Detecting Crowd Features in Video Sequences

Rodolfo Migon Favaretto, Leandro Dihl and Soraia Raupp Musse
Virtual Humans Simulation Laboratory – VHLab, Dept. of Computer Science
Pontifical Catholic University of Rio Grande do Sul, PUCRS – Brazil
{rodolfo.favaretto, leandro.dihl}@acad.pucrs.br and soraia.musse@pucrs.br

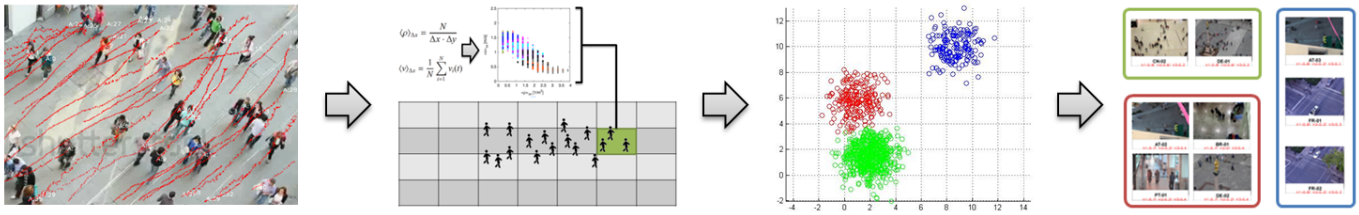


Fig. 1. Illustration of our method: from tracking trajectories (left), the relevant features are extracted using our techniques (second step), data are clustered (third step) and mapped to specific crowd features (right).

***Abstract*—We propose a new methodology to detect social aspects of crowds in video sequences based on pedestrian features, which are obtained through image processing/computer vision techniques. The main idea is to apply and extend the concepts of Fundamental Diagram (FD) with more features, such as grouping and collectivity. Using crowd features we identify the crowd type and the main characteristics. In addition, we also investigated two further results: the visual assessment of people in real video sequences in order to detect crowd characteristics, and the usage of our method to detect similarity of crowds in videos.**

***Keywords*-image processing; fundamental diagrams; classification; crowd analysis.**

I. Introduction

In the last years there is a growing interest in understanding the behavior of crowds in video sequences. This problem is important in many applications, including the safety of pedestrians in complex buildings or in mass events, pedestrian dynamics, computer animation, crowd simulation, virtual reality and games. Many methodologies to detect groups and crowd events have been proposed in the literature and achieved results showing that groups, social behaviors and navigation aspects can be successfully detected in video sequences. For example, counting people in crowds [1], [2], abnormal behavior detection [3], [4], study of social groups in crowds [5], [6], [7], [8], understanding of group behaviors [9] and characterization of crowd features [10]. Most of these approaches are based on individual pedestrian tracking or optical flow algorithms, and in general consider features like speed, directions and distance over time.

However, there is an important attribute that can influence personal behavior, which affects the group that the individual belongs to. Chattaraj et al. [11] suggest that cultural and population differences can produce deviations in speed, density and flow of the crowd. In their work, authors discuss the Fundamental Diagrams (FD) used in planning guidelines [12], [13]. The cultural influence can be considered in crowds attributes as personal spaces, speed, pedestrian avoidance side and group formations [14], and many works ([15], [16], [9], [8]) focus on the identification of groups using computer vision.

In addition to the work on groups and crowd characterization, we are also interested in automatically detecting the type of crowds existent in a video sequence. The main motivation is that we want to work with spontaneous videos (not controlled ones), and for that we need firstly to know the main type of crowd existent in the videos. There is limited research about the types of crowd and crowd membership, and there is no consensus about how to classify the types of crowds. For instance, Momboisse [17] proposed a system that consists of four types: casual, conventional, expressive, and aggressive, while Berlonghi [18] classified crowds as spectator, demonstrator, or escaping, to correlate to the purpose for gathering.

Another approach was proposed by sociologist Herbert Blumer according to emotional intensity. He distinguishes four types of crowds: casual, conventional, expressive, and acting [19]. His system is dynamic in nature. That is, a crowd changes its level of emotional intensity over time, and consequently can be re-classified as any of the four types. Crowds can be active (mobs) or passive (audiences). Active crowds can be further divided into aggressive, escapist, acquisitive, or expressive mobs according Greenberg [20].

In this paper we are interested in automatically detecting crowd features and the type of crowd existent in a given video sequences, which can be used to provide a better understanding of cultural aspects in populations. Furthermore, detecting the type of a crowd in video sequence can be useful to help simulating coherent crowds. Many methods have been

proposed in this area, usually known as data driven crowd simulation [21], [22], [23], [24], [25], [26].

We propose to extend the FD to extract and classify crowd features. According Zhang [27], the FD denotes the relation between pedestrian flow and density and is associated with many qualitative self organization phenomena such as lanes formation and jams. Specifications of various experimental studies, guidelines, and handbooks display substantial differences in maximal flow values and the corresponding densities, as well as the density where the flow vanishes due to overcrowding. In this paper we propose to extend FD data (speeds and densities) using the following crowd attributes: collectivity [10] and population information, namely the total number of people and grouping data (size and number of groups). Results indicate that the proposed extension to FD can be used to better characterize crowds in video sequences.

The main contributions are: a new methodology by extending FD to detect crowd features, that are used to identify the crowd type existent in the video sequence and to analyze similarity between crowds.

This paper is organized as follows. The next section discusses the related work. In Section III, we detail the proposed approach. Section IV shows the experiments performed to evaluate our method. Finally, Section V concludes this article.

II. Related work

The crowd analysis, in general, handle with the detection of the groups of individuals and their trajectories. Here, we cover relevant recent work that focus on analysis of crowd scenes and the identification of group behavior. Sochman and Hogg [28] presented an on-line algorithm for social group inference from trajectories of multiple individuals. The social group inference is formulated as a Social Force Model prediction error minimization. In [29], authors propose an agent-based formulation of pedestrian behavior and a method to estimate hidden personal properties. The work presented by Ge et al. [30] describes an automated pedestrian detection and tracking method that extracts trajectories from video. These methods handle several features to infer the groups. Our idea is to extend the FD to extract and classify crowd features.

The Fundamental Diagrams describes the important relation between density and flow [31] in pedestrian video sequences. The understanding of individual trajectories can arise from studies in pedestrians dynamics and highlight the relationship between crowd density and pedestrian movement.

The concept of the FD is applied in several studies. One example is the detection of cultural differences in crowds. Chattaraj et al. [11] performed a study to verify the cultural influences in individuals trajectories using the FD computed with populations from Germany and Indian. More specifically, they studied the FD using populations with the same size distributed in corridors with two different lengths. The authors observed differences in the estimated minimum personal space for groups, indicating the influence of the cultural differences. Helbing, Johansson and Al-Abideen [32] use the FD for crowd disasters analysis. They presented an algorithm to extract positions and speeds of pedestrians as a function of time and to determine critical crowd conditions, which is important for organization of safer mass events.

The FD can also be used in results of crowd simulators [33], [34]. In the generation of pedestrian trajectories [33], the problem of simulating the movement and behaviors of human-like agents is used in testing and learning phases. The authors proposed an approach based on biomechanical principles and psychological factors. This algorithm exhibits the FD in the crowd movements relative to speed and density relationship. Best et al. [34] proposed an algorithm for density-dependent behaviors in crowd simulation. Their approach aims to generate pedestrian trajectories, being applicable to a large number of GPL multi-agent algorithms that use a combination of local and global planners. Wolinski et al. [35] apply the FD as a macroscopic data metric for evaluation of crowd simulations proposed by his framework.

In this work, we use the FD in addition to other features captured from video sequences, such as collectivity, presence of groups, size of groups and etc.

III. The Proposed Approach

Our approach presents three main modules, as illustrated in Fig. 1: people tracking, statistical data extraction and crowd analysis. The first module (Fig. 1 on the left) is responsible for obtaining the individual trajectories of observed pedestrians in real videos (as an alternative for non-existent scenarios, simulated trajectories can be used). In the second module (Fig. 1 in second and third steps), the statistical information from trajectories is obtained, and finally the last module (Fig. 1 on the right) is responsible for detecting crowd features.

A. Initial detection and tracking

The initial people detection is performed using the work proposed by Viola and Jones [36]. The boosted classifier working with haar-like features was trained with 4500 views of people heads as positive examples, and 1000 negative (datasets CoffeBreak and Caviar Head were used[37]). This detector performs the initial position detection of people based on their heads, which are the input parameters for the next step: tracking. Once the individuals are detected, the trajectories are obtained using the method proposed by Bins et al. [38], which is based on multiple disjoint patches obtained from the target. The patches are represented parametrically by the mean vector and covariance matrix computed from a set of feature vectors that represent each pixel of the target. Each patch is tracked independently using the Bhattacharyya distance [39], and the displacement of the whole template is obtained using a Weighted Vector Median Filter (WVMF). To smooth the trajectory and also cope with short-term total occlusions, a predicted displacement vector based on the motion of the target in the previous frames is also used. The appearance changes of the target are handled by an updating scheme.

The output of tracking phase (illustrated in Fig. 2) is a vector of each person i with the positions $\vec{X}_i^f = (x_i, y_i)$, at each frame f. It should be noticed that tracking is not a contribution

of this work, and any pedestrian/crowd tracking algorithm can be used in this phase.

(a) Initial Detection (b) Tracking

Fig. 2. Tracking phase in our approach: (a) the input image and heads detection, and (b) the people detection and tracked trajectories.

B. Statistical Data Extraction

From data extracted in the first phase, we have following information for each person i, at each frame f:

- *i)* position $\vec{X}_i^f$ of person i (meters);
- *ii)* speed s_i^f of person i (meters/second); and
- *iii)* angular variation α_i^f (degrees) of agent i w.r.t. a reference vector $\vec{r} = (1, 0)$.

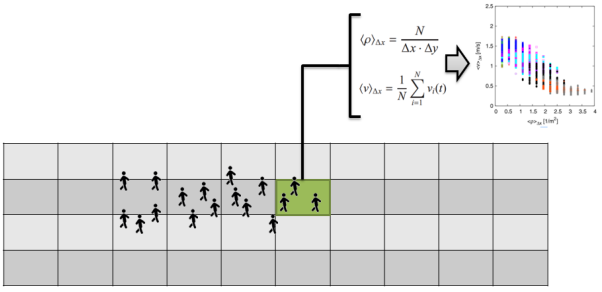
To obtain the desired parameters in the world coordinate system, we computed the planar homography for each video, and transformed the extracted trajectories to the world coordinate system by assuming that the head position is on the ground place ($z = 0$). Since our videos are close to top-view, this assumption does not produce large errors in the projection. Even for videos that are not really top-view, we assume that the error are not impacting the crowd features of interest, as discussed in the next section.

Then, we compute the following parameters for each pair of agents i and j: $s(s_i, s_j)$, $o(\alpha_i, \alpha_j)$ and $d(\vec{X}_i, \vec{X}_j)$, where $s(s_i, s_j)$, $o(\alpha_i, \alpha_j)$ are the differences of speed and orientation and $d(\vec{X}_i, \vec{X}_j)$ is the Euclidean distance between the two individuals. We use the notion of distances based on the "*Proxemics*" described by Hall [40] to define that two agents belong to the same group according to three tests: If $d(\vec{X}_i, \vec{X}_j) \leq 1.2$meter and $o(\alpha_i, \alpha_j) \leq 15°$ and $s(s_i, s_j) < \beta \max\{s_i, s_j\}$, where $\beta = 5\%$ was empirically defined. Based on these rules, agents are grouped in pairs. In the next step, we check which pairs have one individual in common, and merge them into larger groups. This process is performed until the group formation does not share individuals, i.e. they are disjoint. In the first moment, such established groups are nominated *Temporary Groups*. These groups keep temporary if they stay stable (without inputs or outputs of agents) during at most 10% of total frames of video. After this period, if they keep the group structure, they are classified as *Permanent Groups*. Then, for each frame f, we calculated the following data:

- Number G^f of existing groups;
- Number ν_G^f of individuals that belong to any group;
- Number ξ^f of people that do not belong to any group (i.e. are alone in the scene);

In order to have information for each processed video k, we computed the average for all frames, obtaining G_k, ν_k, ξ_k. In addition we compute τ_k, which is the percentage of frames in video sequence k that contains at least one group.

Next section presents our extension to the commonly used FD, as previously discussed. Indeed, we propose to extend data normally used in FD (density $\times$ flow) to the following crowd features (densities, speeds and collectivity values) at each measurement section (ms), which is a part of the image empirically defined as a region of $6m^2$ ($\Delta x = 3$ and $\Delta y = 2$), as illustrated in Fig. 3. Consequently, at each ms we have the 3 types of data, as described in next sections.



(a) Fundamental Diagrams

Fig. 3. Illustration of image subdivision and ms where collectivity, density and speeds are computed.

1) Density (people/sqm): The density $\langle\Phi\rangle$ is defined as the number of people divided by the area of a measurement section ms at each frame, as defined in [41]:

$$\langle\Phi\rangle_{ms} = \frac{N_a}{\Delta x.\Delta y}, \tag{1}$$

where N_a is the number of people in ms and Δx and Δy are, respectively, the length and width of ms. The average density $Avg\langle\Phi\rangle^f$, at each frame f, is then calculated as follow:

$$Avg\langle\Phi\rangle^f = \frac{1}{N_{ms}} \sum_{o=1}^{N_{ms}} \langle\Phi\rangle_{ms_o}^f, \tag{2}$$

where N_{ms} is the number of measurement sections in which the image (each frame f) was divided into.

2) Speed: The Speed $\langle\Theta\rangle$, also inspired on [41], is the average of the instantaneous velocities s_i of all persons i in a measurement section ms, given by:

$$\langle\Theta\rangle_{ms} = \frac{1}{N_a} \sum_{i=1}^{N_a} s_i, \tag{3}$$

where N_a is the number of people in that measurement section ms. At each frame f, the average speed $Avg\langle\Theta\rangle^f$ is defined as:

$$Avg\langle\Theta\rangle^f = \frac{1}{N_{ms}} \sum_{o=1}^{N_{ms}} \langle\Theta\rangle_{ms_o}^f, \tag{4}$$

where N_{ms} is the number of measurement sections in frame f.

In addition to $Avg\langle\Phi\rangle^f$ and $Avg\langle\Theta\rangle^f$, we also computed the collectivity in the video sequence.

3) Collectivity: The Collectivity $\langle\Psi\rangle$, computed for each pair of people, was inspired on [10]. However, since we want to know this parameter for each agent at each time step (without considering a path in front of each one, because we compute this in real-time), we do not consider the path similarity. The collectivity is calculated as a decay function of $\varpi(i,j) = s(s_i, s_j).w_1 + o(\alpha_i, \alpha_j).w_2$, considering s and o respectively the speed and orientation differences between two people i and j, and w_1 and w_2 are constants that should regulate the offset in meters and radians. We have used $w_1 = 1$ and $w_2 = 1$. So, values for $\varpi(i,j)$ are included in interval $0 \leq \varpi(i,j) \leq 4.34$.

The collectivity of a specific measurement section ms is calculated as follow:

$$\langle\Psi\rangle_{ms} = \frac{1}{N_a^2} \sum_{i=1}^{N_a} \sum_{j=1}^{N_a} \gamma e^{(-\beta\varpi(i,j)^2)}, \tag{5}$$

where again N_a is the number of people in that measurement section ms, $\gamma = 1$ is the maximum collectivity value when $\varpi(i,j) = 0$, and $\beta = 0.3$ is empirically defined as decay constant. Hence, $\langle\Psi\rangle_{ms}$ is a value in the interval $[0; 1]$. The average collectivity $Avg\langle\Psi\rangle$ at each frame f is given by:

$$Avg\langle\Psi\rangle^f = \frac{1}{N_{ms}} \sum_{o=1}^{N_{ms}} \langle\Psi\rangle^f_{ms_o}, \tag{6}$$

where N_{ms} is the number of measurement sections of the video.

Finally, we have for each video k a vector $\vec{V_k}$ of extracted data where $\vec{V_k} = [G_k, \nu_k, \xi_k, \tau_k, Avg\langle\Phi\rangle_k, Avg\langle\Theta\rangle_k, Avg\langle\Psi\rangle_k]$. Each element of $\vec{V_k}$ is quantized into three values: small, medium and high values, in order to provide crowd classifications in video sequences. In addition to such information used in the quantization process, we computed also the standard deviation for all average values ($Std\langle\Phi\rangle, Std\langle\Theta\rangle, Std\langle\Psi\rangle$), which were also quantized. The quantization process is performed through clustering, as explained next.

4) Clustering: In order to quantize each element in $\vec{V_k}$, we used K-means clustering [42]. More precisely, we select $k = 3$ to generate exactly three clusters, and apply K-means to each individual element of $\cup_k \vec{V_k}$, i.e. the collection of vectors considering all analyzed video sequences. The class centroids are then used to quantize each element of $\vec{V_k}$ into low, medium and high values, namely S_0, S_1 and S_2. For example, a given video sequence k can have few groups ($G_k \in S_0$), a higher number of individuals ($\nu_k \in$ to S_2), and a medium value of individuals grouped ($\xi_k \in$ to S_1).

C. Mapping Crowd Features

This step is responsible for mapping the computed features into crowd characteristics. The list we are interested is detailed as follows:

- Crowd type: (Casual, Conventional, Demonstrator);
- Presence of groups, if $G > 0$;
- Size of groups: three levels of interactions (no grouping, medium, high), based on ν_G^f;
- Crowd density: three levels of crowd density (low, medium, high), based on $Avg\langle\Phi\rangle_k$; and
- Crowd interaction: three levels of interactions amount (low, medium, high), based on $Avg\langle\Psi\rangle_k$.

Our goal is to characterize crowds based on such aspects. The last four of them map directly to specific elements of $\vec{V_k}$, as explained above. The first one (crowd type), however, is related to a more subjective classification and not directly to measured data. To tackle this first aspect, we proposed some hypothesis, which are based on the studies presented by Momboisse [17] and Berlonghi [18], and detailed next:

- Hypothesis 1: *Casual crowds* should have low or medium density of people, low or medium speed, small or medium quantity of groups, low collectivity and low frequency of groups.
  - $Avg\langle\Phi\rangle_k \in S_0$ or S_1
  - $Avg\langle\Theta\rangle_k \in S_0$ or S_1
  - $G_k \in S_0$ or S_1
  - $Avg\langle\Psi\rangle_k \in S_0$
  - $\tau_k \in S_0$
- Hypothesis 2: *Conventional crowds* should have medium or high density of people, people in similar and high speeds, very small groups, most part of people walk alone and medium or high collectivity.
  - $Avg\langle\Phi\rangle_k \in S_1$ or S_2
  - $Std\langle\Theta\rangle_k \in S_0$
  - $G_k \in S_0$
  - $Avg\langle\Psi\rangle_k \in S_1$ or S_2
  - $\xi_k \in S_2$
- Hypothesis 3: *Demonstrator crowds* should have high density, small speeds, high collectivity, big groups and high frequency of groups
  - $Avg\langle\Phi\rangle_k \in S_1$ or S_2
  - $Avg\langle\Theta\rangle_k \in S_0$
  - $G_k \in S_2$
  - $Avg\langle\Psi\rangle_k \in S_1$ or S_2
  - $\tau_k \in S_2$

The high-level characterization of the crowd type given above provides the expected values for the elements of $\vec{V_k}$ in each type (casual, conventional, demonstrator). To use it in a practical system, we actually compute a weighted sum of the elements in $\vec{V_k}$, in which the weights (scores) are based on the three hypothesis provided above. They were obtained empirically, and are defined in Table I. Then, each video k has a final score for each hypothesis, and the crowd type is assigned based on the hypothesis with the highest score. The next section shows and discusses experimental results.

IV. Experimental Results

We evaluated our technique running some experiments. Initially, we performed a survey to assess the people understanding as a function of visual video information. We want to find out if numerical measured data, e.g. size of

(a) Video 1 (b) Video 2 (c) Video 3 (d) Video 4 (e) Video 5 (f) Video 6
(g) Video 7 (h) Video 8 (i) Video 9 (j) Video 10 (k) Video 11 (l) Video 12
(m) Video 13 (n) Video 14 (o) Video 15 (p) Video 16 (q) Video 17

Fig. 4. Representative frames from all videos used in our tests.

TABLE I
COMPUTED SCORES FOR EACH HYPOTHESIS ACCORDING TO THE SUBSETS OF $\vec{V_k}$ FOR VIDEO k.

Feature	Cluster	Hyp. 1	Hyp. 2	Hyp. 3
Density:	low	0.2	0	0
$Avg\langle\Phi\rangle$ and	medium	0.1	0.1	0.1
G	high	0	0.2	0.2
Avg Speed:	low	0	-	0
$Avg\langle\Theta\rangle$	medium	0.1	-	0.1
	high	0.2	-	0.2
Std Speed:	low	-	0.2	-
$Std\langle\Theta\rangle$	medium	-	0.1	-
	high	-	0	-
Collectivity:	low	0.2	0	0
$Avg\langle\Psi\rangle$	medium	0.1	0.1	0.1
	high	0	0.2	0.2
Grouped people:	low	0	-	0
ν	medium	0.1	-	0.1
	high	0.2	-	0.2
Non-grouped people:	low	-	0	-
ξ	medium	-	0.1	-
	high	-	0.2	-
Group	low	0.2	0	0
frequency: τ	medium	0.1	0.1	0.1
	high	0	0.2	0.2

groups of people, can be perceived in short sequences. The survey was composed by 17 videos and the subjects were invited to answer 5 questions about each video. The videos illustrated people walking or standing in several situations. We used videos from different countries obtained from various databases available on internet [10], [43], [44] and also filmed by the authors. Before asking the subjects, we presented some concepts that should help subjects to answer the questions, as briefly presented as follows:

- **Casual crowds** [19]: Are relatively large gatherings of people who happen to be in the same place at the same time; if they interact at all, it is only briefly. People in a shopping mall or a subway car are examples of casual crowds. Other than sharing a momentary interest, such as a clown's performance or a small child's fall, a casual crowd has nothing in common.
- **Conventional crowds** [19]: Are made up of people who come together for a scheduled event and thus share a common focus. Examples include religious services, graduation ceremonies, concerts, and college lectures. Each of these events has pre-established schedules and norms. Because these events occur regularly, interaction among participants is much more likely; People leaving events or environments can also be examples.
- **Demonstrator crowds** [18]: Are crowds who often have a recognized leader, organized for a specific reason or event, to picket, demonstrate, march, or chant.

The questions for each video from our dataset (see Fig. 4) are of multiple choice. Follow the questions and possible answers:

1) In your opinion, which of the following best describes the crowd type in above video?
   a) Casual Crowd;
   b) Conventional Crowd;
   c) Demonstrator Crowd;
   d) None of them;
   e) I don't know.
2) About groups (people walking together, to the same direction, with similar velocities), do you think the major part of people are grouped or alone in this video?
   a) Grouped;
   b) Alone;
   c) I don't know.
3) About the size of the groups (quantity of people by group), which of the following you might noticed?
   a) Small groups (group $<$ 3 people);
   b) Big groups (group $\geq$ 3 people);
   c) There are no groups;
   d) I don't know;
4) About the crowd density (quantity of people by square meter), which of the following you might noticed?

a) Low density;
b) Medium density;
c) High density;
d) I don't know.

5) In the video above, might you noticed any interaction between people?
   a) No, there are no interactions;
   b) Yes, few interactions;
   c) Yes, many interactions;
   d) I don't know.

The survey was answered by 10 people resulting in 850 responses (85 answers from each subject). The results were used to validate our approach of group definition and group features. Since each data in our method was clustered in 3 levels, we mapped the answers to the possible levels. We considered correct when the major part of subjects answer in accordance with the higher hypothesis. Fig. 5 shows the correctness of the method in comparison with the people answers.

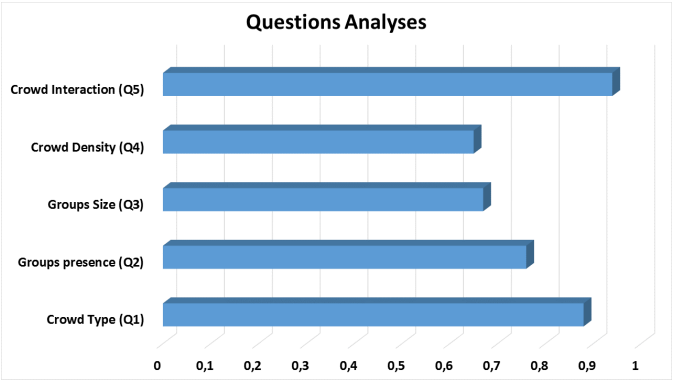


Fig. 5. Correctness of our method when compared to the subject answers.

Regarding Q1, although the correctness rate is not bad (88%), some videos were misclassified, in our opinion, by the subjects. Indeed, the classification is not obvious, mainly because the difference between the crowd types (casual x conventional) and (conventional x Demonstrator) is not very clear in the video sequences.

Concerning Q2 (76%) the question seems difficult to really measure the reality. For example, 70% of answers chosen option <Alone> for Video 5 (see Fig. 4), and numerically it is corrected (*Total number of people in the videos:* 21, *Number of Groups:* 4, *Mean people by group:* 2.25 and *Number of non-grouped people:* 12.). Our method selects <MEDIUM> for this metric, which is also correct, but not correspond to the performed question (Grouped or alone?).

We expected that Q4 (65%) was easily to visually assess the information about crowd density, however it seems to be a problem of concepts of low and medium densities, mainly when the crowd is medium density in comparison with low and high.

Concerning Q3 (67%), our method presented some errors in groups detection. For instance, in Video 15 (see Fig. 4), although people are not grouped, they spend some little time close to each other due to probable environment characteristics (door from where people are arriving and directions where they are going). In this case, the method detected groups, but subjects perceive that people are not really grouped.

Finally, participants achieved higher correction rates in Q5 (94%), indicating that people can visually assess the interaction among people.

A. Finding Similar Crowds in Videos

We also have evaluated our approach using the hypothesis values to investigate if we can detect similar crowds in the videos sequences. Each video k is represented for $H_{1,k}$, $H_{2,k}$ and $H_{3,k}$ values, as detailed in Section III and Table I. Moreover, we assumed that such hypothesis depict well each video. In order to calculate the distance between two videos from a set K of videos, we use the Mahalanobis distance. In statistics, the Mahalanobis distance is a distance measure introduced by Chandra Mahalanobis [45]. It is based on correlations between variables which different patterns. Then, we considered $\vec{H}_k$ as a vector of $(H_{1,k}, H_{2,k}, H_{3,k})$ for video k. Given $\vec{H}_k$ and $\vec{H}_m$ representing the hypothesis of two videos k and m, the dissimilarity measure between them is given by

$$D(\vec{H}_k, \vec{H}_m) = \sqrt{(\vec{H}_k - \vec{H}_m)^T S_K^{-1} (\vec{H}_k - \vec{H}_m)}, \quad (7)$$

where S_K is covariance matrix of set K.

The similarity among crowds were tested in 33 short sequences (17 of them are illustrated in Fig. 4) [1]. Using Equation 7, we computed all possible (n^2) combinations, where n is the number of videos in K set. Firstly, we present in Fig. 6 the most similar video(s) with each one (some of them present the same Malahanobis distance). Indeed, this graphic is a matrix where rows and columns represent the video index. The plot highlights videos which distances D, between the hypothesis values $(\vec{H})$, are smaller.

For instance, videos 16 and 20 (highlighted in blue ellipses in Fig. 6) are reciprocally the most similar with each other, according to the proposed metric. Indeed, the blue ellipse on the left of Fig. 6 represents the most similar crowd with video 16, i.e. video 20. In a reciprocal way, the blue ellipse on the right presents the most similar crowd with video 20, i.e. video 16. Similarly, the most similar crowds in videos 4, 6, 21 and 28 are highlighted with orange ellipses.

To provide a qualitatively assessment of this result, Fig. 7 illustrates frames of such both videos, that really seem similar. In addition to our analysis, we observed the similarity in other numerical groups information, as following described in Table II.

As highlighted in Fig. 6, videos 4, 6, 21 and 28 also present smaller Malahanobis distance among them. Fig. 8 illustrates four frames of these videos where it is possible to remark the visual similarity.

[1] Index of videos do not correspond to the set with 17 sequences.

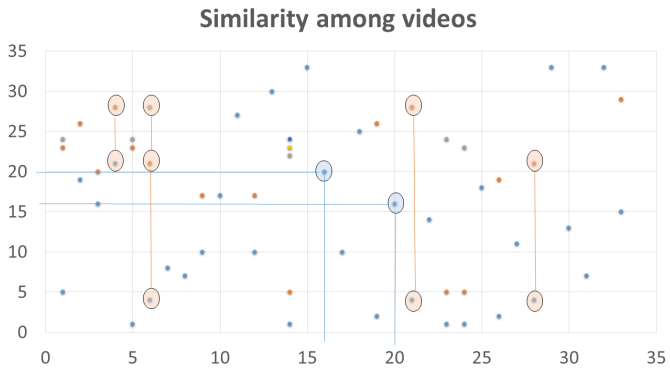
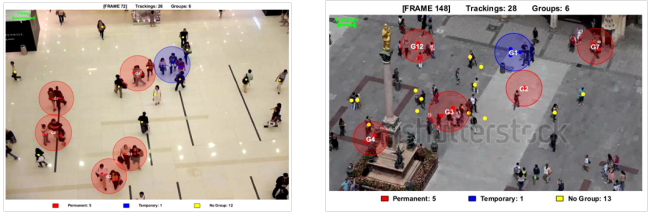


Fig. 6. Similarity among videos. Blue ellipse on the left define the most similar crowd with video 16, i.e. video 20. In a reciprocal way, the blue ellipse on the right presents the most similar crowd with video 20, i.e. video 16. Similarly, the most similar crowds in videos 4, 6, 21 and 28 are highlighted with orange ellipses.



(a) Video 16 (b) Video 20

Fig. 7. Frames of videos 16 and 20. Red circles describe permanent groups while blue circles illustrate temporary groups. Yellow dots represent non-grouped people.

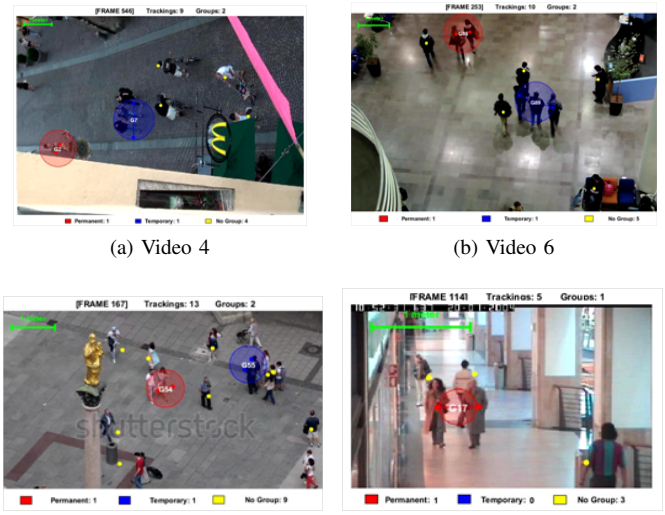
V. Conclusion

In this paper, we introduced an extension to FD to extract crowd characteristics in videos. We included the concepts of collectivity and group information to characterize crowds. We evaluated the technique comparing the results of our method with subjects answers in a survey and interesting analysis could be made, also based on how people visually understand the crowd features. In addition, we also used our metric to find out similarity between crowds in videos. Results seem promising when observing quantitative and also qualitative data.

A limitation is certainly the number of answers we had (850 for 10 persons). We intend to address this aspect in a near future. Also, the increasing of the number of videos for similarity analysis is our intention. This method could not be compared with others in literature because we did not find any technique with the same goal. As described in introduction and related works, many methods exist to detect many features, but the main type of crowds and comparison among crowds in video sequence was not found in literature, as far as we know.

TABLE II
Video information, for $k = 16$ and 20.

	Video 16	Video 20
Total number of people	26	28
Number of Groups	6	6
Mean people by group	2.33	2.55
Number of non-grouped people	12	13



(a) Video 4 (b) Video 6

(c) Video 21 (d) Video 28

Fig. 8. Frames of videos 4, 6, 21 and 28. Red circles describe permanent groups while blue circles illustrate temporary groups. Yellow dots represent non-grouped people.

Acknowledgment

The authors would like to thank FAPERGS, CNPq, CAPES in Brazil. This research is also supported by Office of National Research Global (USA).

References

[1] A. Chan and N. Vasconcelos, "Bayesian poisson regression for crowd counting," in *12th IEEE ICCV*, Sept 2009, pp. 545–551.
[2] Z. Cai, Z. L. Yu, H. Liu, and K. Zhang, "Counting people in crowded scenes by video analyzing," in *9th IEEE ICIEA*, June 2014, pp. 1841–1845.
[3] E. Ermis, V. Saligrama, P. Jodoin, and J. Konrad, "Motion segmentation and abnormal behavior detection via behavior clustering," in *15th IEEE ICIP*, Oct 2008, pp. 769–772.
[4] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *CVPR 2010*, June 2010, pp. 1975–1981.
[5] F. Solera, S. Calderara, and R. Cucchiara, "Structured learning for detection of social groups in crowd," in *10th IEEE AVSS*, Aug. 2013.
[6] J. Shao, C. Loy, and X. Wang, "Scene-independent group profiling in crowd," in *IEEE CVPR*, June 2014, pp. 2227–2234.
[7] L. Feng and B. Bhanu, "Understanding dynamic social grouping behaviors of pedestrians," *IEEE STSP*, vol. 9, no. 2, pp. 317–329, March 2015.
[8] A. Chandran, L. A. Poh, and P. Vadakkepat, "Identifying social groups in pedestrian crowd videos," in *ICAPR*, Jan 2015, pp. 1–6.
[9] B. Solmaz, B. E. Moore, and M. Shah, "Identifying behaviors in crowd scenes using stability analysis for dynamical systems," *IEEE PAMI*, vol. 34, no. 10, pp. 2064–2070, Oct. 2012.
[10] B. Zhou, X. Tang, H. Zhang, and X. Wang, "Measuring crowd collectiveness," *IEEE PAMI*, vol. 36, no. 8, pp. 1586–1599, Aug 2014.
[11] U. Chattaraj, A. Seyfried, and P. Chakroborty, "Comparison of pedestrian fundamental diagram across cultures," *Advances in Complex Systems*, vol. 12, no. 03, pp. 393–405, 2009.

[12] U. Weidmann, "Transporttechnik der fussgnger," Institut fr Verkehrsplanung, ETH Zurich, Tech. Rep., 1993.

[13] A. I. Predtechenskii V. M., Milinskii, *Planning for foot traffic flow in buildings*. New Delhi : Amerind,, 1978.

[14] N. Fridman, G. A. Kaminka, and A. Zilka, "The impact of culture on crowd dynamics: An empirical approach," in *AAMAS*, Richland, SC, 2013, pp. 143–150.

[15] B. Zhan, D. Monekosso, P. Remagnino, S. A. Velastin, and L. Xu, "Crowd analysis: a survey," *MVA*, vol. 19, no. 5-6, pp. 345–357, 2008.

[16] W. G., R. T. C., and B. R., "Vision-based analysis of small groups in pedestrian crowds," *IEEE PAMI*, vol. 34, no. 5, pp. 1003–1016, 2012.

[17] R. Momboisse, *Riots, revolts, and insurrections*. C. C. Thomas, 1967.

[18] A. E. Berlonghi, "Understanding and planning for different spectator crowds," *Safety Science*, vol. 18, no. 4, pp. 239 – 247, 1995, engineering for Crowd Safety.

[19] D. Kendall, *Sociology in Our Times*. Cengage Learning, 2016.

[20] M. S. Greenberg, *Mob Psychology*. John Wiley & Sons, Inc., 2010.

[21] K. H. Lee, M. G. Choi, Q. Hong, and J. Lee, "Group behavior from video: A data-driven approach to crowd simulation," in *Proceedings of the 2007 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, ser. SCA '07. Aire-la-Ville, Switzerland, Switzerland: Eurographics Association, 2007, pp. 109–118.

[22] A. Lerner, Y. Chrysanthou, and D. Lischinski, "Crowds by example," *Computer Graphics Forum*, vol. 26, no. 3, pp. 655–664, 2007.

[23] M. Paravisi, A. Werhli, J. C. S. Jacques Jr., R. Rodrigues, C. J. A. Bicho, and S. R. Musse, "Continuum crowds with local control," in *Proceedings of the 2008 - Computer Graphics International 2008*, ser. CGI'08, 2008.

[24] M. Hu, S. Ali, and M. Shah, "Learning Motion Patterns in Crowded Scenes Using Motion Flow Field," in *Proc. International Conference on Pattern Recognition (ICPR)*, 2008.

[25] J. Pettré, J. Ondřej, A.-H. Olivier, A. Cretual, and S. Donikian, "Experiment-based modeling, simulation and validation of interactions between virtual walkers," in *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, ser. SCA '09. New York, NY, USA: ACM, 2009, pp. 189–198.

[26] J. C. S. J. Junior, S. R. Musse, and C. R. Jung, "Crowd analysis using computer vision techniques," *IEEE Signal Processing Magazine*, vol. 27, no. 5, pp. 66–77, Sept 2010.

[27] J. Zhang, "Pedestrian fundamental diagrams: Comparative analysis of experiments in different geometries," Dr., Universitt Wuppertal, Jlich, 2012, universitt Wuppertal, Diss., 2012.

[28] J. Sochman and D. C. Hogg, "Who knows who - inverting the social force model for finding groups," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, Nov 2011, pp. 830–837.

[29] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg, "Who are you with and where are you going?" in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, June 2011, pp. 1345–1352.

[30] W. Ge, R. T. Collins, and B. Ruback, "Automatically detecting the small group structure of a crowd," in *Applications of Computer Vision (WACV), 2009 Workshop on*, Dec 2009, pp. 1–8.

[31] A. Seyfried, M. Boltes, J. Kähler, W. Klingsch, A. Portz, T. Rupprecht, A. Schadschneider, B. Steffen, and A. Winkens, *Pedestrian and Evacuation Dynamics 2008*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, ch. Enhanced Empirical Data for the Fundamental Diagram and the Flow Through Bottlenecks, pp. 145–156.

[32] D. Helbing, A. Johansson, and H. Z. Al-Abideen, "Dynamics of crowd disasters: An empirical study," *Phys. Rev. E*, vol. 75, p. 046109, Apr 2007.

[33] S. Narang, A. Best, S. Curtis, and D. Manocha, "Generating pedestrian trajectories consistent with the fundamental diagram based on physiological and psychological factors," *PLoS ONE*, vol. 10, no. 4, 2015.

[34] A. Best, S. Narang, S. Curtis, and D. Manocha, "Densesense: Interactive crowd simulation using density-dependent filters," in *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, ser. SCA '14. Aire-la-Ville, Switzerland, Switzerland: Eurographics Association, 2014, pp. 97–102.

[35] D. Wolinski, S. J. Guy, A.-H. Olivier, M. Lin, D. Manocha, and J. Pettré, "Parameter estimation and comparative evaluation of crowd simulations," *Computer Graphics Forum*, vol. 33, no. 2, pp. 303–312, 2014. [Online]. Available: https://hal.inria.fr/hal-01059493

[36] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE CVPR*, vol. 1, 2001, pp. I–511–I–518 vol.1.

[37] D. Tosato, M. Spera, M. Cristani, and V. Murino, "Characterizing humans on riemannian manifolds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1972–1984, 2013.

[38] J. Bins, L. L. Dihl, and C. R. Jung, "Target tracking using multiple patches and weighted vector median filters," *MIV*, vol. 45, no. 3, pp. 293–307, Mar. 2013.

[39] K. Fukunaga, *Introduction to statistical pattern recognition (2nd ed.)*. San Diego, CA, USA, 1990: Academic Press Professional, Inc., 1990.

[40] E. Hall, *The Hidden Dimension*, ser. A Doubleday anchor book. Anchor Books, 1990.

[41] J. Zhang, W. Klingsch, A. Schadschneider, and A. Seyfried, "Transitions in pedestrian fundamental diagrams of straight corridors and t-junctions," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2011, no. 06, p. P06004, 2011.

[42] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, L. M. L. Cam and J. Neyman, Eds., vol. 1. University of California Press, 1967, pp. 281–297.

[43] R. Fisher, *CAVIAR: Context Aware Vision using Image-based Active Recognition*, 2016 (accessed May 13, 2016), http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.

[44] M. Rodriguez, J. Sivic, I. Laptev, and J.-Y. Audibert, "Data-driven crowd analysis in videos," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.

[45] P. C. Mahalanobis, "On the generalised distance in statistics," in *Proceedings National Institute of Science, India*, vol. 2, no. 1, Apr. 1936, pp. 49–55.