Neural Comput & Applic
DOI 10.1007/s00521-016-2282-z

IBPRIA 2015

Cross-layer classification framework for automatic social behavioural analysis in surveillance scenario

Eduardo M. Pereira[1] · Lucian Ciobanu[1] · Jaime S. Cardoso[1]

Received: 1 October 2015 / Accepted: 23 March 2016

Abstract The increasing demand for human activity analysis in surveillance scenarios has been triggered by the emergence of new features and concepts to help in identifying activities of interest. However, the characterisation of individual and group behaviours is a topic not so well studied in the video surveillance community due to not only its intrinsic difficulty and large variety of topics involved, but also because of the lack of valid semantic concepts that relate human activity to social context. In this paper, we address the topic of social semantic meaning in a well-defined surveillance scenario, namely shopping mall, and propose new definitions of individual and group behaviour that consider environment context, a relational descriptor that emphasises position and attention-based characteristics, and a new classification approach based on mini-batches. We also present a wide evaluation process that analyses the sociological meaning of the individual features and outlines the performance impact of automatic features extraction processes into our classification framework. We verify the discriminative value of the selected features, state the descriptor performance and robustness over different stress conditions, confirm the advantage of the proposed mini-batch classification approach which obtains promising results, and outline future research lines to improve our novel social behavioural analysis framework.

Keywords Individual profiles (IP) · Group behaviours (GB) · Relational descriptor · Group dynamics · Social psychology

1 Introduction

The increasing research in video surveillance has been demanding the monitoring of complex individual and collective human activities that express the sociological context of a scene, a topic not extensively studied in the literature. In fact, automatic behaviour understanding from video is a very complicated problem. It comprises several hierarchical layers of processing, from low-level features to high-level semantics interpretation. The reduction in this gap is still a challenge in many applications. Mid-level descriptors are often used to bridge this gap, since they intend to robustly represent spatiotemporal relationships between features and objects, including people, to discriminatively detect actions and events, and form atomic elements of a complex activity.

Spatiotemporal trajectory representations are gaining increased attention in surveillance scenarios to analyse human activity and detect abnormal events. However, the research community has been focusing on solving technical problems associated with multitracking techniques and on encoding trajectory-based features to detect individual atomic actions. Some approaches combine scene and object features with trajectory-based descriptors to detect event primitives [26], while others aggregate interactions measures and cues to analyse small groups of pedestrians [10]. None of them explore the integration of scene objects with individual related features to classify individual and collective behaviour. Modelling human activity within a sociologically principled way has an undeniable value for

✉ Eduardo M. Pereira
ejmp@inesctec.pt

[1] INESC TEC, Campus da FEUP, Rua Dr. Roberto Frias, 4200 - 465 Porto, Portugal

Published online: 09 April 2016

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.

Our aim is to address the problem of automatic behaviour understanding in surveillance scenarios. In this way, we have proposed higher levels of semantic concepts that translate relational connections among people in groups considering their characteristics within the scene and environmental context. We have stated that those concepts could be adapted to known state-of-the-art semantic annotations [23]. A video surveillance data set was extended with low-level detection and tracking information and with the proposed high-level concepts, which conveys a novel sociological perspective in the video surveillance community. We have also proposed a descriptor that considers relational information between an individual and the scene, objects of interest, and among individuals themselves. It represents sociological cues by contextual position and attention-based features, which are temporally sampled over a key-point trajectory scheme of multiple scales and are concatenated in histograms. The strength and effectiveness of such a representation were tested with manually annotated information in our previous work [23, 24], proving its discriminative power and its capability to describe higher abstraction terms for action context.

In this paper, we extended our previous work in several directions. We integrate an automatic procedure to extract the relational features, we evaluate their local and global impact on the classification process, and we inspect their individual relevance for social analysis. We also formulate the classification process in terms of small mini-batches through the trajectory, instead of the whole trajectory, to determine the discriminative power of the descriptor in a short spatiotemporal span and to simultaneously detect the switches between continuous behaviours, namely individual profiles (IPs) and group behaviours (GBs), while classifying the detected segments.

The paper's outline is as follows: In Sect. 2, we survey the related work. Next, in Sect. 3, we present the relevant theoretical concepts behind the proposed semantic concepts and the annotation process. A description about the main framework steps is presented next, in Sect. 4. The experimental setup and results are reported in Sect. 5. Finally, we formulate the conclusions and future work in Sect. 6.

2 Related work

The automatic recognition of social interactions in video is usually achieved by a system that extracts low-level information, followed by a classification stage. In computer vision, this 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 provide insight into how actions and reactions define social behaviour in surveillance scenarios [28].

Regarding the feature extraction phase, trajectory-based dynamics provide intrinsic features that can be used to build useful representations to analyse several application-driven interests such as scene topology, event detection, social interpretation, and activity classification.

A common practice is to use trajectory information to model a scene by a topographical map composed of nodes, which are the areas of interest, and edges, which represent the connectivity between those areas and encode the activity of a human. Makris and Ellis [19] classify the areas of interest as entry/exit zones, junctions, intersections, and stop areas, which are defined by trajectory characteristics. Pusiol et al. [26] slow trajectory points define individual topologies which are combined to form the general topology. They segment the trajectories by topology affinity and then use that information to build a descriptor composed of primitive events. They reported that the statistical and geometrical structures inferred from the scene model could be used in a feedback loop, in order to filter out false detections or enrich tracking approaches that incorporate scene context.

Trajectory information helps detecting typical and unusual events. Owens and Hunter [21] applied a self-organising feature map neural network, with trajectories encoded as point-based flow vectors, to learn normal trajectories and detect new event-related trajectories. However, such an approach cannot distinguish between new normal paths and abnormal behaviours. Khalid and Naftel [17] solved that problem by using the Fourier coefficient space instead of the trajectory space. Tests were carried out on simple synthetic and manually annotated data since a global Fourier approximation is not appropriate for complex trajectories. Therefore, such an approach would not perform well on real scenarios.

The sequence of trajectories' characteristics is normally used to extract motion patterns that segment the scene into semantic regions. Pereira et.al [22] proposed a motion-based system that represents the spatial and temporal features of the flow in terms of long-range trajectories, which can be used to segment different motion patterns. This approach effectively captures instantaneous changes and long-range motions, but it needs a large temporal interval to integrate and advect the motion. Wang et al. [33] introduced a clustering algorithm that takes into account similarity and confidence measures between trajectories to obtain clusters of different activities. This type of approach largely depends on the distance measures, which also vary depending on the activities being detected. To overcome those limitations, some methods formulate the problem in a probabilistic way. Wang et al. [32] proposed a nonparametric Bayesian framework. The number of clusters for both the observations of an object on a trajectory and the trajectories is simultaneously learned from a dual hierarchical Dirichlet process (HDP). However, since trajectories are modelled as words to be quantised into a codebook, such a representation lacks temporal information.

An extension of activity analysis embeds notions of social psychology, normally applied to discover and characterise groups of people. In the literature, collective behaviour analysis tends to fit into two types of taxonomy: one that considers groups as a collective and homogeneous block where the individual is transformed by the group, the so-called *macroscopic* studies [37], and the other that analyses groups as the composition of individual agents that interact with each other and with the environment, the *microscopic* approaches [12]. The latter type of approach considers cues such as relational connections among people, the focus of attention of each person, geometric scene constraints, and proxemics-based distances.

For our specific scenario, *microscopic* studies are more suitable but their formulation is not sufficient to derive social semantic behaviour. In fact, they try to simulate pedestrian physical behaviour and infer characteristics about group formation, dispersion, and evolution, but they do not capture individual semantics. Such approaches follow different models such as social force [12], virtual agents [18], and cellular automata [3]. In particular, Chang et al. [8] adopted a probabilistic grouping strategy which uses a pairwise spatiotemporal measure among people. A connectivity graph was built for further segmentation of groups and derivation of individual probabilistic models. Each model considers motion type, related to an atomic action, direction distribution and distance change, related to interactions. However, no object–scene relation was considered, and they did not use relational context to describe individual behaviour. Floor fields models [3] effectively aid tracking in crowd scenes, but local attractive and repulsive forces have only physical meaning. A generalisation of discrete choice models (DCM) to obtain different group structures was presented by Qiu and Hu [27] through the inclusion of relational matrices, but they only presented simulations over synthetic data without inferring any type of semantic behaviour.

Ge et al. [10] 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 ground truth collected from 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 did not make the data set publicly available, and they also did not assign semantic collective behaviour into the social environment. Chamveha et al. [13] demonstrated the importance of attention-based cues on video surveillance scenario, normally used in other domains such as meeting analysis. However, their approach took into consideration 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.

This paper brings several contributions over the state of the art: (1) assignment of both IPs and GBs in social environments, and analysis of their mutual dependence; (2) in-depth validation of the proposed social concepts and discussion of the discriminative value and social meaning of each selected feature; (3) evaluation of the descriptor performance and robustness over different stress conditions; (4) inspection of the performance impact of automatic features extraction processes into our classification framework; and (5) a new classification approach based on mini-batches.

3 Semantic concepts and annotation

3.1 Semantic concepts

The annotation of human non-verbal behaviour should reveal meaningful representations semantically associated with ontological concepts for human activity. The diversity of theories that intend to explain the link between psychophysiological states and human behaviour has been triggering different representation approaches in the literature that take into consideration temporal processes for actions, spatiotemporal relationships between entities, poses, and gestures, among others.

A general view about human activity analysis was presented by Aggarwal and Ryoo [2]. The authors defined a hierarchical approach where semantic levels were related to an increased complexity of human activity categorisation: (1) gestures, elementary movements of body parts such as *raising an arm*; (2) actions, atomic activity composed of temporal sequences of gestures such as *jumping*; (3) interactions, a sequence of single activities between people such as *a person hugging another*; (4) group activities, single or complex activities performed by a conceptual group such as *a group having dinner*. Such levels follow an analogy to grammar-based semantics that can be used to map annotation labels to relational inference models, for instance an action is associated with a verb, a gesture to a phrase where the entity is the body part, etc.

Some works have already defined semantic concepts: the division between the part of actions as objects and the poselets closely related to those actions [36], Allen's temporal predicates were applied to features and entities to model activities with complex structure [29], the definition of topological and directional relations between persons to build context-free grammars [14] and collective context descriptors [9].

Our aim is to add and explore semantics in IPs and GBs, a topic which is under-explored in the literature. In terms of IPs, we follow a grammar-based analogy and present an abstraction layer that can be associated with adjectives, since we are qualifying person's behaviour characteristics. In terms of GBs, we adopt the definition of group dynamics presented by Cartwright and Zander [6] that explains the interdependence degree among individuals and their influence over the group behaviour they belong to.

All the individual and collective behaviours were characterised considering the environment as social context. The following IP concepts were defined:

- *exploring (Exp.)*, when no specific interest is revealed, but movement and gaze are coherent with the scene structure and context;
- *interested (Int.)*, when an interest by an object in the scene is explicitly revealed;
- *distracted (Dist.)*, when no specific interest is revealed which translates into unstructured movement and variability of gaze;
- *disoriented (Dis.)*, when confusion concerning interests is revealed, expressed as high variability of movement and gaze along with an unstructured movement.

In terms of GB concepts, the following were identified:

- *equally interested (EI)*, when a group presents a coherent behaviour, i.e. one of the following conditions is satisfied: (1) individuals show interest for the same object; therefore, all IPs should be *interested*; (2) individuals explore the environment in a similar perspective and in a close position; therefore, all IPs should be *exploring*, their gaze should be similar, and they should be close to each other.
- *balanced interests (BI)*, 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: (1) individuals explore the environment in a similar perspective but not so close to each other; therefore, all IPs should be *exploring*, their gaze should be relatively similar, and they can be slightly separated from each other.
- *unbalanced interests (UI)*, when a group reveals different types of behaviour in the scene at the same time, i.e. the following condition is satisfied: (1) individuals show different individual profiles and the distance among them, and their gaze can vary.
- *chatting (CHAT.)*, when a group can be considered a free-standing conversational group (FCG), i.e. the following condition is satisfied: (1) individuals should be fixed in a position talking with each other (moving individuals while chatting are not considered). By default, all the IPs are considered as *distracted*.

3.2 Data set and annotation

We selected the IIT (Israel Institute of Technology) data set and were granted access by the authors [1]. The data set is composed of several real-life surveillance scenarios such as shopping, the subway, and the street. We chose the shopping mall since its context provides well-defined social behaviours. This scenario comprises three videos, but, at the present time, and due to the intensive manual labour involved, only one video has been annotated (83,155 frames with resolution 512 × 384 @25 fps). The data set, including our annotation, is available upon request.

We were advised by staff of the laboratory of social psychology of the University of Porto[1] during the annotation process. They helped us to analyse and identify the IPs and the GBs. We validated the annotation process considering the sociological objective measure proposed by McPhail and Wohlstein [20], but a complete validation in the field of social psychology would require an intense and continuous observation process of the space. This effort represents a completely new methodology for social annotation of data sets in the field of computer vision.

Table 1 summarises some relevant statistics about the annotation, which was subdivided into two levels: (1) *low-level features*, related to human detection and tracking, trajectories were acquired from a bounding box enclosing an annotated person on each frame. Re-identification was not considered. When a person was strongly occluded (approximately more than half of the body), his/her bounding box was not marked. Also, a full-oriented gaze direction $[0^\circ, 360^\circ]$ was annotated over the person's head. Objects of interest in the scene were marked, namely candy box, toy cars, and electric stairs (see Fig. 1b); (2) *high-level semantics*, related to IPs and GBs labels, where a trajectory and a group of trajectories reveal different profiles and behaviours, respectively. Group formation and dispersion were also marked.

Since we are dealing with position and attention-based features, the trajectories should be projected onto the ground plane to correctly estimate distances and angles of interest. Such a transformation involves camera calibration

[1] Faculdade de Psicologia e de Ciências da Educação da Universidade do Porto—http://sigarra.up.pt/fpceup.

Table 1 Data set statistics

Frames annotated	Annotation duration	Elapsed time (IP)	Elapsed time (GB)	IPs distribution	GBs distribution	Average individuals per frame	Average individuals per group
80,894	02:22:49	203.5 (s) Dist.	30.7 (s) EI	869 Total	255 total	3.5	1.8
(97.3) %	(hh:mm:ss)	35.3 (s) Exp.	23 (s) BI	45 Dist.	193 EI		(max: 9)
		12.8 (s) Int.	100.3 (s) UI	776 Exp.	27 BI		
		4.2 (s) Dis.	83.7 (s) CHAT.	41 Int.	28 UI		
				7 Dis.	7 CHAT.		

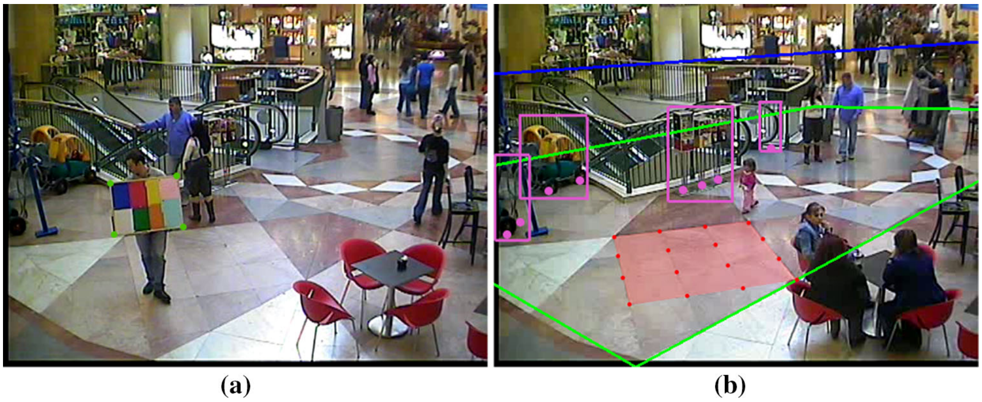


Fig. 1 **a** Detected chessboard points for camera calibration; **b** *horizontal vanishing line* (*blue*), ground plane's projection area (*green*), ground points (*red*) to calculate scale factors and re-projection errors, and objects of interest (*purple*) (colour figure online)

and geometry reconstruction steps (see Fig. 1a, b for camera calibration and ground plane projection, respectively). For more details, refer to Pereira et al. [23].

4 Proposed framework

In our previous work [23, 24], we considered the problem of temporally segmenting the behaviour was solved by using the manually annotated data and focusing on the classification task. However, in practice, individual and group behaviours can evolve in time, leaving us with the challenge of simultaneously detecting the transitions between different behaviours and classifying the detected segments. We address this issue by working with mini-batches, short enough so that the assumption of constant behaviour holds and long enough to enclose sufficient discriminative information. Indeed, we extend the classification process to work with mini-batches through the trajectory, instead of the whole trajectory, to evaluate the sampling and encoding strength of the proposed descriptor for the detection of IPs and GBs. A deeper analysis of the position and attention-based features in terms of their sociological meaning is conducted, and technical issues related to the bag-of-features (BoF) approach such as sampling, pooling, and feature matching techniques are examined. Finally, automatic procedures for the extraction of low-level information, such as tracking and gaze estimation, are included in the framework and their impact is measured.

4.1 General view

Our framework extracts temporally different features through the video. As prior scene knowledge, we mark the position of the objects of interest, previously identified in the annotation process. In each frame, pedestrians are detected and tracked, their gazes estimated, their distances to the closest object of interest computed, and the angle between their gaze and the direction of movement of their neighbours determined. This information is then encoded by our descriptor. In this work, we automate all the features extraction process and, despite not being interested on the automatic detection of transitions between behaviours, we inspect the capacity of the mini-batch approach to help in the detection of transitions among different behaviours. We should highlight that for the analysis of GBs, we considered as solving the problem of group formation and dispersion; therefore, we used the starting and the ending frame of each group from the annotated data, and we know which individuals belong to each group. A preliminary measure that represents the detection of the correct switch between consecutive behaviours is evaluated at extreme (start and end) mini-batches of the corresponding behaviours.

4.2 Automatic features extraction

Computer vision techniques for extracting low-level features from videos are usually exploited in a bottom-up way and are used for further high-level inference. In our case, we exploit two basic components: (1) tracking, to provide the trajectory of each individual; (2) head pose estimation, to obtain an approximation of the individual gaze.

4.2.1 Pedestrian tracking

For automatic tracking of pedestrians, a feasibility study was conducted at first, aiming to identify requirements and potential limitations. Based on recent surveys [30, 35], several promising state-of-the-art algorithms were considered, such as multiple-instance learning (MIL) [4], boosting [11], MedianFlow [16] and Track Learn Detect (TLD) [15]. Among these, only two were actually evaluated (boosting and MedianFlow), given the technical issues of the available implementation of MIL (memory leaking) and the unsuitability of TLD for video surveillance scenarios. Boosting has the advantage of online training, and its trade-off between performance accuracy and computational time is controlled by the features used on the appearance model. For its turn, MedianFlow bases its contribution on the penalisation of inconsistent trajectories taken from forward–backward error propagation.

These tracking algorithms were integrated into our framework, and some suitable metrics were adopted for performance comparison, Multiple Object Tracking Precision (MOTP) and Multiple Object Tracking Accuracy (MOTA), from [5].

4.2.2 Gaze estimation

For automatic gaze estimation, the method presented by Chamveha et al. [7] was adopted. It relies on unsupervised learning of head orientations preceded by several pre-processing tasks: (1) tracking with the head detector from [25]); (2) walking direction estimation (polyline simplification using the Douglas–Peucker algorithm and line fitting); (3) outlier segment rejection rules; (4) selection of representative images through Mahalanobis distance; (5) oversampling for handling imbalanced data. For the tracker, we selected the boosting tracking algorithm, and for the head detection we integrated the fastHOG library [25] into our framework. The fine-tuned parameters of the head detection technique are summarised in Table 2, where ε is the threshold for the Douglas–Peucker algorithm, and $\tau_n, \tau_l, \tau_{\text{var}}$ are the thresholds for rules no. 2, 3, 4, respectively, for outlier segment rejection.[2]

Table 2 Recalibrated parameters of method [7] for our scenario

ε	τ_n	τ_l	τ_{var}
20	0.5	10	2750

Figure 2 presents the eight head orientations (classes) considered for gaze estimation and an example for each class. More representative images, which result from the preprocessing steps and further used as training data, are illustrated in Fig. 3. Note that the head images are converted to grey scale, normalised, and resized to 20 × 22 pixels.

4.3 Relational descriptor

We model human behaviour in terms of space layout, social environment, and non-verbal behavioural interactions. Such social signalling constraints involve attention and position-based cues, which were extracted spatially at each key-point trajectory and temporally at each frame.

In terms of IPs, the following features were taken:

- *Angular direction change*, α_{si}, is the angular variation of movement of the individual between consecutive sampling times (in our case, consecutive frames).
- *Distance of interest*, d_{io}, expresses the distance between individual position and the object of interest.
- *Direction of interest*, β_{gi}, which is the gaze direction.
- *Velocity*, v_i, expresses the instantaneous velocity of the individual at the sampling time.

In terms of GBs, our selection was inspired by the feature-based study of Chamveha et al. [13]. Our aim is to simplify feature identification and collection while keeping global discriminative value. This process is expressed by the number of features as well as the number of measurements required to acquire a complete feature. For instance, Chamveha et al. [13] identified four attention-based and five position-based features, and all their measurements, except two, were collected over pairwise individual relations. In our case, we only considered five features, and only two of them involve pairwise measurements. Another difference is that in [13] for each feature they consider each single pairwise relation per sampling step, while in our case we compute a single global contribution for each feature per sampling step.

Considering the GBs, the following features were taken at each sampling time:

[2] We thank to the first author of [7], Isarun Chamveha from the Institute of Industrial Science, The University of Tokyo, for helping us to recalibrate the technique for our data set.

- *Average velocity*, $\tilde{v}_g$, is the average of the instantaneous velocities of all individuals within a group.
- *Average distance*, $\tilde{d}_g$, is the average distance between a pair of individuals, considering all the pairwise relations within a group.
- *Velocity variance*, $\mathrm{var}[v_g]$, is the variance of the instantaneous velocities of all individuals within a group.

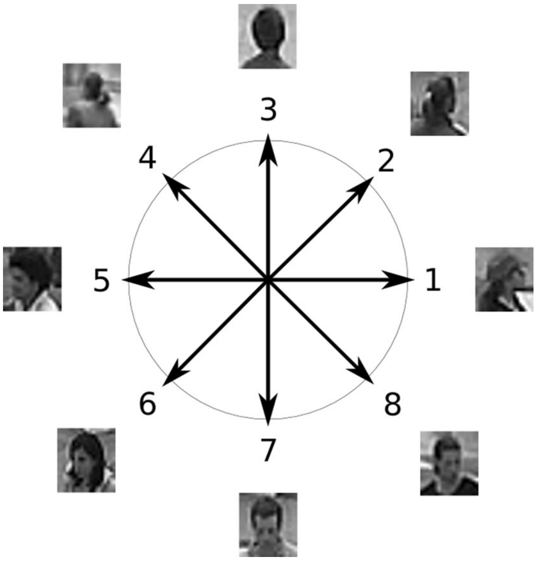


Fig. 2 Head poses divided into discrete classes for gaze estimation

- *Looking at each other*, laeo_g, is a pairwise 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.
- *Profiles*, P_p, reflects the occurrence of IPs within a group. In this case, no global measure per sampling step is computed. All profiles contributions are considered individually.

Our descriptor is inspired by Takahashi et al. [31]. The features extracted during a pre-defined number of frames are collected and encoded into our fixed-length descriptor to be used in a bag-of-features (BoF) approach. The key points along each trajectory are given by

$$\begin{aligned} P_u &= \left[p_u^x,\ p_u^y\right] \\ p_u^x &= \left[p_u^{x,t_1},\ p_u^{x,t_1+1},\ \ldots,\ p_u^{x,t_2}\right], \\ p_u^y &= \left[p_u^{y,t_1},\ p_u^{y,t_1+1},\ \ldots,\ p_u^{y,t_2}\right] \end{aligned} \tag{1}$$

where P_u is a set of points in the $\mathcal{T}_u$ trajectory, p_u^x is the set of its x-coordinates, p_u^y is the set of its y-coordinates, and t_1 and t_2 are the starting and ending frames, respectively.

For each feature, the values collected during a mini-batch are encoded into a multi-scale histogram controlled

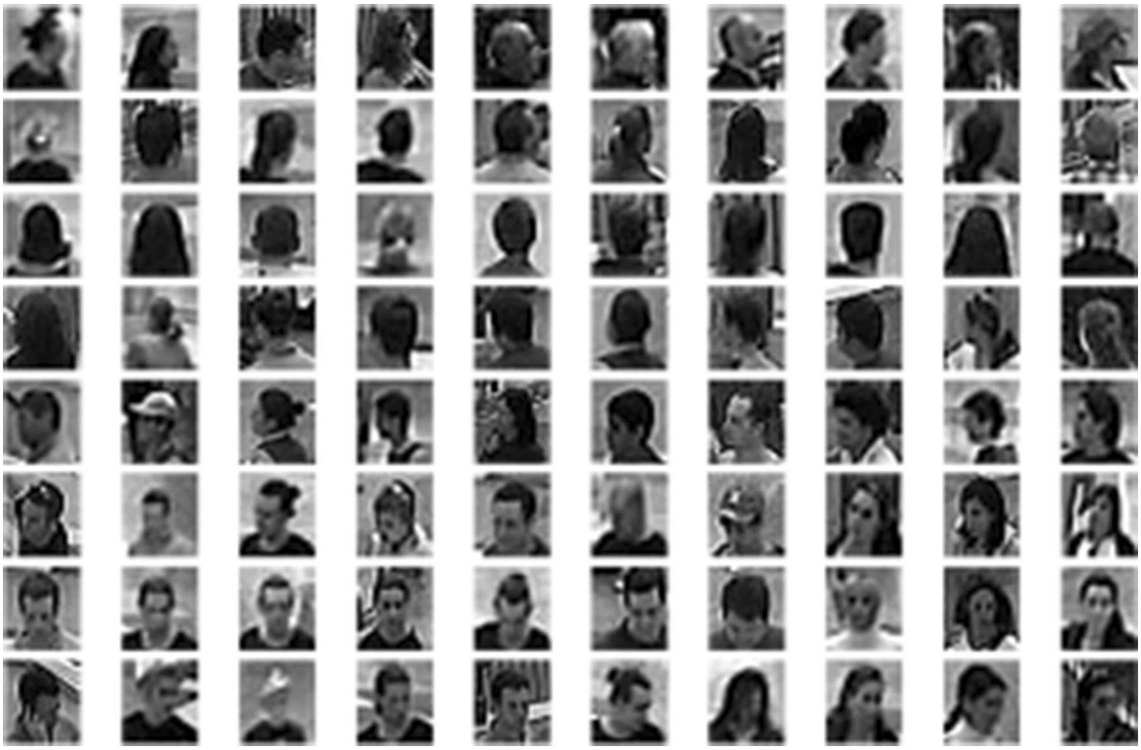


Fig. 3 Representative images generated by the method [7]: a row per class in the exact order of the classes from Fig. 2

by $R \in \mathbb{N}$, the number of granularity levels. Considering the feature f^0 extracted along the trajectory $\mathcal{T}_u$, its multi-scale representation of size R is given by the vector $\mathbf{f_u}^0 = \left[H_u^1, H_u^2, \ldots, H_u^R\right]$, where each entry, H_u^r, is a normalised histogram of 2^{r+1} bins, for each $r = [1, 2, \ldots, R]$. The final descriptor representation for trajectory $\mathcal{T}_u$ is the concatenation of all the multi-scale feature histograms and is given by a fixed-length vector

$$\mathbf{F_u} = \left[\left(\mathbf{f_u}^0\right), \left(\mathbf{f_u}^1\right), \ldots, \left(\mathbf{f_u}^{N-1}\right)\right] \tag{2}$$

4.4 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, and the obtained centres form the vocabulary to be used on further training and classification processes. We trained a multi-class classifier to identify the different IPs and GBs. The sampling follows a key-point trajectory strategy, where each descriptor is extracted (as explained in Sect. 4.3) over a temporal length, τ, expressed in seconds. Each bag is composed by consecutive descriptors and its length is controlled by Γ. In the case of GBs, individual trajectories and gaze orientations within each group are time aligned. The length of the mini-batch, M_i, is given by the number of bags that it might contain (Fig. 4).

The final fixed-length BoF descriptor used for classification can be build at different levels. In this work, we investigate two levels: (1) *coarse mini-batch*, where the whole individual trajectory (IP) or the entire group set (GB) information is used as sample, the fixed-length BoF descriptor is computed from all the bags and the final classification is taken from it; (2) *fine mini-batch*, where the bags are used as individual samples, for each one a fixed-length BoF descriptor is build and a classification is undertaken at the mini-batch level, and the final classification is inferred from a combination of the predicted labels at each mini-batch, either for IP or for GB. On both approaches, the final descriptor vector for each sample is a histogram obtained by nearest cluster counting, which is used as input for an SVM classifier. We adopted a k-fold cross-validation process, maintaining class proportions, to obtain the final classification results for each behaviour. However, in some cases, due to the high class imbalance, we report the classification results under a stratified k-fold cross-validation setting.

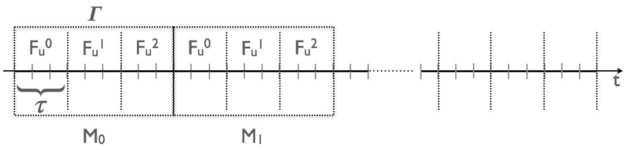


Fig. 4 Key-point trajectory encoding scheme considering descriptor length and bag length

We also investigate two components under the classification framework: (1) feature matching, which is related to the coding step and whose importance relies on a correct cluster histogram matching between the descriptor and the obtained vocabulary; under this component, we also studied the impact of the distance measure; (2) pooling strategy, which is related to the way the encoded features are summarised to form the final descriptor representation and whose combination defines the discriminative power of the descriptor. For the former one, we normalise the individual feature's histograms and the global descriptor histogram. After that, we compute each histogram matching independently and combine the distances on the final descriptor by either the average or the max value. For the latter, we change the temporal length of the bag, Γ, and consider two pooling configurations, average and max, for all the descriptors within each bag.

To inspect feature importance on the final descriptor, we use the relief-F method. However, since the descriptor obtained from the BoF approach is a histogram, it cannot be applied directly because each bin represents a word, which is a combination of several features. In this way, we did a backward procedure starting from the discrete parts of the descriptor (clusters), until the individual feature bins: 1) *cluster ranking*, C_{r_i}, to each cluster was applied the relief-F technique and an importance ranking was obtained; 2) *feature bin ranking*, F_{r_j}, on each cluster the previous step was applied again, resulting in a ranking of bins. Each bin corresponds to an individual feature, described in Sect. 4.3. The final individual feature importance was obtained by

$$F_k = \sum_{i=0}^{C} \sum_{\substack{j=0,\\ l_j=k}}^{B} C_{r_i} \cdot F_{r_j}, \tag{3}$$

where C is the number of clusters, B is the number of bins on each cluster, and the condition $l_j = k$ permits to take into consideration the feature bins that correspond to feature's label k. Inspection of this feature selection process is useful to formulate conclusions about the social meaning of each feature.

5 Results

Under this section, we present results for the entire classification framework. Several steps are evaluated such as feature importance, matching and pooling strategies, a comparison of the mini-batch approach at two levels is conducted, the impact of the automatic feature extraction process over the final results is measured, and the accuracy of the descriptor is estimated considering tracking loss and feature variation. The evaluation is conducted in the novel data set detailed in Sect. 3.2, and the classification results consider three standard parameters: accuracy (A), recall (R), and precision (P); and sometimes, when relevant, the F_1-score (F) is also presented.

5.1 Automatic pedestrian tracking

Considering the selected tracking algorithms in Sect. 4.2.1, namely boosting and MedianFlow, we obtained the results for the measures MOTP and MOTA that clearly show the best performance of the boosting approach (see Table 3). The analysis of Fig. 5 corroborates the previous conclusion, since the density of the trajectories extracted from boosting algorithm is closer to the density of the manual trajectories than the trajectories extracted from the MedianFlow, which exhibit problems related to the temporal continuity of the tracking process. However, we also verified that the boosting algorithm shows some errors derived from tracking loss.

Table 3 Tracking performance (%) of boosting and MedianFlow given by the MOTP and MOTA metrics

	Boosting	MedianFlow
MOTP	12.2	16.8
MOTA	71.3	53.6

For MOTP, the lower is the better, while in MOTA, the higher is the better

Figure 6 illustrates the comparison of both algorithms results with the manual annotation for individual trajectories, where we can clearly see the tracking failures. In fact, despite the recent advances, the state-of-the-art algorithms still underperform in cluttered environments that present many occlusions. Additionally, the trackers in consideration are single track; therefore, they do not take into account the multi-person coexistence and interactions in order to jointly improve the final tracking. We should highlight that no pre-processing technique such as filtering or background subtraction, and post-processing technique like non-maximum suppression or scene knowledge were used to improve the performance of the tracking algorithms. In terms of computational effort, this step is dependent of the tracker algorithm's performance and is conducted for each pedestrian individually. Its execution is slower than real time, around 10 fps.

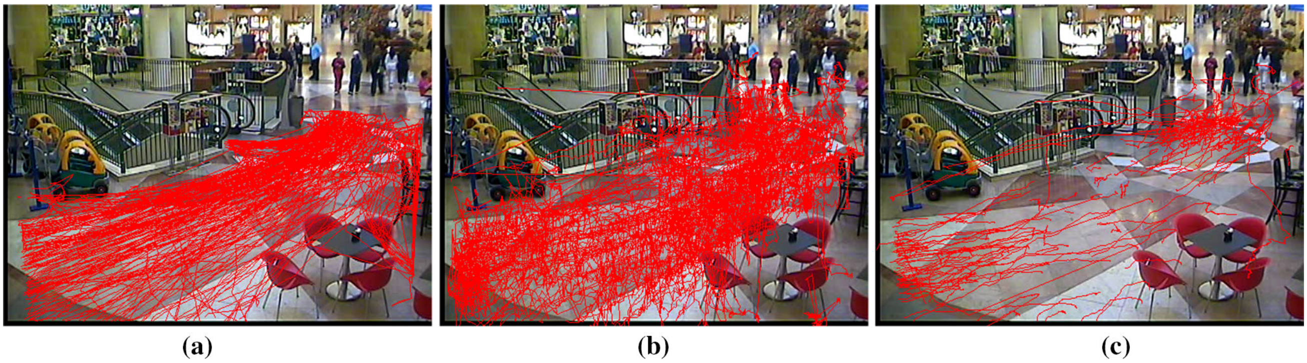


Fig. 5 Subsample (≈25 %) of the trajectories obtained from: **a** manual annotation, **b** boosting algorithm, **c** MedianFlow algorithm

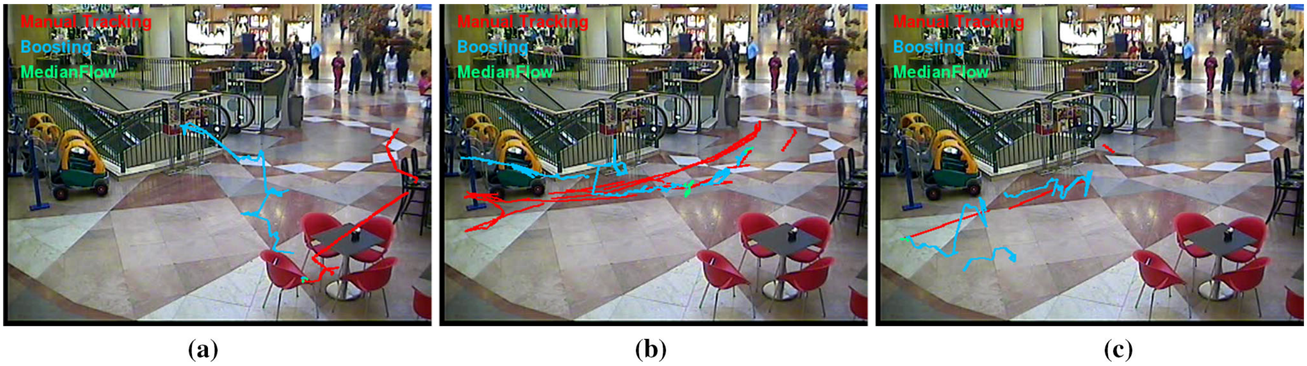


Fig. 6 Example of tracking failures from boosting and MedianFlow algorithms

Table 4 Evaluation of the gaze estimation performance (%) for two tracking data (boosting and GT) and for two evaluation methods (Chamveha et al. [7] and comparing with GT)

Tracking	Metric	Gaze estimation (evaluation from [7])	Gaze estimation (comparing with GT)
Boosting	P	49.0	6.4
	R	50.8	20.4
	A	87.3	84.6
Ground truth (GT)	P	59.5	7.0
	R	54.3	13.9
	A	89.7	84.5

5.2 Automatic gaze estimation

The gaze estimation performance considers two tracking data, from boosting and from ground truth. The evaluation is assessed by two different methods: (1) from Chamveha et al. [7], which considers the auto-determined walking directions as the ground-truth labels of the persons' head orientations; (2) from ground truth (GT), using our annotated data. The results are given in Table 4 and reported on the three standard metrics, namely precision (P), recall (R) and accuracy (A). Figure 7 shows the corresponding confusion matrixes, and Fig. 8 illustrates the number of representative images per gaze orientation index automatically generated. More specifically, we considered eight head orientations (classes), assigned anticlockwise with 45° in between, i.e. first class is "looking right", second is "looking upper-right" and so forth.

As expected, the evaluation method from Chamveha et al. [7] reports significantly better results, as depicted in the third column of Table 4. However, the generated samples are additionally tested with our manual annotation, and the expected deterioration of the results is shown in the fourth column. Finally, Table 4 also shows slightly better results for manual annotation as tracking data, since the tracking loss usually causes less representative images for head orientation and even false positives in some cases.

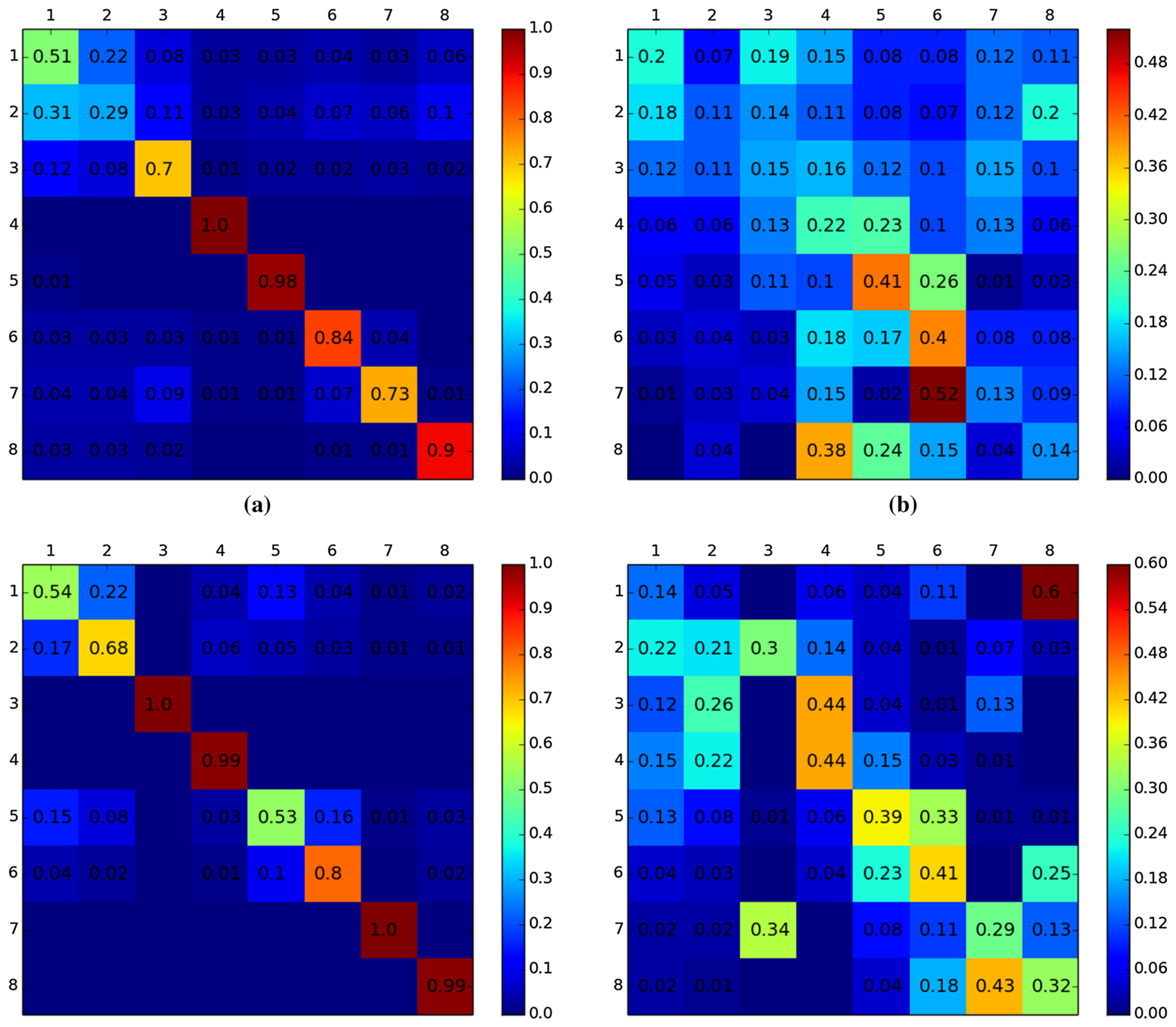
The confusion matrixes from Fig. 7 correspond to the same four situations from Table 4, respectively, and it can be observed the same difference in the results obtained with the two evaluation methods, when comparing Fig. 7a, c with b, d. In particular, the nearly perfect classification of the labels 3, 4, 7, 8 from Fig. 7c is caused by the very few samples for those walking directions, as illustrated in Fig. 8. The work in [7] uses oversampling to handle imbalanced data, but it shows no effect in our data set where, due to environment constrains, the constant flow of pedestrians is restricted to a diagonal path where the 1, 2, 5, 6 walking directions are very frequent (see Fig. 2), causing a severe unbalance of the samples. This step is conducted in offline mode due to the several tasks involved, described in Sect. 4.2.2, specially for the head detection, which is the most demanding task in terms of computational load; for instance, the sample acquisition from the whole video required around one and a half day.

5.3 Impact of automatic feature extraction on classification

Table 5 presents the impact of the boosting algorithm, for the extraction of trajectories, on the overall classification results when compared with the manual annotation, for IPs and GBs (see rows AT-MG and MT-MG, respectively). As expected, the automatic tracking causes a deterioration of the results. Inspecting the second and third rows, we verified that most of the performance drop from the automatic feature extraction is due to the tracking failures (see Fig. 6).

In general, the negative impact is higher on GBs than IPs, which is obvious, since one of the features used to identify a G.B is the inferred label of the IP that can be affected directly by the automatic tracking. We also verified that *BI* and *UI* are the most impaired, which is also expected since their behaviour is highly dependent on their trajectories, while *CHAT.* is mostly affected by detection, since it represents free-standing conversational groups and *EI* is the predominant class with a large number of samples. In terms of IPs, the *Int.* is the most compromised, the *Exp.* is the less affected since it is the most representative class, and the *Dist.* and *Dis.* reveal the largest significative drop since, by definition, they are the most dependent on the trajectory behaviour, specially the *Dis.* which in fact shows the worst result with automatic tracking. Indeed, all the results are affected by the eventual loss of tracks and their random movement later on (see Fig. 6).

Table 5 also presents the impact of gaze estimation, for the extraction of head directions, on the overall classification results w.r.t. the manual annotation, for IPs and GBs (see rows MT-AG and MT-MG, respectively). We verified that all the IPs, with the exception of the *Dis.*, benefit from a small improvement, which can be explained by a regularisation derived from the discretisation of the head directions, thus eliminating some noise from the manual annotation since this process involves some errors due to low image resolution and small size of persons in the scene. In fact, the *Int.* profile is the one that presents the higher improvement, since it is the one with less gaze



	1	2	3	4	5	6	7	8
1	0.51	0.22	0.08	0.03	0.03	0.04	0.03	0.06
2	0.31	0.29	0.11	0.03	0.04	0.07	0.06	0.1
3	0.12	0.08	0.7	0.01	0.02	0.02	0.03	0.02
4				1.0				
5	0.01				0.98			
6	0.03	0.03	0.03	0.01	0.01	0.84	0.04	
7	0.04	0.04	0.09	0.01	0.01	0.07	0.73	0.01
8	0.03	0.03	0.02			0.01	0.01	0.9

(a)

	1	2	3	4	5	6	7	8
1	0.2	0.07	0.19	0.15	0.08	0.08	0.12	0.11
2	0.18	0.11	0.14	0.11	0.08	0.07	0.12	0.2
3	0.12	0.11	0.15	0.16	0.12	0.1	0.15	0.1
4	0.06	0.06	0.13	0.22	0.23	0.1	0.13	0.06
5	0.05	0.03	0.11	0.1	0.41	0.26	0.01	0.03
6	0.03	0.04	0.03	0.18	0.17	0.4	0.08	0.08
7	0.01	0.03	0.04	0.15	0.02	0.52	0.13	0.09
8		0.04		0.38	0.24	0.15	0.04	0.14

(b)

	1	2	3	4	5	6	7	8
1	0.54	0.22		0.04	0.13	0.04	0.01	0.02
2	0.17	0.68		0.06	0.05	0.03	0.01	0.01
3			1.0					
4				0.99				
5	0.15	0.08		0.03	0.53	0.16	0.01	0.03
6	0.04	0.02		0.01	0.1	0.8		0.02
7							1.0	
8								0.99

(c)

	1	2	3	4	5	6	7	8
1	0.14	0.05		0.06	0.04	0.11		0.6
2	0.22	0.21	0.3	0.14	0.04	0.01	0.07	0.03
3	0.12	0.26		0.44	0.04	0.01	0.13	
4	0.15	0.22		0.44	0.15	0.03	0.01	
5	0.13	0.08	0.01	0.06	0.39	0.33	0.01	0.01
6	0.04	0.03		0.04	0.23	0.41		0.25
7	0.02	0.02	0.34		0.08	0.11	0.29	0.13
8	0.02	0.01			0.04	0.18	0.43	0.32

(d)

Fig. 7 Confusion matrixes for gaze estimation (considering eight head orientations—classes—assigned anticlockwise with 45° in between, i.e. first class is “looking right”, second is “looking upper-right” and so forth): for two tracking data (boosting in **a**, **b** and GT in **c**, **d**) and for two evaluation methods (Chamveha et al. [7] in **a**, **c** and comparing with GT in **b**, **d**)

variation, just corroborating the previous conclusion. However, the *Dis.* profile shows a performance drop since, by definition, it should present a high variability in gaze direction, thus affecting its performance due to the imbalance of classes previously stated. Considering GBs, a small degradation is confirmed in all of them.

5.4 Classification

To analyse our descriptor performance, we compare the classification results with a baseline descriptor, which is composed by the same type of features enumerated in Sect. 4.3, but instead of considering a multi-scale histogram from the features, it simply considers the mean, μ, and standard deviation, σ, of each feature, except for the P_p feature. For the case of GBs, we add a state-of-the-art descriptor, referred here as Chamveha, that builds over our multi-scale descriptor formulation but uses the features presented in [13]. Under our experiments, $R = 3$ showed a good trade-off between accuracy and dimensionality length, which leads to a 112 and 116 dimensional descriptor vector for IPs and GBs, respectively.

For exhaustive classification evaluation, we adopted a twofold cross-validation repeated over 10 random iterations. In order to obtain fair results, we kept classes proportions from the original data set for each fold. Since the imbalance among the classes is very high (see Table 1), we randomly replicate the samples by a percentage, ρ, in the training fold, while maintaining fixed the number of samples of the most representative class and increasing proportionally the remaining classes.

5.4.1 Settings

Using the trajectories and gazes manually annotated, we ran experiments over different parameter settings and compared results over an overall F_1-score. The experiments consider several classifiers, replication percentage of the training set, variations on descriptor temporal length τ, bag size Γ, and the size of the vocabulary, K. Several classifiers were tested such as multi-layer perceptrons (MLP), random trees (RTree), gentle AdaBoost, normal Bayes, and Support Vector Machine (SVM), among others. For sake of simplicity, we only report in this work the F_1 score of each classifier for each IP and GB in Table 6 to support our conclusions. We observed that the SVM classifier presents the best overall performance; therefore, a deeper analysis was conducted over different kernels, namely linear, polynomial, RBF and intersection, verifying the best results for the intersection kernel. Several C-SVM values were tested as input, ranging from 2^{-2} to 2^{8}, and chosen the one that optimise the F_1 score. Table 7 summarises the tested parameters and their best values verified empirically. This step is very time efficient. Since the trajectories and gazes were already computed in offline mode and the objects of interest are known, the descriptor's computation is far more faster than real time. Although the offline design of the classifier takes around ten minutes, the prediction operation is also faster than real time.

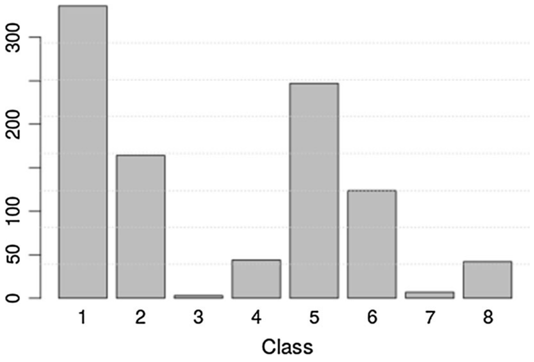


Fig. 8 The obtained number of representative images per gaze orientation index

5.4.2 Feature importance

The proposed backward feature selection technique, explained in Sect. 4.4, permits us to evaluate individual feature importance under the classification framework. Inspection of Fig. 9 shows that for IPs the features have well-defined and layered contributions, highlighting the relevant role of the gaze feature. On the other hand, GBs feature analysis shows a more balanced importance among features, proving their similar importance over the descriptor.

5.4.3 Feature Strategy and Matching Distance

To analyse this component, we compared two histogram matching techniques, namely the average and the max, where the distances between the individual feature's histograms of the final descriptor are computed and the final distance is considered to be the average or the max of them, respectively. In this way, the distances from all clusters are

Table 5 Classification results (%) for all IPs and GBs. considering our descriptor and combinations of manual (M) and automatic (A) feature extraction processes for tracking (T) and gaze (G)

	EI				BI				UI				CHAT.				Avg.			
	P	R	A	F	P	R	A	F	P	R	A	F	P	R	A	F	P	R	A	F
MT-MG	90.8	93.9	87.9	92.3	52.7	56.8	91.9	54.0	56.6	42.7	89.6	46.7	35.6	27.0	95.6	38.9	58.9	55.1	91.2	58.0
MT-AG	90.0	91.7	85.8	90.8	47.9	41.4	91.3	42.8	42.2	41.0	87.7	43.0	27.0	29.0	94.1	29.4	51.8	50.8	89.7	51.5
AT-MG	78.8	82.2	69.3	80.4	14.2	13.6	85.6	18.3	12.4	10.3	81.4	13.9	15.6	8.0	94.8	25.6	30.3	28.5	82.8	34.5
AT-AG	79.2	78.4	67.6	78.7	10.2	11.8	84.1	14.3	14.7	16.0	80.9	15.5	11.2	8.0	93.6	25.5	28.9	28.5	81.6	33.5
	Dist.				Exp.				Dis.				Int.				Avg.			
MT-MG	28.9	44.1	89.4	34.3	95.3	89.8	86.8	92.4	11.6	10.0	98.5	48.8	43.2	54.2	95.7	46.7	44.7	49.5	92.6	55.6
MT-AG	32.4	40.2	90.8	35.4	95.4	90.7	87.8	93.0	12.3	30.0	97.6	25.8	54.6	64.7	96.5	57.0	48.7	56.4	93.2	52.8
AT-MG	17.6	30.1	86.4	21.8	92.4	78.8	75.1	85.0	9.7	22.5	98.2	27.5	7.8	23.6	87.3	11.6	31.8	38.7	86.7	36.5
AT-AG	13.6	23.2	85.3	16.9	92.3	79.0	75.3	85.1	24.6	42.5	89.7	40.5	7.3	21.7	87.4	10.8	34.5	41.6	86.7	38.4

Table 6 Classification results (F_1 score %) of different classifiers for all IPs and GBs

Classifier	EI	BI	UI	CHAT.	Dist.	Exp.	Dis.	Int.	Avg.
Bayes	91.9	51.4	**50.5**	38.4	34.4	89.6	44.7	49.0	56.2
AdaBoost	88.6	40.4	25.2	26.6	**36.7**	66.0	32.4	**55.3**	46.4
MLP	89.8	44.0	32.1	32.5	23.6	91.6	39.2	41.0	49.2
RTree	74.4	37.3	42.3	**41.3**	21.2	**93.1**	38.0	48.1	49.5
SVM	**92.3**	**54.0**	46.7	38.9	34.3	92.4	**48.8**	46.7	**56.8**

Considering our descriptor and the best parameters of our framework, stated in Table 7

Bold values are the best results when comparing all the results related to a specific experience

Table 7 Empirical values for some parameters of our classification framework

ρ (%)	K	τ (s)		Γ	
		IP	GB	IP	GB
15	70	5	1	1	7

stored and a decision is made taking one of both techniques. For evaluation, we considered the F_1-score measure of both matching strategies fixing the pooling scheme.

Inspecting Table 8, we can verify that in general the average matching presents better results than the max. This difference is lower when classifying the IPs, specially when the max is taken, but a drastic performance drop is obtained while classifying the GBs. This can be explained by several factors: (1) more variability, since each group is represented by a global behaviour that depends on the number of individuals within it and their profile's variance; (2) more separability among GBs, by definition two of the classes are highly distinguishable, namely *Exploring* and *Chatting*; (3) similar features contribution for the descriptor, as stated on the previous Sect. 5.4.2; (4) lower correlation among features, which is pretty well represented looking at the values of the correlation distance in Table 8.

Concerning the matching distance, it is obvious that the worse one is the correlation distance, while the remaining present similar performance. Since the average matching technique presents better results for all the classes, we select the intersection distance, which also reveals the best classification. The combination of the histogram intersection measure with the intersection kernel SVM corroborates that the combination of both generate better visual codebooks under unsupervised learning [34].

5.4.4 Pooling strategy

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

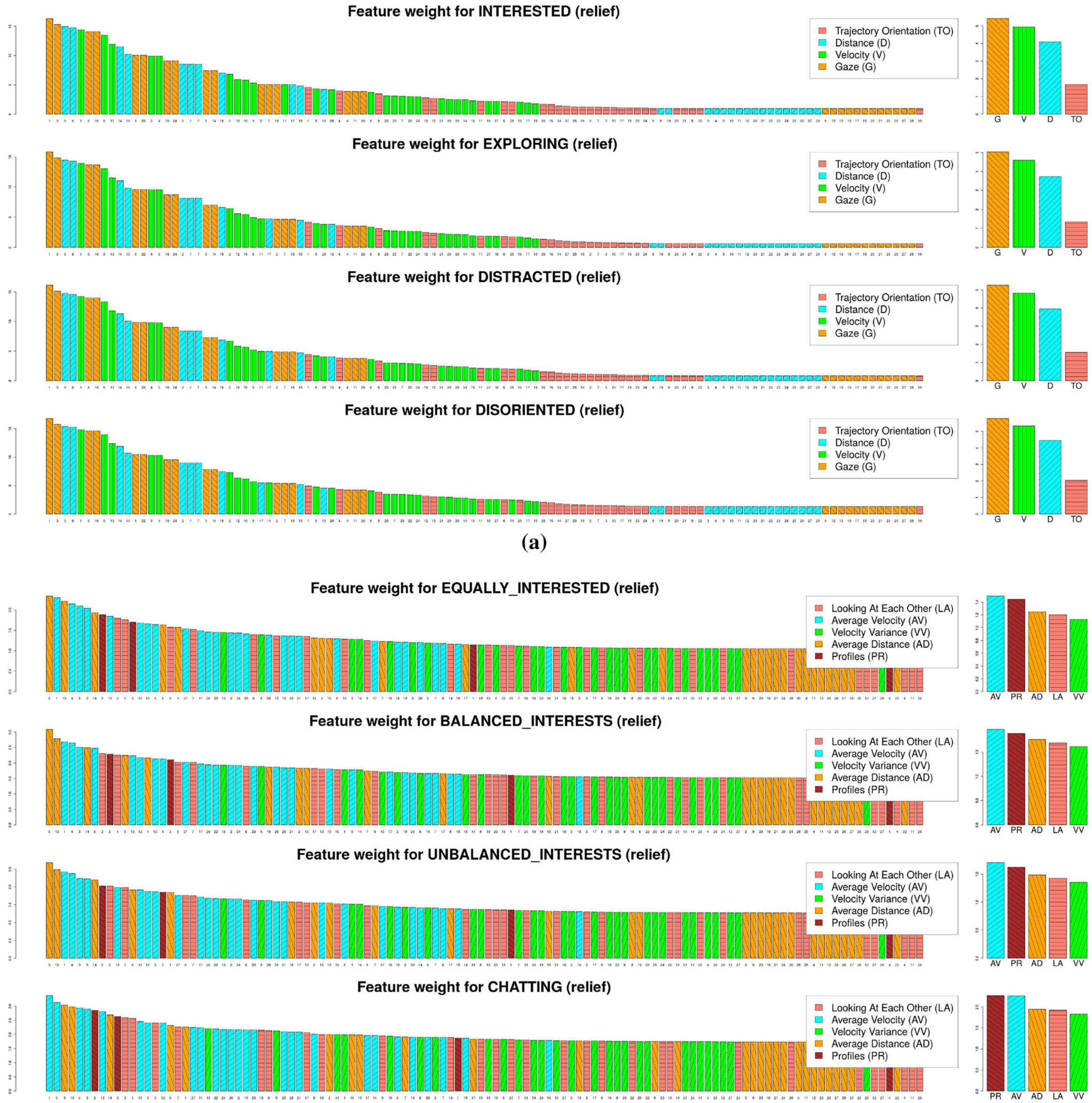
Overall evaluation confirms that average pooling technique performs better than the max pooling technique, specially when classifying the GBs, which makes sense since more information is collected by bag. The difference between both is small, but significant enough to be considered. This leads us to conclude that since each descriptor is extracted by key-point trajectory sampling, all the sampling points are relevant for the final representation of the trajectory.

5.4.5 Descriptor performance

In this section, we simulated two anomalous behaviours that might affect the descriptor performance: (1) tracking loss, where some trajectories' segments were removed; (2) noise variation, where different degrees of noise in terms of the variation of σ were injected into the trajectories and gazes. Both conditions try to simulate tracking and gaze estimation loss and errors. For the training test, we use samples without any kind of perturbation.

From Fig. 10a, b, we verified a slightly decrement of performance. However, what is important to retain is that the fluctuation of performance is higher on IPs than on GBs, which means that GBs are less sensitive to tracking loss and that our descriptor can characterise small temporal segments with a similar performance than the whole set of segments that constitute the IP or GB. Both statements confirm the evidence that will be stated in next Sect. 5.4.6, which shows that the *fine mini-batch* approach presents a performance similar to the *coarse mini-batch* approach for GBs and a slightly worse for IPs. We should also highlight that the simulation of segments loss was done at the level of the bag and not at the level of the key points (tracking level); therefore, despite the remotion of some segments, the remaining ones keep the temporal structure.

The noise variation in gaze and trajectory also affects the performance. Figure 10c, d shows a decreasing function with an initial steepest drop. As expected, the negative impact on GBs is higher than the IPs due to the relational features involved among individuals of the same group.



(a)

(b)

Fig. 9 Feature importance analysis for: **a** IPs; **b** GBs

The decrease in performance could have been even higher if there was not the compensation effect of the average velocity and average distance. As Fig. 9b shows, those two features contribute the most to the recognition of the GBs. The respective average attenuates the fluctuations in velocity and distance, caused by noise variation in trajectory. In fact, such smoothing could be the reason for some oscillations in the performance curve even with the increase in noise.

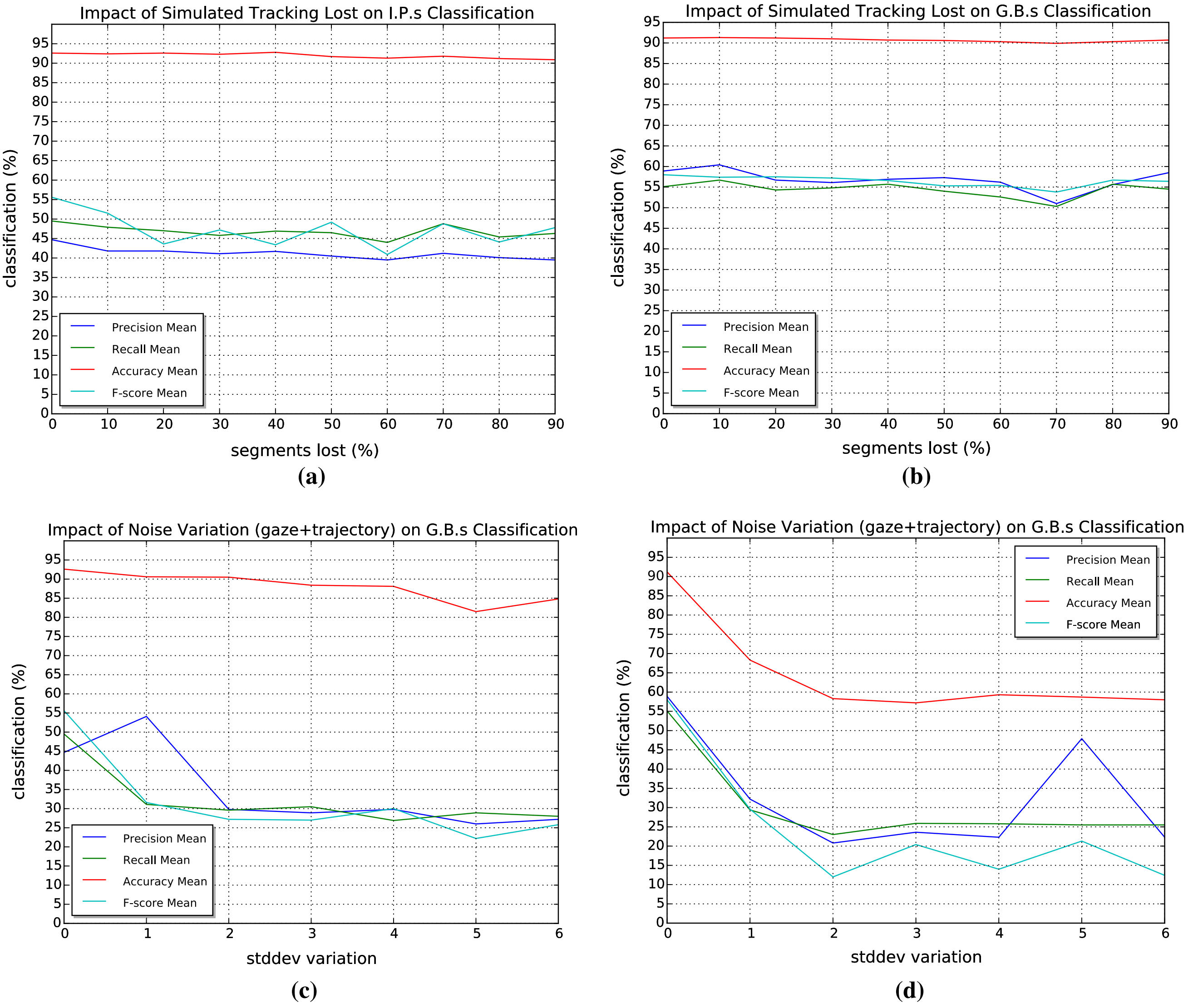
5.4.6 *Mini-batch approaches*

As stated in Sect. 4.4, the *fine mini-batch* approach uses the bags as samples and the final classification result of the whole individual trajectory (IP) or the entire group set (GB) is obtained considering the most predominant label along all the bags. The main theoretical advantage of the *fine mini-batch* over the *coarse mini-batch* approach is its dynamic nature, since if not prior knowledge about the

Table 8 Mean F_1-score (%) of IPs, GBs and overall for the combination of histogram matching, distance measure and pooling configurations, using our descriptor

Matching	Average						Maximum					
Pooling	Avg.			Max.			Avg.			Max.		
Labels	IPs	GBs	All	IPs	GBs	All	IPs	GBs	All	IPs	GBs	All
Intersection	**55.6**	**58.0**	**56.8**	42.7	53.2	48.0	38.1	36.3	37.2	41.4	32.3	36.9
Euclidean	41.7	58.0	49.9	44.3	53.1	48.7	38.4	36.4	37.4	36.2	11.3	23.8
Correlation	40.1	52.4	46.3	42.3	51.7	47.0	39.1	9.6	24.4	37.1	10.8	24.0
Bhattacharyya	43.3	50.7	47.0	44.0	54.2	49.1	38.4	38.0	38.2	39.1	36.5	37.8

Bold values are the best results when comparing all the results related to a specific experience



(a) **(b)** **(c)** **(d)**

Fig. 10 Simulation of the impact of tracking loss (for IP in **a** and GB in **b**) and noise variation in gaze and trajectory (for IP in **c** and GB in **d**) in classification results

starting and ending time of any IP or GB is known, the *coarse mini-batch* approach is useless, while the *fine mini-batch* can provide more detailed information about the temporal behaviour of IPs and GBs. Therefore in this section, we examine the robustness of the *fine mini-batch* level in order to suppress the *coarse mini-batch* level.

Table 9 Classification results (%) for *fine mini-batch* approach

	EI			BI			UI			CHAT.			Avg.		
	P	R	A	P	R	A	P	R	A	P	R	A	P	R	A
MT-MG															
Baseline	68.5	81.7	63.7	18.8	10.3	80.8	23.5	14.4	76.5	10.0	10.4	94.0	30.2	29.2	78.8
Chamveha	91.7	89.4	85.1	**57.5**	55.0	92.2	40.5	48.3	88.2	13.3	21.1	95.2	50.8	53.5	90.2
Our	**93.3**	**90.8**	**87.4**	48.4	**55.1**	**92.5**	**44.3**	**50.7**	**88.4**	**29.0**	**28.4**	**95.4**	**53.8**	**56.3**	**90.9**
AT-AG															
Baseline	46.6	79.5	49.7	**21.1**	8.7	75.1	**30.9**	14.9	71.4	**16.3**	6.4	87.5	28.7	27.4	70.9
Chamveha	**75.6**	80.5	**66.9**	15.1	**21.5**	**87.6**	24.8	16.8	**78.3**	**16.3**	**13.6**	**93.4**	**33.0**	**33.1**	**81.6**
Our	73.4	**81.5**	66.4	12.9	13.9	86.3	28.0	**17.2**	78.1	11.8	6.8	91.6	31.5	29.9	80.6
	Dist.			Exp.			Dis.			Int.			Avg.		
MT-MG															
Baseline	5.7	18.6	89.3	88.7	92.1	82.9	0.0	0.0	**99.7**	25.8	15.2	91.0	30.1	31.5	90.7
Our	**28.0**	**22.5**	**89.7**	**90.7**	**95.1**	**87.2**	0.0	0.0	99.5	**68.4**	**39.1**	**95.2**	**46.8**	**39.2**	**92.9**
AT-AG															
Baseline	3.1	**28.1**	**91.5**	**88.3**	91.3	**81.7**	0.0	0.0	**99.6**	25.9	**9.2**	**88.3**	29.3	**32.2**	**90.3**
Our	**13.4**	9.2	86.9	80.9	**92.7**	76.8	**20.0**	**19.5**	99.4	**26.9**	7.2	86.3	**35.3**	**32.2**	87.4

Bold values are the best results when comparing all the results related to a specific experience

Table 9 shows the classification results for IPs and GBs under both combinations of manual and automatic features extraction. For the GBs and assuming manual annotation, we stated a clear advantage of both multi-scale histogram descriptors, Chamveha and ours, over the baseline. EI and CHAT behaviours are the most well defined. It is expectable that for the EI behaviour the performance difference between the baseline and remaining descriptors to be the smallest one. The overall low performance on CHAT behaviour can be explained by the small number of samples. In overall, our descriptor presents the best results over all the GBs and has a better recall rate that should be emphasised rather than the precision rate for surveillance systems.

Considering automatic features extraction, the baseline performance decreases less than both multi-scale histogram descriptors, which suffer a high marked drop. However, their results are even better than the baseline with manually annotated data. This shows that the discretisation of multi-scale histogram can be affected and confused by tracking and gaze estimation errors. By its way, the small reduction undertaken by the baseline descriptor, which only includes the mean and standard deviation of each feature, proves that our descriptor covers a good selection of discriminative features to describe individual interactions within a group and that a global measurement that includes single occurrences could be representative enough to identify a collective behaviour. Our descriptor was highly affected probably because the tracking and gaze estimation failures introduce noise that is captured by the multi-scale histogram, which is cancelled out by the smoothing of the baseline. The Chamveha descriptor slightly superimposes to our descriptor; therefore, we might conclude that since it has more features they might complement each other in the presence of extraction failure. However, it also sustains the importance of our descriptor sampling strategy as an effective representation over time. In this scenario, the Chamveha descriptor presents a higher overall recall rate.

In terms of IPs analysis, our descriptor presents the best overall result for both manual and automatic feature extraction. However, we observe the same drastic reduction in our descriptor and just a small decay of the baseline while using automatic features. The *Int.* profile is the one with the worst degradation for our descriptor, probably because its dependence with the gaze feature and it is the one which presents the most structured movement aligned with the objects of interest of the scene; therefore, perturbations on tracking and gaze estimation affect its performance.

Table 10 shows a comparison of performance between the *coarse mini-batch* and the *fine mini-batch* approaches. In terms of GBs, both methods obtain similar results, with a slightly higher improvement margin for the *coarse mini-batch* approach while using manual annotation. A deeper analysis of each GB reveals an overall stable result for the *fine mini-batch* approach considering the automatic feature extraction. Therefore, we might conclude that the *fine mini-batch* method is preferred since its performance is not

Table 10 Comparison results (%) between *coarse mini-batch* and *fine mini-batch* approaches, considering our descriptor

	EI			BI			UI			CHAT.			Avg.			
	P	R	A	P	R	A	P	R	A	P	R	A	P	R	A	F
MT-MG																
Coarse	90.8	**93.9**	**87.9**	**52.7**	**56.8**	91.9	**56.6**	42.7	**89.6**	**35.6**	27.0	**95.6**	**58.9**	55.1	**91.2**	**58.0**
Fine	**93.3**	90.8	87.4	48.4	55.1	**92.5**	44.3	**50.7**	88.4	29.0	**28.4**	95.4	53.8	**56.3**	90.9	57.3
AT-AG																
Coarse	**79.2**	78.4	**67.6**	10.2	11.8	84.1	14.7	16.0	**80.9**	11.2	**8.0**	**93.6**	28.9	28.5	**81.6**	33.5
Fine	73.4	**81.5**	66.4	**12.9**	**13.9**	**86.3**	**28.0**	**17.2**	78.1	**11.8**	6.8	91.6	**31.5**	**29.9**	80.6	**33.6**
	Dist.			Exp.			Dis.			Int.			Avg.			
MT-MG																
Coarse	**28.9**	**44.1**	89.4	**95.3**	89.8	86.8	**11.6**	**10.0**	98.5	43.2	**54.2**	**95.7**	44.7	**49.5**	92.6	**55.6**
Fine	28.0	22.5	**89.7**	90.7	**95.1**	**87.2**	0.0	0.0	**99.5**	**68.4**	39.1	95.2	**46.8**	39.2	**92.9**	42.5
AT-AG																
Coarse	**13.6**	**23.2**	85.3	**92.3**	79.0	75.3	**24.6**	**42.5**	89.7	7.3	**21.7**	**87.4**	34.5	**41.6**	86.7	38.4
Fine	13.4	9.2	**86.9**	80.9	**92.7**	**76.8**	20.0	19.5	**99.4**	**26.9**	7.2	86.3	**35.3**	32.2	**87.4**	**42.6**

Bold values are the best results when comparing all the results related to a specific experience

Table 11 Evidence scores (%) for false (EN) and true (EP) detections

	EI		BI		UI		CHAT.		Avg.	
	EN	EP	EN	EP	EN	EP	EN	EP	EN	EP
MT-MG	92.7	97.2	98.1	97.5	89.2	95.8	89.6	88.6	92.4	94.8
AT-AG	94.0	89.7	95.2	94.2	90.6	94.5	89.1	93.9	92.2	93.1
	Dist.		Exp.		Dis.		Int.		Avg.	
MT-MG	89.9	83.3	83.0	98.1	0.0	100.0	79.2	96.4	63.0	94.5
AT-AG	89.5	80.9	89.7	96.2	100.0	100.0	90.6	92.9	92.5	92.5

significantly different than the *coarse mini-batch* approach, while having the advantage of being smoother and more stable.

Considering the IPs, an advantage of the *coarse mini-batch* approach over the *fine mini-batch* is verified when considering the manual data. However, for automatic feature extraction the *fine mini-batch* performs considerable better than the *coarse mini-batch*. In fact, the *fine mini-batch* approach keeps similar classification results with both manual and automatic data. As stated on previous Sect. 5.4.5, the IPs are more sensitive to noise and feature extraction failures than the GBs; therefore, an obvious conclusion is that the *fine mini-batch* approach brings stability to the classification, since the collected information is more compact and consequently the descriptor becomes less error prone.

In order to obtain a measure of confidence for the decision of the *fine mini-batch* approach, two evidence scores are taken. Both consider the ratio between the number of predominant labels and the total number of bags, and they are: (1) EN, which is the evidence of false detections and indicates the level of confidence of false negative occurrences; (2) EP, which is the evidence of true detections and gives the score of true positive occurrences. For an excellent decision process, we expect a high EP and a low EN Inspecting Table 11, and we stated that our descriptor presents high values for both metrics, higher for EP than for EN. This leads us to conclude that the *fine mini-batch* approach keeps a constant behaviour along the entire trajectory, for the IP, or the whole group, for the GB. This regularity among the mini-batches emphasises the discriminative power of this approach.

Since one of the advantages of the *fine mini-batch* approach is its dynamic nature, which can help to determine the switch behaviour between IPs and GBs, we measured the recognition rate at the extreme bags, initial and final, of the IPs and GBs and reported them at Table 12. In general, we stated that the recognition rate is

Table 12 Recognition rate (%) for extreme bags, initial and final, on IPs and GBs

	EI	BI	UI	CHAT.	Avg.
MT-MG	90.6	48.1	46.8	27.3	53.2
AT-AG	66.7	14.6	30.3	11.8	30.9
	Dist.	Exp.	Dis.	Int.	Avg.
MT-MG	25.0	90.2	0.0	70.4	46.4
AT-AG	12.8	79.6	20.0	30.8	35.8

directly related to the number of samples of IPs or GBs, and there is an overall impairment of results for the automatic features extraction, with the exception of the *Dis.* profile.

5.4.7 Sociological meaning of features

In this section, we inspect the sociological meaning of each individual feature within our descriptor and corroborate their importance for final classification as stated in Sect. 5.4.2.

Inspecting Table 13, and considering the GBs, we verified the importance of the individual profiles, P_p. By sociological definition [6], it makes sense since each profile encloses a well-defined behaviour by itself, and the combination of all the individuals is what mainly defines the group behaviour. However, higher values were expected since the definitions of our GBs concepts are highly dependent on individual profiles. This leads us to conclude that the manual annotation should be revised. Indeed, the annotation process is extremely hard not only because of the difficulties associated with low-level features such as gaze (affected by image resolution, camera viewpoint and perspective), but mainly because the decision to choose for the correct IP or GB, which is a subjective process despite the rules imposed (see Sect. 3.1). We also stated that remaining features provide similar performance, confirming the feature importance analysis carried out in Sect. 5.4.2. Just the inclusion of the profiles, P_p, leads to a high performance rate; however, since they are also acquired from a similar classification process, noise and errors are introduced. In general, all the features contribute to the classification process and their mutual combination appears to be the most reliable option. We should highlight that we did not yet ran experiments on GBs classification considering the profiles from automatic IPs classification.

In terms of IPs, also a clear improvement is observed when all the features are aggregated. Indeed, this shows that they complement each other, which validate sociological theories that based their studies on several features to learn complex person behaviour [6]. We verified important observations: (1) for the *Dis.* profile all the contribution comes from the gaze feature, which makes sense since this is the profile with the highest gaze variability; (2) for the *Exp.* profile the predominant benefit comes from the speed, since it is the one that presents the lowest speed and this characteristic can help to discriminate it; (3) for the *Int.* profile the distance to nearest object of interest reveals the highest contribution, which in fact defines this profile; and (4) for the *Dist.* profile all the features assume similar roles, since its behaviour could vary depending on scene context.

As a reference of confidence, we conducted a final experiment to measure the features importance. We

Table 13 Classification results (%) of GBs and IPs considering combination of features within our descriptor, *fine mini-batch* approach and manual annotation data (see Sect. 4.3 for feature list)

	EI			BI			UI			CHAT.			Avg.			
	P	R	A	P	R	A	P	R	A	P	R	A	P	R	A	F
$\tilde{v}_g + \mathrm{var}[v_g]$	76.5	47.8	49.1	7.2	36.8	60.4	34.5	32.5	85.1	34.4	36.0	94.9	38.2	38.3	72.4	35.6
Pp	87.5	0.4	23.6	8.8	**91.8**	19.5	**78.1**	45.6	92.3	**73.1**	**100.0**	**98.5**	**61.9**	**59.5**	58.5	39.9
12345	82.0	39.5	47.1	43.3	11.4	91.0	18.9	**57.4**	41.3	0.0	96.2	0.0	36.1	51.1	44.9	32.8
laeo_g	83.1	65.4	63.2	14.0	35.5	76.0	28.6	39.2	82.4	23.1	17.0	94.6	37.2	39.3	79.1	38.5
All	**90.8**	**93.9**	**87.9**	**52.7**	56.8	**91.9**	56.6	42.7	**89.6**	35.6	27.0	95.6	58.9	55.1	**91.2**	**58.0**
	Dist.			Exp.			Dis.			Int.			Avg.			
α_{si}	28.7	43.3	89.2	94.9	90.7	87.3	0.0	0.0	**98.9**	19.5	22.5	94.3	35.8	39.1	92.5	37.3
d_{io}	16.2	34.6	83.7	93.9	79.6	77.1	0.0	0.0	96.6	**43.2**	**75.3**	95.0	38.3	47.4	88.1	40.1
β_{gi}	14.4	29.8	84.0	92.0	76.4	72.9	10.5	**12.5**	98.7	6.0	19.7	86.7	30.7	34.6	85.6	37.5
v_i	**29.3**	42.3	89.4	**95.3**	**90.9**	**87.8**	0.0	0.0	98.4	24.7	27.3	94.6	37.3	40.1	**92.6**	38.1
All	28.9	**44.1**	**89.4**	**95.3**	89.8	86.8	**11.6**	10.0	98.5	**43.2**	54.2	**95.7**	**44.7**	**49.5**	**92.6**	**55.6**

Bold values are the best results when comparing all the results related to a specific experience

Table 14 Comparison results (%) for k-fold normal (keeping original classes proportions) and stratified cross-validation for our descriptor, *coarse mini-batch* approach and manual annotation

	EI			BI			UI			CHAT.			Avg.			
	P	R	A	P	R	A	P	R	A	P	R	A	P	R	A	F
Stratified	71.4	69.0	84.5	**66.2**	**62.0**	81.7	54.1	**47.0**	75.5	**68.4**	**71.0**	82.7	**65.0**	**62.3**	81.1	**62.1**
Normal	**90.8**	**93.9**	**87.9**	52.7	56.8	**91.9**	**56.6**	42.7	**89.6**	35.6	27.0	**95.6**	58.9	55.1	**91.2**	58.0
	Dist.			Exp.			Dis.			Int.			Avg.			
Stratified	**37.2**	35.0	64.4	47.4	45.0	71.2	**50.0**	**25.0**	74.4	**57.8**	**70.0**	77.5	**48.1**	43.8	71.9	**62.6**
Normal	28.9	**44.1**	**89.4**	**95.3**	**89.8**	**86.8**	11.6	10.0	**98.5**	43.2	54.2	**95.7**	44.7	**49.5**	92.6	55.6

Bold values are the best results when comparing all the results related to a specific experience

considered a twofold stratified cross-validation scheme, where the class proportions are approximately equal on each fold, over 100 random iterations. In this way, we try to measure features importance of balanced data sampling. Table 14 shows that *CHAT* and *Dis.*, the G.B and IP with less samples, respectively, largely improve their performance. We also notice, as expected, that the classes with more samples, namely *EI* and *Exp.*, decrease their classification results. The uniform distribution of samples among the training and testing sets improves, by a significant margin, the overall results for IPs and GBs, which states the importance of the selected features.

6 Conclusions

In this work, we addressed the characterisation of individual profiles and collective behaviours within a social context in a surveillance scenario. For this purpose, we elaborated semantic concepts sustained on social psychology principles and embedded them into the annotation of a novel video surveillance data set for human activity recognition, validated by experts on the sociological field. We extend our previous relational descriptor study to a deep analysis along several steps of the classification framework, propose a new formulation for the classification process, conduct an evaluation of the sociological meaning of the individual features, and outline the performance impact of automatic process into the classification results. In this way, we present a complete automatic framework for analysis of social behaviour.

We have shown that automatic features extraction could impair the classification results, specially from tracking; therefore, pre-processing and post-processing techniques should be included to improve each feature extraction. Also, deep learning techniques could be an interesting line of work to follow in order to suppress the dependency on the correct extraction of low-level *handcraft* features such as tracking. The proposed mini-batch approach with different levels reveals promising performance, and its dynamic behaviour might be a great advantage for a real-time recognition framework of human activity in surveillance scenarios. For future work, such approach could benefit if embedded into temporal models such as HMM (hidden Markov model) to add sequential information, reduce ambiguity, and help to detect automatic switch among consecutive IPs or GBs.

We also verified the discriminative value of single features and their sociological meaning was justified. From that analysis, we state that some features should be emphasised depending on the IP or GB. In this way, the inclusion of a multiple kernel method in the classification could help to improve features importance. The problematic of imbalanced classes should be studied. Future work should also address the annotation process, through the extension to the remaining videos of our data set to validate the assumptions used here in other crowded scenarios, and, probably, preferring a soft-label annotation, in a continuous system coordinates of attributes or a combination of labels such as the ones normally used on affective computing, instead of the current hard-label annotation that might prejudice a correct classification process.

Acknowledgments This work was financed by the ERDF-European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation—COMPETE 2020 Programme within Project POCI-01-0145-FEDER-006961, and by National Funds through the FCT—Fundação para a Ciência e Tecnologia (Portuguese Foundation for Science and Technology) as part of Project UID/EEA/50014/2013, through the Ph.D. Grant reference SFRH/BD/51430/2011 and postdoctoral 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. Adam A, Rivlin E, Shimshoni I, Reinitz D (2008) Robust real-time unusual event detection using multiple fixed-location monitors. IEEE Trans Pattern Anal Mach Intell 30(3):555–560

2. Aggarwal J, Ryoo M (2011) Human activity analysis: a review. ACM Comput Surv 43(3):16:1–16:43. doi:10.1145/1922649.1922653
3. Ali S, Shah M (2008) Floor fields for tracking in high density crowd scenes. In: ECCV, pp 1–14
4. Babenko B, Yang MH, Belongie S (2009) Visual tracking with online multiple instance learning. In: IEEE conference on computer vision and pattern recognition (CVPR), Miami, FL
5. Bernardin K, Stiefelhagen R (2008) Evaluating multiple object tracking performance: the clear mot metrics. J Image Video Process 2008:1:1–1:10. doi:10.1155/2008/246309
6. Cartwright D, Zander A (eds) (1968) Group dynamics: research and theory, 3rd edn. Harper & Row, New York
7. Chamveha I, Sugano Y, Sugimura D, Siriteerakul, T, Okabe T, Sato Y, Sugimoto A (2011) Appearance-based head pose estimation with scene-specific adaptation. In: IEEE international conference on computer vision workshops (ICCV Workshops), 2011, pp 1713–1720. doi:10.1109/ICCVW.2011.6130456
8. Chang MC, Krahnstoever N, Ge W (2011) Probabilistic group-level motion analysis and scenario recognition. In: ICCV, pp 747–754
9. Choi W, Shahid K, Savarese S (2011) Learning context for collective activity recognition. In: CVPR, pp 3273–3280
10. Ge W, Collins RT, Ruback B (2012) Vision-based analysis of small groups in pedestrian crowds. IEEE Trans Pattern Anal Mach Intell 34(5):1003–1016
11. Grabner H, Grabner M, Bischof H (2006) Real-time tracking via on-line boosting. Proc BMVC. doi:10.5244/C.20.6
12. Helbing D, Molnár P (1995) Social force model for pedestrian dynamics. Phys Rev E 51(5):4282–4286. doi:10.1103/physreve.51.4282
13. Chamveha I, Sugano Y, Sato Y (2013) Social group discovery from surveillance videos: a data-driven approach with attention-based cues. In: Proceedings of the British machine vision conference. BMVA Press
14. Jin B, Hu W, Wang H (2012) Human interaction recognition based on transformation of spatial semantics. IEEE Signal Process Lett 19(3):139–142
15. Kalal Z, Matas J, Mikolajczyk K (2011) Tracking learning detection. IEEE Trans Pattern Anal Mach Intell 34:1409–1422
16. Kalal Z, Mikolajczyk K, Matas J (2012) Tracking-learning-detection. IEEE Trans Pattern Anal Mach Intell 34(7):1409-1422
17. Khalid S, Naftel A (2005) Classifying spatiotemporal object trajectories using unsupervised learning of basis function coefficients. In: VSSN '05: proceedings of the third ACM international workshop on Video surveillance and sensor networks, pp 45–52. ACM, New York, NY, USA. doi:10.1145/1099396.1099404
18. Klgl F, Rindsfser G (2007) Large-scale agent-based pedestrian simulation. In: Petta P, Mller JP, Klusch M, Georgeff MP (eds) MATES, Lecture notes in computer science, vol 4687. Springer, pp 145–156
19. Makris D, Ellis T (2005) Learning semantic scene models from observing activity in visual surveillance. IEEE Trans Syst Man Cybern Part B 35(3):397–408
20. McPhail C, Wohlstein RT (1982) Using film to analyze pedestrian behavior. Sociol Methods Res 10(3):347–375
21. Owens J, Hunter A (2000) Application of the self-organizing map to trajectory classification. In: Proceedings of the third IEEE international workshop on visual surveillance (VS'2000), VS '00, pp 77. IEEE Computer Society, Washington, DC, USA
22. Pereira EM, Cardoso JS, Morla R (2015) Long-range trajectories from global and local motion representations. http://arxiv.org/abs/1509.08647
23. Pereira EM, Ciobanu L, Cardoso JS (2014) Context-based trajectory descriptor for human activity profiling. In: Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, IEEE, San Diego, CA, USA, pp 2385–2390
24. Pereira EM, Ciobanu L, Cardoso JS (2015) Social signaling descriptor for group behavior analysis. In: Proceedings of Iberian conference on pattern recognition and image analysis (IbPRIA)
25. Prisacariu V, Reid I (2009) fasthog—a real-time gpu implementation of hog. Technical Report 2310/09. Department of Engineering Science, Oxford University
26. Pusiol G, Bremond F, Thonnat M (2010) Trajectory based activity discovery. In: 7th IEEE international conference on advanced video and signal-based surveillance, Boston, États-Unis
27. Qiu F, Hu X (2010) Modeling group structures in pedestrian crowd simulation. Simul Model Pract Theory 18(2):190–205
28. Rummel RJ (1976) Understanding conflict and war: the conflict helix, vol 2. Sage Publications, Beverly Hills
29. Ryoo MS, Aggarwal JK (2009) Spatio-temporal relationship match: video structure comparison for recognition of complex human activities. In: ICCV, IEEE, pp 1593–1600
30. Smeulders AW, Chu DM, Cucchiara R, Calderara S, Dehghan A, Shah M (2014) Visual tracking: an experimental survey. IEEE Trans Pattern Anal Mach Intell 36(7):1442–1468
31. Takahashi M, Naemura M, Fujii M, Satoh S (2011) Human action recognition in crowded surveillance video sequences by using features taken from key-point trajectories. Comput Vis Pattern Recogn. doi:10.1109/CVPRW.2011.5981713
32. Wang X, Ma KT, Ng GW, Grimson WE (2011) Trajectory analysis and semantic region modeling using nonparametric hierarchical bayesian models. Int J Comput Vis 95(3):287–312. doi:10.1007/s11263-011-0459-6
33. Wang X, Tieu K, Grimson E (2006) Learning semantic scene models by trajectory analysis. In: Leonardis A, Bischof H, Pinz A (eds) ECCV (3), lecture notes in computer science, vol 3953. Springer, pp 110–123
34. Wu J, Rehg JM (2009) Beyond the euclidean distance: creating effective visual codebooks using the histogram intersection kernel. In: IEEE 12th international conference on computer vision, 2009, IEEE, pp 630–637
35. Wu Y, Lim J, Yang MH (2013) Online object tracking: a benchmark. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2411–2418
36. Yao B, Jiang X, Khosla A, Lin AL, Guibas LJ, Li FF (2011) Human action recognition by learning bases of action attributes and parts. In: ICCV, pp 1331–1338
37. Zhou B, Wang X, Tang X (2012) Understanding collective crowd behaviors: learning a mixture model of dynamic pedestrian-agents. In: CVPR, pp 2871–2878