A Shallow Prediction Model with Enhanced Features for Video Time Series

Yifei Zhang\textsuperscript{1} and Hongxia Bie\textsuperscript{2,}\textsuperscript{*}
\textsuperscript{1}Beijing Key Laboratory of Network System and Network Culture, Beijing University of Posts and Telecommunications, Beijing, China
\textsuperscript{2}School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China

*Corresponding author

Abstract. Fast video time series prediction is an important part in video anomaly detection in deep learning framework. Time series deep learning models have good performance in time-series data prediction, such as traffic flow, rainfall and network traffic. The models, such as ConvLSTM, ConvGRU with complex structure and large amount of computation, are difficult to fit the video time series prediction. In this paper, we analyze data characteristics of video sequences and propose a shallow video series prediction model. We not only consider the spatial correlation in continuous video frames, but also investigate the interval video frames, which also maintain spatial correlation. We explore the gray scale pixel values of the interval frames which appear specific trend properties when we connect the context pixels together. Based on the enhanced features, we design a shallow 2D convolutional video time series prediction model with good prediction performance and much less total parameters.

1. Introduction
The prediction of video time series is a major part in surveillance video analysis and anomaly detection. At present, the investigation of the time series prediction\cite{1}\cite{2}\cite{3} is more advanced than the video time series prediction. 1D time series prediction model include shallow conv1d model, Dilated CNN\cite{4}, regression model of Lightgbm\cite{5} and LSTM\cite{6}. Typical Spatio-Temporal data prediction models include ConvLSTM\cite{7}, TrajGRU\cite{8} and ConvGRU\cite{9}, which are based on the continuous property of time series data. Also, the trend and period properties are used in DeepSTCL\cite{10} and ST-ResNet\cite{11}.

For video time series prediction, pixels in the same position from continuous video frames can be considered as 1D data and the video frame as 2D data. But the presented time series prediction models are too complex to be used directly for so large amount of video data. In our research, we analyze the video series data properties. Then, we take advantage of the spatial correlation of continuous video time series, the spatial correlation of interval video frames and the trend property of the gray scale pixel values in interval frames. With these properties as enhanced features, we build a shallow 2D convolutional video time series prediction model. Our experimental results demonstrate that the proposed model has the competitive performance, with much less amount parameters.
2. Feature Analysis of Video Series Data
UCSDped1[12] dataset is employed in this paper. This dataset mainly used for video frame anomaly detection. It contains 34 training videos and 36 testing videos. We train the model by all training videos and test the model by all normal frames in testing videos.

2.1. Feature of Video Series
Gray scale frame of video time series data is shown in figure 1.

![Figure 1. 200 frames from UCSDped1 Train001. 1st: 1st frame; 2nd: 2nd frame; 3rd: 200th frame.](image)

Gray scale pixel values of point (60,100) from UCSDped1 Train001 is shown in figure 2.

![Figure 2. gray scale pixel values curve of point (60,100) from UCSDped1 Train001.](image)

Gray scale block pixel values of area pixels (40, 40)->(42,42) from UCSDped1 Train001 is shown in figure 3.

![Figure 3. block pixel (40,40)->(42,42) value curve from UCSDped1 Train001.](image)

2.2. Enhanced Features from Interval Frames
When we connect the context pixels together, the gray scale pixel values of the interval frames appear specific trend properties. We also consider the statistics result between interval frames. We use the trend property of interval frames and spatial correlation of interval frames as enhanced features.
2.2.1. Trend property. Select the interval frame according to the specified length. One specified length correspond some positions of interval frame. we define them as different trend properties. Finally, fuse all positions of interval frame from every specified length as the trend property features. Compared with the gray value curve of continuous video frame, the new gray value curve has the tendency change. figure 4 shows an example about the trend property of interval frames.

Computation method of the trend property is following: we hypothesis \( \text{seq\_len} \) is the length of continuous video frame. The number of interval is \( n \). \( m \) is the number of trend property features selected from one interval. Especially, \( \kappa \) is even number and \( m \geq 4 \). The amount of frames taken from interval frame is \( m-1 \), the \( m \)-th feature is the average of other frames.

In figure 4, \( n = 4 \), \( m = 4 \), \( \text{seq\_len} = 8 \), when calculate the trend property features with frame 70. Feature 1 is the pixel value of \((70-n*\text{seq\_len})\) frame, which is 62nd, 54th, 46th and 38th frame. Feature 2 is the pixel value of \((70-n*\text{seq\_len}-(m-2)/2)\) frame, which is 61st, 53rd, 45th and 37th frame. Feature 3 is the pixel value of \((70-n*\text{seq\_len}+(m-2)/2)\) frame, which is 63rd, 55th, 47th and 39th frame. Feature 4 is the average pixel value of Feature1, Feature2 and Feature 3.

2.2.2. Spatial correlation. Every specified length interval correspond with some positions of interval frames. We calculate some statistic results from these frames. Finally, fuse all the results from every specified length as the spatial correlation features.

The method of calculate spatial correlation features is following: With the use of \( m-1 \) frames from trend property, besides the \( m \)-th feature. The top \((m-2)/2\) features are subtracted with the \(((m-2)/2+1)\) feature. The last \((m-2)/2\) features also subtracted with the \(((m-2)/2+1)\) feature. Then absolute value of all \( m-2 \) results.

In figure 5, \( n = 1 \), \( m = 4 \), \( \text{seq\_len} = 8 \), when calculate the spatial correlation feature with frame 70. The pixel value of frame \((70-n*\text{seq\_len}-(m-2)/2)\) is \( P_i \), which is 61st frame. The pixel value of

![Figure 4](image_url)

Figure 4. Trend property of pixel point (60, 100) from UCSDped1 Train001. (a):Feature 1; (b):Feature 2; (c):Feature 3; (d):Feature 4.
frame \((70 - n \cdot \text{seq_len})\) is \(P_2\), which is 62nd frame. Feature 5 is \(|P_1 - P_2|\). The pixel value of frame \((70 - n \cdot \text{seq_len} + (m-2)/2)\) is \(P_3\), which is 63rd frame. Feature 6 is \(|P_2 - P_3|\).

**Figure 5.** Interval frame spatial correlation of frame 70 from UCSDped1 Train001 (a): Feature 5; (b): Feature 6.

### 3. Shallow 2D Convolutional Video Time Series Prediction Model

With the spatial correlation of continuous video frames series, which is a given length of video frame series, and the enhanced feature in Section 2. We propose a shallow 2D convolutional video time series prediction model, as shown in figure 6.

**Figure 6.** Shallow 2D convolutional network. FC: Fully-connected; conv: 2D convolutional.

In figure 6, \(X_{\text{time}}\) is continuous video frames series spatial correlation feature, which dimension is \(\text{seq_len} \times \text{height} \times \text{width}\). \(X_{\text{trend}}\) is interval frame trend property feature, which dimension is \((m \times n) \times \text{height} \times \text{width}\). \(X_{\text{corr}}\) is interval frame spatial correlation feature, which dimension is \(((m-2) \times n) \times \text{height} \times \text{width}\). \(\text{seq_len}\) is the length of given continuous video frame. The number of trend property features selected from one interval is \(m\). The number of interval is \(n\). \text{height}\ is image height. \text{width}\ is image width. All pixels are first divided by 255, and the predicted result is multiplied by 255 to get the prediction frame.

We use a convolution with 64 filters of size \(5 \times 5\) convert \(X_{\text{time}}\) to feature \(X_{\text{feature}}\). Then fuse \(X_{\text{feature}}\) feature, \(X_{\text{trend}}\) feature and \(X_{\text{corr}}\) feature as follows:

\[
\text{relu}(x) = \max(0, x) \quad (1)
\]

\[
X_{\text{all}} = \text{relu}(X_{\text{feature}}) + X_{\text{trend}} + X_{\text{corr}} \quad (2)
\]
where $X_{\text{all}}$ is fused by $\text{seq}_\text{len}$ dimension of $\text{relu}(X_{\text{feature}})$, $(m \times n)$ dimension of $X_{\text{cond}}$ and $((m-2) \times n)$ dimension of $X_{\text{corr}}$. ReLU is the activation function. After the Fully-connected layer, tanh is applied to the last layer to cater the output data $y$ to input data label $y_{\text{label}}$. Our model can be trained by minimizing mean squared error between $y$ and $y_{\text{label}}$ to optimize all learnable parameters.

$$\text{mse} = \|y - y_{\text{label}}\|^2$$  \hspace{1cm} (3)

For the shallow 2D convolutional video time series prediction model without enhanced features, only use $X_{\text{cond}}$ for model training and testing.

4. Experimental Results

All our experiments are conducted with 16.0 GB RAM. One GPU (NVIDIA GeForce GTX 1080ti) is equipped and GPU acceleration is adopted with the 10.2 version CUDA.

4.1. Evaluation Metric

The Root Mean Squared Error (RMSE) of each prediction frames and ground truth frames.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i} \sum_{j} (y_{ij} - x_{ij})^2}$$  \hspace{1cm} (4)

Where the number of pixels is $n$, $y_{ij}$ is a pixel value of the prediction frame and $x_{ij}$ is a pixel value of the ground truth frame correspond with the prediction frame. We use the average RMSE of all frames in the test datasets as measured.

4.2. Experiments

We evaluate and compare the model parameter quantity and prediction performance of different state-of-the-art models and our model (as shown in Table 1) with the length of continuous video frames series is 5.

Table 1. Different models comparision.

| model category                      | model         | input length | out length | model params | RMSE  |
|-------------------------------------|---------------|--------------|------------|--------------|-------|
| pixel points series                 | LSTM          | 5            | 1          | 71851        | 3.801 |
| (300*300*1)                         | Dilated CNN   | 5            | 1          | 865          | 6.068 |
|                                    | Lightgbm      | 5            | 1          | -            | 9.292 |
| pixel kernel points series          | LSTM          | 5            | 1          | 76651        | 3.868 |
| (296*296*5)                         | Two Layers Conv1d | 5            | 1          | 12433        | 4.215 |
| pixel frame series                  | ConvLSTM      | 5            | 1          | 1535777      | 3.868 |
| (300*300)                           | ConvGRU       | 5            | 1          | 527681       | 3.915 |
|                                    | One Layer Conv2d(our) | 5            | 1          | 8129         | 4.232 |
| pixel frame series with features    | our           | 5            | 1          | 8147         | 4.067 |
| (300*300)                           |               |              |            |              |       |

For the test dataset of UCSDped1, pixel points series model LSTM has the best prediction effect of RMSE. Compared with the pixel points series model LSTM, our shallow 2D convolutional video time series prediction model parameters are reduced by 88.66% and RMSE is only increased by 10.46%. To the pixel frame series model ConvLSTM, our model decreases the model parameters by 99.47%, only increases the RMSE by 5.1%. To the pixel frame series model ConvGRU, our model decreases the model parameters by 98.46%, only increases the RMSE by 3.9%. To the pixel points series model Dilated CNN, our model decreases the RMSE by 32.98%. Compared with our model without enhanced features, the RMSE of our model has better performance.

5. Conclusions

We present a video time series prediction model, called Shallow 2D Convolutional Video Time Series Prediction Model, by using the spatial correlation of continuous video frames series, the spatial
correlation of interval frames and the trend property of interval frames. The experimental results demonstrate that our model can effectively predict the next frame with a small amount of parameters, and the RMSE of the predicted frames and the real frames is small, which satisfies the needs of tradeoff between speed and accuracy in video time series prediction applications.

References

[1] Zhang, Chuxu, et al. "A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. 2019..

[2] Ienco, Dino, et al. "Combining Sentinel-1 and Sentinel-2 Satellite Image Time Series for land cover mapping via a multi-source deep learning architecture." ISPRS Journal of Photogrammetry and Remote Sensing 158 (2019): 11-22.

[3] He, Zhixiang, Chi-Yin Chow, and Jia-Dong Zhang. "STCNN: A Spatio-Temporal Convolutional Neural Network for Long-Term Traffic Prediction." 2019 20th IEEE International Conference on Mobile Data Management (MDM). IEEE, 2019.

[4] Borovykh A, Bohte S, Oosterlee C W. Dilated convolutional neural networks for time series forecasting[J]. Journal of Computational Finance, Forthcoming, 2018.

[5] Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree[C]//Advances in Neural Information Processing Systems. 2017: 3146-3154.

[6] Gers F A, Schmidhuber J, Cummins F. Learning to forget: Continual prediction with LSTM[J]. 1999.

[7] Xingjian S H I, Chen Z, Wang H, et al. Convolutional LSTM network: A machine learning approach for precipitation nowcasting[C]//Advances in neural information processing systems. 2015: 802-810.

[8] Shi X, Gao Z, Lausen L, et al. Deep learning for precipitation nowcasting: A benchmark and a new model[C]//Advances in neural information processing systems. 2017: 5617-5627.

[9] Ballas N, Yao L, Pal C, et al. Delving deeper into convolutional networks for learning video representations[J]. arXiv preprint arXiv:1511.06432, 2015.

[10] Wang D, Yang Y, Ning S. Deeplstm: A deep spatio-temporal convlstm for travel demand prediction[C]//2018 International Joint Conference on Neural Networks (IJCNN). IEEE, 2018: 1-8.

[11] Zhang J, Zheng Y, Qi D. Deep spatio-temporal residual networks for citywide crowd flows prediction[C]//Thirty-First AAAI Conference on Artificial Intelligence. 2017.

[12] Mahadevan, Vijay, et al. "Anomaly detection in crowded scenes." 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2010.