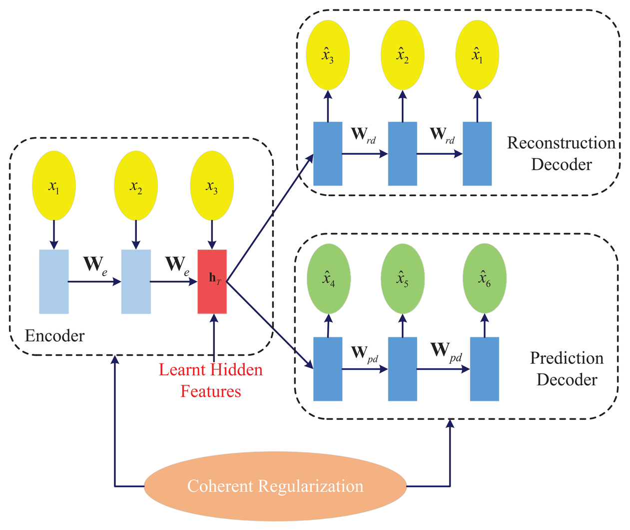


Crowd Scene Understanding with Coherent Recurrent Neural Networks

Hang Su, Yinpeng Dong, Jun Zhu, Haibin Ling, and Bo Zhang
{suhangss, dcsjz, dcszb}@mail.tsinghua.edu.cn, donyp13@mails.tsinghua.edu.cn, hbling@temple.edu

Introduction

Understanding collective behaviors in crowd scenes has a wide range of applications in **video surveillance** and **crowd management** [3]. It has the following challenges:

- Crowd spatio-temporal behavior patterns behave abundantly **nonlinear dynamics**, such as limit cycles, quasi-period and even chaos.
- **Collective effect** (or **coherent motion**), e.g. pedestrians in crowds tend to form coherent groups by aligning with other neighbors.

Contributions

We propose to explore the crowd dynamics with a coherent Long Short Term Memory (LSTM) architecture [4]. 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;
- 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**.

References

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

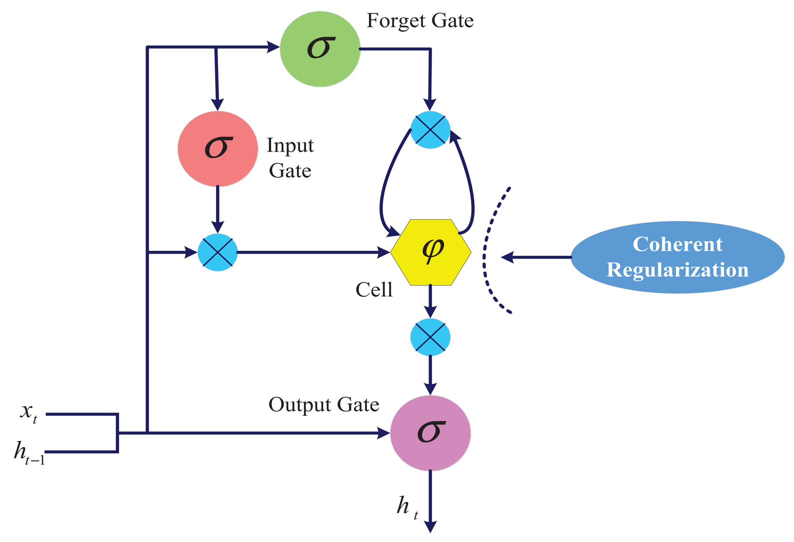
[2] Bolei Zhou, Xiaoou Tang, and Xiaogang Wang: *Scalable Deep Poisson Factor Analysis for Topic Modeling*, In Proceedings of the 32nd International Conference on Machine Learning (ICML), pages 1-8, 2015

[3] 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

[4] Sepp Hochreiter and Jürgen Schmidhuber : *Long short-term memory*, In Neural Computation, pages 1735–1780, 1997

Model Crowd Motions

We use LSTM to model the crowd dynamic. Each LSTM unit has a cell state $\mathbf{c}_t$ which preserves the information.



Update the cell state by incorporating with its neighboring agents by a **coherent regularization** as

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

Coherent Motion

We use the coherent filtering [2] to detect **coherent groups**.

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

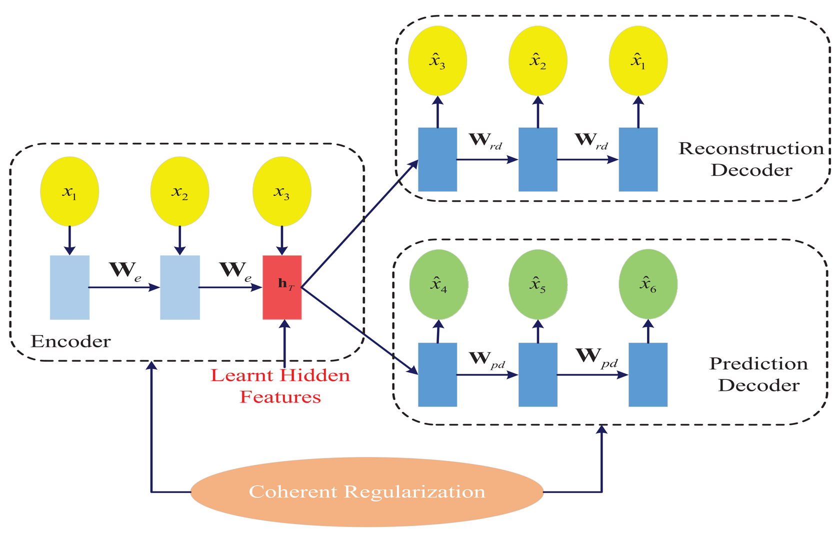
$$\lambda_j(t) = \frac{1}{\mathbf{Z}_i} \exp\left(\frac{\tau_j(t) - 1}{2\sigma^2}\right) \in (0, 1], \qquad (2)$$

and $\tau_j(t)$ is:

$$\tau_j(t) = \frac{\mathbf{v}_i(t) \cdot \mathbf{v}_j(t)}{\|\mathbf{v}_i(t)\| \|\mathbf{v}_j(t)\|} \qquad (3)$$

cLSTM Framework

To learn an informative representation, we take the **"encoder-decoder" approach** [1].



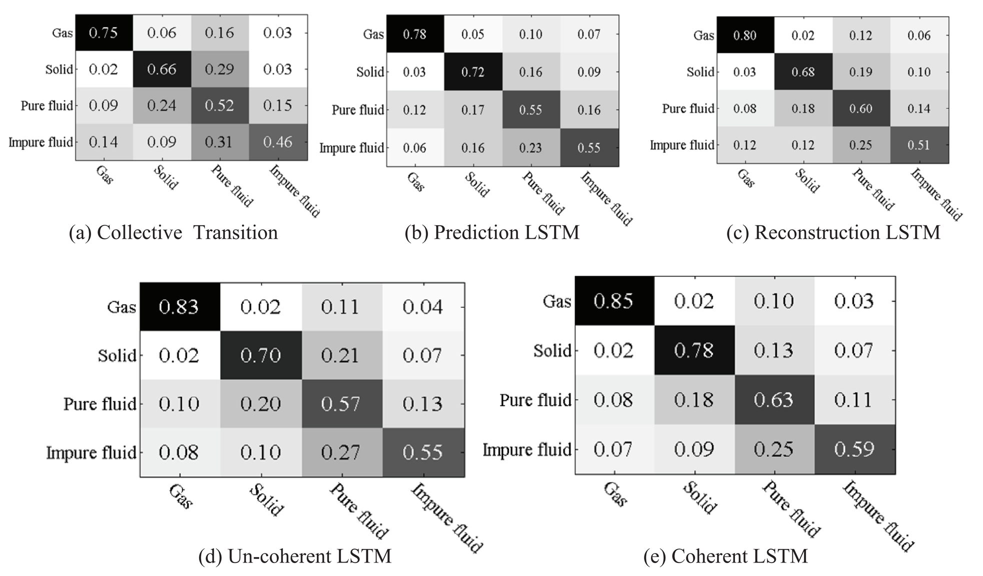
Results

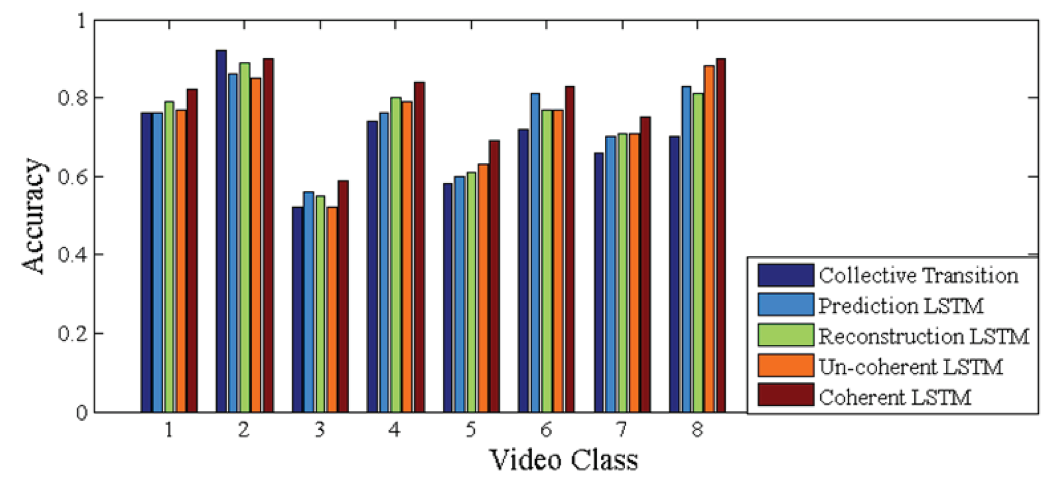
Experiments are conducted on CUHK dataset. Sample results of path forecasting are demonstrated in the figure above. In Table 1, we report the prediction error measured by average pixel distance.

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





Conclusion

- A novel recurrent neural network with **coherent long short term memory unit**;
- Introduce a **coherent regularization** to consider the collective properties;
- **Outperform other methods** in group state estimation and crowd video classification.