Crowd Scene Understanding with Coherent Recurrent Neural Networks

Hang Su, Yinpeng Dong, Jun Zhu

Department of Computer Science and Technology, Tsinghua University

July 12, 2016

Outline

Outline

1. Introduction
2. LSTM Recap
3. Coherent LSTM
4. Experimental Results
5. Conclusion

Background

- Understanding Collective behaviors has a wide range applications in video surveillance and crowd management.

Background

- Understanding Collective behaviors has a wide range applications in video surveillance and crowd management.
- In the real scenes, pedestrians tend to form groups and their trajectories are influenced by others and obstacles.

Background

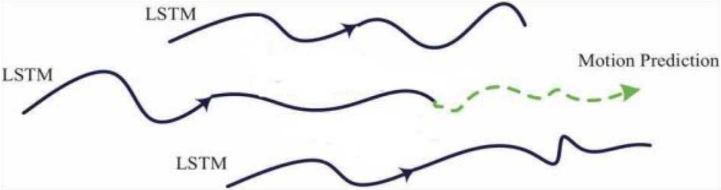
- Understanding Collective behaviors has a wide range applications in video surveillance and crowd management.
- In the real scenes, pedestrians tend to form groups and their trajectories are influenced by others and obstacles.
- The main challenges of crowd motion analysis are *nonlinear dynamics* and *coherent motion*.

Problem Formulation

- Obtain reliable tracklets from each scene using KLT trackers. At any time-instant t, the i^{th} person is represented by his/her coordinate $(\mathbf{x}_i(t), \mathbf{y}_i(t))$.

Problem Formulation

- Obtain reliable tracklets from each scene using KLT trackers. At any time-instant t, the i^{th} person is represented by his/her coordinate $(\mathbf{x}_i(t), \mathbf{y}_i(t))$.
- Predict future trajectories of pedestrians and use extracted hidden features to recognize crowd motions.



Previous Work

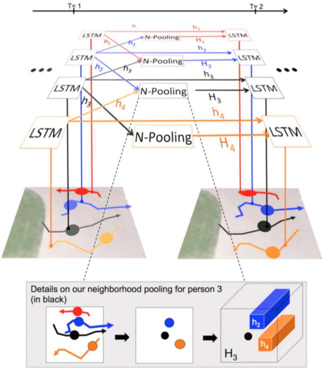
- *Social Force* model
  - Optimize *energy function*
  - Hand-crafted functions
  - Hard to generalize

Previous Work

- *Social Force* model
  - Optimize *energy function*
  - Hand-crafted functions
  - Hard to generalize
- Probabilistic Forecasting
  - *Gaussian Process*

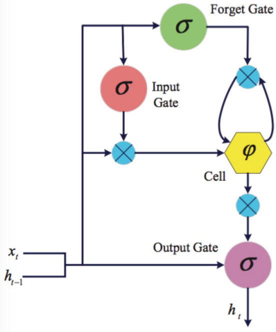
Previous Work

- *Social Force* model
  - Optimize *energy function*
  - Hand-crafted functions
  - Hard to generalize
- Probabilistic Forecasting
  - *Gaussian Process*
- Recurrent Neural Networks
  - N-LSTM [Alahi et al., 2016]

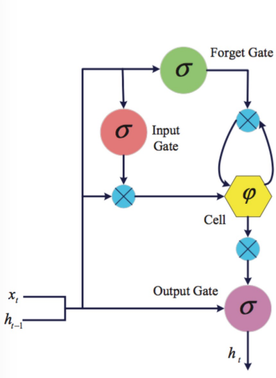


Outline

LSTM



LSTM



- Structure
  - Input / Output / Forget gate
  - Memory state $\mathbf{c}_t$
- Advantage
  - Prevent vanishing gradient problem
  - Nonlinear characteristic
  - Generalization

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

Outline

Why Coherent LSTM?

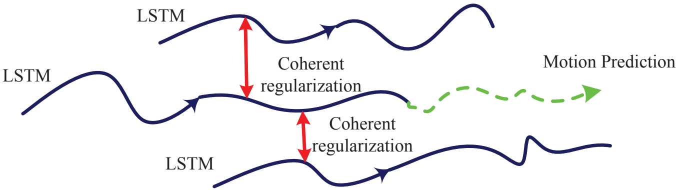
- LSTM can model individual behaviors but can't capture the interaction in a group.

Why Coherent LSTM?

- LSTM can model individual behaviors but can't capture the interaction in a group.
- When the neighboring relationship of individuals remain invariant over time and correlation of their velocities remain high, they tend to have similar hidden state.

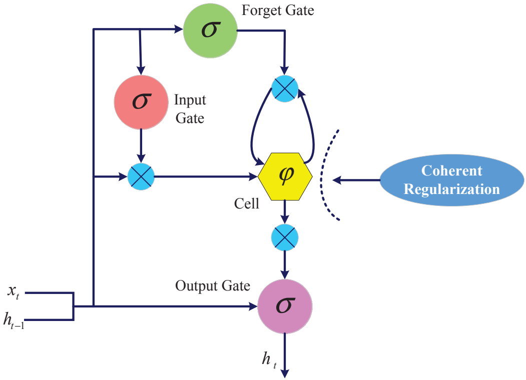
Why Coherent LSTM?

- LSTM can model individual behaviors but can't capture the interaction in a group.
- When the neighboring relationship of individuals remain invariant over time and correlation of their velocities remain high, they tend to have similar hidden state.
- The trajectories of pedestrians not only follow the *old* trend, but also are influenced by *current* environment.



cLSTM Unit

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tanh(\mathbf{W}_{xc}\mathbf{x}_t + \mathbf{W}_{hc}\mathbf{h}_{t-1} + \mathbf{b}_c) + \sum_{j \in \mathcal{N}} \lambda_j(t)\mathbf{f}_t^j \odot \mathbf{c}_{t-1}^j \tag{2}$$



Coherent Motion Modeling

Use coherent filtering [Zhou et al., 2012] [Shao et al., 2014] to discover the coherent group.

Coherent Motion Modeling

Use coherent filtering [Zhou et al., 2012] [Shao et al., 2014] to discover the coherent group.

The dependency relationship between two tracklets within the same group is measured as:

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

Dependency Coefficient

The dependency coefficient between the i_{th} and j_{th} tracklets in Eq. (2) is defined as

$$\lambda_j(t) = \frac{1}{\mathbf{Z}_i} \exp\left(\frac{\tau_j(t) - 1}{2\sigma^2}\right) \in (0, 1] \tag{4}$$

Dependency Coefficient

The dependency coefficient between the i_{th} and j_{th} tracklets in Eq. (2) is defined as

$$\lambda_j(t) = \frac{1}{\mathbf{Z}_i} \exp\left(\frac{\tau_j(t) - 1}{2\sigma^2}\right) \in (0, 1] \tag{4}$$

- $\mathbf{Z}_i$: normalization constant corresponding to the i_{th} tracklet.
- $\lambda_j(t) \simeq \mathbf{Z}_i^{-1}$ if $\mathbf{v}_i(t) \simeq \mathbf{v}_j(t)$ which implies that tracklets i and j are similar.
- Coherent regularization *encourages the tracklets to learn similar feature distributions* by sharing information across tracklets within a coherent group.

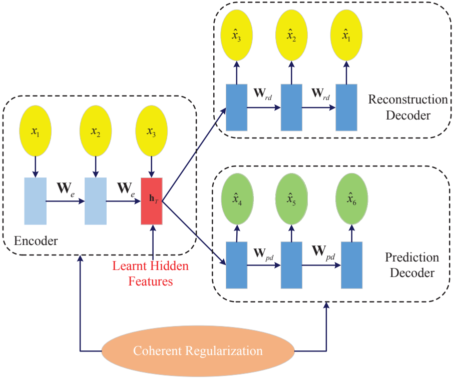
Framework

Unsupervised encoder-decoder cLSTM framework:

$$\mathbf{h}_T = cLSTM_e(\mathbf{x}_T, \mathbf{h}_{T-1}), \tag{5}$$

$$\hat{\mathbf{x}}_t = cLSTM_{dr}(\mathbf{h}_t, \hat{\mathbf{x}}_{t+1}), \text{ where } t \in [1, T], \tag{6}$$

$$\hat{\mathbf{x}}_t = cLSTM_{dp}(\mathbf{h}_t, \hat{\mathbf{x}}_{t-1}). \text{ where } t > T, \tag{7}$$



Crowd Scene Profiling

- Solve critical tasks in crowd scene analysis:
  - Group state estimation
  - Crowd video classification
- Softmax classification using the feature learnt from the unsupervised cLSTM.

Outline

Datasets and Settings

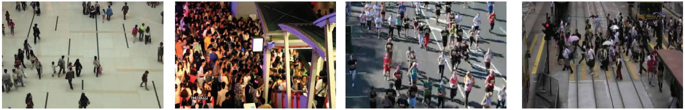
- CUHK Crowd Dataset
  - `http://www.ee.cuhk.edu.hk/~xgwang/CUHKcrowd.html`
  - Scene: streets, shopping malls, airports and parks
  - More than 400 sequences and more then 200,000 traklets
- Settings
  - 128 hidden units in cLSTM
  - 2/3 of tracklets as the input and 1/3 as the predicted tracklets to evaluate the performance.

Future Path Forecasting

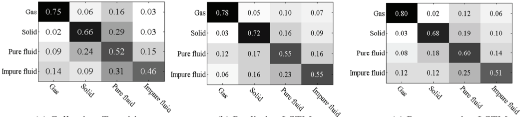
Table 1: Error of Path Prediction(pixels)

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

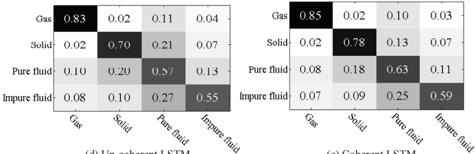
Group State Estimation



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



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

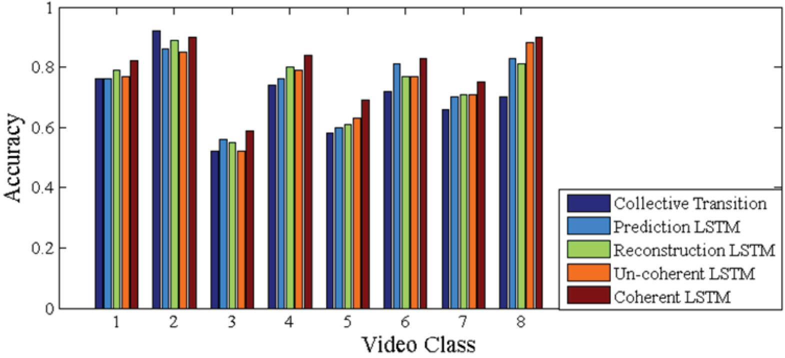


(d) Un-coherent LSTM (e) Coherent LSTM

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.

Crowd Video Classification

All video clips are annotated into 8 classes 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.*



Outline

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.

Thanks for your time!

Questions?