metrics [3], where bounding boxes around the single individuals have been replaced by the convex hull enveloping all the people inside the group [2, 14]. Similar metrics are designed considering the object detection approaches [5, 13, 15]. In these cases, groups are intended as atomic entities, so that the errors of including more people in a group or losing some individuals in a formation are not explicitly modeled.

On the contrary, the metrics proposed in [7] deal with this problem, introducing also the concept of tolerance: in such a formulation, a group is considered as correctly estimated even if some individuals are missing or erroneously included. These metrics are the most used for all the methods that rely on a calibrated cameras and use as input ground position of labelled individuals in the scene; in other words they do not consider any processing on the image level, but assume to have the detections and head pose estimates.

Despite these measures focus on many interesting aspects about the detection of groups, they seem to forget that groups can be different in terms of cardinality, ranging from pairs of people to sets of 7-8 people (after that it is reasonable to talk about crowds [9], which is not the subject of this study). As intuitive by looking at Fig. 1, larger groups of people are harder to be completely individuated than smaller formations: this fact is anyway neglected by

Figure 1. Examples of groups configurations in daylife situations.

the current metrics, that evaluate instead all the detections equally important in the computation of the final scores.

The aim of this paper is to present a set of novel group detection metrics, where the presence of groups of different numbers of people is of primary importance. In particular, the idea is to address the behavior of a group detection approach with respect to specific group cardinalities, individuating the group settings where a given algorithm perform better than another. A first tentative of this kind has been proposed by Setti *et al.* [18], but much more can be done. In facts, an interesting phenomenon that can be captured is the tendency of oversegmenting or undersegmenting a group, and in the detail if this tendency is polarized on particular formation cardinalities (that is, the tendency of systematically capture/break a group of G elements as one or more groups of g individuals). This sort of report will be of sure useful to 1) deeply understanding the nature of a group detection approach; 2) selectively fixing the issues and ameliorating the approach.

The proposed metrics model all of these aspects, and have been applied to most of the group detection approaches whose code is publicly available in the literature, and to different datasets; the results are enlightening, in the sense that many of the above discussed aspects unveil some unexplored characteristics of the approaches, helping the researcher in understanding which methods are more suited to his/her requirements.

The rest of the paper is organized as follows: in Section 2 we discuss the metrics used in the related literature, while in Section 3 we present the new GRODE metrics we propose and Section 4 reports some experiments on a public dataset to validate our claims. Finally Section 5 draws the conclusions of this work.

2. Metrics in the literature

As for the metrics inherited from the tracking/detection fields, Manfredi et al. [13] propose two metrics on the image level considering as main concept the intersection of the group detection convex hulls and the ground truth regions. In particular they define the overlapping ratio (O_r) as:

$$O_r = \frac{\sum_{p=1}^{P} M_{out}(p) \cap M_{gt}(p)}{\sum_{p=1}^{P} M_{out}(p) \cup M_{gt}(p)} \quad (1)$$

where p is a generic pixel and P is the total number of pixels in the image, M_{out} is a mask output by the detector algorithm that is 1 if the pixel belongs to a detected group and 0 otherwise, and M_{gt} is the same for ground truth. They also account precision and recall metrics at object level (P_O and R_O) by considering a group as correctly identified if the spatial overlapping with the ground truth is higher that 50% of the intersection over union. This concept of pixel-wise association of detected and ground truth groups has the main drawback of being very sensitive to camera orientation and occlusions.

A similar metric is used by Choi et al. [5], where, other than considering the intersection over the union ratio, it also forces a detected group to be associated with at most one ground truth group and vice versa. On top of it, the authors use standard precision, recall and F_1 measures. To avoid problems related to misdetections of individuals in the scene, the metric ignores these also from ground truth groups.

Rota et al. [15] use an approach based on identifying interactions between pairs of individuals, therefore the metrics employed are based on the correct identification of each link between pairs of individuals; this is a common choise for clustering problems and is commonly referred as *pairwise loss*. They consider standard precision, recall and accuracy measures where a true positive is a pair of individuals belonging to the same group both for the estimates and ground truth, a false negative is a pair of individuals belonging to different groups both in estimates and ground truth, false positives are pairs of individuals grouped together by the algorithm but in different groups in ground truth, and false negatives vice versa.

The pairwise loss function has the main drawback to be imprecise when dealing with large crowds, due to the quadratic number of connections generated between members. To overcome this problem, Solera *et al.* [19] propose to extend the MITRE loss function [23], commonly used in NLP for the coreference problem, to handle groups; in particular, the MITRE loss is not able to handle singletons. The *GROUP-MITRE loss* is obtained by adding, for each individual, a fake counterpart to which only singletons are connected.

Concerning the concept of tolerance of detection of a

group, in [7] a group is considered as correctly estimated if at least $\lceil (T \cdot |G|) \rceil$ of their members are found by the grouping method, and no more than $1 - \lceil (T \cdot |G|) \rceil$ false subjects are identified, where $|G|$ is the cardinality of the labelled group G, and $T \in]0, 1]$ is the tolerance threshold, typically set as $2/3$ or 1. Standard precision, recall and F_1 measures are then used to compare different methods. Recently, Setti et al. [18] presented a new metric defined as the area under the curve (AUC) in the $F_1 \text{vs} T$ graph with T varying from $1/2$ to 1. This metric is called Global Tolerant Matching score (GTM) and has the value of being independent from the tolerance threshold T. These last metrics are the most used for all the methods that rely on a calibrated cameras and use as input ground position of labelled individuals in the scene; in other words they do not consider any processing on the image level as the previous ones, but assume to have the detections and head pose estimates.

3. The GRODE metrics

The GRODE metrics are all based on a unique data structure, which is the *Histogram of Individuals over Cardinalities* (HIC) matrix. The HIC matrix is computed by considering (over time and over each individual in the scene) the membership of a person of a) a given ground-truth group and b) the detected group in which he/she is currently associated, in relation to the cardinalities of these two groups; in simple words, an individual belonging at a given time instant to a ground-truth group of cardinality i and estimated in a group of cardinality j adds 1 in $HIC(i, j)$, and the process is repeated for all the persons and all the time instants of analysis. The matrix is then normalized over the rows.

Formally, HIC is defined as:

$$HIC(i, j) = \frac{1}{n_j} \sum_{t \in T} \sum_{p \in P} d_{ij}\left(p, G, GT\right) \tag{2}$$

where i and j vary over the group cardinalities in the scene, that is $1 \leq i \leq |g_t|_{max}$ and $1 \leq j \leq |g|_{max}$, with $|g_t|_{max}$ and $|g|_{max}$ the maximum group cardinality of ground truth and estimated groups respectively, t is a time instant – *i.e.* a frame – within the whole set of frames under analysis T, p is a person within the list P of all the detected people at that particular time t, G and GT are the sets of all estimated and groud truth groups at time t, and n_j is a normalization term which sets to 1 the summation over the rows of HIC. Finally, d_{ij} is an accumulation function defined as:

$$d_{ij}\left(p, G, GT\right) = \begin{cases} 1 & \text{if } \exists\, g \in G \ \wedge \ g_t \in GT \ : \\ & \quad p \in g \cap g_t \ \wedge \ |g| = j \ \wedge \ |g_t| = i \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

The ideal detector would only have 1 on the diagonal and 0 everywhere else, while a good balance between upper and lower triangulars means the detector is not prone to merging small groups toghether or splitting big groups into smaller ones. Moreover, a uniform distribution over the diagonal of HIC matrix means the method does not prefer certain cardinalities over the others.

Based on this matrix, we propose a set of scalar metrics to compare performances of different methods on the same dataset.

Thinking about HIC matrix as a confusion matrix in a multi-class classification problem, we can easily define the *cardinality level accuracy* (A) as:

$$A = \frac{\sum_i HIC(i, i)}{\sum_i \sum_j HIC(i, j)} \tag{4}$$

and we can also compute the standard pattern recognition metrics *precision*, *recall* and F_1 for each cardinality C:

$$Pr(C) = \frac{HIC(C, C)}{\sum_i HIC(i, C)} \tag{5}$$

$$Re(C) = \frac{HIC(C, C)}{\sum_j HIC(C, j)} \tag{6}$$

$$F_1(C) = 2\frac{Pr(C) \cdot Re(C)}{Pr(C) + Re(C)} \tag{7}$$

As a measure of how a method is able to detect groups of different cardinalities in the same manner, we propose the *cardinality deviation* (D) defined as the standard deviation within the elements on the diagonal of HIC matrix:

$$D = \sqrt{\frac{1}{N} \sum_i \left(HIC(i, i) - \mu\right)^2} \tag{8}$$

where μ is the average of the elements on the diagonal and N is the dimensionality of matrix HIC.

To account the balance of a method between under- and over-segmentation we propose the *upper-lower difference* (UL), defined as the difference between the summation of all the elements of the upper and lower triangular matrices extracted by HIC:

$$UL = \sum_i \sum_{j>i} HIC(i, j) - \sum_i \sum_{j<i} HIC(i, j) \tag{9}$$

and its weighted version (WUL) where elements far from the diagonal are weighted more then the closer ones:

$$\begin{aligned} WUL = \ & \textstyle\sum_i \sum_{j>i} HIC(i, j) \cdot (j - i) - \\ & - \textstyle\sum_i \sum_{j<i} HIC(i, j) \cdot (i - j) \end{aligned} \tag{10}$$

The rational of this measure is that we want to penalise more the situations where, for example, an element which is supposed to form a group of 2 people is assigned to a

formation consisting of 8 elements instead of to one of 3 elements. The reason is that drastically different sized groups are often treated with different models and thus sharing individuals among them could mean a very low specifity.

With this set of metrics we are now able to evaluate group detectors not only in a global way, but analyzing its most important features, *i.e*. the ability to detect groups of different cardinalities with no bias and the balance between over- and under-segmentation.

4. Experiments

In this section we present some experiments to validate our claims in the previous sections and in particular to prove the relationship between the metrics and the behaviour we expect from the detector.

We used in our experiments the publicly available GDet dataset [1]. This dataset is composed by a total of 403 annotated frames acquired by 2 angled-view low resolution cameras (352×328 pixels) in an indoor scenario of a vending machines area where a maximum of 7 people meet and chat while they are having coffee. For ground truth generation, people tracking has been carried out with the particle filter proposed in [12], while head pose estimation is performed afterwards with the method in [20] considering only 4 orientations (front, back, left and right). Table 1 shows the groups' distribution in terms of number of people for group cardinality in GDet dataset.

Cardinality	*1*	*2*	*3*	*4*	*5*	*6*
People	367	394	372	88	175	78

Table 1. Number of individuals for each groups' cardinality in GDet dataset.

We compare seven different state of the art methods: one exploiting the concept of *view frustum* (IRPM), two based on dominant-sets (IGD and GTCG), three different version of Hough Voting approaches (linear, entropic and multi-scale HVFF), and one based on graph-cuts technique (GCFF). Inter-Relation Pattern Matrix (IRPM), proposed by Bazzani *et al*. [1], uses the head direction to infer the 3D view frustum as approximation of the focus-of-attention of an individual; this is used togheter with proximity information to estimate interactions: the idea is that close-by people whose view frustum is intersecting are in some way interacting. Interacting Group Discovery (IGD), presented by Tran *et al*. [21], is based on dominant sets extraction from an undirected graph where nodes are individuals and the edges have a weight proportional to how much people are interacting; the attention of an individual is modeled as an ellipse centred at a fixed offset in front of him, while the interaction between two individuals is proportional to the interserction of their attention ellipses. In [22] the authors develop a game-theoretic framework called Game-Theory for Conversational Groups (GTCG), supported by a statistical modeling of the uncertainty associated with the position and orientation of people. Specifically, they use a representation of the affinity between candidate pairs by expressing the distance between distributions over the most plausible oriented region of attention. Additionally, they can integrate temporal information over multiple frames by using notions from multi-payoff evolutionary game theory. Under the caption of Hough Voting for F-formation (HVFF) we consider a set of methods based on a Hough Voting strategy to build accumulation spaces and find local maxima of this function to identify F-formations. The general idea is that each individual is associated with a Gaussian probability density function which describes the position of the o-space centre he is pointing at. The pdf is approximated by a set of samples, which basically vote for a given o-space centre location. The voting space is then quantized and the votes are aggregated on squared cells, so to form a discrete accumulation space. Local maxima in this space identify o-space centres, and consequently, F-formations. Over the years, three versions of these framework have been presented: in [7] the votes are linearly accumulated by just summing up all the weights of votes belonging to the same cell, in [16] the votes are aggregated by using the weighted Boltzmann entropy function, while in [17] a multi-scale approach is used on top of the entropic version. Finally, Graph-Cuts for F-formation (GCFF), presented Setti *et al*. [18], proposes an iterative approach that starts by assuming an arbitrarily high number of F-formations: after that, a hill-climbing optimisation alternates between assigning individuals to groups using the efficient graph-cut based optimisation, and updating the centres of the F-formations, pruning unsupported groups in accordance with a Minimum Description Length prior. The iterations continue until convergence, which is guaranteed. For each method, the parameters have been set in order to give the best performances on the specific dataset in terms of precision, recall and F_1 metrics as defined in [18].

Figure 2 shows the HIC matrices for all the methods used in this comparison. Simply looking at the matrices, the reader can have an intuition on the behaviour of each method. For instance, IRPM and IGD are expected to perform very well detecting singletons, while HVFF lin is prone to detect small groups of 2 elements. GCFF looks to be the most balanced within the seven methods, while in general all the other six methods show a tendency to oversegment. Moreover, groups of 6 elements are only detected by multiscale HVFF and GCFF. These intuitions will be confirmed by the analysis of numerical scores that follows.

Figure 3 reports the cardinality level accuracy (A) and cardinality deviation (D). As expected, GCFF is the best performing in both scores, with an accuracy of about 0.64 and a cardinality deviation lower than 0.2. Note that low values of D does not mean that the method is well per-

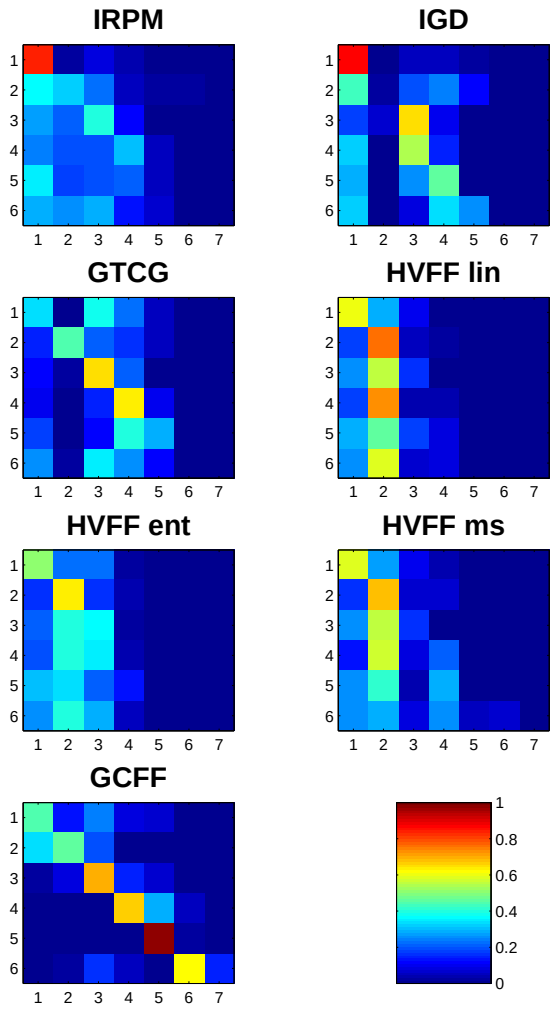

Figure 2. HIC matrices for seven state of the art methods on GDet dataset. (Best viewed in colors)

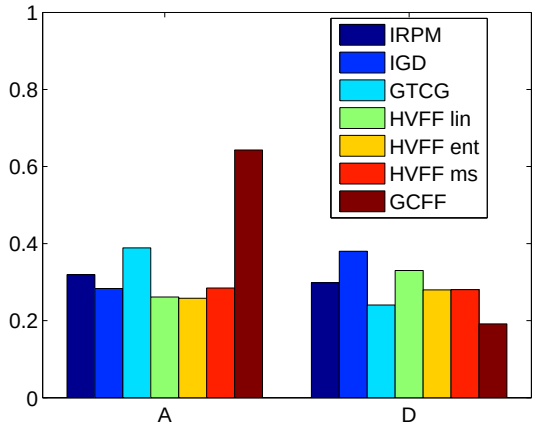

Figure 3. Cardinality level accuracy (A) and cardinality deviation (D) for all the methods on GDet dataset. (Best viewed in colors)

forming in general, but only that it has no preferences for a particular cardinality: for instance IRPM, HVFF ent and HVFF ms have very similar D (less than 1% of difference) but different A, this is due to the fact that IRPM is very good detecting singletons and performs decently for groups with cardinality smaller than 4, HVFF ent performs pretty well for singletons and groups of 2 and 3 people but has no detections for bigger groups, and HVFF ms performs well for singletons and groups of 2 people decreasing the performances with bigger groups; still HVFF ms, together with GCFF, is the only method able to correctly detect groups of 6 individuals in this dataset.

Figure 4 reports the upper-lower difference in its absolute and weighted versions. As already foreseen from the HIC matrices of Figure 2, all the methods except GCFF have a tendency to oversegment, in particular Hough voting methods are performing very poor from this point of view. Note that GCFF and GTCG have very similar UL in absolute value, but the way the two methods generate these values are completely different; indeed, looking at the matrices one can see that GCFF has very few elements off-diagonal, while GTCG has a good balance between over-and under-segmentation. This behaviour can be seen also by analyzing the UL and A together: the best detector would have $UL \simeq 0$ and $A \simeq 1$, while low values of A means the detector is performing bad both in terms of over- and under-segmentation.

Moreover, the weighted version (WUL) is more informative than the absolute one, since it takes into account the ability of a detector to approximately detect correct groups. This effect can be seen from the comparison of HVFF ent and HVFF ms, looking at the absolute difference (UL) the multiscale approach seems to be more prone to over-segmentation than the entropic version, but the entropic method tends to split big groups into small ones while the multiscale only loses some elements in mid-size groups, leading to smaller values of WUL. A similar effect can be seen comparing GTCG and GCFF.

Results in terms of precision, recall and F_1 are reported in Figure 5 and 6. To confirm our intuition from the matrices and the accuracy, also in terms of precision/recall scores the best performing algorithm is GCFF. But, while the average scores only give a general overview on the performaces of each method, the cardinality analysis shown in Figure 6 is much more informative. Other than the best performing in average, GCFF is also the best performing in terms of F_1 score for every single cardinality; moreover, GCFF has a pretty uniform distribution over the different group sizes' for each metric, with a value of 0.45 in the worst case (recall of singletons). Note that all the other methods, with the exception of HVFF ms, have precision, recall (and thus

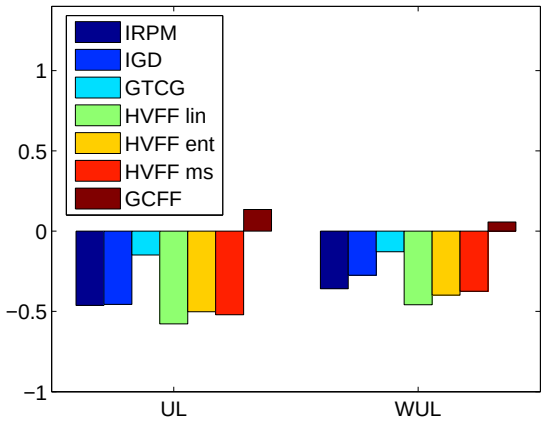

Figure 4. Upper-lower difference (UL) and weighted upper-lower difference (WUL) for all the methods on GDet dataset. (Best viewed in colors)

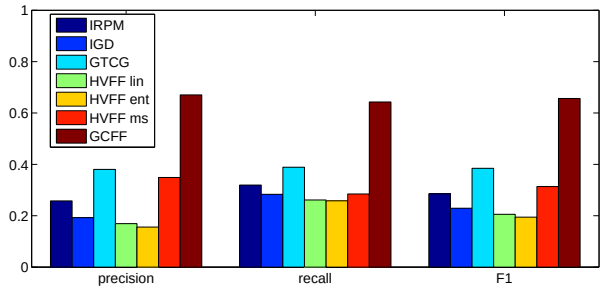

Figure 5. Average precision, recall and F_1 scores for all the methods on GDet dataset. (Best viewed in colors)

also F_1) 0 for cardinality 6 and some of them also for cardinality 5; testifying their inability to detect big groups. On the other hand, most of the methods are pretty good in singletons detection, and most of them perform pretty well in detection of pairs of people (cardinality 2).

5. Conclusions

In this paper we present a novel set of group detection metrics, which are based on evaluating the correct assignment of single individuals to differently sized groups. A bad assignments is not generically contributing to a group detection error, but maps into a data structure which highlights the tendency of over- or under-segmenting a group, other than defining precision and recall score for each group cardinality. At the best of our knowledge, such a level of description for the group detection problem evaluation has never been taken into account. We think that these measures could be used in general for any group detection algorithm, thus promoting a fair comparison in the related literature.

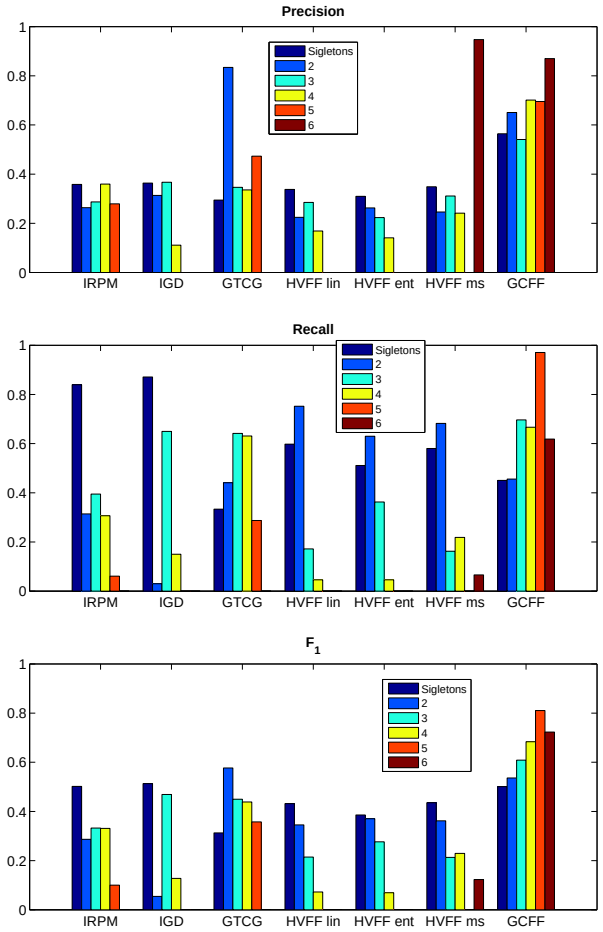

Figure 6. Precision, recall and F_1 scores for each groups' cardinality for all the methods on GDet dataset. (Best viewed in colors)

References

[1] 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, 2013. 1, 4

[2] L. Bazzani, M. Zanotto, M. Cristani, and V. Murino. Joint individual-group modeling for tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 37(4):746–759, 2015. 1

[3] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: the CLEAR MOT metrics. *Journal on Image and Video Processing*, pages 1–10, Jan. 2008. 1

[4] M. Chang, N. Krahnstoever, and W. Ge. Probabilistic group-level motion analysis and scenario recognition. In *International Conference on Computer Vision (ICCV)*, 2011. 1

[5] W. Choi, Y.-W. Chao, C. Pantofaru, and S. Savarese. Discovering groups of people in images. In *European Conference on Computer Vision (ECCV)*, 2014. 1, 2

[6] W. Choi and S. Savarese. A unified framework for multi-target tracking and collective activity recognition. In *European Conference on Computer Vision (ECCV)*, 2012. 1

[7] 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 *British Machine Vision Conference (BMVC)*, 2011. 1, 3, 4

[8] W. Ge, R. T. Collins, and R. B. Ruback. Vision-based analysis of small groups in pedestrian crowds. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 34:1003–1016, 2012. 1

[9] A. P. Hare. Group size. *American Behavioral Scientist*, 24(5):695–708, 1981. 1

[10] H. Hung and B. Kröse. Detecting f-formations as dominant sets. In *International Conference on Multimodal Interaction (ICMI)*, 2011. 1

[11] S. Khamis, V. I. Morariu, and L. S. Davis. A flow model for joint action recognition and identity maintenance. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1218–1225, 2012. 1

[12] O. Lanz. Approximate bayesian multibody tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 28(9):1436–1449, 2006. 4

[13] M. Manfredi, R. Vezzani, S. Calderara, and R. Cucchiara. Detection of static groups and crowds gathered in open spaces by texture classification. *Pattern Recognition Letters*, 44:39–48, 2014. 1, 2

[14] R. Mazzon, F. Poiesi, and A. Cavallaro. Detection and tracking of groups in crowd. In *International Conference on Advanced Video and Signal based Surveillance (AVSS)*, 2013. 1

[15] P. Rota, N. Conci, and N. Sebe. Real time detection of social interactions in surveillance video. In *European Conference on Computer Vision (ECCV)*, 2012. 1, 2

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

[17] F. Setti, O. Lanz, R. Ferrario, V. Murino, and M. Cristani. Multi-scale f-formation discovery for group detection. In *International Conference on Image Procesing (ICIP)*, 2013. 1, 4

[18] F. Setti, C. Russell, C. Bassetti, and M. Cristani. F-formation detection: Individuating free-standing conversational groups in images. *CoRR*, abs/1409.2702, 2014. 1, 2, 3, 4

[19] F. Solera, S. Calderara, and R. Cucchiara. Structured learning for detection of social groups in crowd. In *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2013. 2

[20] D. Tosato, M. Spera, M. Cristani, and V. Murino. Characterizing humans on riemannian manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 35(8):1972–1984, 2013. 4

[21] K. N. Tran, A. Bedagkar-Gala, I. A. Kakadiaris, and S. K. Shah. Social cues in group formation and local interactions for collective activity analysis. In *International Conference on Computer Vision Theory and Applications (VISAPP)*, 2013. 1, 4

[22] S. Vascon, Z. Eyasu, M. Cristani, H. Hung, M. Pelillo, and V. Murino. A game-theoretic probabilistic approach for detecting conversational groups. In *Asian Conference on Computer Vision (ACCV)*, 2014. 1, 4

[23] M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman. A model-theoretic coreference scoring scheme. In *Conference on Message Understanding*, 1995. 2

[24] K. Yamaguchi, A. Berg, L. Ortiz, and T. Berg. Who are you with and where are you going? In *IEEE Conference on Computer Vision and Patter Recognition (CVPR)*, 2011. 1