Crowd Scene Understanding with Coherent Recurrent Neural Networks

Hang Su, Yinpeng Dong, Jun Zhu

May 22, 2016

Outline

Outline

Background

- Crowd scene is the scene of public places where a large group of people who have gathered together such as a university campus or the sidewalks of a busy street.

Background

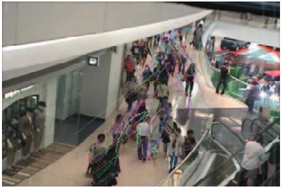
- Crowd scene is the scene of public places where a large group of people who have gathered together such as a university campus or the sidewalks of a busy street.
- Groups are the main entities that make up a crowd.

Background

- Crowd scene is the scene of public places where a large group of people who have gathered together such as a university campus or the sidewalks of a busy street.
- Groups are the main entities that make up a crowd.
- When pedestrians walk in a crowded space, their trajectories are influenced by others and obstacles.

Background

- Crowd scene is the scene of public places where a large group of people who have gathered together such as a university campus or the sidewalks of a busy street.
- Groups are the main entities that make up a crowd.
- When pedestrians walk in a crowded space, their trajectories are influenced by others and obstacles.



Applications

Understanding collective behaviors in crowd scenes has a wide range of applications in

- Video Surveillance

Applications

Understanding collective behaviors in crowd scenes has a wide range of applications in

- Video Surveillance
- Crowd Management

Applications

Understanding collective behaviors in crowd scenes has a wide range of applications in

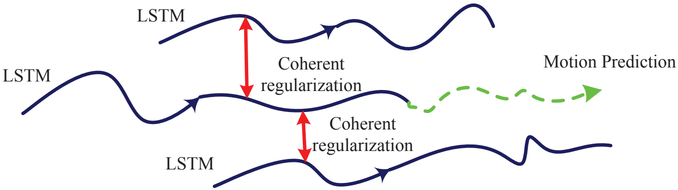
- Video Surveillance
- Crowd Management
- Avoiding Tragic Accidents

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 do other classification tasks.



Challenge

Challenge

- Crowd spatio-temporal patterns behave nonlinear dynamics
    - Limit cycles
    - Quasi-period
    - Chaos

Challenge

- Crowd spatio-temporal patterns behave nonlinear dynamics
  - Limit cycles
  - Quasi-period
  - Chaos
- Collective effect (or coherent motion)
  - Pedestrian tend to form groups
  - Intra-group properties and inter-group properties.

Previous Work

- Traditional approach such as *Social Force* model

Previous Work

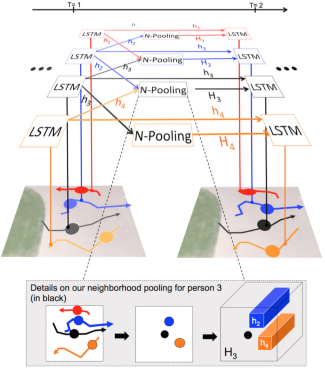
- Traditional approach such as *Social Force* model
  - Optimize *energy function*
  - Hand-crafted functions
  - Hard to generalize

Previous Work

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

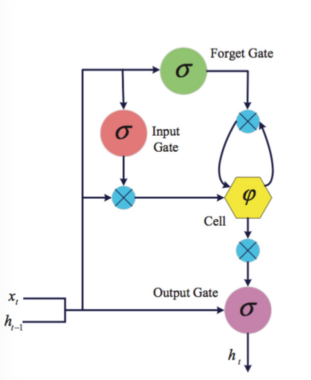
Previous Work

- Traditional approach such as *Social Force* model
  - Optimize *energy function*
  - Hand-crafted functions
  - Hard to generalize
- Probabilistic Forecasting such as *Gaussian Process*
- Recurrent Neural Networks
  - N-LSTM (*CVPR* 2016)

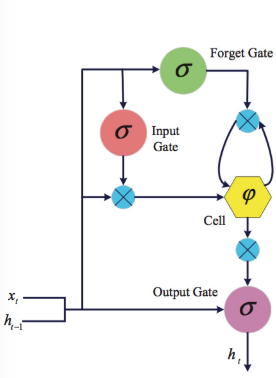


Outline

LSTM

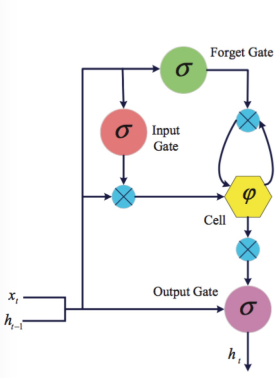


LSTM



- Structure
  - Input / Output / Forget gate
  - Memory state $\mathbf{c}_t$

LSTM



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

LSTM

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

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

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

$$\mathbf{o}_t = \sigma(\mathbf{W}_{xo}\mathbf{x}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{W}_{co}\mathbf{c}_t + \mathbf{b}_o) \tag{4}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \tag{5}$$

Outline

Why Coherent LSTM?

- LSTM can model individual behaviors but dose not capture the interaction of people in a group.

Why Coherent LSTM?

- LSTM can model individual behaviors but dose not capture the interaction of people in a group.
- The individuals are always willing to engage with “seed” groups and form spatially coherent structures.

Why Coherent LSTM?

- LSTM can model individual behaviors but dose not capture the interaction of people in a group.
- The individuals are always willing to engage with “seed” groups and form spatially coherent structures.
- 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?

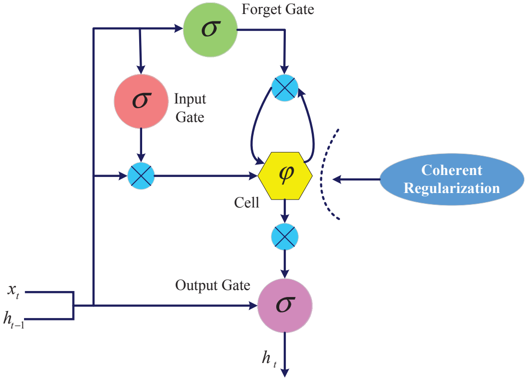
- LSTM can model individual behaviors but dose not capture the interaction of people in a group.
- The individuals are always willing to engage with “seed” groups and form spatially coherent structures.
- When the neighboring relationship of individuals remain invariant over time and correlation of their 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

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

cLSTM Unit

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



Coherent Motion Modeling

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

Coherent Motion Modeling

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

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

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

Dependency Coefficient

The dependency coefficient between the i_{th} and j_{th} tracklets in Eq. (6) is defined as

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

Dependency Coefficient

The dependency coefficient between the i_{th} and j_{th} tracklets in Eq. (6) is defined as

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

- $\mathbf{Z}_i$: normalization constant corresponding to the i_{th} tracklet.

Dependency Coefficient

The dependency coefficient between the i_{th} and j_{th} tracklets in Eq. (6) is defined as

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

- $\mathbf{Z}_i$: normalization constant corresponding to the i_{th} tracklet.
- $\lambda_j(t) \simeq 1$ if $\mathbf{v}_i(t) \simeq \mathbf{v}_j(t)$ which implies that tracklets i and j are similar.

Dependency Coefficient

The dependency coefficient between the i_{th} and j_{th} tracklets in Eq. (6) is defined as

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

- $\mathbf{Z}_i$: normalization constant corresponding to the i_{th} tracklet.
- $\lambda_j(t) \simeq 1$ 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.

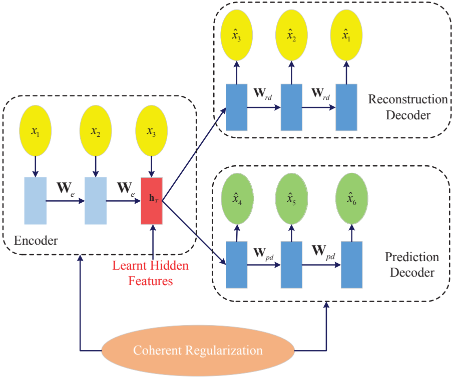
Framework

Unsupervised encoder-decoder cLSTM framework:

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

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

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



Crowd Scene Profiling

Solve critical tasks in crowd scene analysis:

- Estimating group state
  - *Gas, Solid, Pure Fluid and Impure Fluid*
  - Softmax classification using the feature learnt from the unsupervised cLSTM.

Crowd Scene Profiling

Solve critical tasks in crowd scene analysis:

- Estimating group state
  - *Gas, Solid, Pure Fluid and Impure Fluid*
  - Softmax classification using the feature learnt from the unsupervised cLSTM.
- Crowd video classification

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

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

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

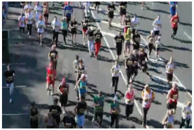
- Gas: Particles move in different directions without forming collective behaviors
- Solid: Particles move in the same direction with relative positions unchanged
- Pure Fluid: Particles move towards the same direction with ever-changing relative positions
- Impure Fluid: Particles move in a pure fluid style with invasion of particles from other groups

Group State Estimation

- Gas: Particles move in different directions without forming collective behaviors
- Solid: Particles move in the same direction with relative positions unchanged
- Pure Fluid: Particles move towards the same direction with ever-changing relative positions
- Impure Fluid: Particles move in a pure fluid style with invasion of particles from other groups

(a) Gas

(b) Solid

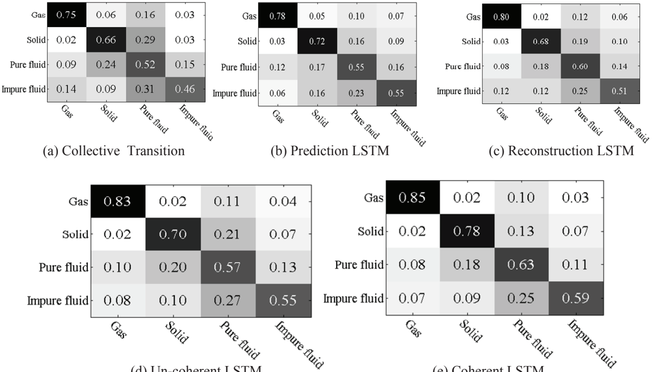


(c) Pure Fluid

(d) Impure Fluid

Group State Estimation

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.



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

(d) Un-coherent LSTM (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.*

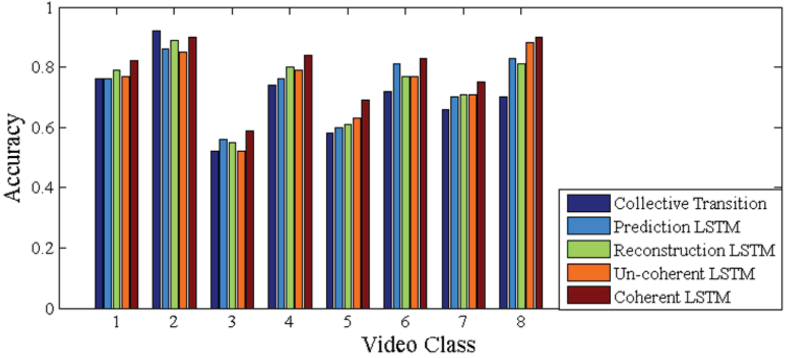
Crowd Video Classification

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*.

	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

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.*



Thanks for your time!

Questions?