Exploring Coherent Motion Patterns via Structured Trajectory Learning for Crowd Mood Modeling

Yanhao Zhang, Lei Qin, *Member, IEEE*, Rongrong Ji, *Senior Member, IEEE*, Sicheng Zhao,
Qingming Huang, *Senior Member, IEEE*, and Jiebo Luo, *Fellow, IEEE*

***Abstract*— Crowd behavior analysis has recently attracted extensive attention in research. However, the existing research mainly focuses on investigating motion patterns in crowds, while the emotional aspects of crowd behaviors are left unexplored. Analyzing the emotion of crowd behaviors is indeed extremely important, as it uncovers the social moods that are beneficial for video surveillance. In this paper, we propose a novel crowd representation termed crowd mood. Crowd mood is established based upon the discovery that the social emotional hypothesis of crowd behaviors can be revealed by investigating the spacing interactions and the structural levels of motion patterns in crowds. To this end, we first learn the structured trajectories of crowds by particle advection using low-rank approximation with group sparsity constraint, which implicitly characterizes the coherent motion patterns. Second, rich emotional motion features are explicitly extracted and fused by support vector regression to reflect social characteristics. In particular, we construct weighted features in a boosted manner by considering the features' significance. Finally, crowd mood is intuitively presented as affective curves to track the emotion states of the crowd dynamics, which is robust to noise, sensitive to semantic shift, and compact for pattern expressions. Extensive evaluations on crowd video data sets demonstrate that our approach effectively models crowd mood and achieves significantly better results with comparisons to several alternative and state-of-the-art approaches for various tasks, i.e., crowd mood classification, global abnormal mood detection, and crowd emotion matching.**

***Index Terms*— Coherent motion pattern, crowd behavior analysis, emotional motion feature, structured trajectory learning (STL).**

I. Introduction

CROWD behavior analysis has attracted ever increasing research attention for various applications, such as video surveillance and beyond. Analyzing the visual crowd behaviors is challenging in recent advances and trends [1], mainly due to the variations, e.g., complex interactions, subtle changes, various semantics, and so on.

Research progress on crowd behavior analysis mainly focuses on designing descriptive motion descriptors, for instance, optical flow (OF) [2], [3], particle flow [4], [5], and local space-time motion pattern [6], which merits in freeing the burden of tracking individuals. To further compensate the lack of high-level semantics, such features are typically combined with some complex models for specific tasks, such as crowd segmentation modeling [4], [7] and crowd collaborating motion estimation [3], [8].

One important task in crowd behavior analysis is to describe the social characteristics, which can be accomplished via understanding the relationship and interaction among individuals, from a mid-level perspective. To construct the useful representations of the crowd behavior, classical sociology [social force (SF) [9], [10]] and dynamics models (sparse reconstruction cost [11] and dynamic texture [12]) with vision-based features are considered. Many works explore the structures of crowds through supervised learning models, such as Bayesian model [13], graphical model [14]–[16], and mixture agents model [17]. By incorporating a data-driven learning method [18], better performance is reported for crowd tracking based on motion patterns. However, the crowd patterns described in these works use only low- or mid-level features, which results in difficulty of digesting the high-level crowd behaviors from a socially holistic perceptive. To encode more semantic information, high-level representations, such as group structure [19], streakline potentials (SPs) [20], and energy potentials [21], are further proposed, in order to discover more insightful structures for crowd analysis. Compared with low-level motion features, mid-level features aim at modeling object interactions while high-level semantic features with rich prior information provide a powerful means toward better describing and gaining a deeper understanding of crowd behaviors. High-level representations integrate physics, psychology, and sociology to increase the descriptive power and diversity.

A. Goal

In this paper, we target at establishing a novel crowd representation called crowd mood. Under the guidance of

Manuscript received October 21, 2015; revised March 27, 2016; accepted July 15, 2016. Date of publication July 20, 2016; date of current version March 3, 2017. This work was supported in part by the National Basic Research Program of China (973 Program) under Grants 2015CB351802 and 2012CB316400, and in part by the National Natural Science Foundation of China under Grants 61402388, 61332016, 61133003, 61390510, and 61572465. This paper was recommended by Associate Editor T. Bouwmans. *(Corresponding authors: Lei Qin; Qingming Huang.)*

Y. Zhang and S. Zhao are with the School of computer science, Harbin Institute of Technology, Harbin 150001, China (e-mail: yhzhang@hit.edu.cn; zsc@hit.edu.cn).

L. Qin is with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: lqin@jdl.ac.cn).

R. Ji is with the Department of Cognitive Science, School of Information Science and Engineering, Xiamen University, Xiamen 361005, China (e-mail: rrji@xmu.edu.cn).

Q. Huang is with the School of computer science, Harbin Institute of Technology, Harbin 150001, China, and also with the University of Chinese Academy of Sciences, Beijing 100190, China (e-mail: qmhuang@jdl.ac.cn).

J. Luo is with the Department of Computer Science, University of Rochester, Rochester, NY 14627 USA (e-mail: jiebo.luo@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2016.2593609

psychological research [22], [23], crowd mood attempts to project different types of statistical motion features into the emotion space, and further assign emotion labels to crowd motion behaviors. Therefore, in this paper, we try to answer *What is the crowd mood? How to describe the mood of crowd motion behavior?* and *What is the relationship between different crowd moods?* To the best of our knowledge, this is the first work on modeling the emotion states of crowds for behavior analysis.

B. Inspirations

Zeitz *et al.* [22] defined the crowd mood as: more practically, the term crowd mood has become an accepted measure of probable crowd behavior outcomes. This is particularly true in the context of crowds during protests or riots, where attempts have been made to identify factors that lead to a change of mood that may underpin more violent behavior. It is further admitted in [22] that crowd mood hails from the crowd type and is more of a psychosocial descriptor of crowd behaviors. Crowd behavior is the demonstrable factor that requires assessment and monitoring to underpin management actions. Inspirations also come from modeling of flocking behavior [24] that are controlled by three simple rules, i.e., separation, alignment, and cohesion. These principles make the flock extremely realistic, creating complex motion and interaction. The basic model is further extended to incorporate the effects of fear and transmit emotion between animals [25].

As mentioned earlier, crowd mood can be regarded as a high-level representation that essentially captures the status of crowd behaviors. It is, therefore, a natural question that the detection of crowd mood, if not impossible, can provide a substantial benefit toward the analysis of crowd behavior. The key barrier lies in the design of discriminative spatial–temporal representation to precisely capture the crowd mood from video sequences.

More particularly, we resort to analyzing the coherent motion patterns to describe crowd mood. First, as a crucial basis, coherent motion patterns are profiling the most representative sketch of crowds that serve as the dominant visual features, which benefits the descriptive ability of crowd mood. Considering the definition of crowd mood, various emotions of the crowd reflect different patterns of motion, which shows implicit correlations between coherent motion patterns and the types of crowd mood. Second, coherent motion patterns are basic elements within crowds that describe both social attributes and conventions of crowd dynamics [10], [21], [22]. Emotional information can be interpreted as social attributes, which can reflect the changes of crowd behaviors. Similarly, the interactions between individuals tend to activate crowd emotion. As such, crowd mood is a demonstrable aspect of groups' feelings because of its capacity to explore semantic patterns from coherent motion patterns. Third, statistic and optical measures on the coherent motion patterns (variance, entropy, heterogeneity, and saliency) [2], [3] are often studied to describe the crowd dynamics, which could bridge the gap of low-level features and crowd emotional states. In our context, crowd mood would be a meaningful semantic measure of crowd behavior, derived from coherent motion pattern that can also draw the discriminative characteristics and result in greater understanding of crowd behavior.

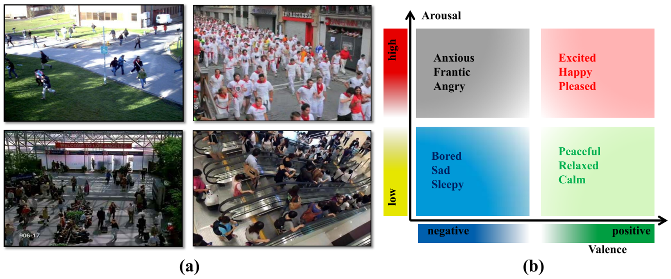


Fig. 1. What is the crowd mood of different crowd motion behaviors? As samples shown in (a), we propose a novel representation to define and describe the mood of crowd motion behaviors in videos. To uncover the relationship between crowd motion patterns and emotion, we adopt the Thayer's A-V emotion space [26] [as shown in (b)] to recognize the ambiguity of emotion states and describe basic emotion classes in terms of A and V. The emotion states of crowd motion patterns intuitively show that different motion patterns reflect different crowd moods and they can transit from one to another. More details can be found in text.

C. From Motion to Emotion

Basic types of crowd motion can reflect representative emotions of the crowd [22]. The mood of crowds at mass gatherings has been measured with a simple matrix involving the study of movement, participation, and behaviors [23]. In order to capture the subtle emotion changes that crowds exhibit in terms of the motion patterns, the emotion state is characterized in numerous latent dimensions [26], [27]. Thayer [26] proposed the arousal–valence (A-V) emotion plane as a 2D emotion space for automatic and continuous emotion recognition, which has been popularly applied to the domain of emotion computation [28]–[30]. Arousal describes the level of emotion activation and characterizes the strength of emotional responses from passive to active. Valence represents the type of emotional happiness, which demonstrates a degree of emotional response from negative to positive. As shown in Fig. 1(b), different sections of the A-V emotion space refer to the emotion from active-positive to passive-negative. The emotion plane can be viewed as a continuous space and each point of the plane can be recognized as an emotion state.

Fig. 1(a) also shows that crowd mood can be evoked from different crowd motion behaviors in various scenes and contexts. For instance, motion patterns of pedestrians in the shopping mall can reflect relaxed, parade ceremony, which makes the crowd excited and active, and anxious or angry that occurs when the crowd is in a scattering and fleeing pattern. Accordingly, to capture the subtle crowd emotion changes, this paper focuses on recognizing crowd mood associated with emotion classes in the A-V emotion plane. In this way, a continuous perspective of A-V model can be successfully adopted to model the emotional responses of the crowds as used in [31] and [32].

D. Our Method

Our method consists of three components (shown in Fig. 2). The first component is structured trajectory learning (STL)

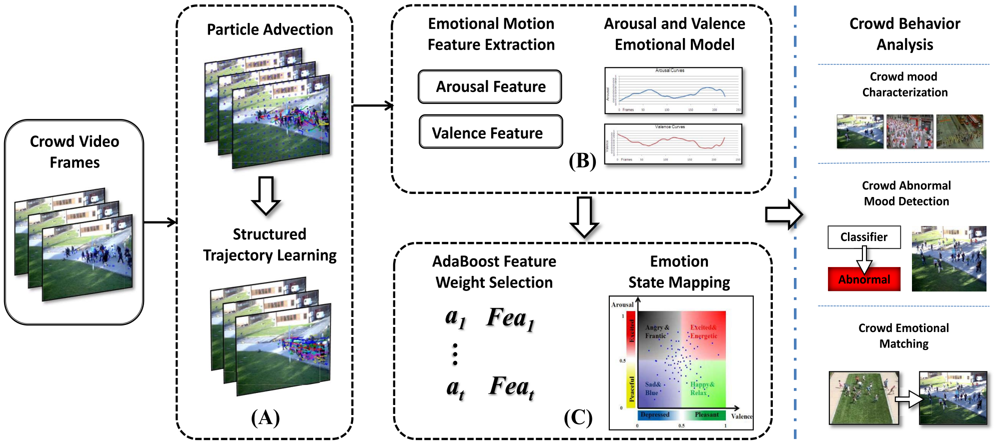


Fig. 2. Framework of the proposed crowd mood model for crowd behavior description and analysis. The crowd mood representation consists of three components. (A) STL. (B) A-V motion feature extraction. (C) Crowd mood modeling.

to extract typical coherent motion patterns of the crowds. We utilize a low-rank learning method to obtain compact and coherent trajectories with a group sparsity constraint incorporating spatial information. The second component is A-V motion feature extraction to extract features for the A-V dimensions by statistical analysis of the motion patterns. Here, we adopt the Thayer's A-V emotion plane [26] to express the emotion states of the crowd behavior. Emotional motion features are extracted corresponding to A and V dimensions. The third component is crowd mood modeling to model crowd mood by mapping the features into the emotion states. By utilizing a boosting approach, the features become more discriminative under the learned importance weights. The learned weighted features are subsequently fed into a support vector regression (SVR)-based model to produce the emotion output. Crowd mood is intuitively presented as affective curves to track the emotion states of the crowd dynamics, which can represent and identify the crowd behaviors and be applied to crowd analysis tasks. The outline of this paper is organized to review related works and summarize social emotional principles in Sections II and III. Sections IV-A–IV-C correspond to components (A)–(C) in Fig. 2, respectively. Experiments and analysis are discussed in Section V. Section VI concludes this paper.

II. Related Work

Crowd mood is proposed as a high-level representation for crowd behavior analysis. Specifically, we explore discriminative and compact emotional representation of crowds via STL. We also point out the works close to ours in the emotion modeling, which covers the relations and differences with others. In this section, related works to the above two directions are reviewed.

A. Representation of Crowds

Efficient representation is the foundation for modeling crowd behaviors. The main difficulty lies in how to effectively describe the structure of crowd motion patterns and represent the typical classes of the crowd. Hierarchical representation of targets from groups to individuals in crowds is proposed, combining motion and appearance saliency features [33]. This paper differs from the works that directly describe the motion patterns of crowd, such as spectral clustering on the OF field [8], direction and velocity in the motion flow [3], crowd kinetic features [2], coherent structure of fluid dynamics [34], or deformable shape matching for crowd modeling [35]. This paper is similar in describing the motion of crowds by introducing the structural contexts [36], [37] or high-level concepts [20], [21], yet from an emotion modeling point of view. In the term of describing relationships and interactions in crowds, crowd mood emphasizes imposing different social hypotheses to model emotion states of crowd dynamics, which also differs from the sociology model proposed in [9] and [10]. We also learn the coherent motion pattern of crowds in low-rank subspace. The technique of low-rank approximation has been widely used to solve computer vision problems, such as motion estimation [3], background subtraction [38], and so on, due to its powerful capability to analyze large-scale data. This paper is especially related to the approaches that learn the structure of the original data [39]. We expand more suitable priors that not limited to low rank for the STL.

B. Emotion Modeling

Emotion modeling and recognition draws extensively from disciplines, such as vision, psychology, or cognitive science. A significant amount of recent works explore the behavioral characteristics involving different types of signals. Wang and Cheong [40] develop a systematic approach to address emotional understanding in films. A number of effective audiovisual cues are formulated to help bridge the emotional gap. Zeng *et al.* [41] propose audiovisual emotional recognition on 11 emotional states for human–computer interaction applications. In [42], facial expressions are recognized through a spatio-temporal approach and mapped to levels of interest using emotional space.

Different from the works using behavioral modalities with various emotional models, we are interested in the holistic crowd mood in videos rather than the observer's personal mood induced by the perceived audiovisual stimuli. Most relevant efforts attempt to estimate positive or negative emotions of a crowd based on simulated data using dynamic probabilistic models, corresponding to behavioral changes under different encountered situations [43], [44]. This paper quantifies the emotion responses of the crowd based on dimensional emotion plane [26], which conceptualizes human emotions by defining where they lie in two/three dimensions. So far, the most popular psychological model in the affective content analysis is the A-V model [Fig. 1(b)] proposed in [26], which is known as the dimensional affective model. It is also commonly employed for emotion state representation and modeling in the domain of emotional perception analysis. Specifically, the model utilizes two components, i.e., A and V, to represent affective responses. By dividing abstract emotional states into two components, the A-V model has some advantages: 1) complicated emotions can be expressed by combining A and V in various ways and 2) the model is very generic, which not only bridges between the emotions and affective features, but could also be extended to different video genres with different A and V description algorithms. Due to such important features, the dimensional affective model is utilized in this paper for crowd mood modeling and representation. Representative works in [30], [45], and [46] express the emotion states of video segments by the linear feature combination of A and V values. Unlike [47] that exploits [26] for affective content analysis to infer the observer's mood, we emphasize detection of coherent motions and extracting emotional features to characterize the emotions that are actually evoked by the entire crowd.

C. Our Contributions

Compared with the previous works on learning and modeling of crowd behaviors, we make the following major contributions.

1) We propose a novel crowd behavior representation called crowd mood, which considers the social emotional principles by statistical analysis of underlying characteristics of crowds. It provides an emotional description that socially reveals the high-level semantic information for crowd behaviors.
2) We put forward an STL approach in low-rank space with group sparsity constraint, which utilizes alternating greedy optimization to explore coherent motion pattern.
3) We investigate rich emotional motion features that are weighted by Adaboost, and map them in the emotion space by SVR.
4) Crowd mood provides a unified solution for representing the semantics of crowd behaviors and intuitively tracks the emotion states of the crowd dynamics as affective curves, which achieves promising and robust performances on several crowd applications in a variety of crowd scenes.

III. Social Emotional Principles

Recent studies [22], [48], [49] on crowd behavior have confirmed the effectiveness of using psychological theories in recognizing a variety of crowd events. Several reoccurring themes have been revealed. In [22], crowd mood and motion behavior were acknowledged as complex phenomena influenced by social conditions, spectator personalities, and situational changes of the environment. These properties indicate that social emotional principles share joint characteristics with social motion behaviors, allowing us to analyze the implicit correlation between emotion states and crowd behaviors.

In the crowd, people are driven by the SF of decisions regarding destinations or desired directions, which is influenced by the environment as well as the interactions between other people. This is a common characteristic of social behavior [9], [10], [21]. As discussed in [21], the interaction potential between individuals is also explicitly presented as the crowd emotion. From the intuitive observations of social behavior, we make three reasonable hypotheses to sum up the social emotional principles.

A. Spacing

People have a general awareness that prevents them from colliding with each other in the crowd. This process can be modeled by an energy potential minimization while trying to maintain desired speed and motion direction [21]. The theory of flocking behavior [24] also confirms that self-propelled entities exhibit collective motion in crowds with basic spacing rules. These are all related to the spacing interaction. Individuals in the crowd modify the existing norms within the neighborhood spacing. The crowd gradually builds up the reacting emotion due to this spacing interaction.

B. Structure

Based on the self-categorization theory and social identify theory [48], collective behavior and social influence are only possible on the basis of shared self-categorization or shared sense of identity. That means people with the same goal or the same motion direction tend to keep the motion consistent. They rarely repel each other while self-organizing into a structured crowd, e.g., in a marathon race game or groups of friends. The crowd emotion of groups is pleasant in such structural motion. A chaotic situation in the unstructured motion pattern causes the crowd to generate depressed and bored moods without sharing the sense of identity [22]. The structural level of motion pattern affects the emotion states in terms of this hypothesis.

C. Statistics

It is clear that the motion pattern is an impacting factor that describes the mood in the crowded scenes. In particular, we obtain the OF as the basic motion patterns from the sample frames in PETS2009 [50]. Fig. 3 shows a sample frame along with the OF (visualization in HSV space). We also show the comparison on the probability density and cumulative distribution between the magnitude of OF and pixels values. Intuitively, it shows that the coherent motion pattern has sparse and structural statistical characteristics, which share

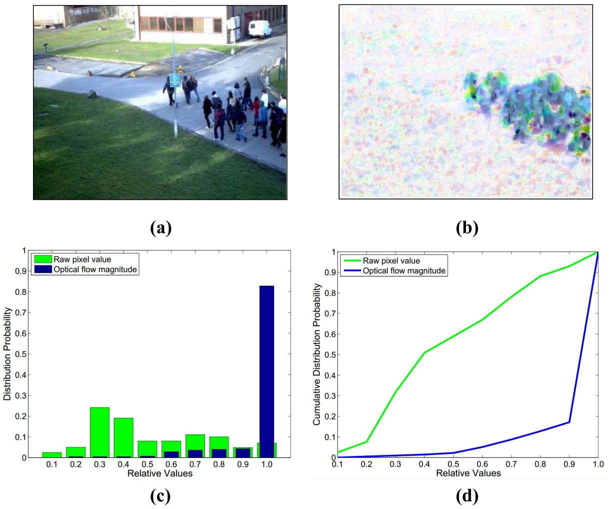


Fig. 3. Motion distribution of the crowd motion patterns shows the sparse and structural statistical characteristics, which is synonymous with Laplace distribution. (a) Video frame. (b) Visualized OF. (c) Probability density of OF magnitude (blue region) and frame pixels (green region). (d) Corresponding cumulative distribution of (c).

a significant similarity with the Laplace distribution. Based on the statistical observations, we discover: 1) motion value is sparse in crowds, which means the sketch of the motion is only located in a sparse portion of the scene and 2) high motion value tends to be located in the patterns with the structural property. Since the crowd videos are full of noisy movements and lack of discriminative appearance cues, the statistical property of the coherent motion pattern is useful for quantitatively analyzing the crowd mood.

From emotion perspective, crowd mood can be reflected by the motion patterns of crowd behaviors, which can be detected by integrating different statistical feature channels. Therefore, we construct the effective motion feature responses based on social emotional principles in order to capture the semantic crowd motion. Different from the previous works focusing on spatial–temporal features or exemplars of the motion patterns, our representation tries to mine coherent motion patterns beyond the particle flow [5] and distinguish them in the low-rank subspace. The process is conducted based on the mentioned observations of the motion patterns. Crowd mood is thus proposed as a statistical representation that accounts for spacing and structure of crowd motion behavior. To this end, compact and discriminative trajectories that reflect the coherent motion are required in order to represent the basic and structured patterns as a first step.

IV. Crowd Mood Representation

A. Structured Trajectory Learning

In a dense crowd scene, the particle advection scheme [10], [20] is utilized to model crowd motion behaviors by treating the individuals as particles. Meanwhile, the particles are moving with the OF field that reflects the property of the continuous evolution in the crowd motion. To start with, a homogeneous grid of particles is placed on the frame with an adaptive scale setting by investigating the crowd density. Subsequently, along with a bilinear interpolation of the OF field, velocities for each particle are computed using the fourth-order Runge–Kutta algorithm. The particles will follow the trajectories in a fluid flow under the guidance of the average neighborhood.

Notations: Given a video sequence, k particles trajectories are obtained over T frames during the particle advection scheme. The trajectory of a particle p_i in the flow field consists of T coordinates

$$p_i = \left[x_i^1, y_i^1, x_i^2, y_i^2, \ldots, x_i^T, y_i^T\right] \in \mathbb{R}^{1\times 2T} \tag{1}$$

where x_i^t and y_i^t denote the position vectors of the particle i at time t, which are obtained from the OF. The collection of k trajectories in the dynamic motion can be represented by the constructed trajectory matrix

$$M_o = \begin{pmatrix} p_1 \\ \vdots \\ p_k \end{pmatrix}^T = \begin{pmatrix} x_1^1, y_1^1 & \ldots & x_1^T, y_1^T \\ \vdots & \ddots & \vdots \\ x_k^1, y_k^1 & \ldots & x_k^T, y_k^T \end{pmatrix}^T \in \mathbb{R}^{2T\times k}. \tag{2}$$

Note that M_o contains all of the coordinates information in a data matrix, in which the trajectory matrix is the typical reflection of the dynamic motion of the crowd behaviors. We then extract coherent motion from M_o, which covers a wide range of the flow fields and is more descriptive for representing crowd behaviors. However, noises exist in video frames, which may be caused by illumination changes, distortions of video, or background motions. Thus, motion patterns can be divided into foreground coherent motions and noises that are caused by the elements above. Here, we target at decomposing motion patterns into representative coherent patterns and sparse noises, which can be formulated as the following:

$$M_o = M_{\text{coherent}} + M_{\text{noise}} \tag{3}$$

where M_o is the original trajectory matrix, M_{coherent} denotes the matrix of coherent trajectories, and M_{noise} refers to the noise trajectory matrix. Obviously, it is difficult to find proper M_{coherent} or M_{noise} without any constraint. We handle this issue by proposing a low-rank approximation with group sparsity constraint, which incorporates three effective priors to robustly solve the problem, as follows.

Low-rank prior: Typical coherent motion subspaces can be learned to represent the sketch of crowds. It inspires us to compute a low-rank approximation to the coherent trajectory matrix with representative subspaces.

Group sparsity prior: Another important prior is motivated by the fact that the noises of trajectories are occupying structural portions of the scenes. Either background movements or illumination changes occur as the structure of groups, which means the noise patterns are in a high correlation within the sparse groups.

Independency prior: We further constraint that the motions of different trajectories are independent of each other, since the movement of groups in crowd flow may not smoothly transfer to each other.

Based on the above priors and notations, we formulate our objective function as the following equation:

$$\tilde{M} = \underset{M_{\text{coherent}}}{\arg\min} \left(\|M_o - M_{\text{coherent}} - M_{\text{noise}}\|_F^2\right)$$
$$\text{s.t. } \operatorname{rank}(M_{\text{coherent}}) \leq a, \quad \lambda \sum_{i=1}^{p} \|i(M_{\text{noise}})\|_2 \leq bk \tag{4}$$

where M_{coherent} denotes the representative coherent motion patterns following a low-rank a constraint. $i(.)$ denotes the ith group sparsity structure, which encourages sparsity across the groups rather than encouraging sparsity within the groups. M_{noise} satisfies the group structural sparsity constraint, which ensures that the noises are spatially sparse, so as to guarantee that elements in each group unit are either zero or nonzero together. The number of noise trajectories should be smaller than a certain ratio b of all trajectories k, where b controls the sparsity of noise trajectories. The number of trajectory groups is up to bk. $\|.\|_F^2$ is the Frobenius norm. λ refers to the balance parameter. The formulation takes the advantages of low rank and group sparsity, which provides an important clue for modeling typical coherent motion patterns as well as the structural information of the noise motion patterns.

1) Greedy Optimization: Solving (4) is not a convex optimization problem with respect to M_{coherent} and M_{noise}. Alternating greedy optimization is utilized to solve two unknown matrices by using two steps iteratively.

Step A has the following form:

$$\tilde{M}^A_{\text{coherent}} = \underset{M}{\arg\min} \left(\left\|M^A_o - M^A_{\text{coherent}}\right\|_F^2\right)$$
$$\text{s.t. } \operatorname{rank}(M_{\text{coherent}}) \leq a \tag{5}$$

where $M^A_o = M_o - M_{\text{noise}}$. For Step A, we use the Colibri [51] to make a fast and space-saving low-rank approximation to update M_{coherent} while keeping M_{noise} fixed. To solve (5), Colibri algorithm iteratively finds a nonredundant basis as exemplars, while eliminating linearly dependent columns to obtain time and memory efficiency. Specially, M_{coherent} can be decomposed into three parts as the following:

$$M_{\text{coherent}} = \text{LUR} \tag{6}$$

where L refers to a representative motion subspace by selecting independent columns from M_{coherent} that also satisfies the independency prior. R is computed by $L^T M_o$. U is generated by minimizing $\|M_o - \text{LUR}\|_F^2$.

For Step B, we assign M_{coherent} as (6) and keep it fixed. Then, M_{noise} is updated by

$$\tilde{M}^B_{\text{noise}} = \underset{M}{\arg\min} \left(\left\|M^B_o - M^B_{\text{noise}}\right\|_F^2\right)$$
$$\text{s.t. } \lambda \sum_{i=1}^{p} \|i(M_{\text{noise}})\|_2 \leq bk \tag{7}$$

where $M^B_o = M_o - M_{\text{coherent}}$. In optimization, the iteration reaches convergence when the matrix M_{coherent} becomes stable. For simplicity, we estimate the approximate group structure by using threshold to preserve bk groups with the largest values. The implementation of structured trajectories learning is presented as Algorithm 1.

Algorithm 1 Structured Trajectory Learning

```
Input:Original trajectories $M_o$;
Initialize $L$, $U$, $R$ by Colibri Decomposition from $M_o$;
while $\left\|M^i_{coherent} - M^{i-1}_{coherent}\right\| > \varepsilon$ and $i < i_{\max}$ do
    Step A:
    Fix $M^{i-1}_{coherent} = L_{i-1}U_{i-1}R_{i-1}$;
    Update $M^i_{noise} = M_o - M^{i-1}_{coherent}$;
    Sort the $S(m) = ||M^i_{noise}(m,:)||_2$ by descend;
    if $S(m) < bk$ then
        S
    end
    et $M^i_{noise}(m,:) = 0$  Step B:
    Fix $M^i_{coherent} = M_o - M^i_{noise}$;
    Update $M_i$ by decomposing into $L_i$, $U_i$, $R_i$ using Colibri;
    Iteration number $i++$;
end
Output:Structured trajectories $M_{coherent}$.
```

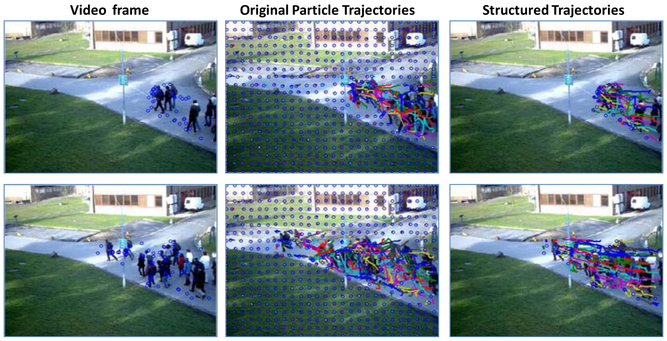


Fig. 4. Comparison between our proposed structured trajectories and the original particle trajectories for the 50th frame (top row) and 70th frame (bottom row). First column: video frame in PET2009 S3 [50]. Second column: original particle trajectories. Third column: structured trajectories.

We test our structured trajectories (rank = 85) by comparing them with original trajectories (rank = 284) for 288 × 384 pixels video frames. As visualized in Fig. 4, our trajectories can reflect structured and coherent motion patterns of the crowds and are basically robust to the various noises, such as illuminations, distortions, and background movements. In most cases, especially in the crowd scene, the structured trajectories can capture typical coherent motion patterns that cover most individual movements with less noise. By the means of STL, we obtain relatively pure subspaces that reflect the coherent motion patterns for the crowd emotional motion feature extraction.

B. Arousal–Valence Motion Features

After obtaining the compact and structured M_{coherent}, statistical motion features (shown in Table I) are extracted from M_{coherent} to model the statistical and spacing hypotheses. Basic coherent motion features of time t from trajectory i include orientation and magnitude obtained from M_{coherent}-based feature channels, listed as

$$\text{Mtn}^t_i(x_i, y_i) = \sqrt{\left(x^t_i - x^{t-1}_i\right)^2 + \left(y^t_i - y^{t-1}_i\right)^2} \tag{8}$$

$$\text{Ori}^t_i(x_i, y_i) = \arctan\left(\frac{x^t_i - x^{t-1}_i}{y^t_i - y^{t-1}_i}\right) \tag{9}$$

TABLE I
EMOTIONAL MOTION FEATURES AND RELATED COMPONENTS

Component	Extracted motion features
Arousal	Kinetic Energy, Magnitude Entropy, Orientation Entropy, Magnitude Variance
Valence	MLCM Energy, MLCM Correlation, MLCM Homogeneity, Orientation Variance

where x_i^t and y_i^t are the position of the trajectory p_i. We obtain the magnitude $\mathrm{Mtn}_i^t(.)$ and orientation $\mathrm{Ori}_i^t(.)$ from trajectory p_i at the time t. We build the orientation histogram to calculate the orientation variance and orientation entropy based on $\mathrm{Ori}_i^t(.)$ with N bins in π ($N = 36$ in this paper)

$$H = \{h_i\}_{i=1}^{N}. \tag{10}$$

We further obtain the orientation variance and orientation entropy from the orientation histogram H, which are shown as the following equations:

$$\mathrm{Ori\_Var} = \mathrm{std}(H) \tag{11}$$

$$\mathrm{Ori\_Ent} = -\sum h_i \log h_i. \tag{12}$$

Arousal dimension captures the strength of the emotion state, while valence is the reflection of the types of emotion states [26]. Motion intensity and uncertainty features are closely related to arousal components, which typically describe the strength of motions. The features we used for arousal include: kinetic energy [2], magnitude entropy, orientation entropy, and magnitude variance [3]. The values of arousal features are increasing as the growth of the motion intensity, which reflect the strength of the moods. In order to capture the types of motion patterns and underlying interactions between groups, we extract texture statistics features to represent flow patterns. Magnitude related features are calculated based on Mtn_i^t. Particle magnitudes are mapped to the image plane using bilinear interpolation to obtain the magnitudes distribution values. The level of values is specified as eight to quantize the magnitude value set Mag.

To that effect, magnitude level co-occurrence matrix (MLCM) is created, which measures the quantity of co-occurring magnitude values in the discrete-valued magnitude set Mag. The definition is as follows:

$$\mathrm{MLCM}(i, j) = \sum_{p \in I} \theta_p \tag{13}$$

where p denotes a point of frame I, $p' = p + \delta$ is an offset point of p, and

$$\theta_p = \begin{cases} 1 & \text{if } [\mathrm{Mag}(p)] = i \text{ and } [\mathrm{Mag}(p')] = j \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

where θ_p records general message of adjacent spacing and amplitude of variation in the $\mathrm{MLCM}(i, j)$. Magnitude textural statistics calculations use the contents of the MLCM that provides a measure of the variations in intensities. It reflects the motion structure characteristics with location distributions and can be used to distinguish different types of motion behaviors.

Algorithm 2 AdaBoost Feature Weight Selection

```
Input:Feature array $fea^{(k)}(fea_1^{(k)},...fea_N^{(k)})$ and ground truth array $G^{(k)}$, $k = 1...M$ of $M$ video frames.;
Initialize the importance weight $D_i = \{w_i\}_{i=1}^{N}$ as $\frac{1}{N}$;
for $k = 1$ to $M$ do
    for $i = 1$ to $N$ do
        Get weak classifier $h_i^k$:$(fea_i^{(k)} \rightarrow \{0, 1\})$ with the lowest error in the pool $\epsilon_k = \mathrm{Pr}_{i \sim D_i}$;
        Choose voting weight for the classifier $h_i^k$:
        $a_k = \frac{1}{2}\ln(\frac{1-\epsilon_k}{\epsilon_k})$;
        Update importance weights:
        $D_i^{k+1} = D_i^k \exp(-a_k(G^{(k)} - 0.5)(h_i - 0.5))$;
    end
end
Output:Voting Weight $a_k$.
```

Practically, we utilize the second-order statistical measurement to express the magnitude of textures. By quantitative observations, MLCM energy, MLCM correlation, and MLCM homogeneity are relevant and descriptive enough to capture the structures of different flow fields and obtain satisfied results. The detailed implementation of calculations is shown in the following:

$$\mathrm{MLCM\_Egy} = \sum_{i,j} \mathrm{MLCM}(i, j)^2 \tag{15}$$

$$\mathrm{MLCM\_Cor} = \sum_{i,j} \frac{(i - \mu_i)(j - \mu_j)}{\sigma_i \sigma_j} \mathrm{MLCM}(i, j) \tag{16}$$

$$\mathrm{MLCM\_Hom} = \sum_{i,j} \frac{1}{1 + (i - j)^2} \mathrm{MLCM}(i, j)^2 \tag{17}$$

where $\mu_{i,j}$ and $\sigma_{i,j}$ denote the means and variances for rows and columns, respectively. The values of textural features are at relatively low values when motion structures are inhomogeneous with high local intensity variation, which represents a chaotic motion pattern with unpleasant crowd mood. Oppositely, the high value expresses homogeneous patterns with pleasant mood. The feature response is in agreement with consistency and exclusion structure hypotheses. Besides the MLCM texture features, we also add the orientation variance to describe the valence component. Finally, we employ Gaussian normalization to normalize these features into (0, 1).

C. Crowd Mood Modeling

1) Adaboost Feature Weight Selection: Due to the complexity of crowd behaviors and various types of motion patterns, different motion features make relatively different contributions to describing the moods of the crowd. Thus, we conduct an implicit manner of boosting to seize essential properties and balance the weights between features. The ground truth of the emotion states is divided into a weight selection part and model training part. The details of obtaining the ground truth will be listed in Section V-A. The boosting-based feature weight selection process can be summed up as Algorithm 2.

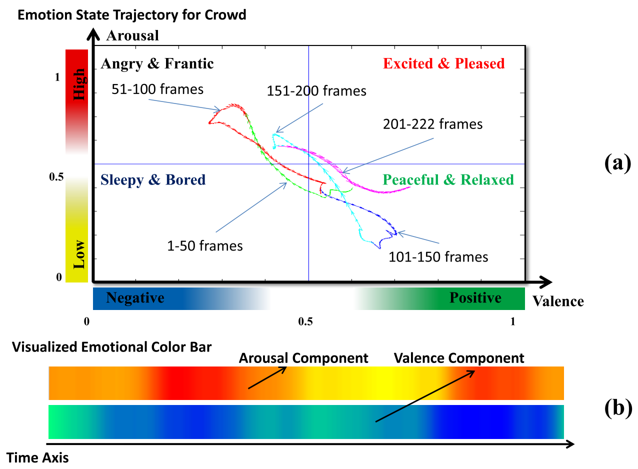


Fig. 5. Crowd mood curve (PET2009 S3 14–16). (a) Emotion state trajectory for the crowd. (b) Visualized emotional color bar for the components.

2) Emotion State Regression: In our case, we utilize the value of fea$_i^{(k)}$ itself as the weak classifier h_i^t that are all normalized to (0, 1), where i and k are the indices of the features and frames. After extracting emotion features from the crowd frames, the feature with selected weight $a_k \times$ fea$^{(k)}$ will be regarded as the balanced feature to be mapped into the corresponding emotion states. In other words, we map the input as a multidimensional vector into the output as an emotional value between (0, 1) that could be solved as a regression problem. Various regression methods, such as least square, logical regression could, therefore, be applied. Due to its fast implementation and better perdition on the unseen data, SVR provides a better solution for the regression problem. In our case, the A and V components are calculated with ϵ-SVR [radial basis function (RBF) kernels], which provides an efficient solution for the regression problem. A and V models are trained independently with the corresponding features and the ground truth to map the multidimensional vectors to (0, 1) following the steps in [30], [46], and [47]. By the means of regression, the features are mapped into the related emotion states through A and V models.

3) Crowd Mood Curve: We divide the space into four quadrants in which different regions represent different emotion states of the crowd behaviors. The mapped emotion state of each crowd video frame can be denoted as a 2D point in the A-V space. We smooth the emotion points by quadratic fit to avoid the errors and roughness caused by the regression. We demonstrate the crowd mood curve result for PETS2009 S3 14–16 [50] [Fig. 5(a)] with the visualized A-V color bar [Fig. 5(b)]. This affective curve presents the emotion state of the crowd intuitively. In Fig. 5(b), gradual changes of the color indicate the evolution change of the emotion components. Note that the crowd state of the video first starts from relaxed to anxious (1–50 frames), then returns back to peaceful (51–150 frames), and finally to frantic again (151–200 frames). From the crowd mood curve, we can intuitively observe the evolution of the crowd motion behavior and predict the changes of crowd emotions. Since the emotion states of crowd are abstract, the crowd mood curve is sensitive and descriptive enough to allow crowd analysis for practical use, because such curve can intuitively reveal subtle changes of emotion states.

V. Experiments

In this section, three groups of quantitative comparisons are included to evaluate the effectiveness of crowd mood representation, including: 1) characterizing different mood types (in Section V-B), which aims to describe the effects of each part to test the discriminations of our representation for distinguishing the moods of crowds; 2) detecting abnormal mood (in Section V-C), which equates to evaluating the capability of capturing the semantic changes of crowd abnormal behavior; and 3) the performance of crowd emotional matching (in Section V-D) that investigates how the visual content affect the crowd mood and the robustness for the visual changes. At the beginning, we provide the details of the data set and ground truth for our tasks. The techniques of event recognition and abnormal detection have been researched for years. However, none of them were directed at the mood of the crowd. Moreover, the proposed representation involves coherent motion pattern extraction and feature weight selection for modeling the emotional factors. Therefore, the discovered crowd mood is basically generated from the motion patterns. In other words, our representation operates on a motion clue basis rather than a multiple clues basis. Admittedly, a multimodal approach maybe more sufficient and robust due to multiple clues, and a visual content-based method derived from the motion patterns should always be easily applicable to any crowd video.

A. Database and Ground Truth

In order to evaluate the effectiveness of crowd mood, we conduct three experiments, including crowd mood characterization experiment, global abnormal mood (GAM) detection, and crowd emotional matching. In the particle advection scheme, every five pixels, we set a particle in the OF field and the length of the trajectory T to ten frames. A total of approximately 100 videos are retained for all experiments. Note that we evaluate the proposed approach using clips containing dense crowds rather than small groups of individuals, because motion patterns are the most significant clue for analysis. Our database contains UCF data set [10], UMN data set [52], and PET2009 [50] for specific crowd tasks. To our knowledge, these are the most frequently used and representative data sets for crowd behavior analysis. Some samples of labeled results are shown in Fig. 6. Apparently, it would be more informative to model the crowd mood if we make use of the appearances, expressions, and sounds in crowds. However, most data sets for crowd analysis are missing audio information and it is difficult to recognize faces under the resolution. In general, it is unrealistic to expect an automatic algorithm to provide the emotion of the whole crowd due to complex semantics about the scene, event, and environment setting. Given the challenges and ambiguities, we believe a third party of ground truth would be valuable. To address this problem, the third-party observers are made to record the common response and understanding of the crowd. Therefore, the third-party ground truth could represent the emotional information

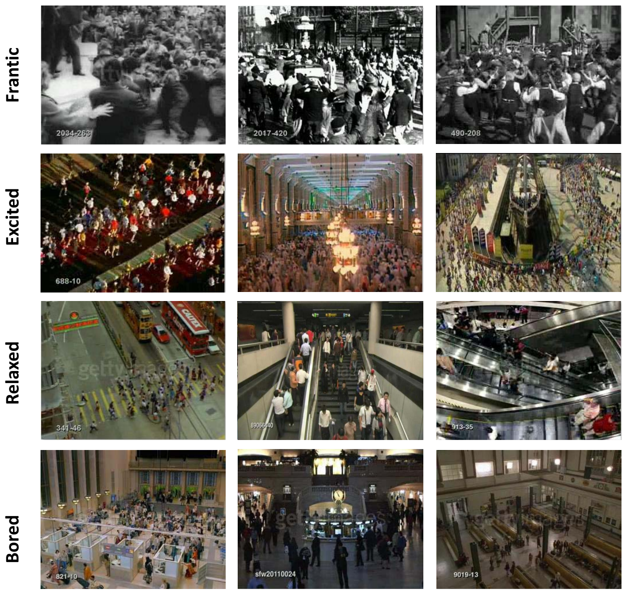


Fig. 6. Examples of the ground-truth labeling. First row: *Frantic* cases. Second row: *Excited* cases. Third row: *Peaceful* cases. Fourth row: *Bored* cases.

of the crowd mood in a video. In a sense, crowd mood is aimed at providing a quantitative measure of crowd emotional information in different crowd scenes associated with a video. Besides, the third-party ground truth is commonly conducted by the algorithms, which can be compared or judged for effectiveness, e.g., for key frames extraction [53], sentiments analysis [54], and emotion recognition [30], [45], [46]. We note that it is generally possible for the third-party observers to determine the mood of a crowd from visual cues, as shown by the user study we conducted.

1) Subjective Test: The purpose of the subjective test is to provide the ground truth of each video as the A-V values. To accomplish this task, we invited ten volunteers (six men and four women) into our research laboratory to manually label the mood of crowd videos. Each of them is asked to watch video samples from the aforementioned crowd database and to label the A-V values from 0 to 1 in ten levels. They were required to choose A and V component values to describe the emotion states that they saw. The ground truth is generated from the volunteers' voting results by averaging the A-V values of labeling. For the ground truth of the discrete crowd mood class, it is inferred by which quadrant the A-V value falls into. Note that the subjects were asked to label the emotion based on their feelings of what the whole crowd was trying to evoke, rather than the emotion that the subjects perceived during the test. Since our crowd mood model is developed to indicate visual motion type and strength through a coordinate in the emotion plane, it is more natural and adequate that the A-V values are correspondent with the crowd's evoking emotion. When the subjects labeled the crowd mood the first time, we needed to inform them of the essence of the crowd mood model, the purpose of the experiment, and several rules of the subjective test, which follow this paper [46].

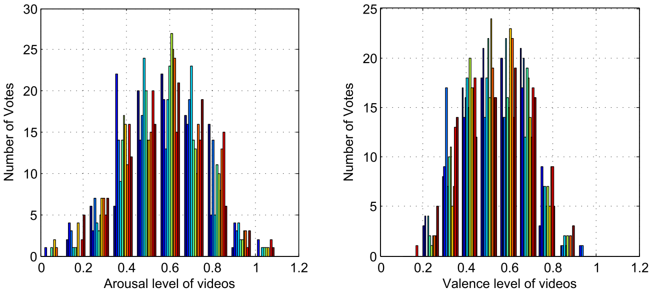


Fig. 7. Histogram of vote for the A and V level of all the videos. Different color stands for the volunteers.

1) We showed the observers crowd videos, which have different levels of strengths and types of the crowd mood. The evoking crowd mood indicated from the motion patterns rather than the perceived emotion of the personal feeling from the observers should be labeled.
2) We would like to have the volunteers' general assessment in response to scene, motion, and context of the crowd. We did not attempt to ignore the influences of the voice or expression (if possible to obtain in the crowd) even though the related features were not yet considered.
3) There was no time limitation during the labeling process. The subjects were allowed to watch the video as many times as they wanted to ensure the labels truly reflected the crowd evoking emotion.

2) Ground-Truth Consistency: Due to the quality of the ground truth being central to the system performance, we also evaluated the consistency of the labeled ground-truth data. To identify the emotional types, we only need the average labeling to provide the generalized labels in four emotional quadrants, including *Frantic*, *Excited*, *Relaxed*, and *Bored*. In this paper, the given A-V values to the videos by subjective test were mapped into the A-V emotion plane. The histogram for the votes of the volunteers is shown as Fig. 7. Ground-truth consistency of labels given by observers for the same video was evaluated based on standard deviation distribution of the labeled values [30], [46]. The larger the standard deviation, the less representative the ground truth is. The reason is that the ground truth is obtained by averaging the labels of the subjects. The histograms of the standard deviation for A and V values in the subjective test are shown in Fig. 8 (left). Note that most standard deviations of the labeled values are approximately 0.1. The maximum values are less than 0.25 and 0.2 for A and V, respectively. Furthermore, we also conduct the pair validation as a reliability study and compute the absolute difference between a pair test during a span of time of the same subject for the same video. The histograms of the absolute difference for A and V values are shown in Fig. 8 (right). Most of the values fall around 0.1, which indicates that the labels given by the subjects were reliable. Even if there is a little fluctuation, the labeled results still remained in the same quadrant by the averaging operation upon all the subjects. For further consideration, the mean standard deviation and the absolute difference evaluation are listed in Table II. We can see that the mean standard deviation of valence is smaller than that of arousal. The mean absolute difference of arousal is a little larger than valence but still remaining small. A reasonable explanation is that negative and

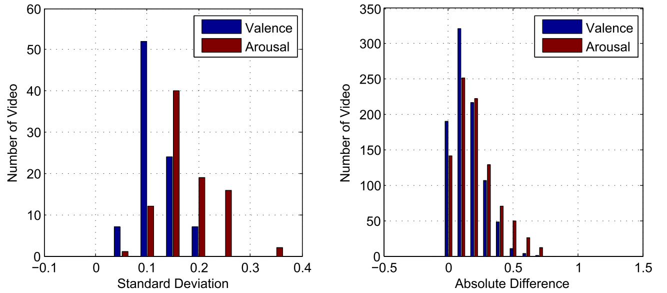


Fig. 8. Left: histogram of standard deviations for A and V of the subjective test. Right: histogram of absolute difference for A and V of pair validation in the test and retest reliability study.

TABLE II
MEAN STANDARD DEVIATION AND ABSOLUTE DIFFERENCE OF THE LABELED RESULTS

Mean Standard Deviation		Mean Absolute Difference	
Valence	Arousal	Valence	Arousal
0.1012	0.1473	0.1302	0.1650

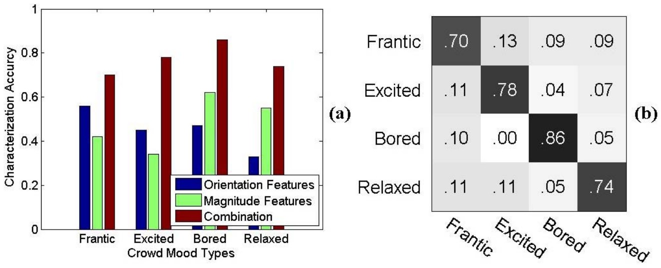
positive emotions maybe more sensitive for humans to discriminate compared with low and high strength [46]. Overall, the results demonstrate that the ground-truth labels are highly consistent as reliable data for the following experiments.

B. Crowd Mood Characterization

To evaluate the effectiveness of our proposed crowd mood on crowd mood characterization,[1] we validate our proposed crowd mood method (combining the orientation and magnitude features, see details in Section IV-C) and two alternative methods: 1) orientation feature-based method, in which only the orientation feature is used; and 2) magnitude feature-based method, in which only magnitude features are used. We predefine four representative moods for evaluating the effectiveness of our proposed crowd mood method, including *Frantic*, *Excited*, *Relaxed*, and *Bored*.[2] We investigate these vital factors in modeling crowd mood: 1) performance of different feature channels; 2) STL; and 3) Adaboost feature weight selection (AFWS).

1) UCF Data Set: This data set consists of 90 videos, which vary in view, resolution, and duration of high-density crowds. The labeled results are as follows—*Frantic*: 23 videos, *Excited*: 26 videos, *Bored*: 21 videos, and *Relaxed*: 20 videos.

2) Evaluation Protocol: We select and label 40 videos (10 videos for each mood) of PETS2009 data set [50] as our training set, and train the SVR models for crowd mood characterization and other experiments. We also train the orientation feature-based model and the magnitude feature-based method using the same configuration. The crowd mood of whole sequence is generated by averaging the coordinates of crowd mood curve. The insights of the results are investigated as follows.



	Frantic	Excited	Bored	Relaxed
Frantic	.70	.13	.09	.09
Excited	.11	.78	.04	.07
Bored	.10	.00	.86	.05
Relaxed	.11	.11	.05	.74

Fig. 9. Evaluating results of crowd mood characterization. (a) Characterizing accuracy using different emotional motion feature channels. Left: orientation-based feature channels. Middle: magnitude-based feature channels. Right: their combination. (b) Final confusion matrix of characterization.

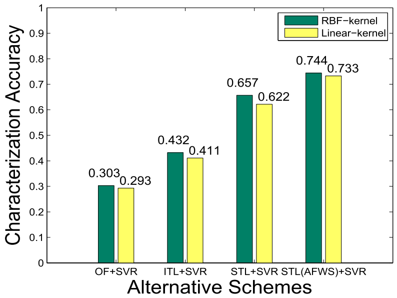


Fig. 10. Quantitative accuracy comparison for crowd mood characterization by alternative schemes. Note that STL and AFWS can make contributions to the characterization improvement compared with the baseline method. The combined STL (AFWS) achieves the best performance.

1) *Effects From Different Feature Channels:* Fig. 9(a) shows the evaluation of results of alternative methods for crowd mood characterization. All three methods achieve more than 40% mood characterization accuracy, and our proposed crowd mood achieves overall 74.4% accuracy and outperform the other two methods. This demonstrates that the structured motion field features and crowd mood model are effective for characterizing crowd mood. Different types of feature channels contribute to the final improvement, which also verifies that our regression fusion method is effective. Our results can successfully infer different types of crowd moods in video sequences with naive operation and achieve impressive accuracy. It further indicates that additional information can compensate for errors in distinguishing events and that more types of mood can be estimated. Fig. 9(b) shows a confusion matrix between different mood categories. Our method effectively classifies different moods, while confusion occurs between category pairs, e.g., *Relaxed*/*Frantic* and *Relaxed*/*Excited*, which have very similar components. According to our observation, *Excited* and *Frantic* overlap the most with each other. Some results are predicted on the boundary line of the emotional quadrants, which causes confusion between these pairs.
2) *Effects From STL:* The analysis of emotion modeling [47] has concluded that ϵ-SVR is excellent in regression for emotional factors with error tolerance. To test the efficiency of each scheme, we evaluate a comparison experiment with two common kernels for

[1] To the best of our knowledge, this paper is the first attempt for representing and characterizing the mood of crowds.

[2] In this paper, we only select this four representative moods to carry out the experiments, and it is easy to expand to more moods.

mood characterization

$$\text{Linear: } K(x, y) = x^T y$$
$$\text{RBF: } K(x, y) = e^{-\gamma \|x-y\|^2}$$

where x and y are the feature values and $\gamma > 0$. As shown in Fig. 10, we replace the STL with the OF and inconsistent trajectory learning that directly derived from particle advection to obtain the motion features. The results show that SVR with linear kernel is less accurate due to unweight motion features, partially because there are relatively few training videos to map the multidimensional features to an emotional value. Meanwhile, RBF as a nonlinear kernel maps training motion features to the high-dimensional space, which leads to better accuracy. Nevertheless, due to that STL already obtains the coherent motions for distinguishable representations, there is not much difference between the accuracies generated by the linear and nonlinear kernels. The result decreases to about 30% and 43% when the motion features obtained by the OF and inconsistent trajectories. This reveals the fact that our structured trajectories capture the coherent motion patterns in a consistent particle manner, which makes it superior to the basic OF and nonstructured trajectories.

3) *Effects of AFWS:* The final accuracy of the AFWS-based approach achieves 74.4% accuracy as in Fig. 10, which gains approximately a 10% increase. The scheme of AFWS uses each feature as a weak classifier to persist the most discriminative motion pattern information. The weighted features show the relative importance in describing the emotional components. The performance comes from the utilities of boosted feature weights on the feature construction, which make it possible to select the optimal combination to represent motion patterns.

C. Global Abnormal Mood Detection

We also conduct a set of experiments for GAM detection to validate the effectiveness of crowd mood for capturing the semantic changes of the crowd behavior. We define the *Frantic* mood as the abnormal mood, which represents the abnormal emotion states of the crowd in a panic escaping situation. The goal of GAM detection is to find the abnormal mood frame, which is caused by a burst of motion and different from the neighbor frames. Similar to most global abnormal event detection work on crowd behavior [6], [10], [12], [21], we follow the same experiment configuration. To that end, crowd mood directly reflects how the crowd behaviors change, which is a natural advantage.

1) GAM on UMN Data Set: In this experiment, we utilize the frame-level measurement [11] to test GAM detection on the UMN data set [52].

a) UMN data set: It consists of 11 clips of crowded escape video events acquired in three different scenarios, including both indoor and outdoor scenes. Each video begins with normal behaviors and ends with panic escaping. All the video frames are resized to 120 × 160 pixels for computation cost.

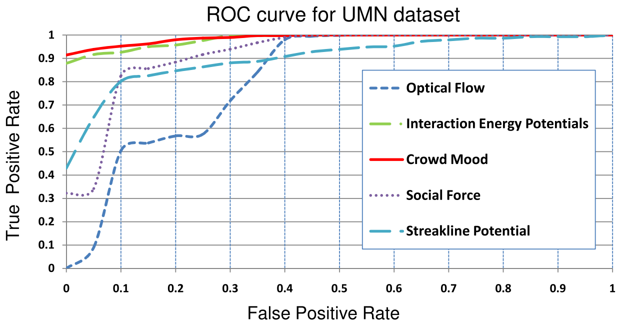


Fig. 11. ROC curves of abnormal detection in UMN data set compared with the other state-of-the-art methods.

TABLE III
COMPARISON OF HIGH-LEVEL METHODS IN UMN DATA SET

Method	CM	IEP [21]	SF [10]	SP [20]	OF
AUC	**0.993**	0.985	0.96	0.90	0.86

TABLE IV
COMPARISON WITH OF IN PETS2009 DATA SET

Method	OF S1	CM S1	OF S2	CM S2
AUC	0.8834	**0.9480**	0.9801	**0.9940**

b) Measurement: The emotional motion features are extracted from the structured trajectories. We get the mapped A and V curve to represent the overall intensity and types of the abnormality in the frames. Each frame is classified as normal or abnormal mood by the type of *Frantic*. The performance is validated by plotting the receiver operating characteristic curve (ROC) curves obtained by a continuous threshold.

c) Insights: Fig. 11 shows the performances of ROC in the experiments compared with other state-of-the-art high-level modeling methods. The ROC curves listed for comparison are directly obtained from paper [10], [20], [21]. The quantitative results are shown in Table III. The results show that our method (CM) can achieve better performance over available state-of-the-art methods, including interaction energy potentials [21] (IEP), SF [10], SP [20], and OF. It comes from the fact that our features not only capture the velocity strength, but also identify the global velocity distribution, which is competitive to improve the performance on GAM detections.

2) GAM on PETS 2009 Data Set: We give GAM detection results obtained on PETS2009 S3 data set following the same experiment configuration with UMN. This data set contains visual content from walking to escaping, which is suitable for testing the sensitivity of crowd mood on detecting the subtle and evolutive changes.

a) Evaluation criterion: To verify the discriminative capability for evolutive changes, we use $f(\phi) = (1 + e^{\phi})^{-1}$to compute an activation scoring function of ours (CM) and OF. And $\phi(\text{CM}) = A \times (1 - V)$ and $\phi(\text{OF}) = \arg\max M_{xy}$, where A and V represent the value of A and V for frame-level crowd mood. M_{xy} is the OF magnitude at pixel (x, y). We use continuous threshold as binary classifier by varying the parameter to generate ROC curves.

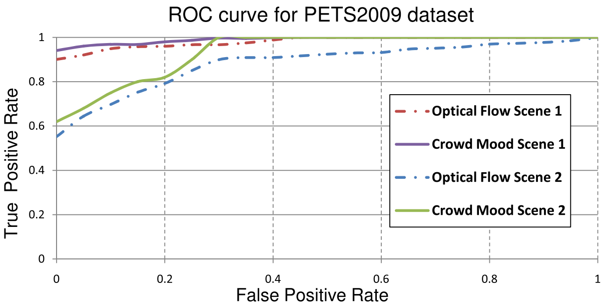


Fig. 12. ROC curves of abnormal detection in PETS2009 data set in two scenes compared with the OF.

b) Case study: Fig. 12 shows the ROC curves in the two scenes with the quantitative results of the comparisons in Table IV, which indicates our approach (CM) outperforms the baseline OF method. Scene 1 (S1), which is more challenging than scene 2 (S2), contains gradual traction from normal to abnormal activity. Our method focuses on the issue of modeling the intensity as well as the patterns of crowd motion, which is more sensitive to gentle changes of abnormality. Our representation is capable of detecting the changes with well mapping dimensions, which makes the normal and abnormal motion pattern more distinctive and sensitive. In addition, based on statistical features of coherent motion pattern, the desired abnormal moods with the structure variations are captured with much more robustness by our approach, which gain higher performance than the OF.

D. Crowd Emotional Matching

1) Task: In this experiment, we consider the emotional matching for the global crowd pattern to investigate the robustness of crowd mood for various conditions, including the changes of view, illumination, occlusion, and so on. It is informative to give the insights into how the visual changes affect the crowd mood representation. The crowd emotional matching focuses on the goal to match the emotion of crowd pattern sequence in a certain video segment with other similar pattern sequences.

2) Data Set: Different crowd patterns could be identified, making use of the emotional curve cues. To validate our idea of the emotional matching, we still conduct experiments on the PETS 2009 S3. We denote crowd sequences 14–16, 14–27, 14–31, and 14–33 as Seq1–Seq4, respectively. So we have four sequences and each has four views. Each view of sequences is trained as a pattern by our emotional model to match the other sequences.

3) Evaluation Protocol: The crowd pattern under different views for the same sequence should have the similar emotion state curves. Fig. 13 shows the visualization of the similarity matrix results across four views. Each view sequence is tested to match the other view sequences using 1-nearest neighbor method. As per our expectations, the views, which have the largest similarity with views in the same sequence, are regarded as a positive match. A dynamic time warping (DTW) sequence matching method is employed to calculate the similarity distance between pairwise emotional curves.

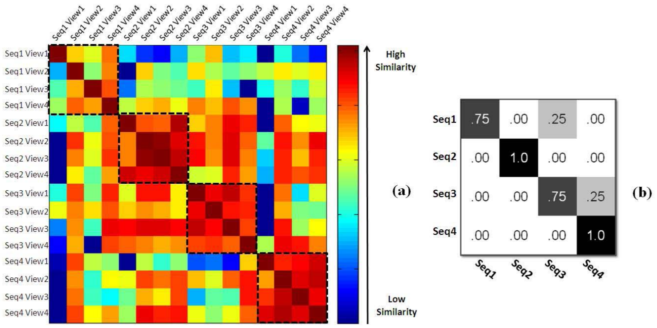


Fig. 13. Crowd pattern matching results. (a) Similarity matrix for four sequences across four views. (b) Confusion matrix for four sequences with a positive match if they have the largest similarity within the same sequence.

4) Discussion: Consequently, it is interesting to note that our method performs much better within the same sequence while indicates discriminative across different sequences. Fig. 13(b) shows the confusion matrix corresponding to the average matching results in Fig. 13(a). It is noticeable that Seq1 and Seq4 share slight similarity with Seq3. This is because Seq3 contains common patterns from walking to splitting with smooth transitions, which would easily capture similarity with others by the DTW algorithm. However, matching results seem pretty promising with an overall accuracy rate 87.5%, which is calculated from the confusion matrix. The result shows our contribution in employing crowd mood curve to capture the semantic shift of global crowd patterns. Note that, crowd mood produces a more invariant representation by affective curve, which demonstrates the robustness across different views and perspectives, due to: 1) structured trajectory reduces motion noise influence by considering motion patterns in the low-rank subspace and 2) motion statistics compensates the occlusion and illumination errors in dense crowd in terms of importance feature selection. Overall, the results reveal the high robustness of our representation, which is tolerant to complex visual changes. It shows that crowd mood can distinguish different motion pattern sequences under the view and illumination changes. The experiment indicates that crowd mood is capable of leveraging global motion pattern as well as transitions between them. It provides a compact representation for analyzing crowd motion behaviors. It greatly facilities the understanding of the crowd motion behavior, and can be served as indicators of transfer and evolution among the crowd patterns.

VI. Conclusion

In this paper, we have presented a novel high-level crowd motion behavior representation, crowd mood, to provide a different viewpoint through which to describe the high-level semantic information about crowd behaviors. We have developed a structured trajectories learning method to extract coherent motion features (the components of crowd mood) that effectively captures compact and discriminative motion patterns. Boosting makes the motion features more discriminative in describing the emotional components.

In contrast to existing features defined in crowd motion modeling, our crowd mood is derived from social emotional evidences and explicitly characterizes types and structures of crowds from an emotional perspective. The crowd mood curve institutively shows the changes of emotion states in dimensional emotion plane to demonstrate the transfer and evolution of crowd patterns. A complete solution has been proposed to detect the sketch of crowds by structured trajectories and quantitatively measure emotional properties of crowds. The preliminary evaluations have shown promising performance in crowd analysis tasks. Crowd mood offers a promising tool for representing the semantics of the crowd behaviors, and is applicable to many crowd tasks. As a first research attempt to leverage crowd moods, we hope it can inspire more opportunities and extensions for better understanding crowd behaviors. In the future, we intend to incorporate other modalities, when available, to further improve the performance of crowd mood analysis.

References

[1] M. S. Zitouni, H. Bhaskar, J. Dias, and M. E. Al-Mualla, "Advances and trends in visual crowd analysis: A systematic survey and evaluation of crowd modelling techniques," *Neurocomputing*, vol. 186, pp. 139–159, Apr. 2016.

[2] Z. Zhong, W. Ye, S. Wang, M. Yang, and Y. Xu, "Crowd energy and feature analysis," in *Proc. ICIT*, 2007, pp. 144–150.

[3] C. Ihaddadene and N. Djeraba, "Real-time crowd motion analysis," in *Proc. ICPR*, 2008, pp. 1–4.

[4] S. Ali and M. Shah, "A Lagrangian particle dynamics approach for crowd flow segmentation and stability analysis," in *Proc. CVPR*, 2007, pp. 1–6.

[5] S. Ali and M. Shah, "Floor fields for tracking in high density crowd scenes," in *Proc. ECCV*, 2008, pp. 1–14.

[6] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *Proc. CVPR*, 2009, pp. 1446–1453.

[7] D. Lin, E. Grimson, and J. Fisher, "Learning visual flows: A Lie algebraic approach," in *Proc. CVPR*, 2009, pp. 747–754.

[8] E. L. Andrade, S. Blunsden, and R. B. Fisher, "Hidden Markov models for optical flow analysis in crowds," in *Proc. ICPR*, 2006, pp. 460–463.

[9] D. Helbing and P. Molnár, "Social force model for pedestrian dynamics," *Phys. Rev. E*, vol. 51, no. 5, p. 4282, 1995.

[10] A. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proc. CVPR*, 2009, pp. 935–942.

[11] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *Proc. CVPR*, 2011, pp. 3449–3456.

[12] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proc. CVPR*, 2010, pp. 1975–1981.

[13] 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. Mach. Intell.*, vol. 31, no. 3, pp. 539–555, Mar. 2009.

[14] J. Kim and K. Grauman, "Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates," in *Proc. CVPR*, 2009, pp. 2921–2928.

[15] B. Zhou, X. Wang, and X. Tang, "Random field topic model for semantic region analysis in crowded scenes from tracklets," in *Proc. CVPR*, 2011, pp. 3441–3448.

[16] T. Hospedales, S. Gong, and T. Xiang, "Video behaviour mining using a dynamic topic model," *Int. J. Comput. Vis.*, vol. 98, no. 3, pp. 303–323, 2012.

[17] B. Zhou, X. Wang, and X. Tang, "Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents," in *Proc. CVPR*, 2012, pp. 2871–2878.

[18] M. Rodriguez, J. Sivic, I. Laptev, and J.-Y. Audibert, "Data-driven crowd analysis in videos," in *Proc. ICCV*, 2011, pp. 1235–1242.

[19] W. Ge, R. T. Collins, and B. Ruback, "Automatically detecting the small group structure of a crowd," in *Proc. WACV*, 2009, pp. 1–8.

[20] B. Mehran, R. E. Moore, and M. Shah, "A streakline representation of flow in crowded scenes," in *Proc. ECCV*, 2010, pp. 439–452.

[21] Q. Cui, Q. Liu, M. Gao, and D. N. Metaxas, "Abnormal detection using interaction energy potentials," in *Proc. CVPR*, 2011, pp. 3161–3167.

[22] K. M. Zeitz, H. M. Tan, M. Grief, P. C. Couns, and C. J. Zeitz, "Crowd behavior at mass gatherings: A literature review," *Prehospital Disaster Med.*, vol. 24, no. 1, pp. 32–38, 2010.

[23] K. M. Zeitz, C. J. Zeitz, and P. Arbon, "Forecasting medical work at mass-gathering events: Predictive model versus retrospective review," *Prehospital Disaster Med.*, vol. 20, no. 3, pp. 164–168, 2005.

[24] C. W. Reynolds, "Flocks, herds and schools: A distributed behavioral model," *ACM SIGGRAPH Comput. Graph.*, vol. 21, no. 4, pp. 25–34, Jul. 1987.

[25] C. Delgado-Mata, J. I. Martinez, S. Bee, R. Ruiz-Rodarte, and R. Aylett, "On the use of virtual animals with artificial fear in virtual environments," *New Generat. Comput.*, vol. 25, no. 2, pp. 145–169, 2007.

[26] R. E. Thayer, *The Biopsychology of Mood and Arousal*. London, U.K.: Oxford Univ. Press, 1989.

[27] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *Int. J. Synth. Emotions*, vol. 1, no. 1, pp. 68–99, 2010.

[28] M. A. Nicolaou, H. Gunes, and M. Pantic, "A multi-layer hybrid framework for dimensional emotion classification," in *Proc. ACM MM*, 2011, pp. 933–936.

[29] D. Wu, T. D. Parsons, E. Mower, and S. Narayanan, "Speech emotion estimation in 3D space," in *Proc. ICME*, 2010, pp. 737–742.

[30] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 2, pp. 448–457, Feb. 2008.

[31] P. J. Lang, "The emotion probe: Studies of motivation and attention," *Amer. Psychol.*, vol. 50, no. 5, pp. 372–385, 1995.

[32] J. A. Russell, A. Weiss, and G. A. Mendelsohn, "Affect grid: A single-item scale of pleasure and arousal," *J. Pers. Soc. Psychol.*, vol. 57, no. 3, pp. 493–502, 1989.

[33] M. S. Zitouni, J. Dias, M. Al-Mualla, and H. Bhaskar, "Hierarchical crowd detection and representation for big data analytics in visual surveillance," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2015, pp. 1827–1832.

[34] B. Solmaz, B. E. Moore, and M. Shah, "Identifying behaviors in crowd scenes using stability analysis for dynamical systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 2064–2070, Oct. 2012.

[35] T. Zhou, J. Yang, A. Loza, H. Bhaskar, and M. Al-Mualla, "Crowd modeling framework using fast head detection and shape-aware matching," *J. Electron. Imag.*, vol. 24, no. 2, 2015, Art. no. 023019.

[36] Y. Yuan, J. Fang, and Q. Wang, "Online anomaly detection in crowd scenes via structure analysis," *IEEE Trans. Cybern.*, vol. 45, no. 3, pp. 548–561, Mar. 2014.

[37] M. Thida, H.-L. Eng, and P. Remagnino, "Laplacian eigenmap with temporal constraints for local abnormality detection in crowded scenes," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 2147–2156, Dec. 2013.

[38] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, 2009, Art. no. 11.

[39] J. Sun, Y. Xie, H. Zhang, and C. Faloutsos, "Less is more: Sparse graph mining with compact matrix decomposition," *Statist. Anal. Data Mining*, vol. 1, no. 1, pp. 6–22, 2008.

[40] H. L. Wang and L.-F. Cheong, "Affective understanding in film," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 6, pp. 689–704, Jun. 2006.

[41] Z. Zeng *et al.*, "Audio-visual affect recognition," *IEEE Trans. Multimedia*, vol. 9, no. 2, pp. 424–428, Feb. 2007.

[42] M. Yeasin, B. Bullot, and R. Sharma, "Recognition of facial expressions and measurement of levels of interest from video," *IEEE Trans. Multimedia*, vol. 8, no. 3, pp. 500–508, Jun. 2006.

[43] M. W. Baig, E. I. Barakova, L. Marcenaro, M. Rauterberg, and C. S. Regazzoni, "Crowd emotion detection using dynamic probabilistic models," in *Proc. Int. Conf. Simulation Adapt. Behavior*, 2014, pp. 328–337.

[44] M. W. Baig, E. I. Barakova, L. Marcenaro, C. S. Regazzoni, and M. Rauterberg, "Bio-inspired probabilistic model for crowd emotion detection," in *Proc. Int. Joint Conf. Neural Netw.*, 2014, pp. 3966–3973.

[45] I. Luengo, E. Navas, and I. Hernáez, "Feature analysis and evaluation for automatic emotion identification in speech," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 490–501, Oct. 2010.

[46] C.-H. Wu, W.-L. Wei, J.-C. Lin, and W.-Y. Lee, "Speaking effect removal on emotion recognition from facial expressions based on eigenface conversion," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1732–1744, Dec. 2013.

[47] S. Zhang, Q. Huang, S. Jiang, W. Gao, and Q. Tian, "Affective visualization and retrieval for music video," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 510–522, Oct. 2010.

[48] J. C. Turner, P. J. Oakes, S. A. Haslam, and C. McGarty, "Self and collective: Cognition and social context," *Pers. Soc. Psychol. Bull.*, vol. 20, no. 5, pp. 454–464, 1994.

[49] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current Psychol.*, vol. 14, no. 4, pp. 261–292, 1996.

[50] *PETS2009 Dataset*. [Online]. Available: http://ftp.cs.rdg.ac.uk/PETS2009

[51] H. Tong, S. Papadimitriou, J. Sun, P. S. Yu, and C. Faloutsos, "Colibri: Fast mining of large static and dynamic graphs," in *Proc. ACM SIGKDD*, 2008, pp. 686–694.

[52] *Unusual Crowd Activity Dataset of University of Minnesota*. [Online]. Available: http://mha.cs.umn.edu/movies/crowdactivity-all.avi

[53] J. Luo, C. Papin, and K. Costello, "Towards extracting semantically meaningful key frames from personal video clips: From humans to computers," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 2, pp. 289–301, Feb. 2009.

[54] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proc. ACM MM*, 2013, pp. 223–232.

Yanhao Zhang is currently working toward the Ph.D. degree with the Harbin Institute of Technology, Harbin, China.

His current research interests include computer vision, multimedia understanding, and machine learning, especially focusing on crowd behavior analysis.

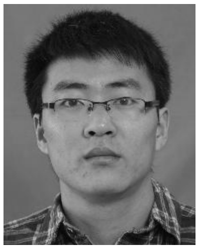
Lei Qin (M'06) received the B.S. and M.S. degrees in mathematics from the Dalian University of Technology, Dalian, China, in 1999 and 2002, respectively, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2008.

He is currently an Associate Professor with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences. He has authored or co-authored over 30 technical papers in the area of computer vision. His current research interests include image/video processing, computer vision, and pattern recognition.

Dr. Qin is a Reviewer of the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and the IEEE TRANSACTIONS ON CYBERNETICS. He has served as a TPC Member for various conferences, including ECCV, ICPR, ICME, PSIVT, ICIMCS, and PCM.

Rongrong Ji (SM'14) is currently a Professor, the Director of the Intelligent Multimedia Technology Laboratory, and the Dean Assistant of the School of Information Science and Engineering with Xiamen University, Xiamen, China. He has authored over 100 papers in international journals and conferences. His current research interests include innovative technologies for multimedia signal processing, computer vision, and pattern recognition.

Mr. Ji is a member of ACM. He has received the ACM Multimedia Best Paper Award and the Best Thesis Award from the Harbin Institute of Technology, Harbin, China. He serves as an Associate/Guest Editor for international journals and magazines, such as *Neurocomputing*, *Signal Processing*, *Multimedia Tools and Applications*, the IEEE MULTIMEDIA MAGAZINE, and *Multimedia Systems*. He also serves as a Program Committee Member for several tier-1 international conferences.



Sicheng Zhao is currently working toward the Ph.D. degree with the Harbin Institute of Technology, Harbin, China.

His current research interests include affective computing, social media analysis, and multimedia information retrieval.

Qingming Huang (SM'08) received the B.S. degree in computer science, and the Ph.D. degree in computer engineering from the Harbin Institute of Technology, Harbin, China, in 1988 and 1994, respectively.

He is currently a Professor with the University of Chinese Academy of Sciences (CAS), Beijing, China, and an Adjunct Research Professor with the Institute of Computing Technology, CAS. He has authored or co-authored over 200 academic papers in prestigious international journals and conferences. His current research interests include multimedia computing, image processing, computer vision, and pattern recognition.

Dr. Huang has served as an Organization Committee Member and a TPC Member for various well-known conferences, including ACM Multimedia, CVPR, ICCV, and ICME.

Jiebo Luo (S'93–M'96–SM'99–F'09) received the B.S. and M.S. degrees in electrical engineering from the University of Science and Technology of China, Hefei, China, in 1989 and 1992, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Rochester, Rochester, NY, USA, in 1995.

He was a Senior Principal Scientist with Kodak Research Laboratories, Rochester, New York, U.S., before joining the Computer Science Department, University of Rochester, in 2011. He has authored over 300 technical papers and holds over 90 U.S. patents. His current research interests include signal and image processing, machine learning, computer vision, social media data mining, and medical imaging.

Dr. Luo is a fellow of the SPIE and IAPR. He has been actively involved in numerous technical conferences, including serving as the General Chair of ACM CIVR 2008 and ACM Multimedia 2018; the Program Co-Chair of ACM Multimedia 2010, the IEEE CVPR 2012, ACM ICMR 2016, and the IEEE ICIP 2017; and the Area Chair of the IEEE ICASSP 2009–2012, ICIP 2008–2012, CVPR 2008/2017, MM 2011/2017, ICCV 2011, IJCAI 2015, and AAAI 2016/2017. He has served on the Editorial Boards of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON MULTIMEDIA (three years), the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, *Pattern Recognition*, *ACM Transactions on Intelligent Systems and Technology*, *Machine Vision and Applications*, and the *Journal of Electronic Imaging*.