Emotion Estimation in Crowds: A Survey

Oscar J. Urizar[1,2], **Emilia I. Barakova**[2], **Lucio Marcenaro**[1], **Carlo S. Regazzoni**[1,3], **Matthias Rauterberg**[2]

[1]Department of Electrical, Electronic, Telecommunications
Engineering and Naval Architecture (DITEN)
University of Genoa - Genoa, Italy

[2]Department of Industrial Design
Eindhoven University of Technology
Eindhoven, Netherlands

[3]Carlos III University of Madrid

Keywords: Crowd emotions survey, collective emotions, emotion estimation, affective models, affective datasets.

Abstract

Emotions play an important role in human behavior, even more so in large congregations of people where emotional states are prompt to be contaged and amplified. This work presents a qualitative systematic review of the literature concerning the estimation of emotions and affects in real-life crowded environments, covering the aspects of methods and datasets. The academic search engine *Scopus* was inquired and the search was limited to publications in the English language addressing any of the aforementioned aspects. The aim of this contribution is to highlight advances, limitation and trends in addressing the estimation of emotions in crowds.

1 Introduction

Emotion is a concept centered on the individual, considered as a process directed to specific events, involving the appraisal of intrinsic features, with influence in multiple bodily systems and a strong impact on behavior; In this sense, collective (crowd) emotions occur when the same event yields a similar appraisal and elicits a common emotion among the members of a crowd. As emotions have a significant influence in the behavior of individuals, they are also essential in understanding the dynamics of a whole crowd. Affective computing (AC) delves in transferring the theoretical knowledge of emotions and affects into systems capable of recognize, model and express such human aspects. In this direction, significant advances have taken place in recent years, mainly focusing on the detection of emotions on single individual, from facial expressions to body language, speech analysis and physiological signals. Further research in the field of AC has started to address the emotions of individuals in groups and crowds; however some of the approaches intended for estimating emotions in individuals are not transferable to groups, emphasizing the need for inventive ways to address this issue. Several issues are introduced when shifting from the estimation of single to collective emotions: which behavioral cues and what sensors are appropriate for inferring emotions in crowds? does a microscopic or a macroscopic level of description better suited? is the emotion of a crowd the sum of the member's emotion or is it something else? to name a few. As cities grow larger and crowds become commonplace, systems capable to identify collective emotions are essential to optimize public spaces, manage crowd flows and ensure safeness.

The academic search engine *Scopus* was inquired, limiting the search to publications in the English language and using the combination of keywords: '*crowd & emotion*', '*collective & emotion*', '*social & emotion*', '*group & emotion*', '*crowd & affect*', '*collective & affect*', '*social & affect*', '*group & affect*', '*crowd & dataset*' and '*emotion & dataset*'. The focus of this qualitative systematic review is on literature dealing with physical crowds, defined here as real-life human groups of any size concurring at a physical location for a significant amount of time. Hence, publications addressing simulated or virtual crowds (e.g. online games/communities, social networks, blogs, etc.) are not considered here. Two main aspects of emotion estimation in crowds are addressed in this paper, namely methods and datasets. Methods are grouped into those dealing with emotion regulation and contagion, and those addressing the estimation of emotions. A representative selection of datasets intended for affective and crowd analysis tasks are examined and a discussion on their applicability to emotion estimation in crowded environments is provided. The aim of this contribution is to examine proposed methods to estimate emotions in crowds, highlighting their capabilities and limitations, and the currently available datasets for such purpose.

The remaining content is organized as follow: Methods applicable for emotion estimation in crowded settings are discussed in section 2. In section 3, datasets and benchmarks available for the evaluation of computational models are presented. Finally, section 4 provides conclusions on the examined methods and datasets along with a discussion on the advances, trends and shortcomings found in the inquired literature.

2 Methods

Several methods addressing emotions have been proposed, some simulate the emotional behavior of crowds, whereas others aim to analyze and understand it from real-world measurements. This survey focuses on the later kind of models, excluding methods that are not applicable to real crowds and focus purely on simulations.

2.1 Emotion Regulation and Contagion

An important aspect in understanding the dynamics of emotions in groups and crowds concerns the problem of whether and how these spread or amplify across individuals in a group. Inspired on neural mechanisms revealed by the recent work of Damasio [1], the authors of [2] and [3] propose ASCRIBE, an agent-based model to describe the interplay of mental states (*emotions*, *beliefs* and *intentions*) of individuals in the decision-making process under stressful situations. ASCRIBE is defined as having an external and internal level of operation; at the external level it incorporates mechanisms for mirroring mental states between individuals, at the internal level it describes how emotions and beliefs affect each other and how both affect a person's intentions. The model was put to the test by simulations and an empirical study case that compared four models and showed the ASCRIBE model to yield higher prediction accuracy. Expanding the concepts applied in the ASCRIBE model, the multi-agent-based model presented in [4] formalizes several concepts of emotion contagion spirals based on fundamental aspects at the individual level: the senders current emotional state and the extent to which the sender expresses the emotion; the strength of the communication channel from sender to receiver; the receiver current emotional state, its openness or sensitivity for the received emotion, bias to adapt emotions upward or downward and tendency to amplify emotions. Although no empiric validation is provided, the model is tested with simulations and mathematical analysis, and it produced interesting emerging patterns identified in psychology literature such as the upward and downward emotion spirals. Further studying the role of emotions in the decision-making process under stressful situations, and with similar concepts to ASCRIBE, an adaptive agent model for affective social decision making is proposed by Manzoor et al in [5] and later extended in [6] to account for emotion regulation and contagion. This model incorporates Hebbian learning principles to adapt the agent's decision-making process, but as found in the experiments conducted, it did not yield significant discrimination in the agent's decisions. Regulation is approached by antecedent-focused strategy (regulation before an emotional response has an effect on behavior), modeling it as a dynamic interaction between internal mental states and contagion is implemented as described in [4].

2.2 Emotion Estimation

Focusing on the task of emotion estimation, the framework introduced in [7] addressed the recognition of individuals' membership and emotions within a group setting by means of multimodal analysis of facial and body expressions. Faces are represented by facial landmark trajectories and extended volume Quantised Local Zernike Moments (QLZM) [8], and encoded into Fisher Vector (FV) [9] representations as input to a Gaussian Mixture Model (GMM) classifier to recognize emotions in arousal and valence dimensions. The framework was tested with a self-collected dataset of 3 groups of 4 individuals each, monitored while watching a movie. The proposed approach vQLZM-FV outperformed the compared methods, namely Facial landmarks, body HOG and body HOF. Although this approach focused on investigating the affective response of individuals while watching long-term videos, it is theoretically applicable to crowds under the assumptions that crowd members' face and body are visible for a long-enough interval and within an acceptable resolution.

The authors of [10] summarize their previous work on three bio-inspired probabilistic algorithms for perception of emotions from crowd dynamics. The first algorithm starts by partitioning the environment using an Instantaneous Topological Map (ITM) and a Dynamic Bayesian Network (DBN) is employed to model conditional interactions occurred in each subregion. these interactions are then converted into super states using a Self-organizing map (SOM) and the occurrence of these super states (events) are encoded by a Gausian mixture model as positive or negative emotions. The second algorithm starts with the detection of events, collecting them over time to obtain behavioral patterns which are then clustered into classes by means of a DBN; the distribution of these classes are modeled using GMM, building one model for positive and one for negative emotions, to finally detect the emotional state by a likelihood ratio test. In the third algorithm the trajectory of single individuals are expressed as transitions (events) between subregions using a DBN and separated models are constructed for the event sequences labeled with a positive or negative emotion, to conclude with a log-likelihood test to determine the emotion according to the movement pattern exhibit by the individual. All three approaches are tested under a simulated scenario, showing the third algorithm to yield the highest emotion prediction accuracy according to the experiments conducted.

The authors in [11] propose a hierarchical Bayesian model aiming to describe the crowd both at the microscopic and macroscopic level. This approach uses pedestrians trajectories to create a topological map by means of a self-organizing map (SOM), dividing the environment into zones. At the microscopic level, the pedestrians trajectories are described as a Markov process transitioning between zones, and behaviors are modeled according to the origin and destination; each behavior is assigned an emotional label (positive, neutral or negative) according to the time required to reach the estimated destination. At the macroscopic level, the crowd is described by a vector state counting the number of people in each zone at a given time, a second SOM clusters and reduce the dimension of the state vectors to describe the dynamics of the crowd as a Markov process; finally the emotion of the crowd is assigned to be the predominant one as displayed by single individuals. This method is validated by a simulated crowd under different levels of crowd density and multiple behaviors.

3 Datasets

A common test bed is essential to measure and compare performance among different methods. This section examines two types of datasets, those designed for affective states recognition and those intended for crowd analysis. The objective is to determine whether the reviewed datasets are suited to test computational models dealing with emotion estimation in crowds, for which a discussion is provided in section 4. The datasets considered in this section are not exhaustive but rather representative of the diversity available for such tasks. Both groups of affective and crowd analysis datasets are listed in tables 1 and 2 respectively.

3.1 Affective Datasets

This subsection considers datasets intended for all kinds of affective tasks, not limiting to those fitted for crowds and using different sensors and ground truth formats in order to provide a comprehensive view of the available options.

Delving in the task of detecting emotional states, facial expressions have become a popular choice due to their universality and intrinsic relation to emotions [12]. By means of conventional cameras and in a controlled environment, the datasets CMU [13] [14] and FER-2013 [15] collected static images of facial expressions from participants who were requested to act different emotional states following the discrete emotions scheme [16]. Aiming to simplify the collection of static images and to reach a greater number of participants, the authors of the Gamo dataset [17] made use of a web-based interface were participants play a game by performing specific facial expressions captured by a web camera. Progressing from static images only, the CK+ dataset [18] provides sequences of images where participants enact a series of facial expressions, and emotions are described in terms of facial action units [19].

Expanding the scope of behavioral markers, the dataset CREMA-D [20] contains short videos of participants displaying facial and vocal expressions for the study of multi-modal emotion expression and perception, whereas the dataset LIRIS-ACCEDE [21] [22] goes one step further and captures body expressions.

However, as the authors in [23] argue, using conventional 2D cameras lacks robustness as this kind of cameras are subject to poor illumination and changes in lighting conditions. In response, they propose the use of Kinetic cameras as these are able to capture depth, and produce a dataset containing 3D models of several participants performing multiple facial expressions. Moving from emotions (brief affective states) to moods (long term affective states), the work in [24] introduces the EMMA database which employs both 2D and kinetic cameras, and provides longer intervals of data capture as the dataset is intended for mood recognition.

Focusing on physiological measurements, the DEAP dataset presented in [25] [26] collected the electroencephalogram (EEG), electrooculogram (EOG), Galvanic skin response (GSR), blood volume pressure (BVP), temperature and respiration signals of participants. And a frontal video face was recorded for some of those participants. One-minute long excerpts of music videos were used as the stimulus to elicit emotions along the four quadrants of the arousal-valence plane.

3.2 Crowd Analysis Datasets

The fields of computer vision and crowd analysis are favored with an overgrowing importance and share a common interest in studying crowded environments, and as a result, multiple datasets have been produced. In compiling such datasets, cameras remain to be the preferred sensor for studying crowds due to the already widespread use of surveillance cameras in most public spaces.

Depending on the focus of study, datasets are designed to capture the desired circumstances. The popular dataset PETS 2009 [27] collected image sequences from multiple cameras with the aim to serve as a test bed for algorithms intended for people counting, density estimation, people tracking, flow analysis and event recognition. All the presented situations are mainly poor in terms of emotional behavior, except for the scenario S3 (event recognition) where an evacuation (rapid dispersion) is observed and can be associated to an emotional state of fear. The authors of CAD [28] recreated several normal collective behaviors adding the challenges of change in illumination and wavering trees in the background, however, the captured situations are not representative of any emotional behavior. Taking advantage of the large number of people attending the World Exposition of 2010 in Shanghai, the massive dataset Shanghai Expo 10 [29] was gathered. It provides a large amount of annotations at a regional level denoting crowd density, collectiveness and cohesiveness features under normal situations, but it lacks any relevance for inferring affective states. Focusing on groups and crowds, the authors of [30] present the Atomic Group Action dataset targeting the dynamics of group formation, yet no meaningful emotional behavior is exhibit. Rabiee's dataset [31] provides some emotional-rich situations such as panic and fight, although in a staged way. Finally, the S-hock dataset [32] focuses on the behavior of spectator crowds with rich annotations at the individual level, enabling the addition of further affective annotations although restricted to this type of crowds.

4 Discussion

In developing a well-grounded computational model, some theoretical issues regarding emotions in crowds must be addressed. One is to establish a working definition of what exactly is meant when talking about a crowd (e.g. features, properties) as there is no consensual definition in the fields of psychology and sociology. Another related aspect is to state if a method is limited to function for certain types of crowds, as these emerge in very diverse contexts and exhibiting a wide range of behaviors. Finally, it is important to provide a definition of what is meant by emotion and the emotional theory employed (e.g. discrete emotions, valence-arousal). Is the purpose of the method to estimate the emotion of individuals within a crowd or the emotion of the crowd as a whole? Except for [11],

Dataset	Modality	Sensory Data	Annotations	Naturalness
3D Face Model [23]	Facial expressions	Kinetic camera	Normal, happiness, sadness, surprise, anger	Acted
CK+ [18]	Facial Behavior	Image sequences	Anger, disgust, fear, happiness, sadness, surprise, contempt	Acted
CMU [13] [14]	Facial expressions	Static images	Happiness, sadness, anger, neutral	Acted
CREMA-D [20]	Facial and vocal expressions	Camera	Happiness, sadness, anger, fear, disgust, neutral	Acted
DEAP [25]	Facial expressions, physiological measurements	EEG, EOG, GSR, BVP, temperature, respiration	Valence, arousal, dominance, liking, familiarity	Induced
EMMA [24]	Facial and body expressions	Camera, kinetic camera	Valence, arousal	Induced and acted
FER-2013 [15]	Facial expressions	Static images	Happiness, sadness, anger, surprise, disgust, fear, neutral	Acted
GaMo [17]	Facial expressions	Static Images	Anger, disgust, fear, happiness, neutral, sad, surprise	Acted
LIRIS-ACCEDE [21]	Facial, vocal and body expressions	Camera	Valence, arousal	Acted

Table 1. Affective Datasets

Dataset	Modality	Sensory Data	Annotations	Naturalness
Shanghai Expo 10 [29]	Crowd movement	Camera	crowd density, collectiveness and cohesiveness	natural
Rabiee's [31]	Crowd movement	Camera	Panic, fight, congestion, obstacle, neutral behaviors	Acted
PETS 2009 [27]	Crowd movement	Image Sequences	Pedestrians' bounding box and location	Acted
CAD [28]	Crowd movement	Camera	Bottleneck, departure, lane, arch/ring and blocking crowd behaviors	Acted
S-Hock [32]	Individual behavior	Camera	People detection, head detection, head pose, body position, posture, locomotion, action/interaction, supported team, best action, social relation.	Natural
Atomic Group Actions [30]	Group actions	Camera	Group-group actions (formation, dispersal, movement) and group-person actions (person joining, person leaving)	Natural

Table 2. Crowd Analysis Datasets

the examined methods fail to provide a working definition of a crowd and implicitly indicate the type of crowd addressed by the method. Similarly, none provide a definition of emotion but the majority do indicate the emotional theory either implied or clearly stated. If a method is intended to estimate the emotion of individual crowd members then a definition as provided in the introduction of this paper is adequate; however, if a method aims to estimate the emotion of a crowd as a whole, a more clear definition is necessary.

The examined methods tend to model emotions in crowds either at the individual level (microscopic) or at a global (macroscopic) level. Microscopic models describe collectives at the individual level, modeling the emotions, behavior, actions and decisions of single crowd members. Macroscopic models examine the crowd as a whole and describe it by means of global features, as a single entity evolving over time. On its own, both the microscopic and macroscopic models ignore important aspects of a crowd. Microscopic models are unable to capture the collective aspects of emotions in crowds, whereas macroscopic models fail to describe the interaction among individuals that foment the emergence of emotions. When capturing the affects of a crowd, an appropriate model should be able to depict: (a) the emotion of the individual, (b) the interaction between individuals, and (c) the crowd as a whole [33]. The majority of the examined models take a microscopic approach and focus on individual emotions of the members within the group or crowd, failing to capture the interaction of crowd members and the global essence of collective emotions. A common trait in the examined methods is the use of crowd members ambulatory behavior to infer emotional states. The reason to favor the use of ambulatory behavior over facial or body expressions when estimating emotions in crowds is that it enables the method's applicability to more (but not all) types of crowds, with different density levels and possibly limited visibility of the members face and body.

The choice of sensors and features to be used in order to estimate emotional states is central in discussing datasets suitable for crowds. Due to the challenging circumstances of crowded environments, noninvasive sensors such as surveillance cameras are preferred. Either at the microscopic or macroscopic level, the features used for individual emotion estimation are generally not suited for crowded environments due to multiple reasons: the faces and body of pedestrians are not always visible or can suffer from occlusion, and vocal expressions are easily distorted or impractical to perceive. Under these circumstances, visual information about the movement behavior of individuals becomes the most practical cue to infer emotional states but with the cost of higher uncertainty as the relation between behavior and emotion is highly dependent on the context of the situation. Given the above observations, all the examined affective datasets are rendered inadequate for methods devoted to detecting crowd emotions. The examined datasets intended for crowd analysis provide no annotations or meaningful behaviors in terms of emotions, only the S-hock dataset presents sufficient relevant affective information but is limited to spectator crowds. A dataset well suited for emotion estimation in crowds needs to capture diverse and meaningful behaviors accompanied with well validated affective annotations, ideally for multiple types of crowds and different emotions. The absence of publicly available datasets for emotion estimation in crowds is clearly evidenced in how the existing body of literature is evaluated. Throughout this survey, two main trends were identified: simulations and case-specific footage. Simulations are a practical solution for obtaining experimental data, however its arguable how well such simulations can replicate the complexity of emotional behavior. Case-specific footage is advantageous due to its naturalness but a more thorough evaluation in multiple cases is necessary to prove how well a method can be generalized. The absence of a common dataset prevents proposed methods to be properly evaluated and compared, decelerating further research in this area.

Acknowledgements

This work was partially supported by the Erasmus Mundus joint Doctorate in Interactive and Cognitive Environments, which is funded by the EACEA, Agency of the European Commission under EMJD ICE. Prof. Carlo Regazzoni contributed to this paper during his stay at Carlos III University of Madrid as a visiting professor.

References

[1] A. R. Damasio. The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 1996.

[2] M. Hoogendoorn, J. Treur, C. N. Van Der Wal, et al. Modelling the interplay of emotions, beliefs and intentions within collective decision making based on insights from social neuroscience. *Lecture Notes in Computer Science*, 2010.

[3] T. Bosse, M. Hoogendoorn, M. Klein, et al. Agent-based modelling of social emotional decision making in emergency situations. *Understanding Complex Systems*, 2013.

[4] T. Bosse, R. Duell, Z. A. Memon, et al. Agent-Based Modeling of Emotion Contagion in Groups. *Cognitive Computation*, 2014.

[5] A. Sharpanskykh and J. Treur. An adaptive agent model for affective social decision making. *Biologically Inspired Cognitive Architectures*, 2013.

[6] A. Manzoor and J. Treur. An agent-based model for integrated emotion regulation and contagion in socially affected decision making. *Biologically Inspired Cognitive Architectures*, 2015.

[7] W. Mou, H. Gunes, and I. Patras. Automatic Recognition of Emotions and Membership in Group Videos, 2016.

[8] E. Sariyanidi, V. Dali, S. C. Tek, et al. Local Zernike Moments: A new representation for face recognition. In *Proceedings - International Conference on Image Processing, ICIP*, pages 585–588, 2012.

[9] J. Sánchez, F. Perronnin, T. Mensink, et al. Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, 2013.

[10] M. W. Baig, M. S. Baig, V. Bastani, et al. Perception of emotions from crowd dynamics. *International Conference on Digital Signal Processing, DSP*, 2015.

[11] O. J. Urizar, M. S. Baig, E. I. Barakova, et al. A Hierarchical Bayesian Model for Crowd Emotions. *Frontiers in computational neuroscience*, 2016.

[12] P. Ekman. Facial expression and emotion, 1993.

[13] T. M. U. Mitchell. UCI Machine Learning Repository: CMU Face Images Data Set, 1999.

[14] D. Das and A. Chakrabarty. Emotion Recognition from Face Dataset Using Deep Neural Nets. *IET Computer Vision*, 2016.

[15] I. J. Goodfellow, D. Erhan, P. Luc Carrier, et al. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 2015.

[16] P. Ekman. An argument for basic emotions, 1992.

[17] C. Tsangouri. Towards an In-the-Wild Emotion Dataset Using a Game-based Framework, 2016.

[18] P. Lucey, J. F. Cohn, T. Kanade, et al. The extended cohn-kande dataset (CK+): A complete facial expression dataset for action unit and emotionspecied expression. *Cvprw*, 2010.

[19] P. Ekman and W. V. Friesen. The Facial Action Coding System. *Consulting*, 1982.

[20] H. Cao, D. Cooper, M. Keutmann, et al. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 2014.

[21] Y. b. Baveye, J.-N. Bettinelli, E. Dellandréa, et al. A large video data base for computational models of induced emotion. In *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, pages 13–18, 2013.

[22] Y. Baveye, C. Chamaret, E. Dellandréa, et al. A Protocol for Cross-Validating Large Crowdsourced Data: The Case of the LIRIS-ACCEDE Affective Video Dataset. *Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia*, 2014.

[23] S. Chickerur and K. Joshi. 3D face model dataset: Automatic detection of facial expressions and emotions for educational environments. *British Journal of Educational Technology*, 2015.

[24] C. Katsimerou. Crowdsourcing Empathetic Intelligence : The Case of the Annotation of EMMA Database for Emotion and Mood Recognition. *Acm Tist*, 2016.

[25] M. Soleymani, S. Member, and J.-s. Lee. DEAP : A Database for Emotion Analysis Using Physiological Signals, 2012.

[26] G. Placidi, P. Di Giamberardino, A. Petracca, et al. Classification of Emotional Signals from the DEAP dataset. *Proceedings of the 4th International Congress on Neurotechnology, Electronics and Informatics*, 2016.

[27] J. Ferryman and A. L. Ellis. Performance evaluation of crowd image analysis using the PETS2009 dataset. *Pattern Recognition Letters*, 2014.

[28] M. A. Hassan, A. S. Malik, W. Nicolas, et al. Reliability of bench-mark datasets for crowd analytic surveillance. *2015 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) Proceedings*, 2015.

[29] C. Zhang, K. Kang, H. Li, et al. Data-Driven Crowd Understanding: A Baseline for a Large-Scale Crowd Dataset. *IEEE Transactions on Multimedia*, 2016.

[30] R. J. Sethi. Towards defining groups and crowds in video using the atomic group actions dataset. *Proceedings - International Conference on Image Processing, ICIP*, 2015.

[31] H. Rabiee. Novel Dataset for Fine-grained Abnormal Behavior Understanding in Crowd, 2016.

[32] F. Setti, D. Conigliaro, P. Rota, et al. The S-Hock dataset: A new benchmark for spectator crowd analysis. *Computer Vision and Image Understanding*, 2017.

[33] Cabinet Office. *Understanding Crowd Behaviours*. 2009.

[34] E. I. Barakova, R. Gorbunov, and M. Rauterberg. Automatic Interpretation of Affective Facial Expressions in the Context of Interpersonal Interaction. *IEEE Transactions on Human-Machine Systems*, 2015.