Two-Stream ConvNet for Video-Based Fall Detection

Weixuan Gao, Jie Wu, Qianmin Hu *{gaow, jiewu22, qianmin}@stanford.edu*

Motivation

Falls are one of the most dangerous situations for elderly people, resulting in fatal injuries and loss of independence of the elderly. Video based fall detection has been more and more popular thanks to its broad application.

In this project, we predict the occurrence of fall in video clips with a variety of background settings. Using a two-stream model combining both spatial and temporal information, we achieved an F-1 score of 91.77%.

Data

Data source

- 873 videos from three online datasets: UR Fall Detection Dataset, Le2i Dataset, and Multiple Camera Fall Dataset
- Positive/Negative labels 427:446. Train/Dev/Test 8:1:1.
- Background settings: coffee room, home, lecture room, office.
- Long videos with multiple falls and other activities are cut into shorter clips.

Pre-processing

- Extract frames from videos using FFMPEG
- Data augmentation: flipping, rotation, scaling, etc.

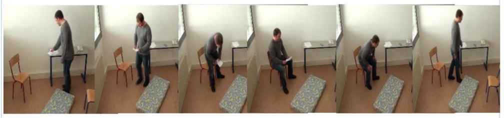


Figure 1: **No Fall**

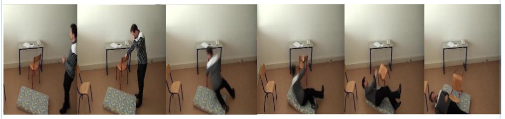


Figure 2: **Fall**

Model

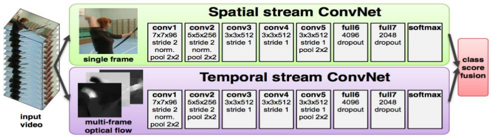


Figure 3: Architecture of two-stream video classification

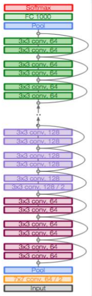


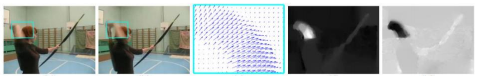
Figure 4: ResNet 152

1. Spatial Stream: Recognize actions from static images

- Model: ResNet 18, 101, 152.
- Input: static video frames
- Pre-trained on ImageNet dataset
- Dropout 0.8
- Model: ResNet 152

2. Temporal Stream: Detect motion with **optical flow**

- Model input: Optical flows, which capture the motion between consecutive frames.
- We use TV-L1 optical flow estimation from OpenCV
- Transform images to optical flow:
  - Input((a) and (b)): two consecutive images $I_0(\mathbf{x})$ and $I_1(\mathbf{x})$, with $\mathbf{x} = (i, j)$ the pixel index.
  - Vector field (c): $\mathbf{u}(\mathbf{x}) = (u_1(\mathbf{x}), u_2(\mathbf{x}))$
  - Output: optical flows: horizontal components of vector field transforms to optical flow (d), vertical components to (e).



(a) (b) (c) (d) (e)

Figure 5: Explain optical flow

3. Merge the two streams

Average the class score of output layer from two streams. Make prediction based on the averaged scores.

Results and Discussion

Results

Model	Dataset	Accuracy	Precision	Recall	F1-Score
Spatial (Resnet-152)	Train	98.29%	94.58%	95.18%	96.26%
	Test	93.51%	92.56%	91.64%	91.77%
Temporal (Resnet-152)	Train	99.99%	100%	99.99%	99.99%
	Test	87.01%	87.60%	84.11%	85.52%
Fusion	Test	91.83%	90.76%	86.40%	88.53%

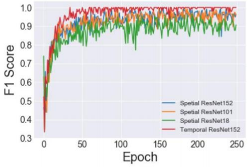


Figure 6: Train set F1 Score

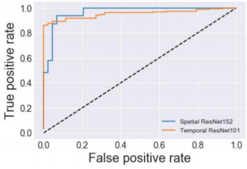


Figure 7: Test set ROC curve

Discussion

- Error Analysis: the models perform not as good (1) when people fall on objects that are not the ground and (2) when the action of falling happens within a very short time in a relatively long video.
- For the protection of the elderly, it is important to detect as many fall occurrences as possible, and to keep false-negative rate low. In real world application, prediction threshold can be set lower than 0.5.

Future work

- Gather more videos where people fall on objects such as coffee table
- Object detection: detect the person before activity recognition.
- Detect multiple fall activities

References

[1] Simonyan, Karen, and Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos." In Advances in neural information processing systems, pp. 568-576. 2014.

[2] Solbach, M. D., & Tsotsos, J. K. Vision-Based Fallen Person Detection for the Elderly. arXiv preprint arXiv:1707.07608. 2017.

[3] J. A. Stevens, P. S. Corso, E. A. Finkelstein, and T. R. Miller. The costs of fatal and non-fatal falls among older adults. Injury Prevention, 12(5):290–295, 2006.