AGORASET: a dataset for crowd video analysis

Pierre Allain, Nicolas Courty, Thomas Corpetti

AGORASET : a dataset for crowd video analysis

Pierre Allain
IRISA, UBS
Vannes, France

Nicolas Courty
IRISA, UBS, and Institute of Automation
Beijing, China

Thomas Corpetti
LIAMA, CNRS
Beijing, China

Abstract

The ability of efficient computer vision tools (detection of pedestrians, tracking, ...) as well as advanced rendering techniques have enabled both the analysis of crowd phenomena and the simulation of realistic scenarios. A recurrent problem lies in the evaluation of those methods since few common benchmark are available to compare and evaluate the techniques is available. This paper proposes a dataset of crowd sequences with associated ground truths (individual trajectories of each pedestrians inside the crowd, related continuous quantities of the scene such as the density and the velocity field). We chosed to rely on realistic image synthesis to achieve our goal. As contributions of this paper, a typology of sequences relevant to the computer vision analysis problem is proposed, along with images of sequences available in the database.

1 Introduction

Crowd phenomena are ubiquitous, and their scale and frequency is likely to increase as human population grows and the urbanization rate raises. Analyzing crowd footages is of prime interest in a variety of problems: it can help to *i)* better understand crowd dynamics or the related behaviors [12], *ii)* develop surveillance or crowd control systems [19], *iii)* design and configure public spaces or as well *iv)* enrich computer animation models used in the context of video games or special effects [7, 17]. Though, as explained by Zhan and colleagues [32], conventional computer vision methods used in the context of tracking fail to analyze crowded situations. Principal reasons arise from several factors. Despite the inherent computational complexity of handling several individuals, the related pixel information is rather poor, and undergo multiple occlusions over time. Moreover, modeling the physical nature of a crowd is a strongly non-trivial task, as it implies inter-indivudal and environments interactions, with pedestrians exhibiting different goals or implied in social interactions.

Both the large applications field and the challenging vision problems have led to the development of an important number of new vision algorithms over the past decades. The related methods either seek to count or track individuals or to detect changes in the crowd flow or abnormal patterns. At the moment, the community is lacking of a common test bed and reference situations (with eventually associated ground truths). Though, it has been already shown in other context such as, for example, object recognition [10], optical flow [4] or articulated motion estimation [27] that datasets greatly stimulates the research field and allows for direct, objective comparisons between state-of-the-art algorithms. Our contribution is in that direction and aims at providing a variety of crowd situations along with their associated ground truths. To account for the different possible representations of crowd phenomena, these ground truths contain: the individual trajectories of each pedestrians in the crowd and the related continuous quantities such as density and dense velocity field.

No real crowd videos were included in the data set. The main reason is that obtaining ground truths for those videos is a very time-consuming task as it requires to label by hand the positions of each individuals in the scene. Moreover, experiments show that manual labeling is prone to errors and can differ between two persons. We instead rely on realistic image synthesis to achieve our goal. This idea is not new and has already been exploited in the context of surveillance of human activity [23, 28, 21]. Still, our data set constitutes the first of this kind devoted to crowd phenomena. Obtaining realistic synthetic crowd videos is in itself a challenge. Regarding crowd video analysis requirements, two major problems are to be taken into account: the visual quality of the images should reflect the diversity that can be observed in real footages and the dynamics of the crowd should be preserved. This last point include both details of single pedestrians gaits and motions, as well as the overall continuum dynamics. In

our work the rendering has been performed thanks to a commercially available renderer and a classical and well established simulation model is used.

The remainder of the paper is decomposed as follows: after a presentation of related work in crowd analysis and synthesis, Section 3 describes the typology of situations available in the dataset along with the reasons for this choice. In a second time, Section 4 describes the production pipeline, with emphasis on both the rendering process and the crowd simulation model used. Section 5 concludes the paper by a discussion on possible exploitation of the dataset as well as future improvements.

2 Crowd video analysis

This section investigates the principal axes carried out by researchers to deal with crowd video analysis. A full review is out of the scope of this paper and for that purpose, we refer the readers to [31] and to associated references. As mentioned above, the inherent diversity and complexity of the behavior of a mass of people makes ambiguous the question of representing and modeling a crowd. In the two last decades, two main strategies have been carried out by the different authors:

1. representation at the *pedestrian* level: each individual is an entity/particle driven in a *Lagrangian* framework. The crowd representation results from the combination of a large number of entities. In the following of this paper we denote as "Lagrangian" such approaches;

2. the crowd is modeled in a *continuous* framework related to some scalar/vectorial characteristic quantities (density or displacement field for instance). The governing equations are represented in an *Eulerian* context and the individual notion vanishes. We denote as "Eulerian" such techniques;

Crowds as a set of individuals. Within this class of method, the scene is modeled as a collection of pedestrians that interact with their environment (obstacles). In general, the associated analysis techniques rely on low-level vision (background subtraction, edge and object detection) that enables a human counting and eventually an action recognition. The former appearance models have been based on low level features like an edge map [6, 25, 30]. This latter is used afterward within a more advanced strategy, like a neural network or a probabilistic tracking approach [15], to segment, regroup and evaluate the number or individuals. In a step forward, some human detection approaches have been defined at an object level. The humans are first detected with an ad-hoc technique of head or body recognition. This relies on a model either based on the appearance of the humans (distribution of some functions based on the luminance) [9, 18, 33], on a 3D body model [33] or on the velocity of entities [24]. An additional post-processing step for classification (clustering, SVM, ...) enables to count and sometimes to characterize the nature of the motion.

Even if some of the mentioned technique exhibit very competitive results, when a large number of pedestrians are present in the crowd (> 50), most of the conventional tracking methods (like Kalman trackers) fail. In such cases, the degradation of the visual features related to single individuals disturbs the analysis. Moreover, the large induced state space yields computationally too expensive problems. In those situations, the Lagrangian approach fails and the analysis of the crowd sequence may amount to the analysis of a crowd flow that have global properties and may be treated as a whole. As an example, one can cite the recent work of Rodriguez and al. [26] who use the global motion of the crowd (the crowd behavior) to help the tracking.

Crowds as a continuous entity. The representation of crowd flows in a pure Eulerian approach has been studied in [14]. In this study, the author creates some links between dense crowds and fluid mechanics laws. Two flow regimes (e.g. high-density and low-density) have been proposed in a complete dynamical model that depends on some objective parameters (optimal orientation to reach the goal, pressure, velocity) and on some more subjective quantities like the crowd comfort. The experimental simulations have been successfully compared to some real scenes. In a context of crowd simulation, the authors in [29] present a real-time crowd model based on continuum dynamics. In [2], the authors have proposed a somehow simpler continuous dynamical model where pedestrians are assumed to reach an objective while interacting together in order to prevent from the formation of too dense areas. A disturbance potential has also been introduced to deal with more subjective interactions inside the flow. This dynamical model is then used as a prior information for analyzing crowded videos in an optimal control theory framework.

On the basis of this constatation that a crowd can be managed with continuous laws, several analysis techniques based on a continuum approach have been proposed [19, 7, 2, 1, 3, 5]. Related methodologies usually tend to solve the different problems of event detection or changes in the flow rate on the basis of the apparent motion (optical flow) estimated on the the whole im-

age from the image luminance. The work of Ali and Shah [1] focuses on segmenting the crowd flow with regions of substantially different dynamics by examining the coherent structures in the flow. In [3], unsupervised feature clustering is used to define normal motion patterns, and Hidden Markov Models are used to detect particular situations. A similar approach has been proposed in [19] where normal and abnormal behaviors are extract from the continuous optical-flow. This displacement indicator is also a prior descriptor to highlight circular and diverging flows in [5]. Finally, the authors in [7] have suggested to use an optical-flow estimation approach devoted to the measurement of fluid flows.

3 Presentation of the dataset

In this first version of the proposed dataset, we have identified seven typical scenes where some crowd behaviors appears. They are schematized in the figure 1. Each scenario topology (also named "environment" or "scene" in the rest of the paper), was designed in accordance to situations often met in crowding issues. They correspond to an evolution on a flow of humans in a free environment (scene #1), in an environment with obstacles (#2 and #3), an evacuation through a door (#4), a dispersion (#5), a rotation (#6) (with an analogy of the famous crowd scene of the pilgrimage in Mekkah) or some crossing flows (#7) (this last case being related to the "unstructured crowd" of [26]).

From the scenes we have depicted in figure 1, several scenarios have been generated. For each environment, two different sequences that correspond to various values of desired velocity (soft and panic) have been generated. In each case, the pedestrian positions are randomly set in a starting area and are not submitted to any motion during $3\,s$. After this delay, the wished direction is included in the simulation model. Concerning the rendering, several videos are also available for a single event. They correspond to various camera parameters and lighting conditions. For now we propose for analysis two camera views : perfect sky, and sided view. And two lighting conditions : shading, and no shading. Thus allowing a variety of rendering realism for a same scenario.

Let us now turn to the production pipeline of the different video of the dataset.

4 Production pipeline

As a first step the crowd simulation model is presented, followed by a short description of the rendering process.

4.1 Crowd simulation model

Because of its compacity and efficiency, we chose to use the model proposed by Helbing *et al.* [11] for crowd simulation. It considers pedestrians as Lagrangian particles carrying individual properties affecting their dynamic through different crowd-related forces. A complete presentation of the model is out of the scope of the paper and the reader interested in the final output data for motion analysis is not necessarily compelled to investigate it. This model has nevertheless been widely used for different purpose involving crowd dynamics as in [22, 20], or revisited as in [16].

As the latter authors, we also chose to slightly revisit the model in order to reach as much as possible visual realism. The original purpose of Helbing *et al.* model is indeed to match different macro data of the crowd like global evacuation time. The model's parameters are thus not specifically calibrated to handle visual realism. The particles are too solid for having the visual impression of people being pressed in case of congestion. In this purpose we modify several parameters which are given in Table 1.

		Pedestrians interaction	Obstacles interaction	*Units*
Contact	k	6×10^3	6×10^3	$kg.s^{-2}$
	κ	12×10^3	12×10^3	$kg.m^{-1}.s^{-1}$
Avoidance	a	600	400	N
	b	0.3	0.3	m

Table 1. Modified parameters of the crowd model proposed by Helbing *et al.* These values lead to more flexible behaviors of individuals.

The sequences to be simulated also need to provide enough variability to show realistic crowd behavior. In this purpose, the dynamic parameters are randomly picked for each pedestrian: in addition to the radius r being set in the interval $[0.2\,m, 0.3\,m]$, the mass m is taken as proportional to the latter with an average of $80\,kg$. The response time of pedestrians is in the interval $0.5\,s \pm 10\%$ while the desired nominal velocity is in $1.2\,m.s^{-1} \pm 10\%$. With these parameters, we assume the crowd is populated by different kinds of persons having little different intentions and capabilities, as in a real crowd flow. As in [13], pedestrians handle differently the social forces wether neighbors are in sight of pedestrians or not. We therefore add it a perception coefficient $\alpha = 1 - \left(\frac{\text{acos}(\mathbf{e}_{in}\mathbf{W}_i)}{\pi}\right)^{\gamma}$, with $\mathbf{e}_{in}$ being the unit vector from pedestrian i to the neighbor n, and $\mathbf{W}_i$ the desired direction to take. This coefficient

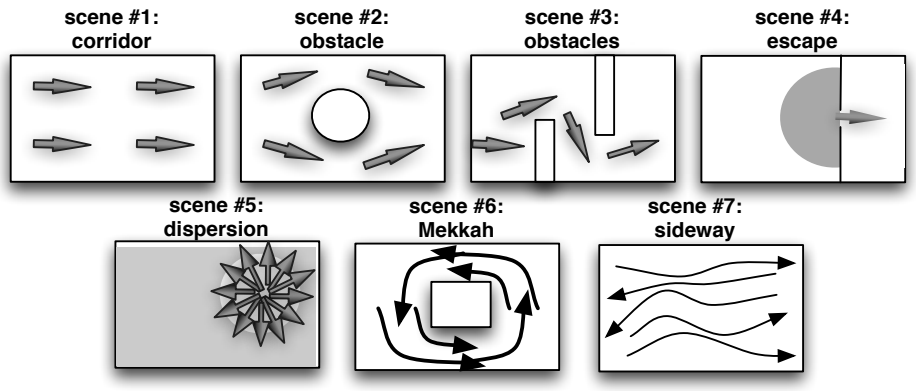

Figure 1. **Scene typology**. The different scenes proposed in our database

also accounts for the anisotropic dynamic behavior of the crowd that we parameterize with $\gamma = 0.7$. In order to provide visual realism in panic situations, we also add a stumble term to the dynamic velocity equation, reading: $-\psi \mathbf{v}_i \mathbf{v}_i^2 g(\mu(r_i + r_n) - d_{in})$, where $\mathbf{v}_i$ is the velocity, d_{in} the distance between the jostling pedestrians, and $g(x)$ a function being zero if $x > 0$ and x otherwise. We set $\psi = 300\, kg.s.m^{-3}$ and $\mu = 1.2$. The pedestrians are also supposed to be reluctant to go backward from their goal. Their speed limitation when going this (wrong) way is then augmented by adding to the dynamic velocity equation the following term: $\nu m_i \frac{\mathbf{v}_i}{\tau_i} g(\frac{\mathbf{v}_i}{\|\mathbf{v}_i\|} \mathbf{W}_i)$, with τ_i being the characteristic reaction time and ν a reluctance coefficient set to 2.

4.2 Rendering process

The output of the simulation model are then used as input for python scripts that automatically generates a 3D scene with human characters. A set of 26 characters were used (see Figure 2, left) in order to guarantee a sufficient local variability in shapes and colors. This number was set experimentally, but it turned out that a lower number of individuals raised notably the probability of having two or more of the same kind of geometric models near each other, with a possibility of disturbing the analysis. A short number of walking motions and idling gaits have been used for each individuals. At runtime, the best motion is chosen with respect to the pedestrian velocity. Here again, the inter-individual diversity is assured by different playback speeds, which prevents from several individuals having the same motion (also known as the clone effects).

For the rendering process we chose the Mental Ray renderer [8]. The mental ray physical sky model was used; it allows to have a natural and intuitive control over the illumination parameters. Most of the scenes were rendered with outdoor lighting conditions. A comparison between a real image extracted from a real video sequence and a rendering of a crowd scene with similar lighting conditions is shown in Figure 2. It illustrates the ability of our rendering pipeline to produce images that qualitatively looks like real ones.

We present in figures 3 some screenshots that rely on simulations #4-1. As for this environment, pedestrians aim at reaching the right part of the scene by crossing a door, under a "normal" pace (fig. 3). The first line of these figures correspond to a streak representation of the pedestrian trajectories whereas on the second line is depicted the rendered scene under a given camera position. From these figures, it is very interesting to observe that the rendering is very realistic. Moreover, the emerging phenomenon highlighted in [11] that correspond to an "arching" or "clogging" effect near the exit appears clearly. This suggests that, in addition to a realistic rendering of the simulations, the crowd behaviors are consistent.

5 Discussion/Conclusion

In this paper we have proposed a dataset of crowd videos with associated ground truths. The sequences have been generated using a simulation model based on Lagrangian forces as proposed by Helbing *et al.*. This model can be easily parametered to simulate several types of crowd behaviors like calm or panic situations. Moreover, a realistic rendering process has enabled to highlight lifelike sequences. Such simulations could allow researchers to work on common crowd sequences and to compare their results in a common framework. Upon publication of this article, the dataset that includes the mentioned video sequences will be available on the following web site: `http:///www-irisa.univ-ubs.fr/Nicolas.Courty/Agoraset`. All the asso-

Figure 2. Crowd Rendering. From left to right: the 26 different avatars used to produce the videos (their choice has been made to exhibit the maximum variability w.r.t. age, sex and cloths style); a real image from a video footage of Shibuya in Tokyo; a crowd rendering with similar day light conditions

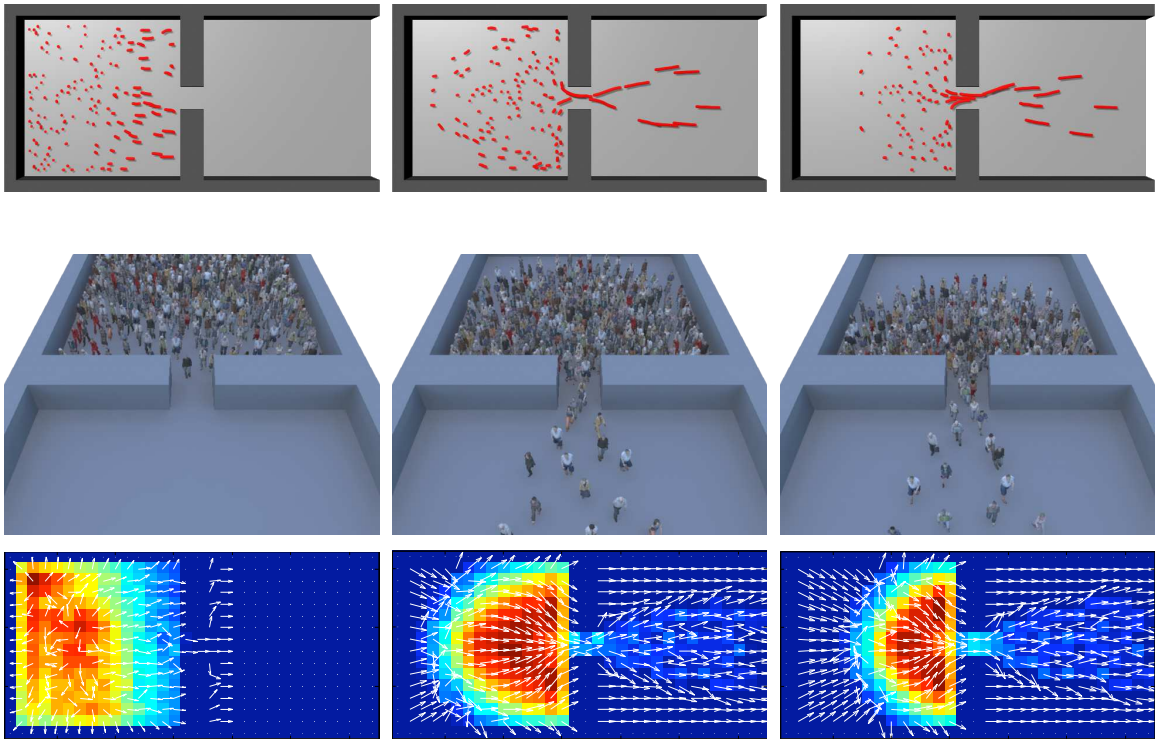

Figure 3. Scenario #4-1 : Evacuation in a normal situation. First strip, Trajectories of one individual out of 3 for a 3 seconds duration. Second strip, screenshots of the corresponding video. Third strip, the related continuous parameters (density and velocity field).

ciated ground truths and routines required for the manipulation of data (C++ and matlab) will also be given. As for the evolution of the dataset, we plan to supply some more sequences, and we may consider opening an evaluation platform, inspired from the one in [4], where some sequences without ground truth will be proposed to the community.

Acknowledgments

This work was partially supported by Region Bretagne (ASFoule project) and a Chinese Academy of Sci-

ences visiting professorship for senior international scientists grant.

References

[1] S. Ali and M. Shah. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *CVPR*, pages 1–6, Minneapolis, Minnesota, June 2007.

[2] P. Allain, N. Courty, and T. Corpetti. Crowd flow characterization with optimal control theory. In *ACCV*, Xi'an, China, 2009.

[3] E. Andrade, S. Blunsden, and R. Fisher. Modelling crowd scenes for event detection. In *ICPR*, pages 175–178, Washington, DC, USA, 2006.

[4] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski. A database and evaluation methodology for optical flow. In *ICCV*, pages 1–8, Rio de Janeiro, Brazil, Oct. 2007.

[5] B. A. Boghossian and S. A. Velastin. Motion-based machine vision techniques for the management of large crowds. In *ICECS*, volume 2, pages 961–964 vol.2, August 2002.

[6] S.-Y. Cho, T. Chow, and C.-T. Leung. A neural-based crowd estimation by hybrid global learning algorithm. *IEEE Tra. on SMC*, 29(4):535 –541, aug 1999.

[7] N. Courty and T. Corpetti. Crowd motion capture. *Computer Animation and Virtual Worlds*, 18(4–5):361–370, 2007.

[8] T. Driemeyer. *Rendering with mental ray*. Springer, New York, 2001.

[9] A. Elgammal and L. S. Davis. Probabilistic framework for segmenting people under occlusion. In *In Proc. of IEEE 8th International Conference on Computer Vision*, pages 145–152, 2001.

[10] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *PAMI*, 28(4):594–611, 2006.

[11] D. Helbing, I. Farkas, and T. Vicsek. Simulating dynamical features of escape panic. *Nature*, 407(1):487–490, 2000.

[12] D. Helbing, A. Johansson, and H. Al-Abideen. Dynamics of crowd disasters: An empirical study. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 75(4):046109, 2007.

[13] D. Helbing and P. Molnar. Social force model for pedestrian dynamics. *PHYSICAL REVIEW E*, 51:4282, 1995.

[14] R. L. Hughes. A continuum theory of pedestrian motion. *Transportation Res. B*, 36(6):507–535, June 2002.

[15] E. Kalogerakis, O. Vesselova, J. Hays, A. Efros, and A. Hertzmann. You'll never walk alone: modeling social behavior for multi-target tracking. In *ICCV*, Kyoto, Japan, 2009.

[16] T. I. Lakoba, D. J. Kaup, and N. M. Finkelstein. Modifications of the helbing-molnár-farkas-vicsek social force model for pedestrian evolution. *Simulation*, 81(5):339–352, May 2005.

[17] K. Lee, M. Choi, Q. Hong, and J. Lee. Group behavior from video: a data-driven approach to crowd simulation. In *ACM SIGGRAPH/Eurographics Symp. on Computer Animation, SCA'07*, pages 109–118, San Diego, California, Aug. 2007.

[18] S.-F. Lin, J.-Y. Chen, and H.-X. Chai. Estimation of number of people in crowded scenes using perspective transformation. *IEEE Tra. on SMC*, 31(6):645–654, 2001.

[19] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *CVPR*, pages 935–942, Los Alamitos, CA, USA, 2009.

[20] M. Moussaïd, D. Helbing, S. Garnier, A. Johansson, M. Combe, and G. Theraulaz. Experimental study of the behavioural mechanisms underlying self-organization in human crowds. *Proceedings of the Royal Society B: Biological Sciences*, 276(1668):2755–2762, 2009.

[21] S. Musse, M. Paravisi, R. Rodrigues, J. J. Jr, and C. Jung. Using synthetic ground truth data to evaluate computer vision techniques. In *Proc. of int. workshop on PETS*, pages 25–32, Rio de Janeiro, Brazil, Oct. 2007.

[22] S. R. Musse, C. R. Jung, J. C. S. Jacques, and A. Braun. Using computer vision to simulate the motion of virtual agents. *Computer Animation and Virtual Worlds*, 18(2):83–93, 2007.

[23] F. Qureshi and D. Terzopoulos. Surveillance in virtual reality: System design and multi-camera control. In *CVPR*, Minneapolis, Minnesota, USA, June 2007.

[24] V. Rabaud and S. Belongie. Counting crowded moving objects. In *CVPR*, pages 705–711, New York, June 2006.

[25] C. S. Regazzoni and A. Tesei. Distributed data fusion for real-time crowding estimation. *Signal Processing*, 53(1):47 – 63, 1996.

[26] M. Rodriguez, S. Ali, and T. Kanade. Tracking in unstructured crowded scenes. In *ICCV*, pages 1–8, Kyoto, Japan, Oct. 2009.

[27] L. Sigal, A. Balan, and M. J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 87(1):4–27, Mar. 2010.

[28] G. R. Taylor, A. J. Chosak, and P. C. Brewer. Ovvv: Using virtual worlds to design and evaluate surveillance systems. In *CVPR*, Minneapolis, Minnesota, USA, June 2007.

[29] A. Treuille, S. Cooper, and Z. Popovic. Continuum crowds. *ACM TOG., Proc. ACM SIGGRAPH 2006*, 25(3):1160–1168, 2006.

[30] D. Yang, H. González-Ba nos, and L. Guibas. Counting people in crowds with a real-time network of simple image sensors. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 122, Washington, DC, USA, 2003. IEEE Computer Society.

[31] B. Zhan, D. Monekosso, P. Remagnino, S. Velastin, and L.-Q. Xu. Crowd analysis: a survey. *Mach. Vision Appl.*, 19(5-6):345–357, 2008.

[32] B. Zhan, D. N. Monekosso, P. Remagnino, S. A. Velastin, and L.-Q. Xu. Crowd analysis: a survey. *Mach. Vis. Appl.*, 19(5-6):345–357, 2008.

[33] T. Zhao and R. Nevatia. Bayesian human segmentation in crowded situations. In *CVPR (2)*, pages 459–466, 2003.