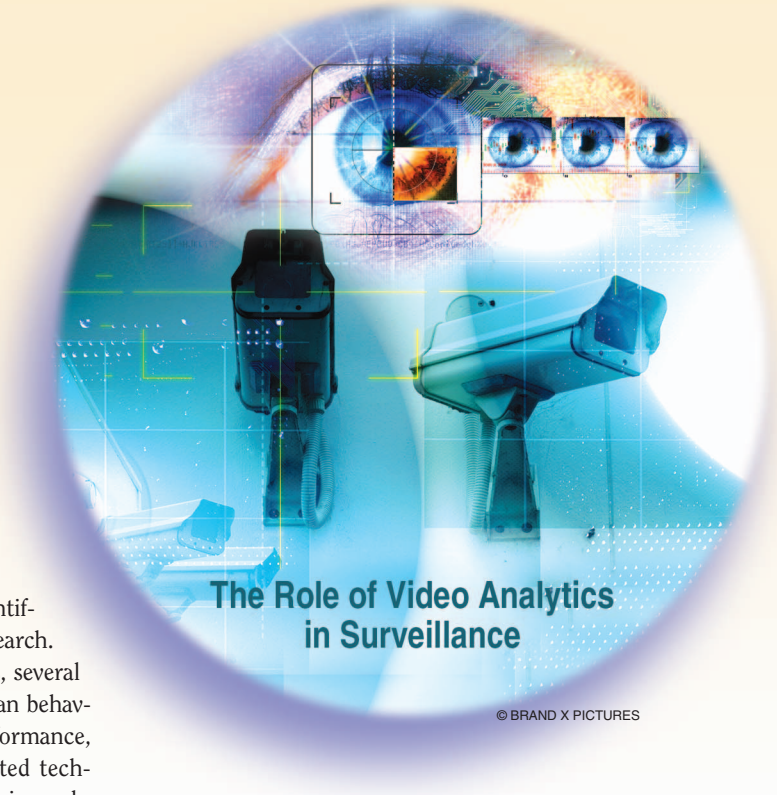
Julio Cezar Silveira Jacques Junior, Soraia Raupp Musse, and Cláudio Rosito Jung

# Crowd Analysis Using Computer Vision Techniques

## A survey




This article presents a survey on crowd analysis using computer vision techniques, covering different aspects such as people tracking, crowd density estimation, event detection, validation, and simulation. It also reports how related the areas of computer vision and computer graphics should be to deal with current challenges in crowd analysis.

### INTRODUCTION AND MOTIVATION

The study of human behavior is a subject of great scientific interest and probably an inexhaustible source of research. With the improvement of computer vision techniques, several applications in this area, like video surveillance, human behavior understanding, or measurements of athletic performance, have been tackled using automated or semiautomated techniques. However, several complex challenges still remain, making this subject relevant in terms of research.

Currently, there are commercial systems developed to track (e.g., www.smarteye.se), recognize (e.g., www.iteris.com), and understand the behavior of a great variety of objects, using one or multiple video cameras, processing the information in one or more computers. Also, there are several scientific papers on the subject (an overview of video-based human motion analysis can be found in [1]).

In terms of recognition, monitoring and behavior analysis of people, an important research topic is their detection/identification, considering that one person could be occluded in many ways (particularly when lateral cameras are employed). In a high-level analysis, one could segment their body parts (e.g., head, face, hands, and arms) that could be used in gesture recognition or machine-human interaction, for example. It is also possible to deal with group behavior, once people can be part of high-level structures, such as groups or crowds [2], [3]. In particular, the behavioral analysis of crowded scenes is of great interest in a large number of applications [4], such as

- *Crowd Management*: It can be used for developing crowd management strategies, to avoid crowd related disasters and insure public safety.
- *Public Space Design*: To provide guidelines for the design of public spaces.
- *Virtual Environments*: It can be used to validate or increase the performance of the mathematical models used in crowd simulations.
- *Visual Surveillance*: It can be used for automatic detection of anomalies and alarms.
- *Intelligent Environments*: It can be used to take a decision on how to split a crowd in a museum, based on their behavior.

For instance, automatic monitoring of crowds exiting stadiums or dense public areas could help the detection of

Digital Object Identifier 10.1109/MSP.2010.937394



1053-5888/10/$26.00©2010IEEE

strangle points, so that safety engineers could suggest modifications in the environment to improve the flow of people. In surveillance applications, the video streams captured by an increasing number of cameras monitoring public spaces must be watched by a limited number of human observers, and computer vision algorithms could be used to detect anomalous events and warn the observers.

**ALTHOUGH ONE MAY THINK THAT A STRAIGHTFORWARD EXTENSION OF TECHNIQUES DESIGNED FOR NONCROWDED SCENES COULD BE SUITABLE FOR DEALING WITH CROWDED SITUATIONS, THAT IS NOT TRUE.**

Despite the potential of crowd analysis applications using computer vision algorithms, most of the existing work for detection/identification of people, groups of people, or even for the estimation of body parts (in a high-level analysis) have been focused on noncrowded situations. As related by Zhan et al. [4], when very crowded video sequences are analyzed, conventional computer vision methods are not appropriate.

Although one may think that a straightforward extension of techniques designed for noncrowded scenes could be suitable for dealing with crowded situations, that is not true. First of all, a crowd is something beyond a simple sum of individuals. The crowd can assume different and complex behaviors as those expected by their individuals. The behavior of crowds is widely understood to have collective characteristics that can be described in general terms. For example, descriptions such as "an angry crowd," or a "peaceful crowd" are well accepted [5]. Further details about crowd dynamics are presented in the section "Crowd Dynamics."

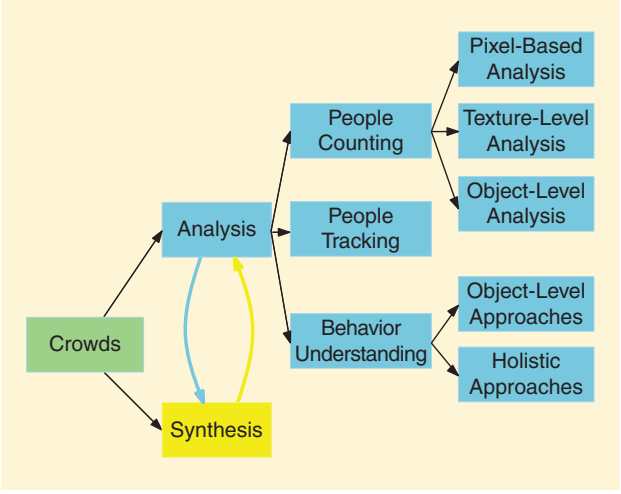
Dealing with crowds also introduces several additional challenges to computer vision techniques used for human behavior understanding. In general, tracking is the first stage in systems for automatic analysis of human behavior. In a crowded situation, it is difficult to segment and track accurately each individual, mostly due to severe occlusions. In fact, when high-density video sequences are employed, the accuracy of traditional methods for object tracking tends to decrease as the density of people increases. In such cases, new methods should be developed to handle these constraints.

For these reasons, an alternative approach is to treat the crowd as a single entity, instead of analyzing each individual in a crowd. This type of analysis is mostly used in two main applications: to measure the level of crowd comfort and to detect some specific events, as described in [6]–[8], e.g., unusual group behavior. In general, the goal of crowd analysis techniques based on computer vision is to extract some kind of information from crowded video sequences that could be used to benefit a large number of applications (e.g., surveillance, design of densely populated public spaces, and safety analysis of crowds exiting sports arenas).

Another important and challenging problem related to the crowd phenomenon is crowd simulation, which relates to the reproduction of realistic crowds based on computer graphics techniques. Animations of crowds find applications in many areas, including entertainment (e.g., animation of large numbers of people in movies and games), creation of immersive virtual environments, and evaluation of crowd management techniques (for instance, simulation of the flow of people leaving a football stadium after a match).

As related by Zhan et al. [4], crowd information can be better exploited to indicate the status of the crowd so that crowd events can be inferred. Crowd models have been built to represent this status, either implicitly or explicitly. Although crowd synthesis and analysis have been done in a relatively independent manner in the past, the integrated use of computer vision and computer graphics is bringing these two problems closer together. For instance, an important problem related to crowd analysis using computer vision is validation. It is not easy to find complete ground-truth data of real crowded scenes (particularly in emergency situations), and data generated by computer graphics techniques could be used to validate/train computer vision algorithms [9]–[11]. On the other hand, crowd simulation models can be more realistic if they rely on information captured from the real world [12]–[14]. For example, in the work of Musse et al. [12], the trajectories of individuals in noncrowded situations are automatically captured using computer vision techniques, and used to generate extrapolated velocity maps that feed the simulation model with main directions and velocities in the filmed environment. Although many crowd simulation methods have been proposed in past years, open challenges still exist due to the complexity of human behavior and several degrees of freedom present in real life scenarios. When using crowd simulation techniques as ground truth for computer vision algorithms, some questions such as "What level of realism can be achieved in crowd simulations?" and "What behaviors can be realistically simulated using crowd simulators?" should be raised.

This work presents a survey on crowd analysis techniques based on computer vision, pointing out some open problems that could be further investigated and future directions. It should be noted that survey papers on crowd analysis have been proposed in the past, such as [4]. However, Zhan et al. [4] reviewed some crowd density estimators and crowd modeling techniques, focusing on object detection, recognition, and tracking in cluttered scenes (that are generally employed in pedestrian detection/tracking with occlusion handling), but they did not specifically tackle the problem of crowd behavior understanding, which is covered in this survey. We also present some advances on people tracking in dense crowds, a taxonomy to categorize the problems of people counting/density estimation and behavior analysis in crowds (summarized in Figure 1), and report the current trends that relate the problems of crowd analysis and synthesis (or simulation).




**[FIG1] Schematic representation of the topics tackled in this survey, and the proposed taxonomy for the people counting and crowd behavior understanding problems.**

## CROWD DYNAMICS

Crowds can be characterized considering four different aspects: i) image space domain (as the main focus of this survey article); ii) sociological domain; iii) level of services, as described by Fruin [15]; and iv) computer graphics domain. In the image space, a common sense is that crowds are identified when the density of people is sufficiently large to disable individual and group identification. This concept has been used in some work in literature, such as [16]. The density of people is used in many decision-taken processes in crowds, as simulation of collision avoidance, for instance.

In the sociological domain, psychologists and sociologists have studied the behavior of groups of people during several years. They have been mainly interested on the effects occurring when people with the same goal become one only entity, named crowd or mass. In this case, people can lose their individualities and adopt the behavior of the crowd entity, behaving in a different way than if they were alone [17]. It means mainly that a collective entity can emerge, depending on many aspects such as people goals, the observed environment, the occurred event, as well as other variables.

Fruin's level of services [15] presents some crowd conditions in terms of the density of people and its temporal evolution, which is very useful in the design of places of public assembly. Fruin presented various levels of service describing characteristics of people in each level, in terms of density, flows and motion of the population.

In the computer graphics domain, many authors have proposed in last years different models of crowd simulation [18]. Some of them can achieve realistic crowd behaviors, but a model where all possible crowd behaviors could be simulated has not been achieved yet. In fact, a big challenge in crowd simulation is the knowledge about real crowds, meaning that many behaviors of real crowds are not yet observed, proved, or still explained. Some characteristics reported in [19] are briefly presented next.

- *Least-Effort Hypothesis*: People try to choose the least-effort route to reach their goals. This issue is closely related to the trajectory of people in real scenarios in crowded scenes, since the trajectories should be minimally changed to avoid collision with others and still cope with the least effort hypothesis.
- *Lane Formation*: It takes less effort for people to follow immediately behind someone who is already moving in their direction than it does to push their own way through a crowd. Emerged as a consequence of the least effort hypothesis, lane formations arise, since people change trajectory whenever they encounter an entity moving in the opposite direction. This action forms chains of entities walking in line, as we would expect.
- *Bottleneck Effect*: This behavior describes a very obvious effect, in which people change velocity as a function of both density of people and restriction in the environment.

Finally, another important aspect to consider in population dynamics is the personal space. Edward Hall [20] proposed the term "proxemics" to describe the social use of space. The proxemics, or personal space, is defined as the area with invisible boundaries surrounding an individual's body. This area represents a comfort zone during interpersonal communication, and can be reduced in specific environments (crowded environments). The classification of interpersonal relationships proposed by Hall (intimate, personal, social, and public) can be explored by crowd analysis algorithms, as done in [2].

## CROWD ANALYSIS

This section tackles three important problems that have been reported in the literature regarding automated analysis of crowded scenes: i) people counting/density estimation models, ii) tracking in crowded scenes, and iii) crowd behavior understanding models. A taxonomy for these problems, as well as some existing techniques are presented next.

### *PEOPLE COUNTING/DENSITY ESTIMATION MODELS*

An important problem in crowd analysis is people counting/density estimation (either in still images or video sequences). For instance, crowd density analysis could be used to measure the comfort level in public spaces, or to detect potentially dangerous situations [15].

There are several models developed to estimate the number of people in crowded scenarios using computer vision techniques. In this work, these models are divided into three categories: i) pixel-based analysis, ii) texture-based analysis, and iii) object-level analysis. A brief description of each category is provided next, along with some representative approaches.

#### PIXEL-LEVEL ANALYSIS

Pixel-based methods rely on very local features (such as individual pixel analysis obtained through background subtraction models or edge detection) to estimate the number of people in a scene. Since very low-level features are used, this class is mostly focused on density estimation rather than precise people counting.

The work described in [5] was one of the pioneers to use computer vision techniques for obtaining automatically some kind of information from crowds. In their work, the authors proposed an approach to estimate the density of the crowd using pixel-level information. They combined a background subtraction technique and detected edges to estimate the density. They assumed linear models to map foreground pixels or edges to the number of people, which were integrated using a Kalman filtering approach to improve the results. Their method also includes geometrical correction due to the perspective of the camera, to account for the fact that the dimension of the same person (in pixels) may change at different locations (the size in pixels decreases as the distance from the camera increases). Clearly, the linear function that maps pixel counts to the number of people in the scene fails when strong occlusions occur, which is common for lateral or oblique camera setups in denser crowds.

**THE GOAL OF CROWD ANALYSIS TECHNIQUES BASED ON COMPUTER VISION IS TO EXTRACT SOME KIND OF INFORMATION FROM CROWDED VIDEO SEQUENCES THAT COULD BE USED TO BENEFIT A LARGE NUMBER OF APPLICATIONS.**

Regazzoni et al. [21] proposed an approach to estimate the crowd density in images. In their work, features extracted from each acquired image (basically the results of an edge detection algorithm and finding vertical edges) are related to the number of people present in the monitored scene by using nonlinear models obtained by means of dynamic programming in an offline training phase. In the operation phase, the detected features are integrated across time through an extended Kalman filter to improve the results. Although they had shown improvement over a competitive approach based on belief Bayesian networks [22], their method was mostly focused on indoor scenes with a limited number of people (up to around 30). Furthermore, the offline training procedure is adjusted to a given camera setup and scenario, and changes require a new training phase.

Cho and collaborators [23] presented a method based on neural networks to estimate the crowd density in subway stations. Their objective was to detect high-density situations and to provide statistical analysis about the flow of people across time for activity planning. They proposed a simple edge detector based on binary image thresholding and explored the length of detected edges as a feature for people counting through a single hidden layer neural network. Their approach was implemented using a hybrid method for global learning that combines the least-squares method with three types of global optimization approaches, defined as follows: random search, simulated annealing, and genetic algorithms. Their results presented approximately 90% of correctness over all tested methods. They also pointed out that the combination of least-square and random search algorithm was the fastest, opposed to the combination of the least-square and simulated annealing (about 100 times slower).

Yang et al. [24] proposed a multicamera-based method to segment people and to estimate the number of people in crowded video sequences. In their work, groups of image sensors were used to segment foreground objects from the background, aggregate the resulting silhouettes over a network, and compute a planar projection of the scene's visual hull. An advantage of their approach is that the system does not compute any feature correspondences across views. Thus, the computation cost increases linearly with the number of cameras. Instead, the authors introduce a geometric algorithm that computes bounds on the number and possible locations of people using silhouettes obtained by each sensor through background subtraction, in each region of the projection. However, in very crowded situations some objects may be completely hidden from all views and therefore impossible to localize individually.

Ma and colleagues [25] proposed a crowd density estimation algorithm for surveillance purposes. In their work, an approach based on pixel counting of foreground objects is used to estimate the crowd density, combined with a projective correction using a calibrated camera. A linear relation between the number of pixels and number of persons was then derived by applying the geometric correction. The density over time is also monitored, aiming to detect some unusual behavior. Their approach suffers the same occlusion problems faced in [5], since a linear model is used to estimate the people count.

Kong and collaborators [7] presented a method based on learning to estimate the number of people in crowds. Edge orientation and the histogram of the object areas (extracted from foreground objects through a background subtraction algorithm) are used as image features. A normalization procedure is performed to account for camera perspective, and the training model used to relate the detected features with the number of people was based on a feed-forward neural network. An interesting characteristic of this approach is the training of normalized features, so that changes in the camera setup do not require a new training phase.

### TEXTURE ANALYSIS

Algorithms that rely on texture analysis explore a coarser grain if compared to pixel-based methods, as texture modeling requires the analysis of image patches. Although this class explores higher-level features when compared to pixel-based approaches, it is also mostly used to estimate the number of people in a scene rather than identifying individuals.

Marana et al. [26] analyzed four methods used in texture analysis and three classifiers to deal with the crowd density estimation problem. Regarding texture analysis, they compared the following four methods: gray-level dependence matrix, straight lines segments, Fourier analysis, and fractal dimension. Regarding the classifiers, they compared the following

three methods: neural network, statistics (Bayesian), and a fitting function-based approach. They found better results when using the gray-level dependence matrix-based method, providing better contrast and homogeneity as texture features, combined with a Bayesian classifier. However, it should be noted that they generated ground-truth information empirically, which could affect the comparison. They estimated the crowd density in one of the five following classes: very low density, low density, moderate density, high density, and very high density. The authors mentioned that the method can not discriminate very well the difference between high and very high densities.

**THE PROBLEM OF VALIDATION IS PARTICULARLY CHALLENGING WHEN DEALING WITH CROWDED SCENES, SINCE GROUND-TRUTHED VIDEO FOOTAGES CONTAINING SPECIFIC ABNORMAL BEHAVIORS IN DENSER CROWDS ARE NOT LARGELY AVAILABLE.**

Wu et al. [27] proposed an approach to estimate the crowd density using support vector machines (SVMs) and texture analysis. In their work, a perspective projection model is adopted to generate a series of multiresolution cells, and the gray-level dependence matrix method is used to extract textural information within these cells. A multiscale texture vector is built, and an SVM is trained to relate the textural features with the actual density of the scene. The authors reported a maximum estimation error for each cell below 5%, and proposed as future work the possibility of including a background subtraction method in the feature extraction stage. One drawback of their approach, however, is the need of retraining the SVM for scenarios with different camera setups, since the density cells are highly dependent on camera parameters.

Rahmalan et al. [6] made a comparison of three techniques used in texture analysis to tackle the crowd density estimation problem. The three analyzed techniques were the gray-level dependence matrix, the Minkowski fractal dimension, and a third one named translation invariant orthonormal Chebyshev moments. The extracted features are classified in a neural network (self-organizing maps), and their analysis indicates that the method based on the Minkowski dimension presented the worst results, while the translation invariant orthonormal Chebyshev moments had the best overall results. However, they found a small difference between the translation invariant orthonormal Chebyshev moments method and the gray-level dependence matrix method, indicating that it should be better investigated in a future work.

Chan et al. [28] developed a crowd counting algorithm based on a texture-based motion segmentation technique and Gaussian process regression. The authors initially segment the crowd into different motion directions using the mixture of dynamic textures, and for each motion cluster they extract segment features (area, perimeter, perimeter edge orientation, perimeter-area ratio), internal edge features (total edge pixels, edge orientation, Minkowski dimension) and texture features (homogeneity, energy, and entropy). These features are normalized to account for camera perspective, and a Gaussian process regression is used to relate the number of people per segment. Chan and Vasconcelos [29] explored a similar idea, but using Bayesian Poisson regression instead (which is more adequate for discrete processes, such as people counting). The regression used in both papers presented good results, but they are dependent on the segmentation step, which may fail for unstructured crowds with erroneous motion.

### OBJECT-LEVEL ANALYSIS

Methods that rely on object-level analysis try to identify individual objects in a scene. They tend to produce more accurate results when compared to pixel-level analysis or texture-based approaches, but identifying individuals in a single image or a video sequence is mostly feasible in lower density crowds. In denser crowds, clutter and severe occlusions make the individual counting problem almost impossible to solve, despite the recent advances of computer vision and pattern recognition techniques.

Lin et al. [30] proposed an algorithm to estimate the crowd density in three stages. In the first one, they searched for objects with head-like contour in the image space, using the Haar wavelet transform. In a second stage, the features of the object are analyzed, using an SVM, aiming to classify it as a head or not. Finally, a perspective transformation is done aiming better estimate the density of the entire crowd.

Zhao and Nevatia [31] proposed a Bayesian approach to segment people in crowds. In their work, several three-dimensional (3-D) human models are used to represent the foreground objects in the scene, and a probabilistic model based on Markov chain Monte Carlo integrates in a Bayesian framework the tracked features, like body shape, people's height, camera model, head candidates, foreground objects, for example. However, in high-density crowds, the full body representation is usually not very useful, as severe occlusions tend to hide most of the body (and part of the head).

Leibe and colleagues [32] proposed a pedestrian detection scheme using a top-down segmentation approach. In fact, they explore a combination of local features (a scale-invariant version of the implicit shape model) and global features (Chamfer distance) to obtain the probability of a person being present, which is measured by comparing small learned image patches of the appearance of humans and their occurrence distribution. Their algorithm can reliably detect and localize pedestrians in relatively crowded scenes and with severe overlaps. However, the lateral camera setup explored in [32] generates too many occlusions (partial or total) in very crowded scenes, which can not be handled adequately in such scenarios.

Rittscher et al. [33] proposed an algorithm for segmenting human figures in video sequences. They try to fit multiple

object hypotheses to explain the occurrence of a set of image features, dealing with occlusions by computing joint image likelihood of multiple objects. The image features are based on the contours of segmented foreground objects, assuming that foreground blobs are available. The joint likelihood is then obtained using the expectation-maximization algorithm. It is interesting to note that explicit camera information is used in [33], which makes the approach suitable for a variety of camera setups. On the other hand, the performance of their approach is highly dependent on the extraction of foreground blobs that generate the image features.

**IN OBJECT-BASED METHODS, CROWD BEHAVIOR UNDERSTANDING IS PERFORMED THROUGH SOME KIND OF SEGMENTATION OR DETECTION OF INDIVIDUALS TO ANALYZE GROUP BEHAVIORS.**

Rabaud and Belongie [34] presented a method to segment individuals in a crowd. They use a feature tracking algorithm, namely the Kanade–Lucas–Tomasi tracker, to detect moving objects in the scene. The tracker is then combined with a temporal and a spatial filter, and a clustering algorithm is used to group similar features into a trajectory, which is related to a single object. The authors validated their results using three data sets, containing ground-truth information generated by a specialist. They also use a video sequence with a crowd of cells, aiming to demonstrate the robustness of the proposed approach to segment individuals in crowds of different entities, but homogeneous among themselves. One clear limitation of this approach relates to stationary crowds, where motion information can not be explored.

Brostow and Cipolla [35] presented an unsupervised Bayesian clustering method to detect independent movements in a crowd. Their hypothesis is that a pair of points that move together should be part of the same entity. An optical flow algorithm combined with an exhaustive search (the search region is defined by ground-plane camera calibration) based on the normalized cross correlation is used to track some image features. An unsupervised Bayesian clustering algorithm then is applied to group such features, aiming to identify each individual moving in the crowd. An interesting characteristic of [35] is that it does not require any training stage or appearance model to track individuals. However, since rigid motion is assumed, the algorithm may fail if strong arm movements are present.

Jones and Snow [36] developed a classifier to detect pedestrians using spatiotemporal information. Their classifier involves three types of features: Haar-like features applied directly at each frame, absolute difference of Haar-like features in adjacent frames, and a shifted difference filter that aims to capture the motion of the pedestrians (eight shifting directions were used). Adaboost is then used to build a soft cascade classifier based on a set of manually labeled training images (in fact, eight classifiers were trained to deal with eight different motion directions). Their approach seems to successfully differentiate pedestrians from vehicles, but tends to fail for relatively dense scenarios. Also, it should be noted that changes in the camera setup that monitors the scene require a new training procedure, since the appearance of pedestrians trained with Adaboost depends on the positioning of the camera.

### *TRACKING IN CROWDED SCENES*

Another important problem related to crowd analysis is people tracking, which consists of identifying the position of the same person in a sequence of frames. The knowledge of individual trajectories in a crowd can be explored to identify main flows of a crowd, or to detect abnormal behaviors. Although there are several approaches for the generic problem of object tracking, clutter and severe partial (or even total) occlusions make individual person tracking a challenging problem in denser crowd. A survey on generic object tracking algorithms along with a taxonomy of approaches can be found in [37], and some references on crowd tracking can be found in [4]. The techniques described next are focused mostly on tracking in higher-density crowds.

It is interesting to note that the problems of people counting and tracking are related, since both of them have the goal of identifying the participants of a crowd. However, the counting problem usually requires only an estimate of the number of people, regardless of their position (and temporal evolution). The tracking problem, on the other hand, involves the determination of the position of each person in the scene as a function of time. Nevertheless, some object-based approaches for people counting described in the section "Crowd Analysis" can be used to initialize tracking algorithms, or even extended to perform both people counting and tracking.

The people-counting approach described in [33] also presented an extension for tracking. In their approach, each track is modeled by a color signature, an appearance template and a probabilistic target mask that is an autoregressive estimate of the foreground information. People walking close to each other are clustered into "group tracks," and individual tracks within the same group are smoothed using a constant velocity Kalman filter. Their approach can handle short term occlusions between isolated tracks, but it tends to fail when the crowd density gets very high.

Ali and Shah [38] proposed an approach for people tracking in structured high-density scenarios. In their approach, each frame of a video sequence is divided into cells, each cell presenting just one particle. A person consists of a set of particles, and each person is affected by the layout of the scene (obstacles and barriers, which are learned automatically), as well as the motion of other people. An interesting aspect of this work is the use of concepts related to crowd modeling (obstacles and relationship with neighbors). On the other hand, since a manual identification of the individuals is required to initialize each track, this approach is not adequate for automatic people counting.

Furthermore, since the cells used for tracking have fixed size, problems may arise in the far field of oblique cameras.

Rodriguez et al. [39] presented a tracking approach to deal with unstructured environments, in which the motion of a crowd appears to be random with different participants moving in different directions over time (e.g., a crossway). They employ the correlated topic model (CTM), which allows each location of the scene to have various crowd behaviors. In their approach, the video sequence is divided into nonoverlapping clips. For each of these clips, the optical flow is computed, and both position and velocity vector are quantized to generate a word in a codebook needed for the CTM. The motion words are assumed to arise from a generative process, whose parameters are estimated using a collection of training video sequences. Their approach is indeed able to deal with very dense crowds, but as the approach described in [38], there is the issue of track initialization.

**A MAJOR CHALLENGE IN CROWD ANALYSIS IS THE GENERATION OF GROUND-TRUTHED IMAGES OR VIDEO SEQUENCES, WHICH CAN BE USED EITHER FOR TRAINING OR VALIDATION PURPOSES.**

### *CROWD BEHAVIOR UNDERSTANDING MODELS*

The behavioral analysis of a crowd is an important topic of research in computer vision. In general, the temporal information is used to estimate the behavior of a crowd in a given environment, such as main directions [40], velocities [5], and unusual motions [25], [35], [8]. A great variety of approaches were proposed in past years to deal with crowd analysis and understanding that could involve researchers from several areas.

The problem of validation is particularly challenging when dealing with crowded scenes, since ground-truthed video footages containing specific abnormal behaviors in denser crowds are not largely available. To overcome this problem, some authors [41], [2] have used crowd simulation algorithms to generate controlled situations with known ground truth to test their algorithms. In fact, concepts related to crowd simulation are also being explored to distinguish normal and abnormal behaviors, as in [42].

As noted in [42], there are two main approaches for crowd behavior analysis. In the "object-based" approach, a crowd is treated as a collection of individuals. On the other hand, "holistic" approaches treat the crowd as a single entity, without the need of segmenting each individual. Clearly, in denser scenes it is very difficult to track individual components in the crowd, and hence the second approach tends to be more appropriate.

### OBJECT-BASED APPROACHES

In object-based methods, crowd behavior understanding is performed through some kind of segmentation or detection of individuals to analyze group behaviors. For instance, the identification of a single person moving against the dominant flow (e.g., one individual trying to enter a sports arena after the match is finished) could be related to a potentially dangerous situation. In a snatch theft, the thief usually approaches the victim from behind, and its automatic detection requires the individual identification of both tracks (thief and victim). Although object-based approaches for crowd behavior understanding allow the detection of high-level events, they face considerable complexity to isolate individuals in denser crowds, being therefore most applicable to low or moderately crowded scenes.

In [2], an algorithm for group detection and classification as voluntary or involuntary based on computer vision was proposed. In their work, a top-down camera is used to track individuals through, and Voronoi diagrams are used to quantify the sociological concept of personal space. The temporal evolution of the Voronoi diagrams is used to identify groups in the scene, and the portion of the personal space within each individual's field of view is used to classify the groups as voluntary or involuntary. An interesting aspect of [2] is the use of sociological aspects to analyze crowds, but one drawback is the need of tracking each individual in the scene, which may be difficult in denser scenes.

Cheriyadat and Radke [40] proposed an approach for clustering a set of low-level motion features into trajectories, similarly to [34] and [35], but using additional rules in the clustering process, such as the dominant movements that are computed based on the longest common subsequences. However, while the goal in [34] and [35] was to identify each member in the scene based on motion cues, the main goal in [40] was to extract dominant motion patterns in a crowded scene. Then, individual motions not coherent with dominant flows could be highlighted and marked as potential unusual behavior. For example, their approach could detect a person making a U-turn in a scenario where most people move along the same direction.

Wang et al. [43] proposed an unsupervised learning framework to model activities and interactions in crowded and complicated scenes. In their work, hierarchical Bayesian models are used to connect three elements in visual surveillance: low-level visual features, simple "atomic" activities, and interactions. Atomic activities are modeled as distributions over low-level visual features, and multiagent interactions are modeled as distributions over atomic activities. These models are learned in an unsupervised way. Given a long video sequence, moving pixels are clustered into different atomic activities and short video clips are clustered into different interactions. Without tracking and human labeling effort, their framework completes many challenging visual surveillance tasks, such as: discovering and providing a summary of typical atomic activities and interactions occurring in the scene; segmenting long video sequences into different interactions; segmenting motions into different activities; detecting abnormality, and supporting high-level queries on activities and interactions.

## HOLISTIC APPROACHES

Instead of tracking individual objects, this top-down view treats the crowd as a single entity, which directly tackles the problem of dense occluded crowds in contrast to the object based methods. Holistic approaches usually try to obtain coarser-level (global) information, such as main crowd flows, and they tend to disregard local information (e.g., a single person moving against the flow, or a party of two that keep together during the hole scene in an unstructured environment).

Davies et al. [5] proposed an approach based on discrete Fourier transform (faster moving objects are associated with high-frequency and nonmoving objects are associated with constant levels), combined with a linear area transform algorithm to distinguish static from moving crowds (a useful indicator of congestion or potential danger). The approach proposed in their work is to measure motion features at pixel or pixel-neighborhood level that are then aggregated to obtain motion properties for larger regions in an image. The aggregated results can then be used to establish overall preferential crowd velocities (direction and magnitude).

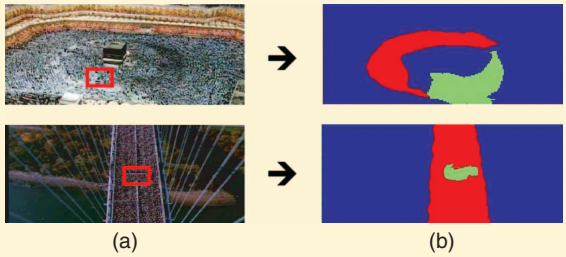
Boghossian and Velastin [16] presented an approach to estimate the paths and main directions of a crowd in a closed-circuit television system using a block-matching algorithm for motion estimation. They also presented an algorithm to detect emergency situations, by performing higher-level processing on the measured flow paths and their directions, which can help surveillance system operators. Aiming to minimize the error over the detected motion, the authors accumulate the previous velocities vectors and apply a filter to them using a predefined neighborhood. The authors mentioned three kinds of detected situations: i) circular flow near to the exits detected in the Hough space, that could indicate traffic jams; ii) crowd flow diverging from a point to all directions, which might indicate a potential danger (such as fights and fire) is manually detected by spotting the diverging crowd flow from that location; and iii) obstacles in the flow paths that might correspond to injured pedestrians or deliberate flow disturbances are detected by performing a region-growing segmentation to group the motion free regions in the scene.

Andrade et al. [41] characterized a "usual behavior" of a crowd based on the analysis of their optical flow, using hidden Markov models (HMMs). In their work, an unsupervised algorithm is used for feature extraction and spectral clustering. They employed a background subtraction model to extract foreground objects, used them as a mask to compute the optical flow. They validated the model with synthetic and real situations. The simulated situations are: unidirectional flow, people evacuating an environment with the exits obstructed, and a crowd moving around a fallen person.

Ali and Shah [8] proposed an approach to segment the flow of a crowd and to detect the crowd's instabilities based on Lagrangian particle dynamics. In their work, an optical flow based method is used to capture the movements of a crowd, which are used to generate a velocity field. In a posterior stage, the authors make use of a numerical integration method to insert particles into this velocity field. The movements of the particles on the velocity field are used to construct a flow, namely finite time Lyapunov exponent, which reveals some coherent structures, namely Lagrangian coherent structures, used in the segmentation stage. The instability detection is detected when the number of segments of some analyzed sequence changes. To validate the method, the authors inserted a perturbation in a part of a known velocity field (like a rotation, for example) and analyzed their segmentation. The red rectangular regions, shown in Figure 2(a), illustrate the detected instabilities captured by their method.

Mehran and his group [42] developed an approach for abnormal crowd behavior detection exploring the social force model, which was originally introduced in [44] to model crowds based on socio-psychological studies. In their method, a set of particles is overlaid to the image, and they are moved according to the computed optical flow. The motion of the particles is then used to estimate the social forces. To detect normal patterns of forces over time, the magnitude of the interaction force vectors are mapped to the image plane, obtaining a force flow. The likelihood force flow is then estimated using the bag of words approach (for normal videos), and a fixed threshold on the estimated likelihood is used to classify each frame as normal or abnormal. An interesting aspect of this work is the exploration of socio-psychological to detect abnormal behavior.

In the work of Kratz and Nishino [45], an approach for anomaly detection in extremely crowded scenes based on spatiotemporal information was presented. The video sequence is initially split into a set of spatiotemporal cuboids, and each cuboid is represented by a 3-D multivariate Gaussian distribution of the spatiotemporal gradients computed within the cuboids. The symmetric Kullback–Leibler (KL) divergence is used to compare the distribution of different cuboids, which are either combined into a single "prototype" (small KL distances) or guide the creation of new prototypes (large distances). To model temporal information, an HMM is built for each "tube" (a temporal collections of cuboids having the same spatial location), in which the possible observations are 3-D Gaussian distributions. Finally, coupled HMMs are built to model the connection that exists between spatially neighboring motion patterns. Despite the good quantitative results reported in [45],




**[FIG2] (a) Example of used scenarios and (b) the respective flow segmentation, used in [8].**

the determination of the size of the cuboids can be challenging, particularly in the far field of oblique camera setups (due to perspective issues).

It should be also noted that papers related to people counting/density estimation could be extended for abnormality detection in crowds. For instance, the crowd density estimation approach proposed in [27] was also explored to detect abnormal situations. The authors argue that density changes may indicate potential danger or emergency in a crowd, and trained a SVM to detect overcrowdedness and excessive emptiness in a local area of the filmed sequence.

**CROWD SYNTHESIS ALGORITHMS COULD ALSO BENEFIT FROM INFORMATION CAPTURED FROM REAL LIFE USING COMPUTER VISION ALGORITHMS.**

## INTERSECTION BETWEEN CROWD SIMULATION AND CROWD ANALYSIS

During the last 15 years, many crowd simulation models have been proposed in the literature [18]. While such research area aims to provide group behaviors in a generic point of view, with main applications in games and simulators, this kind of work can also be used to improve crowd analysis. In fact, computer graphics have been used to train or validate computer vision algorithms [9], [10], and in particular, crowd synthesis approaches have been explored to validate crowd analysis algorithms [11], [41], [2]. Conversely, computer vision algorithms designed for crowd analysis can be used to extract information from real life in an automatic manner, which may reduce manual intervention and help to improve the realism of crowd synthesis algorithms. For instance, the spatial distribution of the crowd in a real scene could be used to initialize a crowd simulator, and the tracked trajectories (or main flows) could be used to guide the motion or virtual agents (as in [13], [12], and [14]).

Next, we present some approaches that either use crowd synthesis to help crowd analysis algorithms, or the other way around. The main goal is to present existing studies as well as open problems in both cases.

### *CROWD SYNTHESIS HELPING CROWD ANALYSIS*

The three main problems described in the section "Crowd Analysis" regarding crowd analysis share a common problem: how to validate the results? For instance, to evaluate the accuracy of a people counting algorithm, it is necessary to know how many people were present in the environment, which is usually done manually. For people tracking, a thorough evaluation would require the labeling of all people in the environment across time, which is a time-consuming and tiresome task. Also, computer vision algorithms that rely on a training stage require a considerable amount of training data in specific situations, which can be difficult to obtain. Next, we present some approaches that explore crowd simulation algorithms to either train or validate crowd analysis techniques.

Andrade et al. [11] presented a method for generating video evidence of dangerous situations in crowded scenes. The scenarios of interest are those with high safety risks such as a blocked exit, collapse of a person in the crowd, and escape panic. Real visual evidence for these scenarios is rare or unsafe to reproduce in a controllable way. Thus, there is a need for simulation to allow training and validation of computer vision systems applied to crowd monitoring.

In [30], an approach for generating ground-truth data using computer graphics was proposed. In their approach, a virtual world containing personlike puppets was created, and it is used to generate a scenario where the features to be found are known (like the number of people, for example). In their experiments, the overall accuracy reported by the authors was around 90–95%. The authors emphasized the great difficulty of validating people counting approach in dense situations using real images, because in such cases, the number of people is usually estimated by a specialist and could probably contain errors.

Jacques, Jr. and collaborators [2] proposed an algorithm for identifying and classifying groups based on individual object tracking. Since the tracking stage is a challenge in very crowded scenes, they used a crowd simulator to virtually reproduce scenarios with bottlenecks (where involuntary groups are supposed to arise) to test their approach in denser scenarios.

Allain and collaborators [46] proposed to characterize crowd flows with optimal control theory, using velocity and a disturbance potential. As a validation strategy, they used a crowd simulation algorithm to generate ground-truthed data. As acknowledged by the authors, the simulation model was totally different than the one used for assimilation, and the results were rendered to be as realistic as possible (including arms/legs movement based on motion capture data).

Despite the successful use of computer graphics techniques to train and/or validate computer vision algorithms in some specific cases [9], [10], using crowd simulation algorithms to help crowd analysis techniques still presents several open problems. Existing crowd simulators are ultimately based on mathematical models that try to reproduce the average behavior of human beings, but are not (and probably will never be) able to simulate unpredictable behaviors. Also, simulating realistically a high-density crowd is still a challenge, particularly regarding collision avoidance. Crowd simulators usually adopt a dot or circle to represent each agent and employ some kind of repulsion force to avoid collisions or interpenetrations. The dot approximation considers a punctual representation of agents (i.e., without area), which is clearly not adequate for high-density situations. The circular approximation does consider an area for each agent, but circles can not be grouped as tightly as real people in high densities. In fact, real people can touch each other and even be deformed due to collisions.

Another important issue is the visual quality of simulation results since the input of crowd analysis algorithms, is a video sequence. There is a great variety of parameters used to model the virtual humans (such as shape, gait, and clothing), the

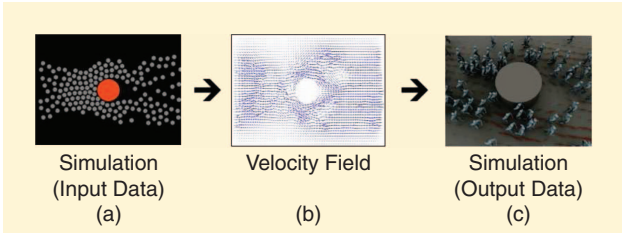
virtual environment (structure, color, and texture), as well as other rendering parameters (lighting and synthetic camera noise and distortion) that impact significantly the visual animation produced by the crowd simulator. For instance, the exact same simulation result could be rendered by setting the colors/textures of the virtual agents different from each other (and different from the background), or by setting all the colors/textures similar to each other. A tracking algorithm could provide very good results for the first scenario and fail in the second one.

### *CROWD ANALYSIS HELPING CROWD SYNTHESIS*

If crowd simulation has been used to help crowd analysis, the reverse is also true. For instance, in several applications of crowd simulation, it is desirable to model a real scenario in the virtual world, and to study how physical changes in the environment would impact the flow of people. For that purpose, it would be important to obtain information about the motion of people in that specific scenario, such as density, main directions, and velocity field. The manual extraction of these data is very time consuming and prone to errors, motivating the use of computer vision algorithms.

Courty and Corpetti [13] presented an approach to capture information from real crowds and feed a crowd simulator. In their work, an optical flow-based method is used to capture the movements of people in a given environment. They generate velocity fields, captured over time, aiming to simulate realistically the behavior of an actual crowd. They simulate a unidirectional flow with a cylindrical obstacle in the middle of the scene, illustrated in Figure 3(a), and capture people movements to generate a velocity field [Figure 3(b)], used to feed the simulation model [Figure 3(c)]. In a real situation, they use two case studies: in a first one, they capture the movement of a crowd, moving in a unidirectional way, without obstacles; in a second one, they capture the movements of a crowd entering in a stadium. The authors mentioned that the results are convincing and pointed out the possibility of using their approach to validate crowd models, providing a quantitative measurement for analysis.

Musse et al. [12] and Paravisi et al. [14] also presented different approaches to capture information from the real world, assisted by computer vision techniques, to improve the realism of crowd simulations. However, in both papers the authors track the individuals trajectories (in noncrowded scenarios), instead of a global movement. The trajectories are clustered into different motion classes, and an extrapolated velocity field is created for each class. These fields are then used to feed the simulation model, informing the direction and velocity at each position. The underlying idea in these two methods is that the desired unconstrained motion of the individuals could be captured from noncrowded scenes, and a crowded version of the scenario could be approximated using the simulation model. It should be pointed out that the motion stimulus in [13] is extracted in a global (holistic) manner, retaining the global flow, while an object-based approach for tracking was employed in [12] and [14], focusing on lower density scenarios. An illustration of the results presented in [12] is shown in Figure 4.



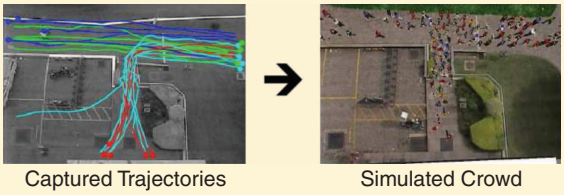

**[FIG3] Example of synthetic scenarios used in [13]. (Figure used with permission.)**

It is important to mention that defining quantitative metrics to evaluate the realism of a crowd simulation result is still an open problem. If the simulation relates to a scenario that exists in real life, some parameters (such as main directions, average speed, and crowd density) in both real and virtual environments could be compared. For example, Musse and collaborators [12] presented some quantitative comparisons between the real and the virtual scenarios, such as the average speed of people. Paravisi et al. [14] proposed a "spatial occupancy map" to measure the spatial distribution of people, which was used for comparison purposes. However, evaluating simulation results of a synthetic scenario that does not relate to a real one is a very complex task. The motion and/or behavior of a crowd can vary depending on several aspects, such as culture and context. For instance, the characteristics of a crowd in an amusement park are probably different from those of a crowd leaving a football stadium.

## FINAL CONSIDERATIONS

In this work, we presented a survey on crowd analysis based on computer vision. This work tackled three important problems in crowd analysis: people counting/density estimation, tracking in crowded scenes, and crowd behavior understanding in a higher-level analysis, like the temporal evolution, main directions, velocity estimations, and detection of unusual situations. Regarding crowd synthesis, the review was focused on crowd models that either use computer vision algorithms to extract real data information, aiming to improve the realism of the simulation, or that are used to train/validate computer vision techniques. Some considerations about these issues are provided next.

The algorithms for people counting were divided into three categories: pixel-based, texture analysis, and object-level analysis. In a nutshell, object-level analysis are adequate for more




**[FIG4] Example of the virtual simulation of a real scenario used in [12].**

accurate counting and localization of people in a scene, since they are based on individual identification. Usually, such class is adequate in low or moderately denser crowds, since occlusions become significant in packed crowds. On the other hands, pixel-based approaches and those that rely on texture analysis explore lower-level features in the image, not trying to identify individuals in a scene. These classes are usually less accurate for people counting, but they tend to work better in very high-density crowds. It is also interesting to note that techniques that somehow explore camera information (such as [33], [28], and [29]) are more flexible regarding the camera setup used to monitor the scene, being also able to correct perspective distortions.

A few algorithms focusing on people tracking in crowded scenes were also analyzed in this article. We believe that approaches that somehow explore the expected behavior of the crowd (as the interactions with obstacles, used in [38]) are promising to deal with severe occlusions and clutter. On the other hand, these approaches may not be adequate to detect unusual behaviors in the crowd (since abnormalities are usually associated with unexpected motion). Despite the advances for individual tracking in highly dense crowds [38], [39] the automatic initialization of each track is still a challenge.

Crowd behavior understanding algorithms were divided into object-based and holistic approaches. The first one relies on the knowledge of the individuals that form the crowd, what is challenging for denser scenarios. Furthermore, the analysis of the possible interactions among the individuals increases the computational cost as the number of people grows. Hence, it is more appropriate for low and moderately crowded scenes. The holistic approach treats the crowd as a single entity, so that the tracking issue is not a challenge. This latter approach is adequate for dense scenes, but usually the detected abnormality can not be as well characterized as object-based methods.

Still regarding crowd behavior understanding, we believe that algorithms that somehow learn normal activities from a set of observations (such as [43], [42], and [45]) are promising. In particular, we think that exploring psycho-social aspects of crowds, as the grouping detection and classification based on personal spaces proposed in [2] or the social force model used in [42], is an interesting trend for abnormality detection in crowded scenes.

Finally, we discuss how computer graphics can help computer vision in crowd's applications and vice versa. A major challenge in crowd analysis is the generation of ground-truthed images or video sequences, which can be used either for training or validation purposes. Manual labeling is a tiresome and time-consuming task. It is also prone to errors and user dependent. An alternative approach to increase the training data set, adopted by some authors [30], [41], [13], is to use synthetic data, where the analyzed features are known. Similarly, images or video sequences generated by computer graphics algorithms can be used to validate computer vision algorithms, as related in [9] and [46].

It should be noted that virtual ground-truth data must be as realistic as possible and try to mimic the same problems that affect the performance of computer vision algorithms in real video sequences (e.g., camera noise, lens distortion, and varying illumination conditions), otherwise the computer vision algorithm could be biased. For example, an interesting experiment to assess the validity of using synthetic ground-truth data for validation would be to compare the results of a given algorithm applied to real video sequences and their synthetic counterparts. If the accuracies on both real and synthetic scenarios are correlated, the hypothesis that validation could be performed using synthetic data would be strengthened. In fact, this kind of comparison was performed in [9], but in the context of background subtraction and head detection in noncrowded scenes. Although the authors used a limited amount of data, results using real and generated ground truth were similar. With the increasing quality of simulation and rendering algorithms, we believe that virtual ground-truth generation may be a good solution to complement manually generated data in both training and validating steps.

Crowd synthesis algorithms could also benefit from information captured from real life using computer vision algorithms. In fact, most existing crowd simulation techniques require some kind of motion stimulus to guide the virtual agents, which can be obtained through tracking algorithms. Such stimulus can be obtained in noncrowded scenes (as in [12]), and used to estimate scenarios in a more crowded scenario, or extracted directly from denser sequences (as in [13]).

In summary, the behavioral analysis of human crowds using computer vision algorithms is, and probably will be for a long time, the focus of attention for several researchers due to the possible potential applications. This problem presents challenges of great complexity that could involve researchers from several areas and backgrounds. In particular, the integration of computer vision and computer graphics is becoming more popular in both crowd analysis and synthesis.

## ACKNOWLEDGMENT

This work was partially supported by the Brazilian research agency CNPq.

## AUTHORS

*Julio Cezar Silveira Jacques Junior* (julio.silveira@pucrs.br) received the B.S. degree in computer science in 2003 from Universidade Luterana do Brasil, and the M.S. degree in applied computer science in 2006, from Universidade do Vale do Rio dos Sinos, Brazil. He is currently a Ph.D. student at Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), where he is also a researcher in the Virtual Human Lab. He is a substitute lecturer at Universidade Federal do Rio Grande do Sul, Brazil. His current research interests include object segmentation/tracking, human motion analysis, and crowd simulations.

*Soraia Raupp Musse* (soraia.musse@pucrs.br) earned a Ph.D. degree in computer science from EPFL in Switzerland in 2000.

Her current research interests include crowd simulation and virtual human animation. She managed some research projects with significant budgets, with private companies (e.g., Petrobras, HP Brazil, and LEGION). She has been a reviewer of many journals, including *IEEE Transactions on Visualization and Computer Graphics* and *IEEE Computer Graphics and Applications*. She is organizing the Brazilian Network on Visualization, specifically in the context of security, where she has applied her efforts in crowd simulation.

***Cláudio Rosito Jung*** (crjung@inf.ufrgs.br) received the B.S. and M.S. degrees in applied mathematics, and the Ph.D. degree in computer sciences, from Universidade Federal do Rio Grande do Sul (UFRGS), Brazil, in 1993, 1995 and 2002, respectively. He is currently an assistant professor at UFRGS. His research interests include image denoising and enhancement, image segmentation, medical imaging, multiscale image analysis, intelligent vehicles, object tracking, multimedia applications and human motion analysis. He is a Member of the IEEE.

## REFERENCES


[1] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Comput. Vis. Image Understand.*, vol. 104, no. 2, pp. 90–126, 2006.

[2] J. C. S. Jacques, Jr., A. Braun, J. Soldera, S. R. Musse, and C. R. Jung, "Understanding people motion in video sequences using voronoi diagrams," *Pattern Anal. Applicat.*, vol. 10, no. 4, pp. 321–332, 2007.

[3] P. Kilambi, E. Ribnick, A. J. Joshi, O. Masoud, and N. Papanikolopoulos, "Estimating pedestrian counts in groups," *Comput. Vis. Image Understand.*, vol. 110, no. 1, pp. 43–59, 2008.

[4] B. Zhan, D. N. Monekosso, P. Remagnino, S. A. Velastin, and L.-Q. Xu, "Crowd analysis: A survey," *Machine Vis. Applicat.*, vol. 19, no. 2, pp. 345–357, 2008.

[5] A. C. Davies, J. H. Yin, and S. A. Velastin, "Crowd monitoring using image processing," *IEE Electron. Commun. Eng. J.*, vol. 7, no. 1, pp. 37–47, 1995.

[6] H. Rahmalan, M. Nixon, and J. Carter, "On crowd density estimation for surveillance," in *Proc. Institution of Engineering and Technology Conf. Crime and Security*, 2006, pp. 540–545.

[7] D. Kong, D. Gray, and H. Tao, "A viewpoint invariant approach for crowd counting," in *Proc. Int. Conf. Pattern Recognition*, 2006, vol. 3, pp. 1187–1190.

[8] S. Ali and M. Shah, "A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2007, pp. 1–6.

[9] S. R. Musse, M. Paravisi, R. Rodrigues, J. C. S. Jacques, Jr., and C. R. Jung, "Using synthetic ground truth data to evaluate computer vision techniques," in *Proc. IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance*, 2007, pp. 25–32.

[10] G. Taylor, A. Chosak, and P. Brewer, "OVVV: Using virtual worlds to design and evaluate surveillance systems," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2007, pp. 1–8.

[11] E. L. Andrade and R. B. Fisher, "Simulation of crowd problems for computer vision," in *Proc. 1st Int. Workshop on Crowd Simulation*, 2005, pp. 71–80.

[12] S. R. Musse, C. R. Jung, J. C. S. Jacques, Jr., and A. Braun, "Using computer vision to simulate the motion of virtual agents," *Comput. Animation Virtual Worlds*, vol. 18, no. 2, pp. 83–93, 2007.

[13] N. Courty and T. Corpetti, "Crowd motion capture," *Comput. Animation Virtual Worlds*, vol. 18, no. 4–5, pp. 361–370, 2007.

[14] M. Paravisi, A. Werhli, J. C. S. Jacques, Jr., R. Rodrigues, A. Bicho, C. Jung, and S. R. Musse, "Continuum crowds with local control," *in Proc. Computer Graphics Int. (CGI'08)*, Istambul, Turquia, June 2008, pp. 108–115.

[15] J. Fruin, *Pedestrian and Planning Design*. Mobile, AL: Elevator World Inc. 1971.

[16] B. A. Boghossian and S. A. Velastin, "Motion-based machine vision technique for the management of large crowds," in *Proc. 6th IEEE Int. Conf. Electronics, Circuits and Systems*, 1999, vol. 2, pp. 961–964.

[17] H. Benesch, *Atlas de la Psychologie*. France: Encyclopdies d'Aujourd'huik, 1995.

[18] D. Thalmann and S. R. Musse, *Crowd Simulation*. New York: Springer-Verlag, 2007.

[19] K. Still, "Crowd dynamics," Ph.D. dissertation, Univ. Warwick, England, 2000.

[20] E. T. Hall, *The Silent Language*. New York: Doubleday, 1959.

[21] C. S. Regazzoni, A. Tesei, and V. Murino, "A real-time vision system for crowding monitoring," in *Proc. Int. Conf. Industrial Electronics, Control, and Instrumentation*, 1993, vol. 3, pp. 1860–1864.

[22] C. Ottonello, M. Peri, C. Regazzoni, and A. Tesei, "Integration of multisensor data for overcrowding estimation," in *Proc. IEEE Int. Conf. Systems, Man and Cybernetics*, Oct. 1992, vol. 1, pp. 791–796.

[23] S.-Y. Cho, T. W. S. Chow, and C.-T. Leung, "A neural-based crowd estimation by hybrid global learning algorithm," *IEEE Trans. Syst., Man, Cybern.*, 1999, vol. 29, no. 4, pp. 535–541.

[24] D. B. Yang, N. Héctor, H. González-Ba, and L. J. Guibas, "Counting people in crowds with a real-time network of simple image sensors," in *Proc. IEEE Int. Conf. Computer Vision*, Washington, DC, USA, 2003, p. 122.

[25] R. Ma, L. Li, W. Huang, and Q. Tian, "On pixel count based crowd density estimation for visual surveillance," in *Proc. IEEE Conf. Cybernetics and Intelligent Systems*, 2004, vol. 1, pp. 170–173.

[26] A. Marana, L. da Costa, R. Lotufo, and S. Velastin, "On the efficacy of texture analysis for crowd monitoring," in *Proc. Int. Symp. Computer Graphics, Image Processing, and Vision (SIBGRAPI'98)*, Washington, DC, 1998, p. 354.

[27] X. Wu, G. Liang, K. K. Lee, and Y. Xu, "Crowd density estimation using texture analysis and learning," in *Proc. IEEE Int. Conf. Robotics and Biomimetics*, 2006, pp. 214–219.

[28] A. Chan, Z. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2008, pp. 1–7.

[29] A. B. Chan and N. Vasconcelos, "Bayesian Poisson regression for crowd counting," in *Proc. IEEE Int. Conf. Computer Vision*, 2009, pp. 1–7.

[30] S.-F. Lin, J.-Y. Chen, and H.-X. Chao, "Estimation of number of people in crowded scenes using perspective transformation," *IEEE Trans. Systems, Man, Cybern. A*, vol. 31, no. 6, pp. 645–654, 2001.

[31] T. Zhao and R. Nevatia, "Bayesian human segmentation in crowded situations," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2003, vol. 2, pp. 459–466.

[32] E. Leibe, B. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Washington, DC, 2005, pp. 878–885.

[33] J. Rittsche, P. H. Tu, and N. Krahnstoeve, "Simultaneous estimation of segmentation and shape," in *Proc. Computer Vision and Pattern Recognition*, Washington, DC, 2005, vol. 2, pp. 486–493.

[34] V. Rabaud and S. Belongie, "Counting crowded moving objects," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2006, pp. 705–711.

[35] G. J. Brostow and R. Cipolla, "Unsupervised Bayesian detection of independent motion in crowds," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Washington, DC, 2006, pp. 594–601.

[36] M. J. Jones and D. Snow, "Pedestrian detection using boosted features over many frames," in *Proc. Int. Conf. Pattern Recognition*, 2008, pp. 1–4.

[37] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Comput. Surv.*, vol. 38, no. 4, 2006, pp. 13.1–13.45.

[38] S. Ali and M. Shah, "Floor fields for tracking in high density crowd scenes," in *Proc. European Conf. Computer Vision*, 2008, pp. II:1–14.

[39] M. Rodriguez, S. Ali, and T. Kanade, "Tracking in unstructured crowded scenes," in *Proc. IEEE Int. Conf. Computer Vision*, Kyoto, Japan, 2009, pp. 1389–1396.

[40] A. M. Cheriyadat and R. Radke, "Detecting dominant motions in dense crowds," *IEEE J. Select. Topics Signal Process.*, vol. 2, no. 4, pp. 568–581, Aug. 2008.

[41] E. L. Andrade, S. Blunsden, and R. B. Fisher, "Modelling crowd scenes for event detection," in *Proc. Int. Conf. Pattern Recognition*, Washington, DC, 2006, pp. 175–178.

[42] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 935–942.

[43] X. Wang, X. Ma, and W. E. L. Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 31, no. 3, pp. 539–555, 2009.

[44] D. Helbing and P. Molnár, "Social force model for pedestrian dynamics," *Phys. Rev. E*, vol. 51, no. 5, pp. 4282–4286, May 1995.

[45] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 1446–1453.

[46] P. Allain, T. Corpetti, and N. Courty, "Crowd flow characterization with optimal control theory," in *Proc. 9th Asian Conf. Computer Vision* (Lecture Notes in Computer Science Series), Xi'an, China, Sept. 2009, pp. 279–290. [SP]