INVESTIGATION IN SPATIAL-TEMPORAL DOMAIN FOR FACE SPOOF DETECTION

Zhonglin Sun, Li Sun, and Qingli Li
Shanghai Key Laboratory of Multidimensional Information Processing,
East China Normal University

1. Introduction

Goal :

Find the best way to perform the end-to-end feature learning for face spoof detection task.

Key issue :

How to exploit the temporal domain and how to combine the temporal domain scheme with CNN for this particular task.

Proposed method:

We compare schemes on the raw data in single stream and fusion methods with optical flow in two streams.

2. Overall Architecture

Two stream framework:

First stream:
single or stacked region(s) of raw images
4 types of temporal models

Second stream:
optical flow in face region

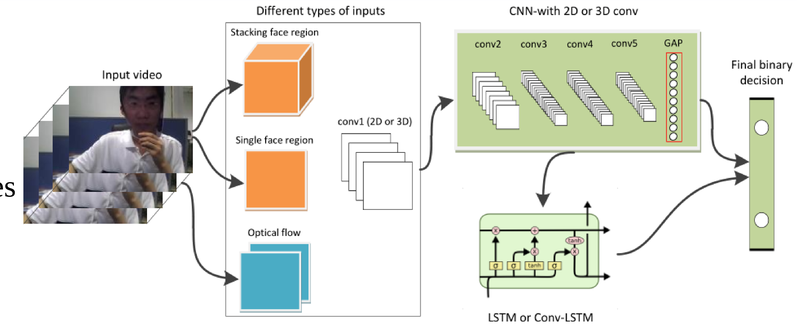

Fig. 1. Overview of different structures to perform feature learning for face spoof detection.

3. INVESTIGATION ON DIFFERENT STRUCTURES IN SPATIAL-TEMPORAL DOMAIN

3.1 2D or 3D convolution

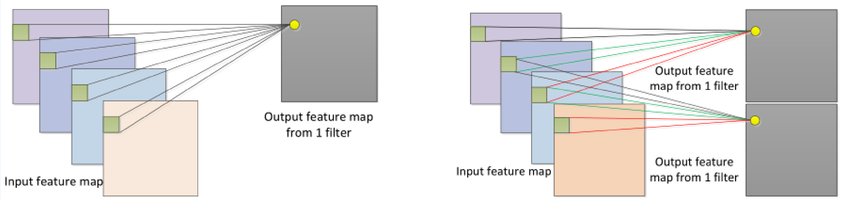

Fig. 2. Comparison between 2D and 3D conv. The normal 2D conv is shown on the left, and 3D conv is shown on the right.

3D[11] conv extends 2D conv by sliding not only in 2D spatial domain, but also along the channel direction.

3.2 LSTM or Conv-LSTM

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}\circ c_{t-1} + b_i) \quad (1a)$$
$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}\circ c_{t-1} + b_f) \quad (1b)$$
$$c_t = f_t\circ c_{t-1} + i_t\circ tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (1c)$$
$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}\circ c_t + b_o) \quad (1d)$$
$$h_t = o_t \circ tanh(c_t) \quad (1e)$$

LSTM takes feature vector x_t from the GAP layer in CNN at time t. $x_t$corresponds only to the image at time t because CNN uses the single face region as its input in this case. The output of LSTM is $h_t$which will be used as a feature vector for the final decision. The cell state is represented by vector c_t .There are also vectors specified by the input, forget and output gate represented by i_t , f_t and o_t. Details of Conv-LSTM can be found in [13].

3.3 Optical flow stream and fusion strategies for two streams

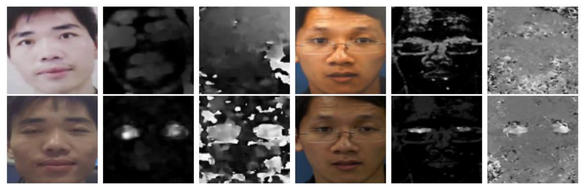

Fig. 3. Optical flow feature demonstration. The first row is the results for attack faces and the second row is for the real one. We show both horizontal and vertical components in the flow.

Fusion strategies
1)Concatenate vectors
2)Train a 4-2 FC layer

Optical flow stream
Similar to CNN-LSTM method

4. Implementation Details

- Number of images in each video clip: 15
- Images are sampled: every 3 frame
- Network: a variant of CaffeNet
- Optical flow: Gunner Farneback's algorithm

- Performance improvement:
  1)Pre-train :parameters need to extend in channel depth direction(CNN-Stacking)
  2)Batch Norm
  3)VGG/ResNet

5. Results and Conclusion

	model	3DMAD	Replay-attack	CASIA
	Spoofnet [21]	0/-	0.70/-	-
	FASNet [7]	0/-	1.20/-	-
	Pluse [22]	7.94/4.71	-	-
	LSTM-CNN [9]	-	-	5.93/5.17
	Multi-cues Integration [23]	0/-	0/-	-/5.83
	Diffusion-based Kernel Matrix [24]	-	4.30/-	-
	Dynamic Texture [25]	-	7.60/-	-/10.00
	Motion Mag [26]	-	1.25/-	-
	Moire pattern [27]	-	3.30/-	0/-
	Colour Texture [3]	-	2.80/0	-/2.10
	Patch and Depth CNN [28]	-	0.72/0.79	2.27/2.67
single	CNN-Stacking	0/0	0.64/3.84	3.72/6.74
	CNN-3Dconv	0/3.30	1.80/3.84	6.51/11.23
	CNN+LSTM	0/0	1.80/2.50	6.51/16.85
	CNN+Conv-LSTM	1.16/3.30	5.13/5.12	14.60/22.40
	CNN-Optical	1.60/0	3.60/11.26	13.84/13.48
fusion	CNN-Stacking	0/0	0.38/2.66	3.49/6.70
	CNN-3Dconv	0/0	2.56/3.77	9.12/13.40
	CNN+LSTM	0/0	1.68/1.28	5.22/14.60
	CNN+Conv-LSTM	0.81/1.66	1.92/6.40	11.44/23.50

Table 1. Comparison of HTER/EER performance on 3 datasets.

	model	replay-attack	CASIA
	Motion [29]	48.28	50.25
	LBP [29]	57.90	47.05
	LBF-TOP [29]	61.33	50.64
	Motion Mag [26]	47.00	50.10
	Spectral cubes [30]	50.00	34.38
	Colour Texture [3]	37.70	30.30
single	CNN-Stacking	41.60	22.72
	CNN-3DConv	49.60	37.74
	CNN+LSTM	42.73	41.10
	CNN+Conv-LSTM	48.70	33.20
	optical flow	30.14	36.80
fusion	stacking	40.40	20.59

Table 2. HTER performance for cross-dataset evaluation.

We compare the performance of the single stream model, named as CNN-Stacking, CNN-3DConv, CNN+LSTM, CNN+ConvLSTM,

With proper fusion scheme, the two stream structure, with its first stream using CNN-Stacking, gives the state-of-the art performance.

6.REFERENCES

[11] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu, "3d convolutional neural networks for human action recognition," IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 1, pp. 221–231, 2013.

[13] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai KinWong, andWang-chunWoo, "Convolutionallstmnet-work: A machine learning approach for precipitation nowcasting," in Advances in neural information processing systems, 2015, pp. 802–810.