PSTNet: Crowd Flow Prediction by Pyramidal Spatio-Temporal Network

Enze YANG†, Shuoyan LIU†, Yuxin LIU†, Nonmembers, and Kai FANG†, Member

SUMMARY Crowd flow prediction in high density urban scenes is involved in a wide range of intelligent transportation and smart city applications, and it has become a significant topic in urban computing. In this letter, a CNN-based framework called Pyramidal Spatio-Temporal Network (PSTNet) for crowd flow prediction is proposed. Spatial encoding is employed for spatial representation of external factors, while prior pyramid enhances feature dependence of spatial scale distances and temporal spans, after that, post pyramid is proposed to fuse the heterogeneous spatio-temporal features of multiple scales. Experimental results based on TaxiBJ and MobileBJ demonstrate that proposed PSTNet outperforms the state-of-the-art methods.

key words: crowd flow prediction, spatial encoding, multi-scale feature extraction, neural architecture search

1. Introduction

With the explosion of population and vehicles in bigalopolis, big issues such as traffic congestion, frequent traffic accidents and air pollution have brought serious influence to the development of society and economics. Urban computing is proposed in [1] to tackle the problems with urban sensing and big data analytics. As a significant topic in urban computing, crowd flow prediction has involved in intelligent transportation and smart city applications.

With great achievement of deep learning algorithms, crowd flow prediction is proved to make breakthrough progresses by convolutional neural network (CNN). According to Zhang et al. [2]. As shown in Fig. 1, crowd inflow and outflow are measured by the number of people that move in or out of a citywide region, Deep-ST is proposed to describe the distribution of spatio-temporal features. Based on Deep-ST, [3] is proposed to model the spatial and temporal properties of crowd traffic. Inspired by ResNet [4], residual units are employed to extract the spatial dependence of city regions.

However, Deep-ST [2] and ST-ResNet [3] extract the external factors by fully connected layers, which is lack of spatial association. Actually, external factors like weather also impact on crowd flow distribution in spatial and temporal domain. Further, feature maps are simply fused by element-wise, which neglect the interaction between components and layers. According to DeepSTN+ [5], multi-scale feature maps contain hierarchal spatio-temporal semantics, the scheme of hand-craft multi-scale feature fusion could be optimized, yet the improvement of prediction accuracy is partly obtained by point of interests information. Additionally, as a hotspot of neural network researches in recent years, remarkable performance is achieved by Neural Architecture Search, NAS [6] in computer vision tasks. Taking advantage of the NAS ideas for spatio-temporal prediction is a field of great prospect.

Among CNN-based methods, the performance of crowd flow prediction is still limited. Hence, a novel framework called Pyramidal Spatio-Temporal Network (PSTNet) is proposed in this letter, which makes a major contribution to research on spatio-temporal prediction by demonstrating:

- In order to enrich the spatial attributes of external features, we introduce spatial encoding to extract the features of external vectors instead of fully connected layers.
- Prior pyramid is proposed to enhance the feature dependence of various spatial distances and temporal spans, which leverages a novel convolution of periodical dilation. A heterogeneous map with rich features is generated by prior pyramid.
- A multi-scale feature fusion scheme adapted to crowd flow prediction is searched in post pyramid, which effectively improves the capability of multi-scale feature fusion.

The proposed PSTNet consists of spatial encoding and two pyramidal architectures. To the best of our knowledge, PSTNet improves the accuracy of crowd flow prediction compared with the state-of-the-art model.
2. Proposed Method

In this section, we focus on the details of our proposed PST-Net for crowd flow prediction, which consists of a structure of spatial encoding and two pyramidal architectures.

2.1 Spatial Encoding

Except for spatio-temporal features, the factors that represent the status and environment have an impact on the distribution of crowd flow. These factors are counted as external vectors indexed by date in crowd dataset TaxiBJ. Due to the limitation of the statistical means, the format of external factors could not represent the specific location like:

- **Weather**: The regional impacts of rainfall or strong wind that affects the distribution of crowd flow.
- **Festival**: During the major holidays like the National Day and the Spring Festival, the crowd flow of places like shopping malls and railway stations could increase significantly.

In order to introduce the spatial attributes of external factors, spatial coding is proposed in Eq. (1):

\[
E = f_w(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)
\]  

where \(f_w(\cdot)\) indicates spatial coding and \(\varepsilon_i\) is the external vector indexed by date. To be specific, external factors are classified by the type of radiation scale. For large-scale external factors like weather, regional spatial initialization is utilized to learn encoding maps, while the external factors of spot region like railway stations in holidays, the encoding maps is initialized by sparse metrics.

2.2 Prior Pyramid

According to [3], ResUnit is leveraged to extract the features of spatio-temporal maps, which integrates the deep features of various time span. However, the dependence between features of layers is limited to the adjacent space by the fixed convolution scale, while the association of long-distance spatial dependence is weakened.

Receptive field is a concept that represents the scope of spatial features. There are two widely accepted strategies to expand the receptive field: i) A large kernel size could increase the mapping area of feature, but it is not an efficient solution because the explosion of the parameters introduced by the large kernel size. ii) Dilated convolution proposed in [7], which is to extend the receptive field, yet the fixed dilation rate of kernel size is insufficient to collect adjacent relationship of crowd flow. Li et al. [8] proposed a novel convolution recently, the idea of periodically switching the dilated ratio of kernel size can effectively deal with the limitation of fixed-size dilated convolution. As shown in Fig. 2, the dilated ratio of each \(K \times K\) kernel is altered by the depth of convolution channel for multi-scale feature extraction.

The dependence of spatial features could be extracted by a periodical kernel with sharing weights, which enables the network to aggregate the rich nearby and remote spatial dependence. According to [8], semantics of multi-scale features can be fused by adjusting the pattern of convolution without introducing additional parameters and computational complexity, which is novel and efficient.

In this letter, periodical convolution is included in prior pyramid, after that, a heterogeneous feature map is preliminarily fused in Eq. (2):

\[
HF = f_{prior}(X^C, X^P, X^T, E)
\]

where \(HF\) indicates the heterogeneous feature map, \(X^C, X^P, X^T\) indicate the feature maps of different time spans. \(E\) represents external feature map of Eq. (1). \(f_{prior}(\cdot)\) is the process of prior pyramid that contains few steps. i) The feature of \(X^C, X^P, X^T\) are extracted via periodical convolution. ii) \(E\) is merged with \(X^C, X^P, X^T\) separately, the sensitivity of external factors to different time spans is enhanced. iii) The feature of \(X^C, X^P, X^T\) are fused after periodical convolution, a heterogeneous map with multiple features is generated.

The proposed prior pyramid consists of a batch of periodically dilated convolution layers, aiming to extract spatial dependency of input grid maps at various time spans. Spatial and temporal feature of crowd flow distribution are closely related, that is due to the data fragmentation and feature merging scheme. Concisely, prior pyramid enriches spatio-temporal feature correlation without introducing additional computing costs, which turns out to be an effective module of feature extraction and fusion.

2.3 Post Pyramid

In major computer vision tasks, feature maps of various scales contain rich features. An acceptable statement is that the feature maps with high resolution have sufficient spatial features, while deep feature maps contain rich semantic features. In order to enrich both of the features, the method of multi-scale fusion is highly researched. In noted [9], feature maps are fused by feature pyramid network, which brings an
idea of network reconstruction and effectively improve the detection accuracy.

As mentioned in [6], neural architecture search is the process of automating architecture engineering, which achieves the optimal performance by searching the network composition in a fixed search space. However, searching the entire network architecture requires a number of computation resources with such a long training period. To deal with this issue, the search task is split into branches. Golnaz G et al. [10] propose a novel neck architecture with scalable search space and cross-scale connections. NAS-FPN leverages the fusion of different layers as the basic unit, so as to determine the specific feature maps used for merging. Covering all possible cross-scale connections, NAS-FPN acquires the optimal architecture for multiple feature fusion.

In our framework, multi-scale feature maps are fused by several pyramid units searched by NAS, the relation of spatio-temporal features are further enriched in various resolution. Proposed post pyramid aims to aggregate the features with multiple spatio-temporal association by an optimal fusion architecture with NAS. Different from NAS-FPN for object detection, the input resolution of the gridded feature map (such as TaxiBJ, the shape of input map is 32 × 32) is much lower than normal pixel-level images. Therefore, the initial merging cell of post pyramid is generated by 2 down-sample layers, which indicates the search task is sustainable. Further, in order to predict crowd flow that aggregates multi-scale spatial features and rich semantic information, the resolution of output feature map is fixed at the same scale as the input.

From the efficient perspective of our framework, the connections between the merging cells that have less contribution are pruned to reduce parameters of the merging unit. The pyramidal block that contains merging cells and cross connections are repeated subsequently, hence the architecture of post pyramid could obtain an efficient multi-feature fusion scheme with less drop of accuracy. The grid map of crowd flow is predicted in Eq. (3):

\[
\hat{Y} = \sigma(f_{con}(f_{post}(HF)))
\]  

where \( \hat{Y} \) indicates the predicted crowd inflow and outflow map, \( \sigma(\cdot) \) is the activation function of \( \text{Tanh} \). \( f_{post}(\cdot) \) represents the post pyramid architecture and \( HF \) is the heterogeneous feature map derived from Eq. (2). Post pyramid fuses the features of multi-scale spatial correlation, which associates nearby crowd flow with distant events. Additionally, the heterogeneous map contains temporal feature from various time spans, the fusion scheme have impact on the distribution of crowd flow in multi-scale spatio-temporal feature.

The PSTNet proposed in this letter is formed by aggregating the structure of spatial encoding, prior pyramid and post pyramid. The overall framework of PSTNet is shown in Fig. 3.

Firstly, the vectors that represent external factors are encoded by a spatial feature map. By initializing the external map with different radiation scales, spatial encoding is leveraged to enrich the composition of crowd flow characteristics. Then, we adopt the structure of prior pyramid to extract multi-scale features of spatio-temporal input gird maps, which associates the spatial dependence and temporal relationship, a heterogeneous feature map is fused by prior pyramid. Finally, post pyramid is proposed to further fuse the multiple features of various attributes, which effectively improves the performance in capturing both spatial and temporal dependency.

3. Experimental Results and Discussion

In this section, we briefly introduce the open source crowd flow prediction datasets, then the evaluation metrics and experimental results are illustrated and discussed.

TaxiBJ is the dataset produced by Zheng Yu et al. in [2]. As a milestone of crowd flow prediction, trajectories are represented as grid maps and external vectors with time stamps. External factors are collected as a set of weather conditions and holidays.

MobileBJ is another dataset collected from the most popular social network vendor in Beijing, China. In the demand of crowd prediction task, the data is transformed as the definition of [2]. To make sure that the fairness of the experiments, external factors are collected as vectors as well.

Our model is trained and evaluated separately on TaxiBJ and MobileBJ. In the training process, PSTNet is optimized via Adam with a weight decay of 0.0001 and momentum of 0.9. The learning rate is set to 0.0001, where the batch-size is 64 on a RTX 8000 graphics card.

To evaluate the crowd flow prediction results, Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are utilized to describe the differences between the predictions and ground-truth as in the Eqs. (4) and (5):
Table 1  Evaluation results on TaxiBJ. Prior represents the prior pyramid, Post is the structure of post pyramid. SE means the implementation of spatial encoding. The proposed PSTNet is a combination of spatial encoding, prior pyramid and post pyramid.

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>HA</td>
<td>45.01</td>
<td>24.48</td>
</tr>
<tr>
<td>VAR</td>
<td>22.88</td>
<td>13.79</td>
</tr>
<tr>
<td>ST-ResNet</td>
<td>18.70</td>
<td>10.49</td>
</tr>
<tr>
<td>DeepSTN</td>
<td>17.04</td>
<td>10.08</td>
</tr>
<tr>
<td>Prior + SE</td>
<td>16.68</td>
<td>9.30</td>
</tr>
<tr>
<td>Post + SE</td>
<td>16.46</td>
<td>9.13</td>
</tr>
<tr>
<td>Prior + Post</td>
<td>15.96</td>
<td>9.02</td>
</tr>
<tr>
<td>Prior + Post + SE (PSTNet)</td>
<td><strong>15.40</strong></td>
<td><strong>8.58</strong></td>
</tr>
</tbody>
</table>

Table 2  Evaluation Results on MobileBJ

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>HA</td>
<td>122.36</td>
<td>50.31</td>
</tr>
<tr>
<td>VAR</td>
<td>61.25</td>
<td>43.78</td>
</tr>
<tr>
<td>ST-ResNet</td>
<td>40.13</td>
<td>25.53</td>
</tr>
<tr>
<td>DeepSTN</td>
<td>37.08</td>
<td>23.94</td>
</tr>
<tr>
<td>Prior + SE</td>
<td>34.92</td>
<td>23.10</td>
</tr>
<tr>
<td>Post + SE</td>
<td>34.03</td>
<td>21.85</td>
</tr>
<tr>
<td>Prior + Post</td>
<td>31.99</td>
<td>20.31</td>
</tr>
<tr>
<td>Prior + Post + SE (PSTNet)</td>
<td><strong>31.32</strong></td>
<td><strong>20.04</strong></td>
</tr>
</tbody>
</table>

\[
RMS\ E = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \| Y_i - \hat{Y}_i \|^2} \quad (4)
\]

\[
MAE = \frac{1}{n} \sum_{i=1}^{n} \| \hat{Y}_i - Y_i \|^2 \quad (5)
\]

where the \( \hat{Y}_i \) indicates the predicted grid map of \( i \)th time interval and \( Y_i \) is the corresponding ground-truth. RMSE represents the differences between the prediction and the ground-truth, which is widely accepted to measure the loss of predicted vectors. Hence, RMSE is also utilized as the loss function for optimization.

Table 1 and 2, show the evaluation results of baselines and our model on TaxiBJ and MobileBJ. The horizontal line across the table indicates the boundary between the baseline models and proposed methods. HA means the result of historical average of crowd flow, while VAR is a method of vector auto-regressive in early years. ST-ResNet [3] represents the model of convolution neural network. DeepSTN is proposed to improve the performance of ST-ResNet, which achieves the state-of-the-art results. Four sets of experiments are conducted to verify the effectiveness of each sub-module of our PSTNet, the results are listed below the horizontal line.

4. Conclusion

A pyramidal network for crowd flow prediction is proposed in this letter, named PSTNet. Spatial encoding is applied to enrich the spatial attributes of external features, while the architecture of prior pyramid is proposed to extract features of spatio-temporal dependence and external factors, after that, post pyramid fuses the multi-scale rich features of the heterogeneous map. According to the experimental results of TaxiBJ and MobileBJ, proposed PSTNet is proved to have a strong ability to predict crowd flow and improves the accuracy of crowd flow prediction compared with the state-of-the-art models.

Acknowledgments

This work was partially support by the fund of China Academy of Railway Sciences (2019YJ114, 2018YJ102).

References