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.

3. INVESTIGATION ON DIFFERENT STRUCTURES IN SPATIAL-TEMPORAL DOMAIN

3.1 2D or 3D convolution

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

\[ i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \]  \hspace{1cm} (1a)

\[ f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \]  \hspace{1cm} (1b)

\[ c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b) \]  \hspace{1cm} (1c)

\[ o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_{t} + b_o) \]  \hspace{1cm} (1d)

\[ h_t = o_t \odot \tanh(c_t) \]  \hspace{1cm} (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

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

3.4 2D and 3D convolution

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

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.

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

Table 2. HTER performance for cross-dataset evaluation.

6. REFERENCES
