# Structured learning for detection of social groups in crowd

**Conference Paper** · August 2013

DOI: 10.1109/AVSS.2013.6636608

CITATIONS
33

READS
382

**3 authors:**

Francesco Solera
Università degli Studi di Modena e Reggio Emilia
**17** PUBLICATIONS **1,377** CITATIONS
SEE PROFILE

Simone Calderara
Università degli Studi di Modena e Reggio Emilia
**118** PUBLICATIONS **2,576** CITATIONS
SEE PROFILE

Rita Cucchiara
Università degli Studi di Modena e Reggio Emilia
**491** PUBLICATIONS **12,646** CITATIONS
SEE PROFILE

# Structured learning for detection of social groups in crowd

Francesco Solera Simone Calderara Rita Cucchiara
DIEF University of Modena and Reggio Emilia
Via Vignolese, 905 - 41100 Modena - Italy
{name.surname}@unimore.it

## Abstract

*Group detection in crowds will play a key role in future behavior analysis surveillance systems. In this work we build a new Structural SVM-based learning framework able to solve the group detection task by exploiting annotated video data to deduce a sociologically motivated distance measure founded on Hall's proxemics and Granger's causality. We improve over state-of-the-art results even in the most crowded test scenarios, while keeping the classification time affordable for quasi-real time applications. A new scoring scheme specifically designed for the group detection task is also proposed.*

## 1. Introduction

Behavior analysis will play a central role in future video surveillance systems as research on this topic has been revealing promising in helping to discover public safety risks or predict crimes. Nevertheless, trying to understand complex interactions in the scene just by looking at each individual separately is unrealistic, due to the inherent social nature of human behavior. This is because those interactions do not occur at an individual level nor at a crowd level, but they typically involve small subsets of people, namely groups. We thus believe future challenges will reside in enhancing action analysis by considering social interactions among small gathering of people sharing a common goal, to this end group detection becomes a mandatory step for modern crowd surveillance systems.

When walking down a street or when going to a public event, we have no doubts about who is there alone or about the people that are there together. This is why we have at our disposal an ensemble of information ranging from verbal exchanges to the understanding of popular culture gestures, such as waving a hand to someone who is at the other side of the plaza. In a typical video surveillance context we can't access this plethora of information, but nowadays we can entrust tracking systems to be able to extract very precise pedestrian trajectories, as emphasized by Fig. 1. Since

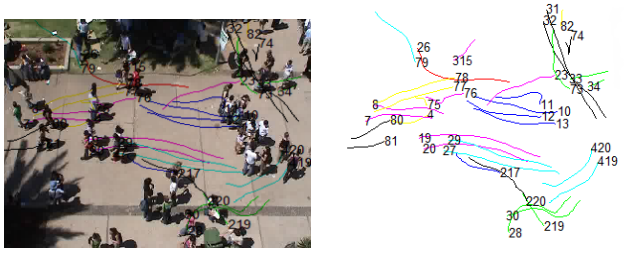


Figure 1: Example of trajectories from a crowded scenario.

collective behavior, at group level, is by definition a dynamic process, trajectories analysis can convey many important cues about group relationships, their formation and temporal evolution. For these motivations we restate the problem of group detection as the one of clustering people trajectories.

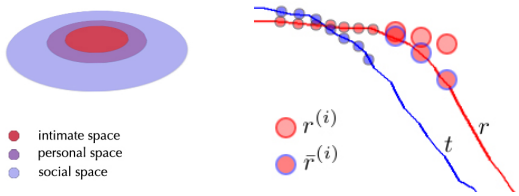
Group detection has been reaching interesting results only recently since the problem involves several challenging aspects of computer vision, ranging from reliable people detection and tracking to complex classifiers. A significant amount of work focused on the performance improvements that can be achieved by tracking algorithms when considering groups as structured entities of the scene [14]. Group tracking can be partitioned according to the availability of tracklets when analyzing the scene. In *group-based* approaches groups are considered as atomic entities in the scene as no higher level information can be extracted neatly, typically due to high noise or high complexity of crowded scenes [21, 6, 12]. Since these models are too simplistic to be used to further infer on groups behavior, *individual-group joint* approaches try to overcome the lack of finer information by hypothesizing trajectories while tracking groups at a coarser level [14, 2]. Conversely, *individual-based* tracking algorithms build up on single pedestrians trajectories, which are the most informative features we can hope to extract in a crowded scene. This kind of approach has been gaining momentum only lately since tracking even in high density crowds is becoming everyday a more feasible task [18]. Pellegrini *et al.* [15] employ a Conditional Random Field (CRF) to jointly predict trajectories and esti-

mate group memberships, modeled as latent variables, over a short time window. Similarly, Yamaguchi *et al*. [22] frame the task of predicting groups, among other modeled behaviors, as minimization of an energy function that encodes physical condition, personal motivation and social interactions features. More recently Chang *et al*. [4] proposed a soft segmentation process to partition the crowd by constructing a weighted graph, where the edges represent the probability of individuals to belong to the same group. Above all we mention Ge *et al*. [8] as we share an agglomerative approach to associate trajectories. They hierarchically merge clusters by evaluating an inter-group closeness measure defined on a combination of proximity and velocity features, stopping when a given condition is met. All these works actually present some drawbacks, ranging from lack of sociological motivation in the choice of features [15] to the naive understanding that groups can be considered as such even without modeling sociological aspects of transitivity [4] or by ignoring it at all [22]. The use of thresholds [8] is also a limiting aspect when dealing with dynamic concepts as groups are.

We propose to employ a supervised hierarchical bottom-up correlation clustering for solving the group detection task when trajectories of pedestrians are available. We overcome most of the limitations of the aforementioned solutions since a learning approach is followed to perform clustering without providing a formal definition of groups, as social studies underlines the lack of a universal theory about their formation. A novel set of features inspired by sociology and econometric is presented and through Structural SVM we learn how to linearly combine them in order to find a distance measure able to explain the concept of groups in any particular scenario. Another contribution resides in the definition of a robust and sociologically motivated Structural SVM loss function that improves the classifier accuracy over state-of-the-art loss functions. Besides, as the group detection task has specific peculiarities often neglected by traditional scoring measures, we suggest the use of our loss function in the evaluation of performances as well.

## 2. Social features and econometric for people trajectories analysis

Pedestrian trajectories are in fact multivariate time series projected on the space plane, still decades of social aggregation theories provide us some useful concepts when approaching the group detection task. Among these, Hall's **proxemics** theory [10] states that social distance between people is reliably correlated with physical distance, and more importantly it highlights the non-linearity of this relation. More formally, the theory defines bubbles around every individual, as depicted in Fig. 2a, where the interaction between pairs of individuals can be classified according to a quantization of their mutual distance into *intimate, personal, social* and *public space*. Note that Hall's work is based only on North American culture so, in order to obtain a continuous and more general measure, we substitute the original quantization with an exponential function and the proxemic score of any two trajectories $r$ and $t$ is computed as follows:




(a) (b)

Figure 2: (a) shows Hall's discrete definition of proxemics. (b) illustrates how taking into consideration past values of the $t$ might improve the prediction on future values of $r$.

$$f_{rt}^{prox} = \frac{1}{\max\{|r|,|t|\}} \sum_{i \in I_{t,r}} e^{-\sqrt{(t_x^{(i)} - r_x^{(i)})^2 + (t_y^{(i)} - r_y^{(i)})^2}} \tag{1}$$

where $I_{t,r}$ is the subset of time instances which restricts the summation to temporal intersection only and the coefficient out of the sum is needed in order for $f_{rt}^{prox}$ to be normalized.

Another intuitive way to recognize whether two pedestrians are walking together is to observe if in any way they are mutually affecting their path. This consideration is supported by a recent work by Couzin *et al*. [5] where is shown that groups are capable of taking united decisions regarding their direction and speed of movement even when only a few members have the information necessary to make such decisions. These concepts can be mathematically modeled through a measure of **causality**. Unfortunately causality isn't uniquely defined, throughout this work we choose to adopt from economics the definition of causality given by Granger [9]. A time series $t$ is said to Granger-cause $r$ ($t \rightarrow_{\mathrm{G}} r$) if it can be shown that the values of $t$ provide statistically significant information about future values of $r$. To infer about Granger causality one first has to test the null hypothesis, *i.e*. that $t \not\rightarrow_{\mathrm{G}} r$, by evaluating the autoregression of $r$:

$$r^{(i)} = a_0 + a_1 r^{(i-1)} + a_2 r^{(i-2)} + \cdots + a_m r^{(i-m)} \tag{2}$$

and compare it with the autoregression augmented with lagged values of $t$:

$$\bar{r}^{(i)} = r^{(i)} + b_0 + b_1 t^{(i-1)} + b_2 t^{(i-2)} + \cdots + b_n t^{(i-n)} \tag{3}$$

where $m$ and $n$ are the number of past observations we choose to take into account for $r$ and $t$ respectively. We call

$r^{(i)}$ the restricted model and $\bar{r}^{(i)}$ the unrestricted one: since the restricted model is contained in the unrestricted one, the latter will always be better to predict data, what we want to estimate is how much better it can get. An illustrated example is shown in Fig. 2b. Let $RSS_{r^{(i)}}$ and $RSS_{\bar{r}^{(i)}}$ be the residual sum of square errors computed by fitting the models on the ground truth series. In a standard $F$-test [9], the null hypothesis is rejected if the $F$ calculated from the data is greater than the critical value of the $F$-distribution for some desired false-rejection probability. Since we are more interested in knowing how much $t \rightarrow_{\mathrm{G}} r$ rather than in if $t \rightarrow_{\mathrm{G}} r$ at all, no threshold is set but $F$ is taken as a measure of direct causality:

$$F_{t \rightarrow r} = \frac{(RSS_{r^{(i)}} - RSS_{\bar{r}^{(i)}})/(n+1)}{RSS_{\bar{r}^{(i)}}/(N-m-m-2)} \tag{4}$$

with $N$ being the number of points used to fit the models to the data. Then our causality feature between two trajectories $r$ and $t$ is defined as follows:

$$f_{rt}^{caus} = \max\{F_{t \rightarrow r}, F_{r \rightarrow t}\}. \tag{5}$$

## 3. Methodology

There are many others of this concepts that could be distilled from social theories, nevertheless no rigorous definition of group has yet been accepted by the sociological community. To be able to grasp the complexity of the task imagine you have at disposal proxemic measures of pedestrians in a low crowded scene, as it could be a common supermarket - here we could surely exploit proxemics to infer the group structure of the crowd. But what about a highly crowded scene such as the exit of a stadium, where all the pedestrians are actually touching each others? Other than crowd density, also the environment conformation, the local culture and many other factors that we do not want to model explicitly make groups a dynamic concept. For this reason we follow a supervised clustering approach: we want

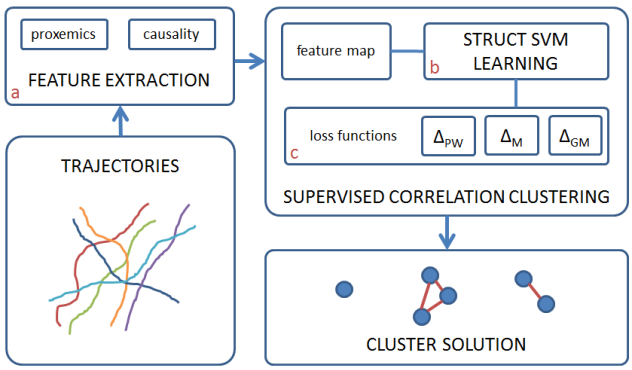



Figure 3: Block diagram of the group detection algorithm.

to learn which features are more significant in each scenario to describe groups. First, social features and causality (Fig. 3.a) are computed as explained in Sec. 2, then we proceed to define a learning framework that perform correlation clustering among trajectories.

We employed a Structural SVM (Fig. 3.b) to learn, from annotated data (*i.e.* where groups had been manually identified), the classification function, detailed in Sec. 4. In the training process, we make use of different loss functions (Fig. 3.c), better explained in Sec. 5, in particular we use a domain specific measure for group comparison in order to improve the classifier performances. The output of our classifier is a partition of the pedestrians set currently found on the scene.

The choice of a correlation clustering approach is motivated by the fact that we have to deal with really heterogeneous information such as space and time, *e.g.* two trajectories could have the same coordinates series but shifted in time. Devising a feature to grasp the relationship between space and time is not a trivial task and it turns out to be more effective to design a feature that describes the similarity between two trajectories than the trajectories themselves. A pairwise feature vector is defined for every couple of trajectories $r$ and $t$ as $\phi_{rt} = \{\phi_{rt}^k\}_k$.

## 4. Structural SVM for correlation social clustering

Correlation clustering [1] takes as input an affinity matrix $W$ where for $W_{rt} > 0$ we say that elements $r$ and $t$ are similar with certainty $|W_{rt}|$, and for $W_{rt} < 0$ we say elements $r$ and $t$ belong to different clusters with certainty $|W_{rt}|$. We can now find the correlation clustering $\mathbf{y}$ of a set of trajectories $\mathbf{x}$ as the one that maximize the sum of affinities for item pairs in the same cluster:

$$\arg\max_{\mathbf{y}} \sum_{y \in \mathbf{y}} \sum_{r \neq t \in y} W_{rt} \tag{6}$$

where we parametrize the affinity between trajectories $t$ and $r$ as the linear combination of the pairwise features $W_{rt} = \mathbf{w}^T \phi_{rt}$. By learning the weight vector $\mathbf{w}$ from the data, we are able to leave the features unchanged but combine them in the most effective way to discriminate groups in the current scenario.

Under these premises, let the input $\mathbf{x}_i$ be a set of trajectories and $\mathbf{y}_i$ its clustering solution it is straightforward to observe that the output cannot be described by a single valued function but is inherently structured. In this context we cannot use neither a traditional classifier nor a regressor. Structural SVM [19] in particular, offer a generalized framework to model and learn structured outputs by solving a loss augmented problem. The classifier learns the mapping $f : \mathcal{X} \rightarrow \mathcal{Y}$ between input space $\mathcal{X}$ and structured output space $\mathcal{Y}$ given a sample of input-output pairs

$S = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)\}$. A discriminant function $F : \mathcal{X} \times \mathcal{Y} \rightarrow \Re$ is defined over the joint input-output space where $F(\mathbf{x}, \mathbf{y})$ can be interpreted as measuring the compatibility of $\mathbf{x}$ and $\mathbf{y}$. From this function the prediction function $f$ results:

$$f(\mathbf{x}) = \arg\max_{\mathbf{y} \in \mathcal{Y}} F(\mathbf{x}, \mathbf{y}; \mathbf{w}) \tag{7}$$

where the maximizer over the label space $\mathcal{Y}$ is the predicted label, *i.e.* the solution of the inference problem. For simplicity we choose to restrict the space of $F$ to linear functions over some combined feature representation $\Psi(\mathbf{x}, \mathbf{y})$. The feature mapping cannot be defined out of the context of the problem, as it is the problem itself that specifies what kind of solution we want, given a particular input. As a matter of fact, following the parametric definition of correlation clustering in Eq. 6 the compatibility of an input-output pair can be defined as

$$F(\mathbf{x}, \mathbf{y}; \mathbf{w}) = \mathbf{w}^T \Psi(\mathbf{x}, \mathbf{y}) = \mathbf{w}^T \sum_{y \in \mathbf{y}} \sum_{r \neq t \in y} \phi_{rt}. \tag{8}$$

The problem of learning in structured and interdependent output spaces can been formulated as a maximum-margin problem. We adopt the $n$-slack, margin-rescaling formulation of [19]:

$$\begin{aligned} \min_{\mathbf{w}, \xi} \quad & \frac{1}{2}\|w\|^2 + \frac{C}{n}\sum_{i=1}^{n} \xi_i \\ \text{s.t.} \quad & \forall i : \xi_i \geq 0, \\ & \forall i, \forall \mathbf{y} \in \mathcal{Y} \backslash \mathbf{y}_i : \mathbf{w}^T \delta\Psi_i(\mathbf{y}) \geq \Delta(\mathbf{y}, \mathbf{y}_i) - \xi_i, \end{aligned} \tag{9}$$

where $\delta\Psi_i(\mathbf{y}) = \Psi(\mathbf{x}_i, \mathbf{y}_i) - \Psi(\mathbf{x}_i, \mathbf{y})$, $\xi_i$ are the slack variables introduced in order to accommodate for margin violations and $\Delta(\mathbf{y}, \mathbf{y}_i)$ is the loss function. Intuitively, we want to maximize the margin and jointly guarantee that for a given input, every possible output result is considered worst than the correct one by at least a margin of $\Delta(\mathbf{y}_i, \mathbf{y}) - \xi_i$, where $\Delta(\mathbf{y}_i, \mathbf{y})$ is bigger when the two predictions are known to be more different.

The quadratic program QP 9 introduces a constraint for every possible wrong clustering of the set. Unfortunately, the number of wrong clusterings scales more than exponentially with the number of items. So we choose to adopt the cutting plane algorithm proposed by Tsochantaridis *et al.* [19] where we start with no constraints, and iteratively find the most violated constraint

$$\hat{\mathbf{y}}_i = \arg\max_{\mathbf{y}} \Delta(\mathbf{y}_i, \mathbf{y}) - \delta\Psi_i(\mathbf{y}) \tag{10}$$

and re-optimize until convergence. Finding the most violated constraint requires to solve the correlation cluster problem, which we know to be NP-hard [1]. Finley and Joachims [7] propose a greedy approximation algorithm which works by initially considering each pedestrian in its own cluster, then iteratively merging the two clusters whose union would produce the worst clustering score.

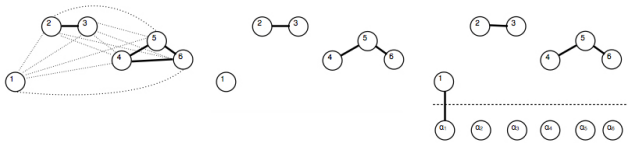
One remarkable aspect of supervised correlation clustering is that it doesn't need a priori model selection, *i.e.* there is no need to know in advance how many groups are present in the scene. Moreover two elements could end up in the same cluster if the net effect of the merging process is positive even if their affinity measure is negative, implicitly modeling the transitive property of relationships in groups which is known from sociological studies [13]: to be considered part of a group you typically will have to be positively connected with at least half of the members.

## 5. Loss function and scoring procedure

The learning ability of the algorithm highly depends on the choice of the loss function since it has the power to force or relax input margins. One common choice for clustering could be to adopt the well-known **pairwise loss** function [17], $\Delta_{PW}(\mathbf{y}, \bar{\mathbf{y}})$, which is defined as the ratio between the number of pairs on which $\mathbf{y}$ and $\bar{\mathbf{y}}$ disagree on their cluster membership and the number of all possible pairs of elements in the set. The main strengths of this measure are its simplicity and the computational efficiency but it tends to be imprecise when dealing with groups in large crowds. This is due to the quadratic number of connections that exist between crowd members, *e.g.* in Fig. 4a for $n = 6$ elements we have 15 links according to $\binom{n}{2}$. For not very crowded scenes this measure seems to behave fine, but beyond some threshold the loss will eventually become insignificant because the number of positive links (among group members, bold links in Fig. 4a) becomes negligible w.r.t. the total number of links.

The problem of clustering trajectories is in many ways similar to the noun-coreference problem [3] in NLP, where nouns have to be clustered according to who they refer to. Above all, the combinatorial number of connections is shared. For this problem, the MITRE score [20] has been identified as a suitable scoring measure. The **MITRE loss**, $\Delta_M(\mathbf{y}, \bar{\mathbf{y}})$, is founded on the understanding that connected components are sufficient to describe dynamic groups. Thus spanning trees instead of complete graphs are used to represent clusters, as depicted in Fig. 4b. Nonetheless this measure is still imprecise when applied to the group detection task since it doesn't deal with singletons which should be considered positively when correctly classified. Our proposed loss function overcomes this limitation.

The **GROUP-MITRE loss** $\Delta_{GM}(\mathbf{y}, \bar{\mathbf{y}})$ ($G$-MITRE), is obtained by adding, for each pedestrian described by the trajectory $r$, a fake counterpart described by $\alpha_r$, as illustrated in Fig. 4c, to which only singletons are connected. Through this shrewdness we can now take into considera-



(a) $\Delta_{PW}$ (b) $\Delta_M$ (c) $\Delta_{GM}$

Figure 4: Example of loss functions computation.

tion singletons as well when computing the loss function. Consider two clustering solutions $\mathbf{y}$, $\bar{\mathbf{y}}$ and an instance of their respective spanning forests $S$ and $R$. The connected components of $S$ and $R$ are identified respectively by the trees $S_i, i = 1, \ldots, n$ and $R_i, i = 1, \ldots, m$. Note that if the number of elements in $S_i$ is $|S_i|$, then only $c(S_i) = |S_i| - 1$ links are needed in order to create a spanning tree. Let us define the partition of a tree $S_i$ with respect to $R$, $p(S_i)$, as a set of subtrees obtained by considering only the links in $S_i$ that are also found in $R$. Besides, if $R$ partitions $S_i$ in $p(S_i)$ subtrees, $m(S_i) = |p(S_i)| - 1$ links are sufficient to restore the original tree. It follows that the recall error for $S_i$ can be computed as the number of missing links divided by the minimum number of links needed to create the spanning tree. Accounting for all trees $S_i$ we define the global recall measure of $S$ as

$$\mathcal{R}_S = 1 - \frac{\sum_{i=1}^{n} m(S_i)}{\sum_{i=1}^{n} c(S_i)} \quad (11)$$

$$= \frac{\sum_{i=1}^{n} |S_i| - |p(S_i)|}{\sum_{i=1}^{n} |S_i| - 1} \quad (12)$$

The precision of $S$ can be computed by exchanging $S$ and $R$, which can be also seen as the recall of $R$ with respect to $S$, guaranteeing that the measure is symmetric. Given the recall $\mathcal{R}$ the loss is defined as

$$\Delta_{GM} = 1 - F_1 \quad (13)$$

$$= 1 - 2\frac{\mathcal{R}_S \mathcal{R}_R}{\mathcal{R}_S + \mathcal{R}_R} \quad (14)$$

where $F_1$ is the standard $F$-score.

## 6. Experimental results

We tested our system on two publicly available datasets, namely the *BIWI Walking Pedestrians* dataset [16] and the *Crowds-By-Examples* (*CBE*) dataset [11]. The *BIWI* dataset records two low crowded scenes, one outside a university, named `eth`, and one, `hotel`, at a bus stop, both shown in Fig. 5, while the *CBE* dataset records a high density crowd video outside another university, `student003` (`stu003`)[1]. As it can be seen from Fig. 5c this dataset provides some real challenge as the density of the pedestrians is significant as well as for the presence of stairs and multiple entry and exit points.

[1]We made social groups annotations for the *CBE* `student003` dataset available at `http://imagelab.ing.unimore.it`.

| Dataset | *BIWI* `eth` | *BIWI* `hotel` | *CBE* `stu003` |
|---|---|---|---|
| Frames | 1448 | 1168 | 541 |
| Pedestrians | 360 | 390 | 434 |
| Groups | 243 | 326 | 284 |

Table 1: Datasets annotations description.

In the test setting we use three minutes of video for training the classifier and the remaining for testing; the trajectories were acquired on a time window of 10 seconds. For each pair of trajectories $r$ and $t$ part of input vector $\mathbf{x}_i$, the feature vector $\phi_{rt}$ is defined as

$$\phi_{rt} = [f_{rt}^{prox}, f_{rt}^{caus}, 1 - f_{rt}^{prox}, 1 - f_{rt}^{caus}]. \quad (15)$$

A peculiar drawback of correlation clustering is in fact that it doesn't specify how to obtain the affinity measure for the $k$-th feature $\phi_{rt}^k$ out of our pairwise similarity feature $f_{rt}^k$. Since the affinity matrix $W_{rt}$ needs to be negative when two trajectories $r$ and $t$ are dissimilar we decided to extend the feature vector by including both similarities and dissimilarities measure for each $f_{rt}^k$.

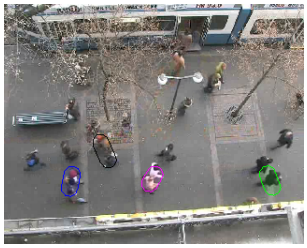
We run tests for every dataset evaluating the impact on performances of the different loss functions discussed in Sec. 5. Results in terms of precision, recall and $F$-score are depicted in Tab. 2. We observe that the gap between the $G$-MITRE and the pairwise loss is wider in the `eth` dataset, where the balance between groups and pedestrian is higher and as such learning to distinguish them become crucial. As already pointed out, the pairwise loss can obtain outstanding performances when the number of pedestrians in the scene is limited, but it becomes ineffective in `stu03`. Our loss function not only outperforms the others when used in the training process, but it also shows to be more robust and turned out to be faster. Moreover we compare our results

| | $\Delta_{PW}$ | | $\Delta_M$ | | $\Delta_{GM}$ | | |
|---|---|---|---|---|---|---|---|
| | $\mathcal{P}$ | $\mathcal{R}$ | $\mathcal{P}$ | $\mathcal{R}$ | $\mathcal{P}$ | $\mathcal{R}$ | $F_1$ |
| `hotel` | 92.7 | 92.0 | 90.1 | 92.7 | 93.4 | 93.7 | **93.6** |
| `eth` | 86.6 | 89.8 | 52.6 | 56.3 | 87.0 | 91.0 | **89.8** |
| `stu003` | 66.6 | 67.5 | 74.6 | 76.1 | 82.3 | 80.1 | **81.2** |

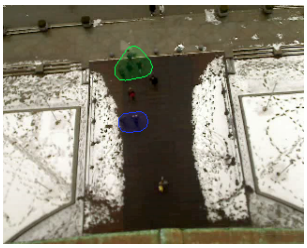
Table 2: Experimental results obtained by varying the loss.

| | our | | [15] | | [22] | |
|---|---|---|---|---|---|---|
| | $\mathcal{P}$ | $\mathcal{R}$ | $\mathcal{P}$ | $\mathcal{R}$ | $\mathcal{P}$ | $\mathcal{R}$ |
| `hotel` | **93.4** | 93.7 | - | - | 91.3 | **95.9** |
| `eth` | **87.0** | **91.0** | - | - | 83.0 | 80.2 |
| `stu003` | **82.3** | 80.1 | 46.0 | **82.0** | 80.5 | 77.0 |

Table 3: Comparison of our method with state-of-the-art.



(a) *BIWI* `hotel`



(b) *BIWI* `eth`

(c) *CBE* `student003`

Figure 5: Note how groups can be detected regardless of the number of pedestrians, shape and structure of the clusters.

with state-of-the-art methods in [15, 22], the quantitative results shown in Tab. 3 indicate that method outperforms all the approaches in most of the proposed videos. This is due to the use of sociological features, supervised learning able to better generalize to previously unseen scenarios and a specifically designed loss function. Our method takes about 1 second to cluster 10 seconds of observed trajectories in an averagely crowded scene. Fig. 5 reports a visual example of an instance of the classifier solutions.

## 7. Conclusions

We propose a new algorithm for the group detection task by reformulating the problem to the one of clustering trajectories and solving it through a parametric correlation clustering trained by a Structural SVM. The algorithm is able to generalize to previously unseen group structures and compositions making it robust and strongly invariant to crowd density and the environment. We also propose our new scoring measure, motivated by both sociological studies and intuition, to be taken as a standard performance indicator for the group detection task.

## References


[1] N. Bansal, A. Blum, and S. Chawla. Correlation clustering. *Machine Learning*, 56:89–113, 2004. 3, 4

[2] L. Bazzani, Cristani, and V. Murino. Decentralized particle filter for joint individual-group tracking. In *CVPR*, 2012. 1

[3] C. Cardie and K. Wagstaff. Noun Phrase Coreference as Clustering. 1999. 4

[4] M. C. Chang, N. Krahnstoever, and W. Ge. Probabilistic group-level motion analysis and scenario recognition. In *ICCV*, pages 747–754, 2011. 2

[5] I. D. Couzin, J. Krause, N. R. Franks, and S. A. Levin. Effective leadership and decision-making in animal groups on the move. *Nature*, 433, 2005. 2

[6] M. Feldmann, D. Fränken, and W. Koch. Tracking of extended objects and group targets using random matrices. *IEEE Trans. Signal Processing*, 59(1):1409–1420, 2011. 1

[7] T. Finley and T. Joachims. Supervised clustering with support vector machines. In *International Conference on Machine Learning (ICML)*, pages 217–224, 2005. 4

[8] W. Ge, R. Collins, and R. Ruback. Vision-based analysis of small groups in pedestrian crowds. *PAMI*, 34, 2012. 2

[9] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969. 2, 3

[10] E. Hall. *The hidden dimension*. Doubleday, 1966. 2

[11] Lerner, Alon, Chrysanthou, Yiorgos, Lischinski, and Dani. Crowds by Example. *Computer Graphics Forum*, 26:655–664, 2007. 5

[12] W. C. Lin and Y. Liu. A lattice-based mrf model for dynamic near-regular texture tracking. *PAMI*, 29(5):777–792, 2007. 1

[13] C. McPhail and R. T. Wohlstein. Using film to analyze pedestrian behavior. *Sociological Methods & Research*, 10(3):347–375, 1982. 4

[14] S. K. Pang, J. Li, and S. Godsill. Models and algorithms for detection and tracking of coordinated groups. In *Aerospace Conference*, pages 1–17, 2008. 1

[15] S. Pellegrini, A. Ess, and L. J. V. Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In *ECCV*, pages 452–465, 2010. 1, 2, 5, 6

[16] S. Pellegrini, A. Ess, K. Schindler, and L. J. V. Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *ICCV*, pages 261–268, 2009. 5

[17] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Jou. of the American Statistical Association*, 66(336):846–850, 1971. 4

[18] M. Rodriguez, I. Laptev, J. Sivic, and J.-Y. Audibert. Density-aware person detection and tracking in crowds. In *ICCV*, pages 2423–2430, 2011. 1

[19] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *International Conference on Machine Learning (ICML)*, pages 104–112, 2004. 3, 4

[20] M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman. A model-theoretic coreference scoring scheme. In *Proc. Conf. on Message understanding*, pages 45–52, 1995. 4

[21] Y. D. Wang, J. K. Wuand, A. A. Kassim, and W. M. Huang. Tracking a variable number of human groups in video using probability hypothesis density. *ICPR*, 2006. 1

[22] K. Yamaguchi, A. Berg, L. Ortiz, and T. Berg. Who are you with and where are you going? In *CVPR*, pages 1345 –1352, 2011. 2, 5, 6