

Social Signaling Descriptor for Group Behaviour Analysis

Eduardo M. Pereira^{1,2(✉)}, Lucian Ciobanu¹, and Jaime S. Cardoso^{1,2}

¹ INESC TEC, Porto, Portugal

² Faculty of Engineering of the University of Porto, Rua Dr. Roberto Frias, 378,
4200 - 465 Porto, Portugal
ejmp@inescporto.pt

Abstract. Group behaviour characterisation is a topic not so well studied in the video surveillance community due to its difficulty and large variety of topics involved, but mainly because the lack of valid semantic concepts that relate collective activity to social context. In this work, our proposal is three-fold: a new definition of semantic concepts for social group analysis considering environment context, a novel video surveillance dataset that conveys a sociological perspective, and a descriptor that emphasises social interactions cues within a group. Promising results were revealed in order to deal with such complex problem.

1 Introduction

Increasing research in video surveillance has been demanding the monitoring of complex human activities related to group of individuals. Such complexification has lead to higher levels of semantic abstraction that translate relational connections among people in groups. Collective structure varies depending on context, but common attention and position-based cues could be used as basis for further mid-level representations that encode relations involving social interaction between individuals within a group. Modelling collective activity within a sociological principled way has an undeniable value for both low-level problems such as pedestrian tracking, and high-level applications such as anomaly detection in security and human behaviour prediction for marketing purposes.

Assuming that the group discovery problem is solved, we focus on group behaviour characterisation. Our proposal build on trajectory data a multi-scale histogram descriptor that combines and accumulates relational position-based and attention-based features. The powerful and effectiveness of such representation was stated on our previous work [1] for the classification of individual profiles. In this work, we extend our study to a more complex problem, the analysis of collective behaviour of small groups in a very specific context, namely shopping-mall. Under the proposed descriptor we inspect the relevance of social signaling features that explore interaction among humans, a topic not so well studied in the video surveillance community.

We validate our approach on a novel video surveillance dataset entirely annotated taking in consideration social-psychological principles. We extend recent

evidence on group activity, which state that individual actions guide recognition of collective activities [2], to a higher semantic perception of group behaviour within social context. Therefore, this work presents the following contributions: (i) a new definition of semantic concepts for social group analysis considering environmental context; (ii) a complete social annotation of a very rich dataset for human activity analysis to detect and classify *individual profiles (I.P.)* and *group behaviours (G.B.)*, that will be publicly released for the research community; (iii) a descriptor that identify meaningful social interactions cues within a group, and aggregates them dynamically over time through a trajectory sampling scheme to robustly discriminate among several collective behaviours.

2 Related Work

Analysing the group structure and extracting its behaviour has important practical applications and has attracted the attention of the research community in surveillance settings. Under computer vision field such problem involves many research topics such as object detection, tracking, action discovery, human-to-human and human-to-object interactions recognition. Such tasks are complex and mutually dependent. Knowing how individuals are related to each other considering space structure and social context could give us the insight of how actions and reactions define social group behaviour [3].

In the literature, collective behaviour analysis tend to fit into two types of taxonomy: the one that considers groups as a collective and homogeneous block where individual is transformed by the group, the so-called macroscopic studies [4], and the one that analyse groups as the composition of individual agents that interact with each other and with the environment, the microscopic approaches [5]. For our specific scenario, microscopic studies are more suitable but their formulation is not enough to derive social semantic behaviour. Such approaches follow different models such as social force [5], virtual agents [6], and cellular automata [7]. In particular, Chang et al. [8] adopted a probabilistic grouping strategy which accounts with a pairwise spatiotemporal measure between people. A connectivity graph was built for further segmentation of groups and derivation of individual probabilistic models. However, no object-scene relation was considered, and they did not use relational context to describe individual behaviour. Floor fields models [7] effectively aid tracking in crowd scenes, but local attractive and repulsive forces have only physical meaning. Generalisation of discrete choice models (DCM) to obtain different group structures was presented in [9] through the inclusion of relational matrices, but they just presented simulations over synthetic data without inferring any type of semantic behaviour.

The work of [2] considered the composition of a crowd by small groups and incorporated a hierarchical clustering technique based on social psychological models. Their results were correlated with a ground truth collected by two sources, namely interviews and real-time observers. To the best of our knowledge this is the closest study that brings together computer vision and sociological fields. However, they not made publicly available the dataset, and they also

lack to assign semantic collective behaviour into social environment. In [10], it was demonstrated the importance of attention-based cues on video-surveillance scenario that normally were used in other domains such as meeting analysis. However, their approach accounts with many features and they did not evaluate the discriminative value and social meaning of each one. We got inspiration from both works and embed social analysis into a robust descriptor formulation.

3 Semantic Concepts and Annotation

Our aim is to add and explore semantics in group behaviours, a topic not so well explored on the literature. We look for a real-life surveillance dataset with large duration, intense activity, and high diversity of semantics in terms of individual and collective activity, in order to extend it for new human activity analysis challenges. We found the IIT (Israel Institute of Technology) dataset and grant permissions from the authors [11]. It is composed by several urban scenarios such as shopping, subway, and street. We chose the shopping-mall since its social context provides more well-defined behaviours. This scenario comprises three videos, but until this moment, due to the intensive manual labor involved, only one video is annotated (83155 frames with resolution 512×384 @25 fps).

We were advised by the lab of social-psychology of the University of Porto¹ during the annotation process. They help us to analyse and identify individual profiles and group behaviours. We follow the definition of group dynamics presented in [12] that explains the interdependence degree among individuals and their influence over the group behaviour they belong to. The complete validation of this work in the field of social-psychology would require an intense and continuous observation process of the same space. However, we validate the annotation process considering the sociological objective measure proposed in [13]. This effort represents a complete new methodology for social annotation of datasets in the field of computer vision.

The annotation was subdivided into two levels: (i) *low-level features*, related to human detection and tracking, trajectories are acquired from bounding box of enclosing person's annotation on each frame. Re-identification was not considered. When a person is partially or fully occluded, his bounding box was not marked. Also, a full-oriented gaze-direction $[0^\circ, 360^\circ]$ was annotated over the person's head; (ii) *high-level semantics*, labels related to I.P.s and G.B.s, where a trajectory and a group could reveal different profiles and behaviours, respectively. There are the following I.P.s: (i) *distracted (Dist.)*, (ii) *exploring (Exp.)*, (iii) *interested (Int.)*, (iv) *disoriented (Dis.)*; and the following G.B.s: (i) *equally interested (E.I.)*, (ii) *balance interests (B.I.)*, (iii) *unbalance interests (U.I.)*, (iv) *chatting (CHAT.)*. Objects of interest in the scene were marked, namely candy box, toy cars, and electric stairs. Table 1 summarises some relevant statistics about the annotation. Please refer to [1] for a more detailed explanation about I.P.s.

¹ Faculdade de Psicologia e de Ciências da Educação da Universidade do Porto - <http://sigarra.up.pt/fpceup>.

Table 1. Dataset statistics.

Frames Annotated	Annotation duration	Elapsed Time (I.P.)	Elapsed Time (G.B.)	I.P.s distribution	G.B.s distribution	Average Individuals per frame	Average Individuals per group
80894 (97.3)%	02:22:49 (hh:mm:ss)	203.5 (s) Dist	30.7 (s) E.I	869 total	255 total	3.5	1.8 (max: 9)
		35.3 (s) Exp	23 (s) B.I	45 Dist	193 E.I		
		12.8 (s) Int	100.3 (s) U.I	776 Exp	27 B.I		
		4.2 (s) Dis	83.7 (s) CHAT	41 Int	28 U.I		
				7 Dis	7 CHAT		

The I.P.s and G.B.s are defined considering the environment as social context. For instance, an individual is considered *distracted* if he is not aware of the environment in that moment. The G.B. concepts are described as:

- *Equally Interested (E.I.)*, when a group presents a coherent behaviour, i.e. one of the following conditions are satisfied: (i) individuals show interest for the same object, therefore all I.P.s should be *interested*; (ii) individuals explore the environment in a similar perspective and in a close position, therefore all I.P.s should be *exploring*, their gaze should be similar, and they should be close to each others.
- *Balance Interests (B.I.)*, when individuals within a group do not reveal the same level of interest but maintain the same behaviour, i.e. the following condition is verified: (i) individuals explore the environment in a similar perspective but not so close to each other, therefore all I.P.s should be *exploring*, their gaze should be relatively similar, and they can be a bit far away from each other.
- *Unbalance Interests (U.I.)*, when a group reveals different types of behaviour in the scene at the same time, i.e. the following condition is satisfied: (i) individuals show different individual profiles and the distance among them, as well as their gaze, can vary.
- *Chatting (CHAT.)*, when a group can be considered a free-standing conversational group (FCG), i.e. the following condition is verified: (i) individuals should be fixed in a position talking with each other (movable individuals while chatting are not considered). By default, all the I.P.s are considered as *distracted*.

4 Proposed Framework

Following our previous work [1], we extended the key-point sampling strategy with multi-scale histogram representation to group behaviour analysis. For this purpose, several attention-based and position-based features were explored in order to obtain discriminative power in terms of classification and meaningful value in terms of social context. The same Bag-of-Features (BoF) approach was considered but we investigated new forms of sampling, pooling and feature matching techniques. For the classification process, we kept the same non-parametric discriminative approach using SVM, while testing different settings.

4.1 Social-Based Descriptor

Our descriptor collects information from key-point trajectory sampling where different features are encoded into a multi-scale histogram controlled by R , the number of granularity levels where the number of bins are given by 2^R , and concatenated to form the final descriptor's histogram. The descriptor also considers different timescales to smooth the gathered information from trajectory's steps. We verified that temporal smoothing did not carried significant difference under our settings probably due to the low spatial complexity and noise associated to the annotated data.

We model group behaviour in terms of space layout, social environment and nonverbal behavioural interactions. Such social signaling constraints involve attention and position-based cues. Inspired by the feature-based study of [10], our aim is to simplify feature identification and collection while keeping global discriminative value. This process is translated by the number of features considered as well as the number of measurements required to acquire a complete feature. For instance, in [10] they identified 4 attention-based cues and 5 position-based cues, and all features measurements, except one, were collected over pair-wise individual relations. In our case, we just considered 5 features, and just 2 of them involve pair-wise measurements. Another difference is that in [10] for each feature they account with each single pair-wise relation per sampling step, while in our case we compute a single global contribution for each feature per sampling step. The proposed social descriptor is composed by the concatenation of the following features:

- *average velocity*, \tilde{v}_g , is the average velocity taken from all the individuals within a group.
- *average distance*, \tilde{p}_g , is the average distance between a pair of individuals, considering all the pair-wise relations within a group.
- *velocity variance*, $\text{Var}[v_g]$, is the variance of the velocity from all the individuals within a group.
- *looking at each other*, $laeo_g$, is a pair-wise relationship and expresses the minimum angle difference between the individual's gaze and the displacement vector between both individual's positions. For each individual, we just considered individuals which fall inside his field of view. This measurement is determined as the mean square error (MSE) of all the differences in order to augment discrepancies.
- *profiles*, P_p , it reflects the occurrence of I.P.s within a group. In this case, no global measure per sampling step is compute. All profiles contributions are considered individually.

4.2 Group Behaviour Classification

The descriptor is fixed-length to be embedded into a BoF classification approach. The codebook was build by running k-means over a subset of the annotated data. The obtained clusters form the vocabulary to be used on further training and classification processes.



Fig. 1. (a) Detected chessboard points for camera calibration; (b) Horizontal vanishing line (blue), ground plane’s projection area (green), ground points (red) to calculate scale factors and reprojection errors, and objects of interest (purple) (Color figure online).

We trained a multi-class classifier to identify the different G.B.s. Each sample is a sequential number of frames with I.P.s and G.B.s labels, individual trajectories and gaze orientations. Each sequence is treated as a bag. On each bag a temporal sampling, τ , is assumed to acquire feature information from key-point trajectory sampling and form a descriptor. The final descriptor vector for each behaviour is a histogram obtained by nearest cluster counting, which is used as input for the SVM classifier.

Under the classification framework we investigate two problems: (i) feature matching, related to the coding step whose importance relies on a correct cluster histogram matching between descriptor and obtained vocabulary, as well as a proper distance measure; (ii) pooling strategy, related to the way of how the encoded features are summarised to form the final descriptor representation, and its relevance pass through the discriminative power of the descriptor. For the former one, we normalize individual feature’s histograms and global descriptor histogram. After that we compute histogram matching independently, and combine distances on final descriptor by either the average or the maximum value. For the latter, we subdivide a bag into temporal gaps, Γ , and considered two pooling configurations, average and max, of such sequences for the entire bag.

We took advantage of our backward feature selection technique proposed in [1] to inspect feature discriminative importance on final descriptor and formulate conclusions about the social meaning of each one.

5 Experimental Results

The manual trajectories and gazes were projected onto the ground plane to correctly estimate distances and angles of interest. We follow the camera calibration and ground-plane projection described in [1]. Figure 1 illustrates some information from calibration, ground plane estimation and manual annotation.

To evaluate our descriptor performance, we compare the classification results with a baseline descriptor and a competitive descriptor, referred here as Chamveha, that builds over our descriptor formulation but uses the features

presented in [10]. The baseline is composed by the same features enumerated on Sect. 4.1, but instead of considering a multiscale histogram based on key-point trajectories, it simply considers the mean (μ) and standard deviation (σ) of each feature, except for the P_p feature. Under our experiments, $R = 3$ showed a good trade-off between accuracy and dimensionality length, which leads to a 116-dimensional feature vector for our descriptor.

For exhaustive classification evaluation, we adopted a 2-fold cross-validation repeated over 100 random iterations. In order to obtain fair results we kept classes proportions from the original dataset for each fold. The evaluation considers three standard parameters: accuracy (A), recall (R), and precision (P). We investigated classification performance over different kernels, namely linear, RBF and intersection, without optimisation of parameters. Apart from performance, we want to analyse feature matching method, distance function, and pooling strategy. We ran experiments over all possible combinations and compare results over an overall F-score of all classes. For sake of simplicity and lack of space we only report in this work the most significative to support our conclusions. In this way, we verified that the intersection kernel SVM and the histogram intersection measure are the best performing alternatives, which corroborates that the combination of both generate better visual codebooks under unsupervised learning [14]. Table 2 summarises the global measure that sustain our conclusions about feature matching and pooling strategy problems.

Table 2. F-score results (%) for the combination of both histogram matching techniques and both pooling configurations.

Matching	Average		Maximum	
Pooling	Avg.	Max.	Avg.	Max.
Baseline	45.1	44.8	44.6	44.3
Chamveha	45.5	45.3	31.8	32.5
Our	54.4	55.7	50.8	53.2

5.1 Feature Matching

For this problem we compare average histogram matching with max histogram matching computed from individual feature’s histogram on final descriptor. In this way, the distances from all cluster centers are stored and a decision is made taking one of both techniques. For evaluation we considered the F-score measure over both pooling strategies fixing each matching technique.

Overall evaluation shows supremacy performance of our descriptor and average histogram matching reveals better results. Inspecting individual classification of each G.B. we can take the following conclusions: (i) in general E.I. presents the highest results for all descriptors, which is expected since it has a large number of samples and it is a well-defined behaviour; (ii) the baseline and Chamveha descriptors reveal problems in recognising the CHAT behaviour ($R \simeq 5\%$), while

our descriptor attains a much higher performance ($R \simeq 75\%$); (iii) maximum matching largely affect our descriptor performance on B.I. behaviour, while average matching brings a minor decrease on U.I. behaviour. This makes sense since B.I. and U.I. are nearly related and are the most difficult ones to recognise and distinguish, therefore an average matching could incorporate contributions from all features; (iv) for the Chamveha descriptor the opposite from the previous conclusion holds. Since this descriptor aggregates more features, there could be redundant information that could confuse the classifier if an average matching is taken. Also CHAT behaviour manifest the same performance decrease, which corroborates our conclusion.

5.2 Pooling Strategy

The objective of pooling strategy is to achieve invariance over possible transformations, provide compact representations and achieve higher performance removing irrelevant information. Indeed pooling strategy could modify the BoF representation. In this way, we investigate if the temporal subdivision of bags and their mode of aggregation affect final performance.

Overall evaluation confirms our expectation. Under our settings, since we are using annotated data, background noise is reduced and features are acquired with high level of confidence. Therefore, we state negligible difference among both pooling techniques. However, we adopted the max pooling technique for further analysis, since it presents a slightly large difference among all descriptors.

5.3 Classification Results

Considering the metrics presented on Table 3, in special the importance of accuracy for classification tasks and the relevance of recall rate for surveillance systems, we highlight conclusions that state the value of our descriptor formulation as well as the pertinence of the selected features to translate social interactions.

Table 3. Classification results (%) for all G.B.s.

	E.I.			B.I.			U.I.			CHAT.			Avg.		
	P	R	A	P	R	A	P	R	A	P	R	A	P	R	A
Baseline	90.9	87.6	82.9	39.5	46.7	90.1	32.4	36.4	86.9	14.6	11.3	94.4	44.4	45.5	88.6
Chamveha	88.4	86.1	79.8	17.4	16.6	87.9	36.2	43.4	87.8	36.2	38.5	95.3	44.6	46.1	87.7
Our	87.0	88.1	79.9	23.9	22.9	88.6	34.6	25.1	88.3	81.5	90.4	98.8	56.8	56.6	88.9

At first glance, the Chamveha descriptor superimposes the remaining descriptors in U.I. behaviour, the most complex one. This reveals that some of its features improve its performance over our descriptor. However, it also sustains the importance of our descriptor sampling strategy as an effective representation over time. E.I. and CHAT behaviours are the most well-defined. It is expectable that for the E.I. behaviour the baseline performs at the same level of remaining

descriptors. Low performance on CHAT behaviour for the baseline descriptor is explained by the lack of information about the individual profiles, which is a feature that our descriptor includes. The high performance of baseline on B.I. probably is due to its simplicity, since the mean and standard deviation of each feature might encompass and describe such behaviour composed by individuals that share common behaviour but present few differences of space interests. Indeed, the high performance of the baseline in E.I. and B.I. behaviours proves that our descriptor covers a good selection of discriminative features to describe individual interactions within a group, and that a global measurement that account with single occurrences could be representative enough to identify a collective behaviour. The high performance of our descriptor on CHAT behaviour proves its versatility over a wide range of collective behaviours. We should emphasise the high contrast between the number of samples per G.B.

The feature importance inspection, illustrated in Fig. 2, clearly shows that all selected features contribute in a balance way for the discriminative power of our descriptor (refer to [1] for a more detailed explanation about this technique).

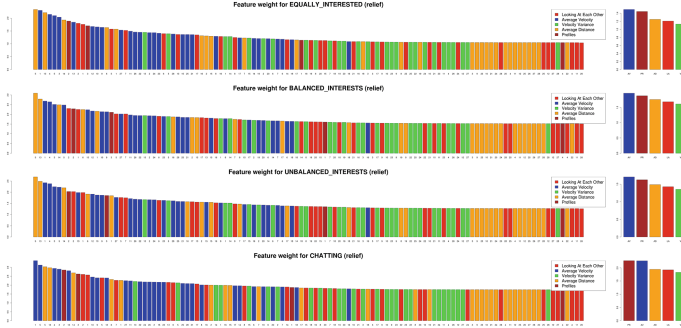


Fig. 2. Features' importance analysis.

6 Conclusions

In this work we addressed the characterisation of collective behaviour within a social context. For this purpose, we elaborated semantic concepts sustained on social-psychology principles and embedded them into the annotation of a novel video surveillance dataset for human activity recognition. Such process was advised by experts on sociological field.

We obtained promising recognition rates for such a complex problem, and validated the scalability of our trajectory-based descriptor to gather meaningful information over time. We also presented a preliminary approach towards the inspection of real sociological meaning of each feature. However, further work should be done over this direction. We are also embedding our framework into a more complete system to account with automatic trajectories and semi-supervised multi-label classification in order to understand better the limitations of our approach.

Acknowledgment. This work is financed by National Funds through the FCT - Fundação para a Ciência e Tecnologia (Portuguese Foundation for Science and Technology) within PhD grant reference SFRH/BD/51430/2011, and post-doctoral grant SFRH/BPD/85225/2012. The authors would like to thank Amit Adam for supplying the video sequences, Kelly Rodrigues and the Social Psychology Research Group of the University of Porto for their scientific advice.

References

1. Pereira, E.M., Ciobanu, L., Cardoso, J.S.: Context-based trajectory descriptor for human activity profiling. In: *Proceedings of IEEE International Conference System, Man, Cybernetics*, San Diego, CA, USA (2014)
2. Ge, W., Collins, R.T., Ruback, B.: Vision-based analysis of small groups in pedestrian crowds. *IEEE Trans. Pattern Anal. Mach. Intel.* **34**, 1003–1016 (2012)
3. Rummel, R.J.: *Understanding Conflict and War: The Conflict Helix*, vol. 2. Sage Publications, Beverly Hills (1976)
4. Zhou, B., Wang, X., Tang, X.: Understanding collective crowd behaviors: learning a mixture model of dynamic pedestrian-agents. In: *CVPR*, pp. 2871–2878 (2012)
5. Helbing, D., Molnár, P.: Social force model for pedestrian dynamics. *Phys. Rev. E* **51**, 4282–4286 (1995)
6. Klügl, F., Rindsfuser, G.: Large-scale agent-based pedestrian simulation. In: Georgeff, M., Klusch, M., Müller, J.P., Petta, P. (eds.) *MATES 2007. LNCS (LNAI)*, vol. 4687, pp. 145–156. Springer, Heidelberg (2007)
7. Ali, S., Shah, M.: Floor fields for tracking in high density crowd scenes. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) *ECCV 2008, Part II. LNCS*, vol. 5303, pp. 1–14. Springer, Heidelberg (2008)
8. Chang, M.C., Krahnstoeber, N., Ge, W.: Probabilistic group-level motion analysis and scenario recognition. In: *ICCV*, pp. 747–754 (2011)
9. Qiu, F., Hu, X.: Modeling group structures in pedestrian crowd simulation. *Simul. Model. Pract. Theor.* **18**, 190–205 (2010)
10. Chamveha, I., Sugano, Y., Sato, Y., Sugimoto, A.: Social group discovery from surveillance videos: a data-driven approach with attention-based cues. In: *Proceedings of the British Machine Vision Conference*, BMVA Press (2013)
11. Adam, A., Rivlin, E., Shimshoni, I., Reinitz, D.: Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE Trans. Pattern Anal. Mach. Intel.* **30**(3), 555–560 (2008)
12. Cartwright, D., Zander, A.: *Group Dynamics: Research and Theory*. Harper & Row, New York (1968)
13. McPhail, C., Wohlstein, R.T.: Using film to analyze pedestrian behavior. *Sociol. Methods Res.* **10**, 347–375 (1982)
14. Wu, J., Rehg, J.M.: Beyond the euclidean distance: Creating effective visual code-books using the histogram intersection kernel (2009)

Pattern Recognition and Image Analysis

7th Iberian Conference, IbPRIA 2015, Santiago de
Compostela, Spain, June 17-19, 2015, Proceedings

Paredes, R.; Cardoso, J.S.; Pardo, X.M. (Eds.)

2015, XVI, 753 p. 275 illus., Softcover

ISBN: 978-3-319-19389-2