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

1 Introduction

2 LSTM Recap

3 Coherent LSTM

4 Experimental Results

5 Conclusion
Outline

1. Introduction
2. LSTM Recap
3. Coherent LSTM
4. Experimental Results
5. Conclusion
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 $(x_i(t), y_i(t))$. Predict future trajectories of pedestrians and use extracted hidden features to recognize crowd motions.
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 $(x_i(t), y_i(t))$.
- 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

1. Introduction

2. LSTM Recap

3. Coherent LSTM

4. Experimental Results

5. Conclusion
LSTM

Structure

Input / Output / Forget gate

Memory state $c_t$

Advantage

Prevent vanishing gradient problem

Nonlinear characteristic

Generalization

$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_x x_t + W_h h_{t-1} + b_c)$ (1)
**LSTM**

- **Structure**
  - Input / Output / Forget gate
  - Memory state $c_t$

- **Advantage**
  - Prevent vanishing gradient problem
  - Nonlinear characteristic
  - Generalization

\[
c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c)
\]  

Hang Su, Yinpeng Dong, Jun Zhu  
July 12, 2016
Outline

1. Introduction
2. LSTM Recap
3. Coherent LSTM
4. Experimental Results
5. Conclusion
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.
\[ c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_{xc} x_t + W_{hc} h_{t-1} + b_c) + \sum_{j \in \mathcal{N}} \lambda_j(t) f^j_t \odot c^j_{t-1} \]
Coherent Motion Modeling

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

\[
\tau_j(t) = \frac{v_i(t) \cdot v_j(t)}{\|v_i(t)\| \|v_j(t)\|} \tag{3}
\]
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:

$$\tau_j(t) = \frac{v_i(t) \cdot v_j(t)}{\|v_i(t)\| \|v_j(t)\|}$$  \hspace{1cm} (3)
The dependency coefficient between the $i_{th}$ and $j_{th}$ tracklets in Eq. (2) is defined as

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

(4)
The dependency coefficient between the $i_{th}$ and $j_{th}$ tracklets in Eq. (2) is defined as

$$\lambda_j(t) = \frac{1}{Z_i} \exp \left( \frac{\tau_j(t) - 1}{2\sigma^2} \right) \in (0, 1]$$  \hspace{1cm} (4)

- $Z_i$: normalization constant corresponding to the $i_{th}$ tracklet.
- $\lambda_j(t) \simeq Z_i^{-1}$ if $v_i(t) \simeq v_j(t)$ 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.
Unsupervised encoder-decoder cLSTM framework:

\[ h_T = cLSTM_e(x_T, h_{T-1}), \]

\[ \hat{x}_t = cLSTM_{dr}(h_t, \hat{x}_{t+1}), \text{ where } t \in [1, T], \]

\[ \hat{x}_t = cLSTM_{dp}(h_t, \hat{x}_{t-1}), \text{ where } t > T, \]
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

1 Introduction

2 LSTM Recap

3 Coherent LSTM

4 Experimental Results

5 Conclusion
Datasets and Settings

- CUHK Crowd Dataset
  - Scene: streets, shopping malls, airports and parks
  - More than 400 sequences and more than 200,000 tracklets

- 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)

<table>
<thead>
<tr>
<th>Kalman Filter</th>
<th>Un-coherent LSTM</th>
<th>Coherent LSTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.32 ± 1.99</td>
<td>6.64 ± 1.76</td>
<td>4.37 ± 0.93</td>
</tr>
</tbody>
</table>
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

1 Introduction

2 LSTM Recap

3 Coherent LSTM

4 Experimental Results

5 Conclusion
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?