Crowd Scene Understanding with Coherent Recurrent Neural Networks *

Hang Su*, **Yinpeng Dong***, **Jun Zhu***, **Haibin Ling**†, **and Bo Zhang***
* Tsinghua National Lab for Information Science and Technology
* State Key Lab of Intelligent Technology and Systems
* Department of Computer Science and Technology, Tsinghua University, Beijing, China
† Department of Computer and Information Sciences, Temple University, USA
*1{suhangss, dongyp13, dcszj, dcszb}@mail.tsinghua.edu.cn, † hbling@temple.edu

Abstract

Exploring crowd dynamics is essential in understanding crowd scenes, which still remains as a challenging task due to the nonlinear characteristics and coherent spatio-temporal motion patterns in crowd behaviors. To address these issues, we present a Coherent Long Short Term Memory (cLSTM) network to capture the nonlinear crowd dynamics by learning an informative representation of crowd motions, which facilitates the critical tasks in crowd scene analysis. By describing the crowd motion patterns with a cloud of keypoint tracklets, we explore the nonlinear crowd dynamics embedded in the tracklets with a *stacked LSTM model*, which is further improved to capture the collective properties by introducing a *coherent regularization term*; and finally, we adopt an unsupervised *encoder-decoder* framework to learn a hidden feature for each input tracklet that embeds its inherent dynamics. With the learnt features properly harnessed, crowd scene understanding is conducted effectively in predicting the future paths of agents, estimating group states, and classifying crowd events. Extensive experiments on hundreds of public crowd videos demonstrate that our method is state-of-the-art performance by exploring the coherent spatio-temporal structures in crowd behaviors.

1 Introduction

Understanding collective behaviors in crowd scenes has a wide range of applications in video surveillance and crowd management [Sulman *et al.*, 2008], especially in present era with recurrent and tragic accidents in populous and diverse human activities. However, a crowd is more than sum of individuals, thus making the vision-related tasks disproportionately difficult along with the crowd scales. The past decade has witnessed a significant progress in crowd scene analysis in learning global motion patterns [Mehran *et al.*, 2010; Wu *et al.*, 2010], modeling local spatio-temporal variations [Kratz and Nishino, 2012; Su *et al.*, 2013], analyzing interactions among individuals [Mehran *et al.*, 2009], profiling group behaviors [Zhou *et al.*, 2011; 2012b], and detecting abnormal crowd behaviors [Solmaz *et al.*, 2012; Mahadevan *et al.*, 2010; Li *et al.*, 2014]. Recently, Li et al. [2015] gave a comprehensive review on the state-of-the-art techniques on crowd scene understanding.

Although various methods have been developed, there is still no publicly accepted framework in understanding the crowd scenes, especially when extreme clutters or severe occlusions occur. One of the essential challenges is that crowd spatio-temporal behavior patterns behave abundantly *nonlinear dynamics*, such as limit cycles, quasi-period and even chaos. This non-linear interaction between individuals always result in various complex, spatio-temporal motion patterns, e.g., the oscillations of the pedestrian flow at bottlenecks [Helbing and Johansson, 2009]. The popular linear dynamic systems in crowd modeling [Lin *et al.*, 2009; Shao *et al.*, 2014] may fail to capture the nonlinear characteristics. Although the nonlinear characteristics of crowd motions investigated in crowd simulation [Massink *et al.*, 2011], few attempts are made in the vision-based crowd motion analysis.

Another challenge in crowd behavior analysis is the *collective effect* (or *coherent motion*) [Zhou *et al.*, 2012a; 2014], e.g., pedestrians in crowds tend to form coherent groups by aligning with other neighbors. Different from the individual motion phenomena, there widely exist various self-organized spatio-temporal patterns even without externally planned or organized, which has been well explained with social force assumption [Helbing and Johansson, 2009]. In this case, methods that do not leverage the coherent characteristics may hinder the capabilities in capturing the inherent crowd dynamics. For instance, crowd features learnt from a multi-task deep architecture [Shao *et al.*, 2015], although more effective than the handcrafted features, suffer from the lack of considering the essential non-linear temporal correlations and coherent motions in crowd behavior analysis. More recently, coherent motions in crowd scenes are detected with a thermal energy field such that predefined activities are effectively recognized [Lin *et al.*, 2016; Wang *et al.*, 2014]. However, it still fails to explore the nonlinear crowd dynamics which hinders the performance for

*This research is partly supported by National Natural Science Foundation of China (Nos. 61571261, 61322308, 61332007), and China Postdoctoral Science Foundation (No. 2015M580099)

complex crowd behaviors.

1.1 Our Proposal

To address the aforementioned challenges, we propose to explore the crowd dynamics with a coherent Long Short Term Memory (LSTM) architecture, which investigates the nonlinearities of crowd behaviors with a stacked LSTM and enforces the consistence of spatio-temporal structures with a coherent regularization.

Recently, deep learning [Schmidhuber, 2015] has achieved state-of-the-art performance in several vision tasks related to visual understanding [Simonyan and Zisserman, 2014] and scene analysis [Ramanathan *et al.*, 2015] partly due to its good capabilities in modeling nonlinearity. It inspires us to explore the nonlinear dynamics of crowd behaviors with a deep architecture model. Specifically, we build our model based on the Long Short Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997] network, which is an improved Recurrent Neural Network (RNN) for temporal data modeling. LSTM has been proven successful on tasks in generating image caption [Vinyals *et al.*, 2015] and video description [Donahue *et al.*, 2015], since it provides a good probability to explore the long term dynamics by overcoming the problem of gradient vanishing and exploding for conventional RNNs [Lipton, 2015]. In order to cope with *nonlinear dynamics in the long term crowd behaviors*, we use a multilayer LSTM network to learn an informative representation of crowd tracklets[1], which are more conservative and less likely to drift than long trajectories.

In order to capture the *coherent spatio-temporal structures*, we further improve the multi-layer LSTM network by introducing a coherent regularization term to model the local spatial and temporal dependency between neighboring pedestrians within a coherent group. The memory unit in LSTM therefore not only stores the dynamic information embedded in the tracklet of its own but also the dynamics of its neighboring agents. The resulting model is denoted as coherent LSTM (cLSTM).

Finally, we adopt an *unsupervised LSTM auto-encoder* framework [Srivastava *et al.*, 2015] to learn a representation to explore the crowd dynamics, such that the tedious efforts in collecting labeled data are significant reduced. Specifically, a stacked cLSTM first encodes the input tracklets into a hidden feature, which is subsequently decoded to reproduce the input tracklets. By exploiting the hidden feature and the coherent regularization, we can extrapolate the past dynamics to the future and forecast the motion beyond what has been observed by recursively unrolling the feature to the future. More critical tasks in crowd scene analysis are also conducted using the learnt representation, including group state estimation and crowd event classification.

In summary, our work differs significantly from the existing studies [Shao *et al.*, 2014; 2015] in that the crowd dynamics are captured via an LSTM model, which is "deep in time" and can identify informative structures in time domain. To the best of our knowledge, this study is a first attempt that investigates the non-linear characteristics of crowd motion patterns with LSTM. Our main contributions are:

- We propose to investigate the crowd dynamics with a stacked LSTM model, such that the complex and nonlinear crowd motion patterns are well captured;
- To consider the collective properties in crowd motion patterns, we propose to improve LSTM by introducing a coherent regularization which encourages a consistent spatio-temporal hidden feature;
- Finally, we adopt the hidden features learnt from the coherent LSTM to critical tasks in crowd scene analysis, including future path prediction, group state estimation, and crowd behavior classification. Experiments demonstrate state-of-the-art performance of our method.

2 Model Crowd Motions with cLSTM

We consider to describe the crowd motion pattern with a set of tracklets $\{\mathbf{x}_t\}$ due to its explainable semantics in crowd behaviors [Li *et al.*, 2015], as illustrated in Fig. 1. In this section, we aim to extract informative features from these tracklets to explore the crowd dynamics, which facilitates the subsequent tasks in crowd scene analysis. To this end, we first introduce the basic coherent LSTM unit which updates its memory with the tracklet of its own together with its neighboring agents; afterwards, we step into details of the coherent regularization term and describe the LSTM models by stacking the coherent LSTM unit.

Figure 1: Samples of tracklets for crowd motion patterns.

2.1 Coherent Long Short Term Memory Unit

LSTM has a powerful capability in modeling the nonlinear dynamics for sequential data [Lipton, 2015; Greff *et al.*, 2015]. As illustrated in Fig. 2, each LSTM unit has a cell as a memory, which maintains its state $\mathbf{c}_t$ at time t by regulating the information flow into/out of the LSTM unit with non-linear gates [Greff *et al.*, 2015].

Specifically, the state of a cell is controlled through sigmoidal gates including an input gate $\mathbf{i}_t$ which takes activation from the current data point $\mathbf{x}_t$ and the hidden layer at the previous time step $\mathbf{h}_{t-1}$ as

$$\mathbf{i}_t = \sigma(\mathbf{W}_{xi}\mathbf{x}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{W}_{ci}\mathbf{c}_{t-1} + \mathbf{b}_i), \quad (1)$$

and a forget gate $\mathbf{f}_t$ which enables the cell to reset its state as

$$\mathbf{f}_t = \sigma(\mathbf{W}_{xf}\mathbf{x}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{W}_{cf}\mathbf{c}_{t-1} + \mathbf{b}_f), \quad (2)$$

where $\mathbf{W}$s are the weight matrices with a proper size; $\mathbf{b}$s are bias vectors. Note that all the $\mathbf{W}_{c\bullet}$ matrices corresponding to

[1] A tracklet is a fragment of a trajectory obtained by a tracker, e.g., a KLT keypoint tracker [Baker and Matthews, 2004], which starts when a novel key-point is detected and terminates when ambiguities arise.

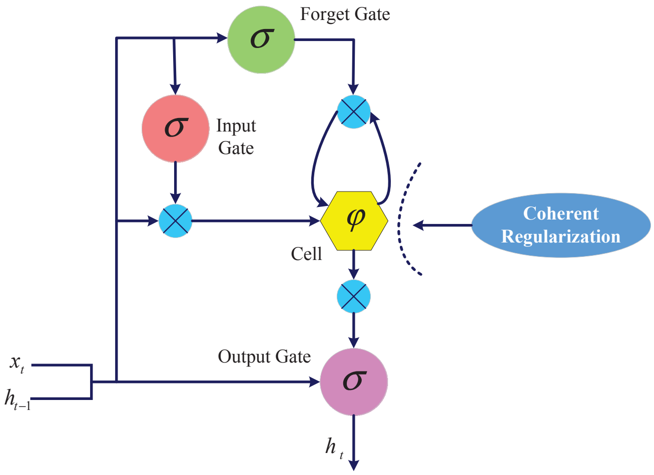


Figure 2: A diagram of a coherent LSTM unit.

the cell states are diagonal, whereas the rest are full matrices. The total input at the input terminal is passed through the tanh non-linearity and multiplied by the activation of the input gate $\mathbf{i}_t$, which is then added to the cell state $\mathbf{c}_{t-1}$ multiplied by the forget gate's activation $\mathbf{f}_t$ as

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tanh(\mathbf{W}_{xc}\mathbf{x}_t + \mathbf{W}_{hc}\mathbf{h}_{t-1} + \mathbf{b}_c), \quad (3)$$

where $\odot$ is the element-wise multiplication. The final output from the LSTM unit $\mathbf{h}_t$ is computed by multiplying the updated cell state that passed through a tanh non-linearity with an output gate's activation as

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t), \quad (4)$$

where $\mathbf{o}_t$ is the output gate as

$$\mathbf{o}_t = \sigma(\mathbf{W}_{xo}\mathbf{x}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{W}_{co}\mathbf{c}_t + \mathbf{b}_o). \quad (5)$$

The output gate $\mathbf{o}_t$ controls how much of the memory cell should be transferred to the hidden feature. Compared with the conventional RNN, the additional cells in LSTMs sum the activities over time. Such strategy avoids a quick gradient vanishing and enables the LSTMs to learn extremely complex and long-term temporal dynamics in crowd behavior analysis.

Different from the individual behaviors, there widely exist coherent motion phenomena [Zhou *et al.*, 2012a] in crowds, since the individuals are always willing to engage with "seed" groups and form spatially coherent structures. Therefore, we propose to improve the conventional LSTM by taking the neighboring tracklets into account. The intuition behind the model is *if the dynamics of two tracklets are coherent in the spatial and temporal domain, i.e., when the neighboring relationship of individuals remains invariant over time or the correlation of their velocities remains high, they tend to have similar hidden states.* To this end, we propose to update the memory unit by incorporating its own state together with its neighboring agents with a coherent regularization as

$$\begin{aligned}\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \sum_{j\in\mathcal{N}} \lambda_j(t)\mathbf{f}_t^j \odot \mathbf{c}_{t-1}^j \\ + \mathbf{i}_t \odot \tanh(\mathbf{W}_{xc}\mathbf{x}_t + \mathbf{W}_{hc}\mathbf{h}_{t-1} + \mathbf{b}_c),\end{aligned} \quad (6)$$

where $\mathcal{N}$ denotes the set of neighboring tracklets within a coherent group; $\mathbf{f}_t^j$ and $\mathbf{c}_{t-1}^j$ are corresponding to the forget gate and cell state for LSTM in the coherent group; and $\lambda_j(t)$ weights the dependency between the tracklets, as detailed below.

2.2 Coherent Motion Modeling

In this section, we first investigate the dependency between agents in coherent groups, which are discovered using the coherent filtering [Zhou *et al.*, 2012a], as illustrated in Fig. 3. The coherent keypoints with similar motion patterns and tendencies are marked with the same color.

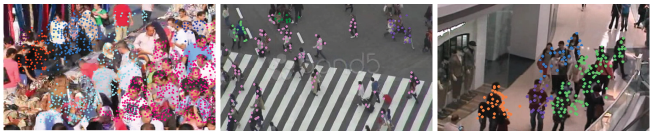


Figure 3: Group detection using coherent filtering [Zhou *et al.*, 2012a], in which different groups are indicated with different colors (best viewed in color version).

The dependency relationship between two tracklets within the same group is measured with their pairwise velocity correlations as

$$\tau_j(t) = \frac{\mathbf{v}_i(t) \cdot \mathbf{v}_j(t)}{\|\mathbf{v}_i(t)\|\|\mathbf{v}_j(t)\|}, \quad (7)$$

where $\mathbf{v}_i(t)$ and $\mathbf{v}_j(t)$ are the velocities of the i_{th} and j_{th} tracklets, respectively. The dependency coefficient between the i_{th} and j_{th} tracklets in Eq. (6) is defined as

$$\lambda_j(t) = \frac{1}{\mathbf{Z}_i} \exp\left(\frac{\tau_j(t) - 1}{2\sigma^2}\right) \in (0, 1], \quad (8)$$

where $\mathbf{Z}_i$ is the normalization constant corresponding to the i_{th} tracklet. $\lambda_j(t)$ tends to be 1 when the tracklets i and j are similar, and decreases when the tracklets become different. In this case, our model with coherent regularization encourages the tracklets to learn similar feature distributions by sharing information across tracklets within a coherent group.

In order to model the long term crowd dynamics, additional depths are also added to LSTMs by stacking them on top of each other, i.e., using the output of the LSTM in the $(l-1)_{\text{th}}$ layer as the input to the LSTM in the l_{th} layer, as illustrated in Fig. 4.

3 Crowd Scene Analysis based on cLSTM

In this section, we describe an unsupervised encoder-decoder framework using the coherent LSTM to generate a hidden feature, which embeds the inherent characteristics for each tracklet. It shares similar ideas to that of auto-encoders [Vincent *et al.*, 2010] such that the parameters are optimized by minimizing the difference between the reproduced and target sequences. Critical tasks in crowd scene analysis is conducted based on the hidden features, e.g., future path forecasting, group state estimation and crowd behavior classification.

To learn an informative representation, we take the "encoder-decoder"approach inspired by [Donahue *et al.*,

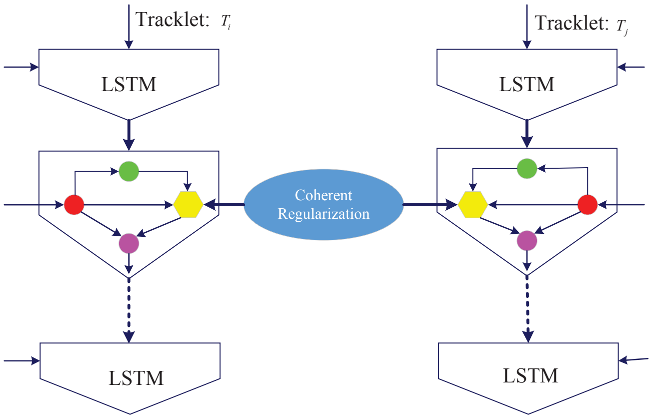


Figure 4: We stack a series of coherent LSTM units to capture the time-varying nonlinear crowd dynamics by mapping tracklets in a coherent group to similar hidden features, which is conducted by incorporating a coherent regularization.

2015], which consists of an encoder coherent LSTM and a decoder coherent LSTM, as is shown in Fig. 5. The encoder coherent LSTM runs though a tracklet to come up with a hidden feature, which is decoded to produce a target sequence with a decoder coherent LSTM. Note that our model allows to reproduce the input tracklets, and also provides a method to forecast the unseen future paths by exploiting the hidden features.

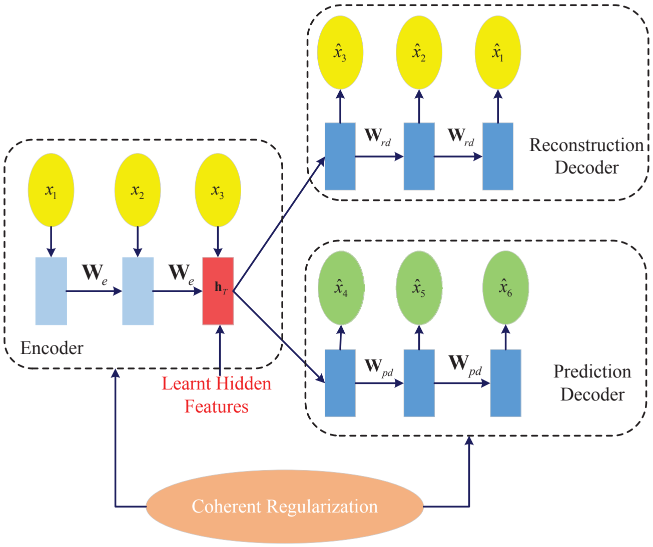


Figure 5: The encoder-decoder framework with coherent LSTM. The encoder generates a hidden feature of each tracklet, and the decoder reproduces the input tracklets and predicts the future paths.

At the encoding stage, we obtain representation vectors for tracklets with the encoder coherent LSTM following Eq. (4) as

$$\mathbf{h}_T = cLSTM_e(\mathbf{x}_T, \mathbf{h}_{T-1}), \tag{9}$$

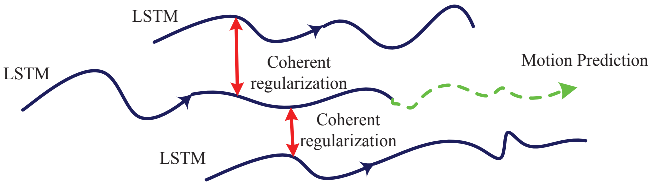


Figure 6: Prediction of future path with coherent regularization, in which the blue paths are reliable tracklets obtained from KLT trackers and the green ones are the forecasted paths from the cLSTM.

where $cLSTM_e$ is an encoding operation to map the input tracklet to a hidden feature; $\mathbf{x}_T$ and $\mathbf{h}_{T-1}$ are the input tracklets at time step T and hidden feature vectors at the previous time step T–1, respectively.

As for the reconstruction decoder, the coherent LSTM generates a set of estimated tracklet $\hat{\mathbf{x}}$ for each tracklet, which reproduces the input tracklets but in a reverse order to avoid the long range correlations as

$$\hat{\mathbf{x}}_t = cLSTM_{dr}(\mathbf{h}_t, \hat{\mathbf{x}}_{t+1}), \text{ where } t \in [1, T], \tag{10}$$

where $cLSTM_{dr}$ decodes the representation recursively to reproduce the input tracklets; $\mathbf{h}_t$ is the feature that is derived from the hidden features $\mathbf{h}_T$ in Eq. 9. The parameters $\{\mathbf{W}s, \mathbf{b}s\}$ are optimized by minimizing the reconstruction error between the input tracklets and the reproduced ones when training the model.

The prediction decoder is similar to that of the reconstruction decoder, except that the decoder LSTM extrapolates future paths. Specifically, the prediction is implemented by unrolling the hidden feature by

$$\hat{\mathbf{x}}_t = cLSTM_{dp}(\mathbf{h}_t, \hat{\mathbf{x}}_{t-1}), \text{ where } t > T, \tag{11}$$

where $cLSTM_{dp}$ is a decoding operation to predict the future path of an agent derived from the hidden feature by taking the coherent motion into account. During the training stage, tracklets in the training dataset $\{\mathbf{x}_t\}_{t=1}^{T}$ are divided into two segments as $\{\mathbf{x}_t\}_{t=1}^{T_0}$ and $\{\mathbf{x}_t\}_{t=T_0+1}^{T}$. The former segments are input to the encoder to learn a hidden feature, which is used to predict the latter segment by minimizing the difference between the original and estimated tracklets. The parameter T_0 is modified to adjust the length ratio between the segments, which enhances inherent evolutionary dynamics are well captured.

The success of path prediction lies in the facts that the hidden states generated from the encoder captures the dynamics of tracklets to forecast the future within a coherent group, which is further enhanced by incorporating with the coherent regularization, as the schematic shown in Fig. 6.

In summary, we capture the inherent crowd dynamics by mapping the tracklets to hidden features with the cLSTM model. With the features properly harnessed, we reconstruct the input tracklets as well as predict the future paths. A reconstruction-oriented encoder would suffer from the tendency to memorize the inputs, and the future predictor would suffer from the tendency to ignore the initial frames because of the more significant impacts from the last few frames.

Therefore, the proposed method encourages a more inherent feature by mutually enhancing each other, such that the features maintain the dynamics embedded in *the whole sequences* by *not just memorizing* the information.

3.1 Crowd Scene Profiling

With crowd dynamics properly captured by our coherent LSTM model, the critical tasks in crowd scene analysis can be conducted based on the learnt features, e.g., understanding group states and recognizing crowd behaviors.

The states of groups with coherent spatio-temporal structures are generally recognized as *Gas, Solid, Pure Fluid and Impure Fluid* [Shao *et al.*, 2014], which is related to multiple social psychological and physical factors, e.g., crowd density, goals, interactions between group members, etc. In this section, we implement group state estimation with the learnt representation since it embeds dynamics of coherent groups. Specifically, we feed the feature learnt from the unsupervised cLSTM to a softmax classifier, which infers the group states based on the dynamics of tracklet within a coherent group as

$$\mathcal{L} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T}\sum_{c=1}^{C}\left\{\mathbf{1}\{y_i(t)=c\}\log\frac{e^{\boldsymbol{\eta}_c^T\mathbf{h}_i(t)}}{\sum\limits_{c=1}^{N_c} e^{\boldsymbol{\eta}_c^T\mathbf{h}_i(t)}}\right\}, \quad (12)$$

where $\mathbf{1}(\cdot)$ is an indicator function with value 1 if the predicate is true otherwise 0; $\boldsymbol{\eta}_c$ is a parameter to weight the hidden feature corresponding to class c; $\mathbf{h}_i(t)$ is the hidden features corresponding to the i_{th} tracklet in the group at time step t; N denotes the total number of tracklets in a coherent group; and T is length of tracklets. The term $\log\sum\limits_{c=1}^{N_c} e^{\boldsymbol{\eta}_c^T\mathbf{h}_i(t)}$ normalizes the distribution to guarantee the probability condition. A reliable inference is conducted by maximizing the softmax regression.

Besides estimating the state of each individual group, we also implement the holistic crowd video classification by training another softmax classifier over the sequential hidden features of all the tracklets in a crowd video, producing a distribution over the holistic crowd behavior categories. The success of complex crowd behavior recognition lies on the facts that composing deep layers of cLSTMs can result in a powerful capability in exploring the nonlinearities and coherent spatio-temporal structures in crowd motions.

4 Experimental Results

In this section, we demonstrate the effectiveness of the features learnt from our algorithm on three critical applications in crowd scene analysis: pedestrian future path prediction, group state estimation, and crowd behavior classification. Evaluations are conducted on the CUHK Crowd Dataset [Shao *et al.*, 2014], which includes crowd videos with different densities and perspective scales in many environments, e.g., street, airports, etc. The ground truth of keypoint tracklets, group state, and crowd video classification are also available for the dataset. It consists of more than 400 sequences, and 200,000+ tracklets in total.

In each experiment, we construct a coherent LSTM with 128 hidden units, such that the input tracklets are mapped to 128-dimensional hidden features. Similar as the work in [Shao *et al.*, 2014], we randomly select half of the tracklets in the sequences for training and the remaining for testing. When optimizing the parameters in predicting the future paths, we divide each tracklet into two segments, and use the hidden features learnt from the first segments (e.g., 2/3 of each tracklet) to predict the latter segments (e.g., the rest 1/3 tracklet).

4.1 Future Path Forecasting of Pedestrians

We first test the performance of our framework on path forecasting by comparing our approach with a baseline method of Kalman filter, which implements path prediction by incorporating a linear dynamic model with the uncertainties of the current state, which cannot capture the nonlinear characteristics of the complex crowd motions and leverage information of the coherent groups. We also conduct path prediction with a variant of the proposed cLSTM model by neglecting the coherent regularization between neighboring agents, which is denoted as Un-coherent LSTM. In each experiment, we take a fragment of tracklets as the input (e.g., 2/3 of each tracklet in this paper), reconstruct them and then generate the rest of the predicted tracklets (e.g., 1/3 of each tracklet) to evaluate the performance.

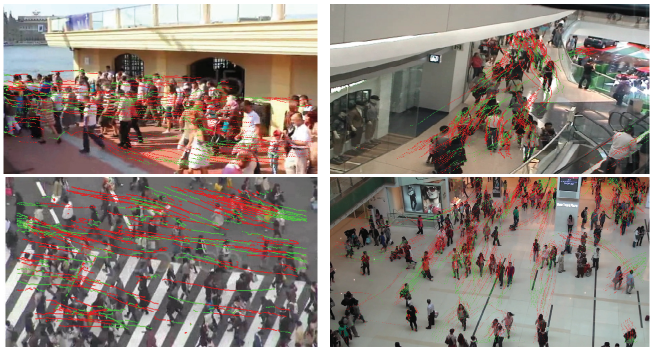


Figure 7: Prediction of future paths with coherent regularization, in which the red paths are reliable tracklets obtained from KLT trackers and the green ones are the forecasted paths from the cLSTM.

Sample results are demonstrated in Fig. 7, in which the red tracklets are the paths obtained with the KLT tracker [Baker and Matthews, 2004] that are followed by green curves of tracklets generated with our cLSTM prediction model. The results demonstrate that our algorithm is capable to capture the inherent dynamics of each tracklet, which reflects the tendency of neighboring agents by taking the coherent regularization into account. Since the near future paths are well predicted, it provides a good probability to prevent the serious incidents before they would emerge.

In Table 1, we report the quantitative performance of path forecasting in terms of prediction error, which measures the average distance between the ground-truth tracklets and the estimated paths in terms of pixel as unit. It shows that our approach significantly outperforms the alternative methods. Compared with the baseline method of Kalman filter, our cLSTM captures the inherent nonlinearity of the crowd behavior such that the complex crowd motions are well predicted with a high precision; moreover, our method also benefits from the coherent regularization due to the collective properties in moving crowds.

Table 1: Error of Path Prediction

Kalman Filter	Un-coherent LSTM	Coherent LSTM
9.32 ± 1.99	6.64 ± 1.76	4.37±0.93

4.2 Group State Estimation

Group states well reflect the characteristics of a crowd, which are useful in various applications. In the CUHK Crowd Dataset [Shao *et al.*, 2014], groups are classified into four states as Gas[2], Solid[3], Pure Fluid[4] and Impure Fluid[5]. Samples of groups with different states are illustrated in Fig. 8.

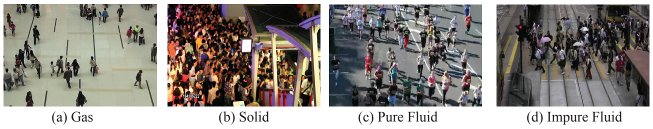
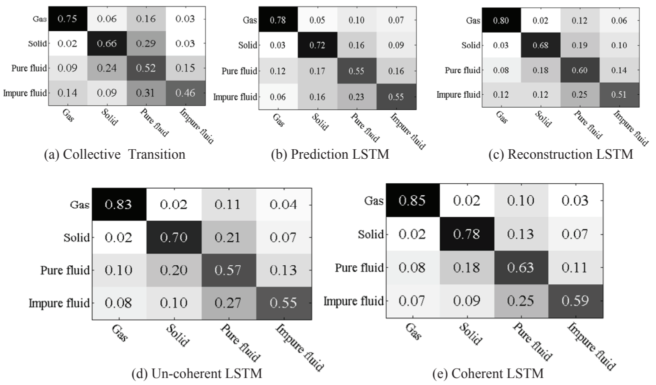

(a) Gas (b) Solid (c) Pure Fluid (d) Impure Fluid

Figure 8: Samples of groups undergoing different states.

In this section, we train a softmax classifier using the hidden features learnt by our *cLSTM*, and then implement the group state estimation. As a comparison, we also conduct group state estimation by feeding the descriptors learnt from variants of the proposed cLSTM model, including the *prediction LSTM* by only targeting at predicting future paths, *reconstruction LSTM* by neglecting the prediction components, and *un-coherent LSTM* by neglecting the coherent regularization. Besides, we also report the results based on a *collective transition* [Shao *et al.*, 2014], which is the state-of-the-art algorithm in group state estimation.

The quantitative evaluation in terms of confusion matrix is reported in Fig. 9. Obviously, our proposed algorithm with *coherent LSTM* (Fig. 9 (e)) outperforms the alternative methods. As a baseline method, group state estimation based on *collective transition* [Shao *et al.*, 2014] (Fig. 9(a)) explores the crowd dynamics via a linear transition matrix, which is not valid for groups with complex motion patterns, e.g., when groups undergo an impure fluid state. The performance also degrades when we neglect the reconstruction or predication components for *prediction LSTM* (Fig. 9 (b)) or *reconstruction LSTM* (Fig. 9 (c)), since the *prediction LSTM* tends to maintain information of the last few frames rather than the whole tracklets, and the *reconstruction LSTM* suffers from the tendency to memorize the tracklets rather than explore the inherent dynamics. When we implement group state estimation without taking the collective properties into account using the *un-coherent LSTM* (Fig. 9(d)), the performance is significantly inferior to our cLSTM method especially for groups with organized structures, e.g., the group undergoing a solid or fluid state.



(a) Collective Transition (b) Prediction LSTM (c) Reconstruction LSTM

(d) Un-coherent LSTM (e) Coherent LSTM

Figure 9: Confusion matrices of estimating group states using different methods: (a) collective transition [Shao *et al.*, 2014]; (b) prediction LSTM; (c) reconstruction LSTM; (d) un-coherent LSTM; and (e) coherent LSTM. See text for the description of each method.

4.3 Crowd Video Classification

Finally, we demonstrate the effectiveness of our method in classifying crowd video dependent on the holistic crowd behaviors in a scene. In CUHK Crowd Dataset [Shao *et al.*, 2014], all video clips are annotated into 8 classes which are commonly seen in crowd videos as 1) *Highly mixed pedestrian walking*; 2) *Crowd walking following a mainstream and well organized*; 3) *Crowd walking following a mainstream but poorly organized*; 4) *Crowd merge*; 5) *Crowd split*; 6) *Crowd crossing in opposite directions*; 7) *Intervened escalator traffic*; and 8) *Smooth escalator traffic*.

Similar as the implementations of group state estimation, we conduct crowd video classification by feeding the representations, which are learnt by our *coherent LSTM*, to a softmax classifier. As comparisons, crowd video classification is also implemented with variants of the proposed cLSTM method, including *prediction LSTM*, *reconstruction LSTM*, and *un-coherent LSTM* (See Section 4.2 for details). Besides, we also compare with *collective transition* [Shao *et al.*, 2014]. In Fig. 10, we report a confusion matrix in crowd video classification based on our cLSTM, which demonstrates that videos are classified into specific categories with high qualities.

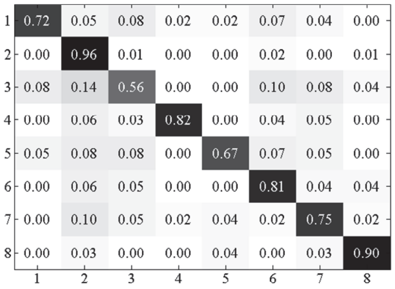
In Fig. 11, we demonstrate the accuracy of crowd video classification on each class by various methods as mentioned above. We can see that the *coherent LSTM* outperforms the alternative methods in general, since it captures the nonlin-

[2] Gas: Particles move in different directions without forming collective behaviors

[3] Solid: Particles move in the same direction with relative positions unchanged

[4] Pure Fluid: Particles move towards the same direction with ever-changing relative positions

[5] Impure Fluid: Particles move in a pure fluid style with invasion of particles from other groups



	1	2	3	4	5	6	7	8
1	0.72	0.05	0.08	0.02	0.02	0.07	0.04	0.00
2	0.00	0.96	0.01	0.00	0.00	0.02	0.00	0.01
3	0.08	0.14	0.56	0.00	0.00	0.10	0.08	0.04
4	0.00	0.06	0.03	0.82	0.00	0.04	0.05	0.00
5	0.05	0.08	0.08	0.00	0.67	0.07	0.05	0.00
6	0.00	0.06	0.05	0.00	0.00	0.81	0.04	0.04
7	0.00	0.10	0.05	0.02	0.04	0.02	0.75	0.02
8	0.00	0.03	0.00	0.00	0.04	0.00	0.03	0.90

Figure 10: Confusion matrix of crowd video classification based on coherent LSTM. See text for the name of each class.

ear dynamics embedded in the complex crowd motion by taking into the coherent motion at the same time, especially for crowds with high nonlinearity, e.g., highly mixed pedestrian walking. Our method significantly outperforms the *collective transition* method [Shao *et al.*, 2014], which fails to explore the nonlinear characteristics of crowd motion and meanwhile neglects the coherent spatio-temporal structures in crowd videos. Note that the videos of *Crowd walking following a mainstream and well organized* are well recognized by collective transition, since crowd motions of these videos satisfy the linear dynamics assumption of collective transition in most cases.

Our cLSTM also benefits from exacting hidden features by incorporating the reconstruction with the prediction tasks together. Obviously, neither *prediction LSTM* or *reconstruction LSTM* fails to capture the inherent dynamics of overall crowd motion, which degenerates the performance in video classification. Besides, the effect of coherent regularization is also essential in the scenarios when the collectiveness within a crowd are significant, e.g., crowd merge or split.

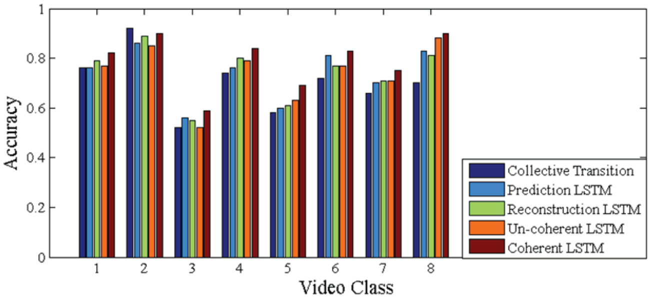


Figure 11: Per-class accuracy comparison of crowd video classification using different methods. See text for the name of each class.

5 Conclusions

We present a novel recurrent neural network with coherent long short term memory (cLSTM) units to understand crowd scenes. To address the nonlinear dynamics in complex crowd scenes, we propose to map a set of tracklets that describe the crowd motion patterns to hidden features with LSTM, which can keep track of an input tracklet. To consider the collective properties of moving crowds, we introduce a coherent regularization such that the memory units are updated by taking into account dynamics of tracklets with coherent motions, which encourages the learnt hidden features to have consistent spatial and temporal structures. With the inherent dynamics properly harnessed, our algorithm also provides a method to predict the possible future paths, which is of great importance in crowd management to prevent serious accidents before they emerge. Extensive experiments on real-world dataset demonstrate that our method outperforms its variants and alternative method in group state estimation and crowd video classification.

References

[Baker and Matthews, 2004] Simon Baker and Iain Matthews. Lucas-kanade 20 years on: A unifying framework. *International Journal of Computer Vision*, 56(3):221–255, 2004.

[Donahue *et al.*, 2015] Jeffrey Donahue, Anne Hendricks, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2625–2634, 2015.

[Greff *et al.*, 2015] Klaus Greff, Rupesh Kumar Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. *arXiv preprint arXiv:1503.04069*, 2015.

[Helbing and Johansson, 2009] Dirk Helbing and Anders Johansson. *Pedestrian, crowd and evacuation dynamics*. Springer, 2009.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[Kratz and Nishino, 2012] Kratz and K. Nishino. Tracking pedestrians using local spatio-temporal motion patterns in extremely crowded scenes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, pages 987 – 1002, May 2012.

[Li *et al.*, 2014] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(1):18–32, 2014.

[Li *et al.*, 2015] Teng Li, Huan Chang, Meng Wang, Bingbing Ni, Richang Hong, and Shuicheng Yan. Crowded scene analysis: A survey. *Circuits and Systems for Video Technology (TCSVT), IEEE Transactions on*, 25(3):367–386, 2015.

[Lin *et al.*, 2009] Dahua Lin, E. Grimson, and J. Fisher. Learning visual flows: A lie algebraic approach. In *Computer Vision and Pattern Recognition, IEEE Conference on (CVPR)*, pages 747 –754, June 2009.

[Lin *et al.*, 2016] Weiyao Lin, Yang Mi, Weiyue Wang, Jianxin Wu, Jingdong Wang, and Tao Mei. A diffusion and clustering-based approach for finding coherent motions and understanding crowd scenes. *IEEE Transactions on Image Processing*, 25(4):1674–1687, 2016.

[Lipton, 2015] Zachary C Lipton. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*, 2015.

[Mahadevan *et al.*, 2010] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1975–1981, 2010.

[Massink *et al.*, 2011] Mieke Massink, Diego Latella, Andrea Bracciali, and Jane Hillston. Modelling non-linear crowd dynamics in bio-pepa. In *Fundamental Approaches to Software Engineering*, pages 96–110. 2011.

[Mehran *et al.*, 2009] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *Computer Vision and Pattern Recognition, IEEE Conference on (CVPR)*, pages 935 –942, June 2009.

[Mehran *et al.*, 2010] Ramin Mehran, Brian E. Moore, and Mubarak Shah. A streakline representation of flow in crowded scenes. In *European Conference on Computer Vision (ECCV)*, pages 439–452, 2010.

[Ramanathan *et al.*, 2015] Vignesh Ramanathan, Kevin Tang, Greg Mori, and Li Fei-Fei. Learning temporal embeddings for complex video analysis. *arXiv preprint arXiv:1505.00315*, 2015.

[Schmidhuber, 2015] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

[Shao *et al.*, 2014] Jing Shao, Chen Change Loy, and Xiaogang Wang. Scene-independent group profiling in crowd. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2227–2234, 2014.

[Shao *et al.*, 2015] Jing Shao, Kai Kang, Chen Change Loy, and Xiaogang Wang. Deeply learned attributes for crowded scene understanding. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 1–9, 2015.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems (NIPS)*, pages 568–576, 2014.

[Solmaz *et al.*, 2012] Berkan Solmaz, Brian Moore, and Mubarak Shah. Identifying behaviors in crowd scenes using stability analysis for dynamical systems. *Pattern Analysis and Machine Intelligence (TPAMI), IEEE Transactions on*, pages 2064 –2070, Oct. 2012.

[Srivastava *et al.*, 2015] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. Unsupervised learning of video representations using lstms. *arXiv preprint arXiv:1502.04681*, 2015.

[Su *et al.*, 2013] Hang Su, Hua Yang, Shibao Zheng, Yawen Fan, and Sha Wei. The large-scale crowd behavior perception based on spatio-temporal viscous fluid field. *Information Forensics and Security, IEEE Transactions on*, 8(10):1575–1589, 2013.

[Sulman *et al.*, 2008] Noah Sulman, Thomas Sanocki, Dmitry Goldgof, and Rangachar Kasturi. How effective is human video surveillance performance? In *Pattern Recognition, 19th International Conference on (ICPR)*, pages 1–8, 2008.

[Vincent *et al.*, 2010] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research (JLMR)*, 11:3371–3408, 2010.

[Vinyals *et al.*, 2015] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, 2015.

[Wang *et al.*, 2014] Weiyue Wang, Weiyao Lin, Yuanzhe Chen, Jianxin Wu, Jingdong Wang, and Bin Sheng. Finding coherent motions and semantic regions in crowd scenes: A diffusion and clustering approach. In *European Conference on Computer Vision (ECCV)*, pages 756–771. 2014.

[Wu *et al.*, 2010] Shandong Wu, Brian E. Moore, and Mubarak Shah. Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 2054 –2060, June 2010.

[Zhou *et al.*, 2011] Bolei Zhou, Xiaogang Wang, and Xiaoou Tang. Random field topic model for semantic region analysis in crowded scenes from tracklets. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, June 2011.

[Zhou *et al.*, 2012a] Bolei Zhou, Xiaoou Tang, and Xiaogang Wang. Coherent filtering: detecting coherent motions from crowd clutters. In *European Conference on Computer Vision (ECCV)*, pages 857–871, 2012.

[Zhou *et al.*, 2012b] Bolei Zhou, Xiaogang Wang, and Xiaoou Tang. Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 2871 –2878, June 2012.

[Zhou *et al.*, 2014] Bolei Zhou, Xiaoou Tang, Hepeng Zhang, and Xiaogang Wang. Measuring crowd collectiveness. *Pattern Analysis and Machine Intelligence, IEEE Transactions on (TPAMI)*, 36(8):1586–1599, 2014.