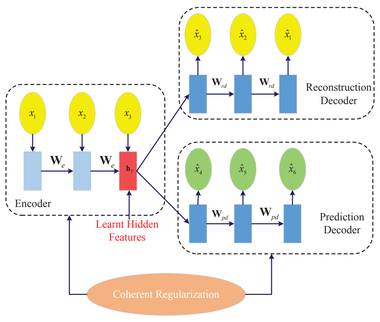
Crowd Scene Understanding with Coherent Recurrent Neural Networks

Hang Su, Yinpeng Dong, and Jun Zhu
{suhangss, dcsjz}@mail.tsinghua.edu.cn, donyp13@mails.tsinghua.edu.cn



Introduction

Understanding collective behaviors in crowd scenes has a wide range of applications in video surveillance and crowd management, 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.

Another challenge in crowd behavior analysis is the *collective effect* (or *coherent motion*), 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.

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.

Contributions

To address the aforementioned challenges, we propose to explore the crowd dynamics with a coherent Long Short Term Memory (LSTM) architecture. Our main contributions are:

- We propose to investigate the crowd dynamics with a stacked LSTM model [1], 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.

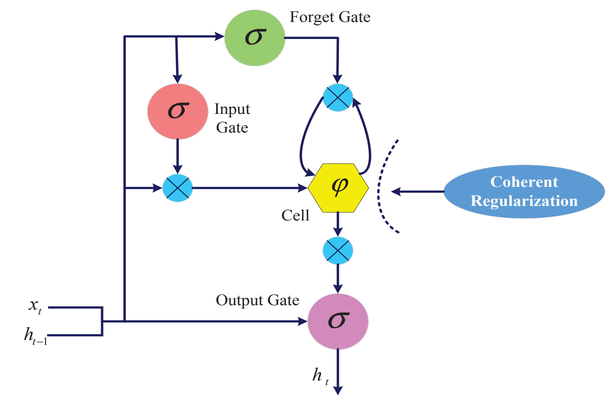
References

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

[2] Zhou, Bolei and Tang, Xiaoou and Wang, Xiaogang: *Scalable Deep Poisson Factor Analysis for Topic Modeling* In Proceedings of the 32nd International Conference on Machine Learning (ICML), pages 1-8, 2015

Model Crowd Motions

We use LSTM to model the crowd dynamic. Each LSTM unit has a cell or memory unit, which maintains its state $\mathbf{c}_t$.



Due to the interaction between different pedestrians, we propose to update the memory unit by incorporating its own state together with its neighboring agents with a coherent regularization as

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

Coherent Motion

We investigate the dependency between agents in coherent groups, which are discovered using the coherent filtering [2].

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

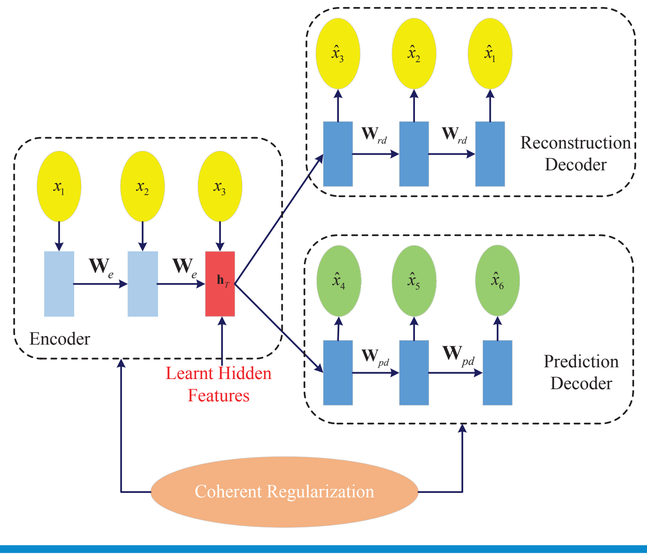
$$\tau_j(t) = \frac{\mathbf{v}_i(t) \cdot \mathbf{v}_j(t)}{\|\mathbf{v}_i(t)\| \|\mathbf{v}_j(t)\|}, \quad (2)$$

The dependency coefficient between the i_{th} and j_{th} tracklets in Eq. (1) is defined as

$$\lambda_j(t) = \frac{1}{\mathbf{Z}_i} \exp\left(\frac{\tau_j(t) - 1}{2\sigma^2}\right) \in (0, 1], \quad (3)$$

cLSTM Framework

To learn an informative representation, we take the "encoder-decoder" approach which consists of an encoder coherent LSTM and a decoder coherent LSTM.



Results

We first test the performance of our framework on path forecasting. Sample results are demonstrated in the figure below, in which the red tracklets are the paths obtained with the KLT tracker that are followed by green curves of tracklets generated with our cLSTM prediction model.

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.

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

We demonstrate the effectiveness of our method in estimating group state and classifying crowd video dependent on the holistic crowd behaviors in a scene. In the CUHK Crowd Dataset, groups are classified into four states as Gas, Solid, Pure Fluid and Impure Fluid and 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*.

We train a softmax classifier using the hidden features learnt by our *cLSTM*, and then implement the group state estimation and crowd video classification.