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

MODEL CROWD MOTIONS
We use LSTM to model the crowd dynamic. Each LSTM unit has a cell state $c_t$ which preserves the information.

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

$$c_t = f_t \odot c_{t-1} + \sum_{j \in N} \lambda_j(t)f_t \odot c_{j-1}^t$$

$$+ i_t \odot \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c)$$

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

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

and $\tau_j(t)$ is:

$$\tau_j(t) = \frac{v_i(t) \cdot v_j(t)}{||v_i(t)|| ||v_j(t)||}$$

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>
<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>

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.