Social Groups Detection in Crowd through Shape-Augmented Structured Learning

Francesco Solera and Simone Calderara

DIEF University of Modena and Reggio Emilia, Italy
francesco.solera@gmail.com, simone.calderara@unimore.it

Abstract. Most of the behaviors people exhibit while being part of a crowd are social processes that tend to emerge among groups and as a consequence, detecting groups in crowds is becoming an important issue in modern behavior analysis. We propose a supervised correlation clustering technique that employs Structural SVM and a proxemic based feature to learn how to partition people trajectories in groups, by injecting in the model socially plausible shape configurations. By taking into account social groups patterns, the system is able to outperform state of the art methods on two publicly available benchmark sets of videos.

Keywords: group detection, proxemic theory, Structural SVM.

1 Introduction

Group detection in video surveillance systems is profoundly motivated by behavior analysis and security issues in enhancing scene understanding, as the truth is that many interesting behaviors don't occur at an individual level but are the results of complex interactions between individuals in specific subsets of the crowd, namely groups. It is known, in fact, that the existence of groups highly influences the behavior of the individuals as it is at this level that people start to experiment structured interactions [1]. While there isn't a general solution to the problem of locating groups, data driven approaches have recently been obtaining interesting results, mainly motivated by the improvements in tracking performance that can be achieved when considering groups as structured entities of the scene [2]. Group tracking can be partitioned according to the availability of tracklets. In *group-based* approaches [3][4][5] groups are seen as the atomic entities to look for, with the major drawback of creating models that are too simplistic and cannot be used to further infer on groups behavior. On the other hand, *individual-based* tracking algorithms [6][7][8] focus on single pedestrians trajectories, which are the most informative features we can extract from a crowded scene. This latter approach has been gaining momentum only lately as tracking even in dense crowds is becoming everyday a more feasible task [9].

The common drawback of all these approaches lays in their scarce consideration of decades of sociological theories which can though provide many brilliant insights. Methods that rely exclusively on data assume to have at disposal a

A. Petrosino (Ed.): ICIAP 2013, Part I, LNCS 8156, pp. 542–551, 2013.

(a) (b) (c)

Fig. 1. Example of small groups in crowd

highly descriptive and complete dataset, which is quite an unrealistic assumption to state when considering dynamic concepts as groups and crowds are. As opposite, we believe social dynamics can help to deeply understand the group formation phenomenon. In particular, despite the well-assessed theory about collective crowd will, new results are underlining that people in groups tend to produce predefined shapes while maintaining their identity and goals [1], as can be observed in Fig. 1b. Ge *et al.* [8] actually present a statistical shape analysis method to analyze the spatial position of all group members jointly and estimate the typical group formations of walking pedestrians; but then they don't take advantage of these configurations to improve their group detection method.

Given the aforementioned sociological breakthrough we devise a new algorithm for group detection based on structured learning, which let us incorporate shapes structure evaluation in the optimization process yielding to more sociologically plausible predictions. We propose to train a supervised hierarchical bottom-up correlation clustering when trajectories of pedestrians are available. As we want to focus on the sociological side of the problem, we employ a feature founded on Hall's *proxemic distance theory* [10] and through Structural SVM we decide, based on previously annotated scenes, how this feature can be significant in explaining the concept of groups in any given scenario.

2 Structured Output Learning for Group Detection

Pedestrian trajectories encode many sociological and physical information about the way people interact. If two pedestrians have diverging trajectories it's very unlikely that they were on the scene together and, at the same time, in a group of friends everyone will likely have very similar and compact trajectories over a generic period of observation. Starting from this consideration we reformulate the problem of finding groups in the scene as the one of clustering trajectories, or partitioning the set of those. The solution of a clustering algorithm is a collection of members' assignments, thus conventional classifiers aren't suitable to deal with the combinatorial size of this problem output space. Building up on the work of Finley and Joachims [11], we propose a structured supervised algorithm able to learn how to partition crowds using sociologically founded features between pairs of trajectories. The method is summarized in the scheme of Fig. 2.

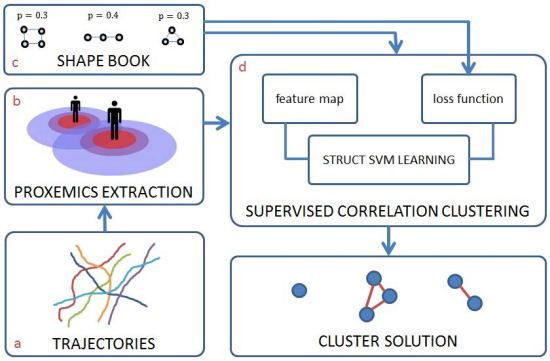


Fig. 2. Block diagram illustrating our group detection algorithm

2.1 Hall's Proxemic Distance: A Sociological Feature

An important relation between space and social interaction has been first formulated in Hall's proxemic theory and has already been applied to people trajectories in order to better understand diadic interactions [12]. Proxemic theory states that the social distance between people is reliably correlated with their physical distance, and more important it tells us that this relation is not linear. Intuitively, the theory defines circles around every individual and the interaction between pairs of individuals is classified according to which circle they mutually reside in, as depicted in Fig. 2b. The result is a non linear quantization of their distance in intimate, personal, social and public space. We generalize the original quantization by approximating it with an exponential term. The proxemic score of any two trajectories r and t is computed as

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

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

2.2 Supervised Correlation Clustering through Structural SVM

While Hall's proxemic theory is a valuable instrument which can be exploited in order to better understand group dynamics, it is not sufficient to solve the problem of their detection. It is possible to grasp the complexity of the task by considering a highly crowded scene where all the pedestrians are touching each others, here distance is much less discriminant. Other than crowd density, also the environment conformation, the local culture and other factors that cannot be modeled explicitly make groups a challenging concept to define. For this reason we adopt a supervised clustering approach in order to learn how proxemic

distance can be significant to describe groups in different scenarios. In [13] we prove our framework is also able to balance the contributions of multiple features useful to describe crowded scenarios, while here we only employ Hall's distance as we rather investigate the importance of group patterns in the detection task.

Since information about groups doesn't reside in the trajectories, but in the spatiotemporal relationships they are engaged in, we can't apply standard clustering techniques. **Correlation clustering** [14] operates exactly in this scope, where we don't want to describe the elements themselves but rather their pairwise relationships, computed as in Eq. 1 for the special case of Hall's proxemic distance. Formally, correlation clustering takes as input an affinity matrix $W = \{W_{rt}\}_{rt}$ 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 $|W_{rt}|$. Our model aims to parametrize $W_{rt} = \mathbf{w}^T \phi_{rt}$ in $\mathbf{w}$ so that we can fix the feature but still be able to adjust its importance for as much discriminative information it can provide on the current scenario. By defining our feature vector to be $\phi_{rt} = [f_{rt}^{prox}, 1 - f_{rt}^{prox}]^T$ we can create negative entries in W_{rt} representing unlikely pairs of individuals. The correlation clustering $\bar{\mathbf{y}}$ of a set of trajectories $\mathbf{x}$ is the configuration that maximizes the sum of affinities for item pairs in the same cluster:

$$\bar{\mathbf{y}} = \arg\max_{\mathbf{y}} \sum_{y \in \mathbf{y}} \sum_{r \neq t \in y} W_{rt} = \arg\max_{\mathbf{y}} \sum_{y \in \mathbf{y}} \sum_{r \neq t \in y} \mathbf{w}^T \phi_{rt} \tag{2}$$

Given the parametric model of the correlation clustering, the weight vector $\mathbf{w}$ can be learned using structured learning. **Structural SVMs** [15] offer a generalized framework to model and solve structured output problems by learning a mapping $f : \mathcal{X} \to \mathcal{Y}$ between input space $\mathcal{X}$ and structured output space $\mathcal{Y}$ given a sample of input-output pairs $S = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)\}$. Recall $\mathbf{x}_i$ is a set of trajectories and $\mathbf{y}_i$ is a clustering solution for $\mathbf{x}_i$. In contrast to standard multiclass classification, where a different prediction function for each class is learned independently, we define a discriminant function $F : \mathcal{X} \times \mathcal{Y} \to \Re$ over the joint input-output space where $F(\mathbf{x}, \mathbf{y})$ can be interpreted as a compatibility measure between $\mathbf{x}$ and $\mathbf{y}$. We remark that $F(\mathbf{x}, \mathbf{y})$ 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, the parametric formulation of correlation clustering presented in Eq. 2, implicitly defines the compatibility of an input-output pair. We can thus restrict the space of F to linear functions over some combined feature representation $\Psi(\mathbf{x}, \mathbf{y})$, yielding to $F(\mathbf{x}, \mathbf{y}; \mathbf{w}) = \mathbf{w}^T \Psi(\mathbf{x}, \mathbf{y})$. From Eq. 2 it follows

$$\Psi(\mathbf{x}, \mathbf{y}) = \sum_{y \in \mathbf{y}} \sum_{r \neq t \in y} \phi_{rt}. \tag{3}$$

Given F we define the prediction function $f(\mathbf{x}) = \arg\max_{\mathbf{y} \in \mathcal{Y}} F(\mathbf{x}, \mathbf{y}; \mathbf{w})$. According to Finley and Joachims [11] the problem of finding f can be restated as a n-slack margin rescaling maximum-margin problem:

$$\begin{aligned} \min_{\mathbf{w},\xi} \quad & \frac{1}{2}\|w\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & \forall i : \xi_i \geq 0, \\ & \forall i, \forall \mathbf{y} \in \mathcal{Y} \backslash \mathbf{y}_i : \mathbf{w}^T \delta\Psi_i(\mathbf{y}) \geq \Delta(\mathbf{y}, \mathbf{y}_i) - \xi_i, \end{aligned} \tag{4}$$

where $\delta\Psi_i(\mathbf{y}) = \Psi(\mathbf{x}_i, \mathbf{y}_i) - \Psi(\mathbf{x}_i, \mathbf{y})$, ξ_i are the slack variables introduced in order to accommodate for margin violations and $\Delta(\mathbf{y}, \mathbf{y}_i)$ is the loss function.

The optimization problem of Eq. 4 introduces a constraint for every possible wrong clustering of the set. Since the number of wrong clusterings scales more than exponentially with the number of items, we choose to employ the *cutting plane* algorithm [15] that, starting with no constraints, aims at iteratively finding the most violated one $\hat{\mathbf{y}}_i = \arg\max_{\mathbf{y}} \Delta(\mathbf{y}_i, \mathbf{y}) - \mathbf{w}^T \delta\Psi_i(\mathbf{y})$ and re-optimize until convergence. Finding the most violated constraint requires to solve the correlation clustering problem, which we know to be NP-hard [14]. Finley and Joachims [11] propose a greedy approximation which works by initially considering each element in its own cluster, then iteratively merging the two clusters whose union would produce the worst clustering score.

The learning ability of the algorithm strongly depends on the choice of the loss function in Eq. 4, as it has the power to force or relax input margins. Given the analogy between trajectories clustering and the noun-coreference problem [17], we adopt the **MITRE** score [18] from NLP in defining the loss of Eq. 4:

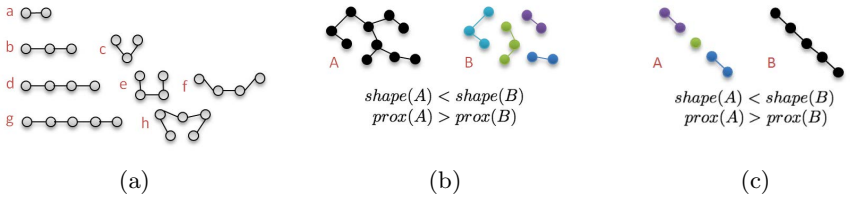
$$\Delta(\mathbf{y}_i, \mathbf{y}) \equiv \Delta_M(\mathbf{y}_i, \mathbf{y}) = 1 - F_1 \tag{5}$$

where F_1 is the harmonic mean of precision and recall computed as in [18].

3 Shape-Augmented Learning

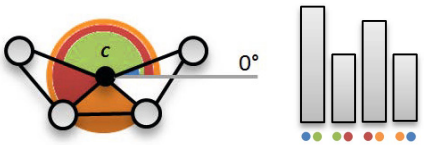
The spatial organization of pedestrians inside groups tends to obey to patterns that facilitate social interactions and verbal communications, while trying to avoid collisions with in-group members and out-group pedestrians. This patterns are highly dependent on the crowd density, the environment conformation and the group speed and it's not completely clear yet how these elements are all correlated together. Nevertheless recent behavior analysis research has been focusing on the formalization of such concepts. Moussaïd *et al.* [1] show from real-world data that up to 70% of observed pedestrians in a commercial street are walking in groups and provide a distribution of those groups size. Despite the different experimental scenario, Bandini *et al.* [16] provide empirical results about the frequency of group patterns which converge accordingly to the work of Moussaïd *et al.* [1]. As a consequence, we think it does make sense to consider some pattern more probable than others while searching the crowd for groups.

A peculiar consideration emerging from these studies [16] states that in dense scenarios, people tend to rearrange in formations which allow to feel compact and protected from out-group individuals, Fig. 3b (case B), while if space is available people prefer to reside in a line-like shape, Fig. 3c (case B). In both



(a) (b) (c)

Fig. 3. (a) reports the most frequent group patterns. (b) and (c) show two scenarios where proxemic measure would produces socially implausible grouping predictions.



(a) (b)

Fig. 4. The process employed in order to obtain a shape histogram (b) from a given group pattern (a)

cases, minimizing mutual distances among members (*i.e.* exploiting proxemics) is not sufficient to produce socially plausible groups, Fig. 3b and 3c (cases A).

3.1 Modeling Shapes

In order for a group-shape descriptor to be considered effective, it should be invariant under translation, rotation and scale, as relative distances are already taken into account by pairwise features. Moreover it should be robust against small variations, *i.e.* small differences in the group configuration should correspond to small differences in the shape description.

We begin by finding the mean point C between group members and starting from the closest one and following a counter clock-wise direction, we keep track of the angular position of each member w.r.t. C, as shown in Fig. 4a. This step guarantees that the scoring measure will be invariant under translation and rotation. In order to ensure scale invariance as well, we choose to neglect the distance between group members focusing only on their sequential angular distance, *i.e.* we only measure the angular distance from one member to the next one, yielding to the histogram reported in Fig. 4b. Note that every histogram obtained as described above is characterized by a number of bins equal to the number of group members and a normalization value of a turn. Shapes can now be compared using a trivial histogram intersection measure.

3.2 Shape-Aware Loss Function

Given the social groups configurations depicted in Fig. 3a and their relative shape histograms obtained as explained in Sec. 3.1, we propose to build a codebook

of preferred shapes, *shapebook*, where each configuration is associated with an a-priori probability of occurrence extracted from sociological studies [1][16]. The configurations listed in Fig. 3a are only a subset of all the possible ones, but we remark that a group can still be detected as such, even if it lays in a configuration that we do not model. It will simply be considered less probable.

Given a possible clustering solution $\mathbf{y}$, we can measure how much the detected groups are sociologically conceivable by (i) finding, for each frame f in the time window T and for the configuration $s_{y,f}$ of each group $y \in \mathbf{y}$ the most similar shape $\bar{s}_{y,f}$ among the ones described in the shapebook, (ii) evaluating the similarities of their histograms through histogram intersection and by (iii) assigning a score proportional to the probability of the group pattern $\bar{s}_{y,f}$ and its similarity with pattern $s_{y,f}$. The process is synthesized in the following formula:

$$\Delta_S(\mathbf{y}) = 1 - \frac{1}{|T|} \sum_{y \in \mathbf{y}} \sum_{f \in T} [s_{y,f} \cap \bar{s}_{y,f}]\, p(\bar{s}_{y,f}). \tag{6}$$

As shapes embody our a-priori knowledge of the problem, we want to force them into the learning framework. To our knowledge, there isn't any explicit proposal in literature focused on the idea of encoding a-priori data in Structural SVM. Moreover, since we let correlation clustering define the joint feature representation, we employ pairwise features which cannot capture global level information such as the shape of the group they are in. Due to the impossibility of including shapes at a feature level, we consider the loss function level, where information about all individuals in the scene can be accessed simultaneously. By pursuing this idea, we define the *shape-aware* loss function $\Delta(\mathbf{y}_i, \mathbf{y})$ as

$$\Delta(\mathbf{y}_i, \mathbf{y}) \equiv \lambda \Delta_M(\mathbf{y}_i, \mathbf{y}) + (1 - \lambda)\Delta_S(\mathbf{y}). \tag{7}$$

By replacing the loss function in the constraints of Eq. 4 with the one of Eq. 7, we obtain

$$\begin{aligned} \min_{\mathbf{w},\xi} \quad & \frac{1}{2}\|w\|^2 + \frac{C}{n}\sum_{i=1}^{n} \xi_i \\ \text{s.t.} \quad & \forall i : \xi_i \geq 0, \\ & \forall i, \forall \mathbf{y} \in \mathcal{Y} \backslash \mathbf{y}_i : \mathbf{w}^T \delta\Psi_i(\mathbf{y}) - (1 - \lambda)\Delta_S(\mathbf{y}) \geq \lambda \Delta_M(\mathbf{y}, \mathbf{y}_i) - \xi_i. \end{aligned} \tag{8}$$

The new set of constraints state that it doesn't matter if the margin of the feature space isn't optimally maximized as the classifier can also count on the fact that poorly structured solutions will be highly penalized. This allows $\mathbf{w}$ to be slightly less fit on the data, enhancing the generalization capability of the algorithm.

Given the optimization problem of Eq. 8, questions arise on the correctness and on the convergence of classical algorithms to solve this unconventional SSVM problem. We modified the original cutting plane algorithm [15] and solved the dual form of the quadratic problem of line 11 in Alg. 1 through the Sequential Minimal Optimization (SMO) approach, inspired by the work of Lee and

Jang [19]. Let $\mathcal{L}_P$ be the Lagrangian of the problem in Eq. 8, by differentiating it and back-substituting we obtain its dual counterpart $\mathcal{L}_D$, as in Eq. 9.

$$\begin{aligned} \max_{\alpha,\beta} \quad & \mathcal{L}_D(\alpha,\beta) = -\frac{1}{2}\sum_{i,\mathbf{y}\neq\mathbf{y}_i}\sum_{j,\bar{\mathbf{y}}\neq\mathbf{y}_j}\alpha_{i\mathbf{y}}\alpha_{j\bar{\mathbf{y}}}\delta\Psi_i(\mathbf{x}_i,\mathbf{y})\delta\Psi_j(\mathbf{x}_j,\bar{\mathbf{y}}) \\ & + \sum_{i,\mathbf{y}\neq\mathbf{y}_i}\alpha_{i\mathbf{y}}\Delta(\mathbf{y}_i,\mathbf{y}) + \sum_{i,\mathbf{y}\neq\mathbf{y}_i}\beta_{i\mathbf{y}}(1-\lambda)\Delta_S(\mathbf{y}) \\ \text{s.t.} \quad & \forall i : \sum_{\mathbf{y}\neq\mathbf{y}_i}(\alpha_{i\mathbf{y}}+\beta_{i\mathbf{y}}) = \frac{C}{n}, \quad \forall i, \forall \mathbf{y} : \alpha_{i\mathbf{y}} \geq 0, \beta_{i\mathbf{y}} \geq 0. \end{aligned} \tag{9}$$

The maximum of $\mathcal{L}_D$ can be found by differentiating with respect to both α and β, resulting in the identity $\lambda\Delta_M(\mathbf{y}_i,\mathbf{y}) = \mathbf{w}^T\delta\Psi_i(\mathbf{x}_i,\mathbf{y})$. It can be shown that by exploiting this equivalence, the SMO step update results as in line 8 of Alg. 2.

Algorithm 1. Cutting Plane Shape

Input: $\{(\mathbf{x}_i,\mathbf{y}_i)\}_n, C, \epsilon$
Output: $\mathbf{w}$
1: $S_i := 0, \forall i \in \mathbb{N}_n$
2: **repeat**
3: **for** $i := 1 \to n$ **do**
4: $H(\mathbf{y}) = \Delta(\mathbf{y}_i,\mathbf{y}) - \mathbf{w}^T\delta\Psi_i(\mathbf{x}_i,\mathbf{y})$
5: $\hat{\mathbf{y}}_i = \arg\max_{\mathbf{y}} H(\mathbf{y})$
6: $\xi_i = \max\{$
7: $\max_{\mathbf{y}\in S_i}(1-\lambda)\Delta_S(\mathbf{y}),$
8: $\max_{\mathbf{y}\in S_i} H(\mathbf{y})\}$
9: **if** $H(\hat{\mathbf{y}} > \xi_i + \epsilon$ **then**
10: $S_i \leftarrow S_i \cup \{\hat{\mathbf{y}}\}$
11: $\alpha \leftarrow$ opt. dual over $S = \bigcup_i S_i$
12: **end if**
13: **end for**
14: **until** no S_i changes during iteration

Algorithm 2. SMO Shape

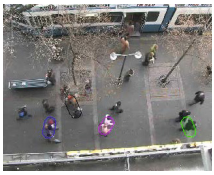
Input: $\{(\mathbf{x}_i,\mathbf{y}_i)\}_n, S, \alpha, C$
Output: α
1: $\mathbf{w} := \sum_{i,\mathbf{y}\neq\mathbf{y}_i}\alpha_{i\mathbf{y}}\delta\Psi_i(\mathbf{x}_i,\mathbf{y})$
2: **repeat**
3: **for all** $(\mathbf{x}_i,\hat{\mathbf{y}}) \in S$ **do**
4: **if** $(\mathbf{x}_i,\hat{\mathbf{y}})$ violates KKT **then**
5: $s := \frac{\lambda\Delta_M(\mathbf{y}_i,\hat{\mathbf{y}}) - \mathbf{w}^T\delta\Psi_i(\mathbf{x}_i,\hat{\mathbf{y}})}{\|\delta\Psi_i(\mathbf{x}_i,\hat{\mathbf{y}})\|^2}$
6: $s_{\text{clip}} \leftarrow \min\{s, \frac{C}{n} - \sum_{\mathbf{y}\neq\mathbf{y}_i}\alpha_{i\hat{\mathbf{y}}}\}$
7: $s_{\text{clip}} \leftarrow \max\{s_{\text{clip}}, -\alpha_{i\hat{\mathbf{y}}}\}$
8: $\alpha \leftarrow \alpha + s_{\text{clip}}$
9: $\mathbf{w} \leftarrow \mathbf{w} + s_{\text{clip}}\delta\Psi_i(\mathbf{x}_i,\mathbf{y})$
10: **end if**
11: **end for**
12: **until** no $\alpha_{i\hat{\mathbf{y}}}$ changes during iteration

Eq. 10 provides the KKT conditions of line 4 of Alg. 2 that have to be met in order for the algorithm to converge to an optimal solution.

$$\begin{aligned} \alpha_{i\mathbf{y}} = 0 \quad &\Rightarrow \quad \mathbf{w}^T\delta\Psi_i(\mathbf{x}_i,\mathbf{y}) \geq \lambda\Delta_M(\mathbf{y}_i,\mathbf{y}) \\ 0 < \textstyle\sum_{\mathbf{y}\neq\mathbf{y}_i}\alpha_{i\mathbf{y}} < \frac{C}{n} \quad &\Rightarrow \quad \mathbf{w}^T\delta\Psi_i(\mathbf{x}_i,\mathbf{y}) = \lambda\Delta_M(\mathbf{y}_i,\mathbf{y}) \\ \textstyle\sum_{\mathbf{y}\neq\mathbf{y}_i}\alpha_{i\mathbf{y}} = \frac{C}{n} \quad &\Rightarrow \quad \mathbf{w}^T\delta\Psi_i(\mathbf{x}_i,\mathbf{y}) \leq \lambda\Delta_M(\mathbf{y}_i,\mathbf{y}) \end{aligned} \tag{10}$$

4 Experimental Results

We tested our system on two publicly available datasets, namely the *BIWI Walking Pedestrians* dataset [6] and the *Crowds-By-Examples* (*CBE*) dataset [20]. The former 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 latter records a high density crowd video outside another university, `student003` (`stu003`). As it can be seen from Fig. 1, `stu003` dataset provides some real challenge as the



(a) *BIWI* `hotel`

(b) *BIWI* `eth`

(c) *CBE* `student003`

Fig. 5. Visual results on test videos[1]

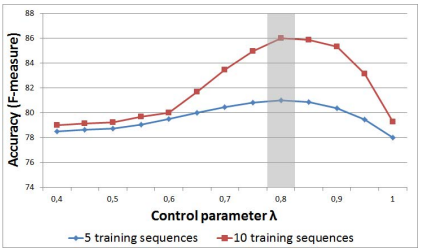


Fig. 6. Test results for different values of λ on `stu003`

Table 1. Performance comparison in terms of precision and recall computed according to the MITRE [18]

	our $\lambda = 0.8$		[6]		[7]	
	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{R}$
`hotel`	**93.2**	93.7	-	-	91.3	**95.9**
`eth`	**88.4**	**91.6**	-	-	83.0	80.2
`stu003`	**85.9**	**86.3**	46.0	82.0	80.5	77.0

density of the pedestrians is significant as well as for the presence of stairs and multiple entry and exit points. In the test setting we trained the classifier with one minute and two minutes of crowd videos; the trajectories were acquired on a time window of 10 seconds with no overlap. We evaluate the impact on performances of taking into account group shapes. The control parameter λ of Eq. 7 can be seen as a trade-off between how correct the solutions must be (according to the training data) and how valuable are to be considered the structured patterns within those solutions. In particular, Fig. 6 shows the accuracy trend obtained by varying λ in the `stu003` dataset, with a significant improvement at $\lambda = 0.8$. Table 1 highlights the performance gain is more relevant in this particular dataset as the density of the crowd doesn't allow the feature alone to easily separate groups, suggesting it is at this level that shapes become actually incisive. Moreover we compare our results with state-of-the-art methods [6][7] and the quantitative results shown in Tab. 1 indicate that our method outperforms all the approaches in most of the proposed videos. This is due to the use of a sociological feature, 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 the classifier solutions.

5 Conclusions

We proposed a method for detecting small groups of pedestrian in crowd by employing supervised structured learning and sociological theories. In particular,

[1] See more video examples at `http://imagelab.ing.unimore.it`.

we devised a method to encode common groups shapes as a-priori knowledge. Results prove the effectiveness of adopting a social perspective on the task as our method outperforms current state of the art work. This project was supported by Modena local police and Softech ICT center of Regione Emilia Romagna.

References

1. Moussaïd, M., Perozo, N., Garnier, S., Helbing, D.: The Walking Behaviour of Pedestrian Social Groups and Its Impact on Crowd Dynamics. In: PLoS ONE, vol. 5 (2010)
2. Pang, S.K., Li, J., Godsill, S.J.: Models and Algorithms for Detection and Tracking of Coordinated Groups. In: Aerospace Conference, pp. 1–17 (2008)
3. Feldmann, M., Fränken, D., Koch, W.: Tracking of Extended Objects and Group Targets Using Random Matrices. IEEE Trans. Signal Processing 59 (2011)
4. Lau, B., Arras, K., Burgard, W.: Multi-Model Hypothesis Group Tracking and Group Size Estimation. I. J. Social Robotics 2, 19–30 (2010)
5. Lin, W.C., Liu, Y.: A Lattice-Based MRF Model for Dynamic Near-Regular Texture Tracking. PAMI 29, 777–792 (2007)
6. Pellegrini, S., Ess, A., Van Gool, L.: Improving Data Association by Joint Modeling of Pedestrian Trajectories and Groupings. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 452–465. Springer, Heidelberg (2010)
7. Yamaguchi, K., Berg, A.C., Ortiz, L.E., Berg, T.L.: Who Are You with and Where Are You Going? In: CVPR, pp. 1345–1352 (2011)
8. Ge, W., Collins, R.T., Ruback, R.B.: Vision-Based Analysis of Small Groups in Pedestrian Crowds. PAMI 34, 1003–1016 (2012)
9. Rodriguez, M., Laptev, I., Sivic, J., Audibert, J.-Y.: Density-Aware Person Detection and Tracking in Crowds. In: ICCV, pp. 2423–2430 (2011)
10. Hall, E.T.: The Hidden Dimension. Doubleday (1966)
11. Finley, T., Joachims, T.: Supervised Clustering with Support Vector Machines. In: ICML, pp. 217–224 (2005)
12. Calderara, S., Cucchiara, R.: Understanding Dyadic Interactions Applying Proxemic Theory on Videosurveillance Trajectories. In: CVPRW, pp. 20–27 (2012)
13. Solera, F., Calderara, S., Cucchiara, R.: Structured learning for detection of social groups in crowd. In: Proc. of Advanced Video and Signal-Based Surveillance (2013)
14. Bansal, N., Blum, A., Chawla, S.: Correlation Clustering. Machine Learning 56, 89–113 (2004)
15. Tsochantaridis, I., Hofmann, T., Joachims, T., Altun, Y.: Support Vector Machine Learning for Interdependent and Structured Output Spaces. In: ICML (2004)
16. Bandini, S., Gorrini, A., Manenti, L., Vizzari, G.: Crowd and Pedestrian Dynamics: Empirical Investigation and Simulation. In: Proc. of the Measuring Behavior (2012)
17. Cardie, C., Wagstaff, K.: Noun Phrase Coreference As Clustering. In: Proc. of the Empirical Methods in NLP and Very Large Corpora (1999)
18. Vilain, M., Burger, J., Aberdeen, J., Connolly, D., Hirschman, L.: A Model-Theoretic Coreference Scoring Scheme. In: Conf. on Message Understanding (1995)
19. Lee, C., Jang, M.-G.: Fast Training of Structured SVM Using Fixed-Threshold Sequential Minimal Optimization. Etri Journal 31, 121–128 (2009)
20. Lerner, A., Chrysanthou, Y., Lischinski, D.: Crowds by Example. Computer Graphics Forum 26, 655–664 (2007)