# A long short-term memory convolutional neural network for first-person vision activity recognition

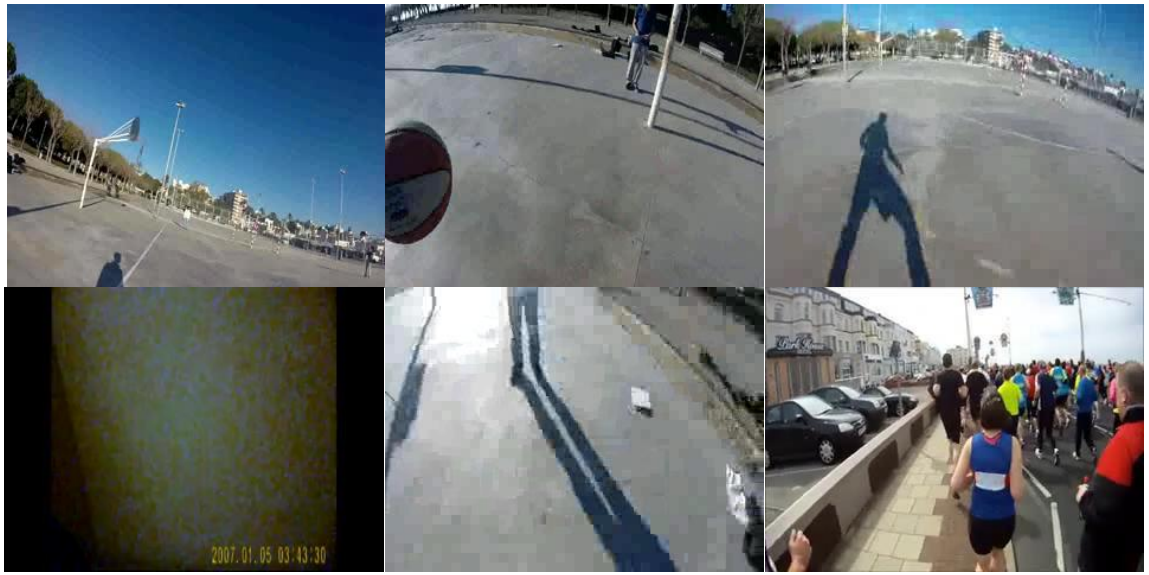
Girmaw Abebe and Andrea Cavallaro

Published in: ICCV Workshop on Assistive Computer Vision and Robotics (ACVR2017)

Centre for Intelligent Sensing
Queen Mary University of London

# Introduction

- Proprioceptive activities involve muscles and joints, e.g., cardiovascular and weight-lifting activities
- Characteristics: motion blur, self occlusion, illumination changes, outlier motion



# Motivation

- Continuous recognition of the activities
  - Requires multi-stage temporal encoding
- Integration of existing motion representation [Abebe2016CVIU,2017NC]
  - Reduce the complexity of deep frameworks
- Transfer knowledge from image-datasets
  - Reduce the need of large datasets

# Challenges

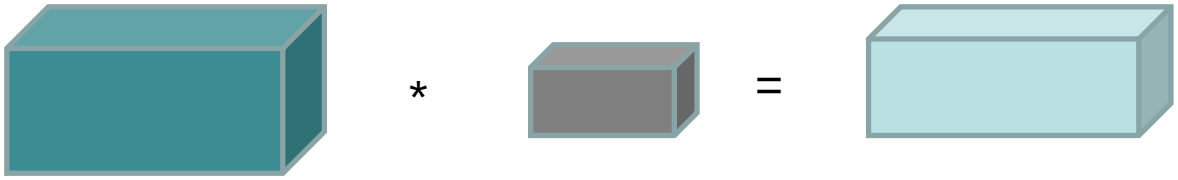
- Deep networks for video-based applications involve
  - High-input data dimension
  - Complex network
  - Large data size requirement
  - Lack of long-term temporal encoding

# Background

- A few works in first-person vision setting  focused on object-interactive activities e.g., cooking [singh2016CVPR, ma2016CVPR]
  - Multi-stream networks that encode objects and local motions
- Few existing methods related to proprioceptive activities [ryoo2015CVPR, poleg2016WACV]
- Times-series gradient pooling (TGP) of features that can also be extracted from deep networks[Ryoo2015CVPR]
  - The deep features do not encode motion and high-feature dimension

# Learning temporal representation

- 3D Convolutions (C3D) [tran2015ICCV, poleg2016WACV]
  - Complex network
  - Early temporal suppression [poleg2016WACV]
  - Very short temporal duration (0.64 seconds) [tran2015ICCV]




Video 3D filter Map

# Learning temporal representation

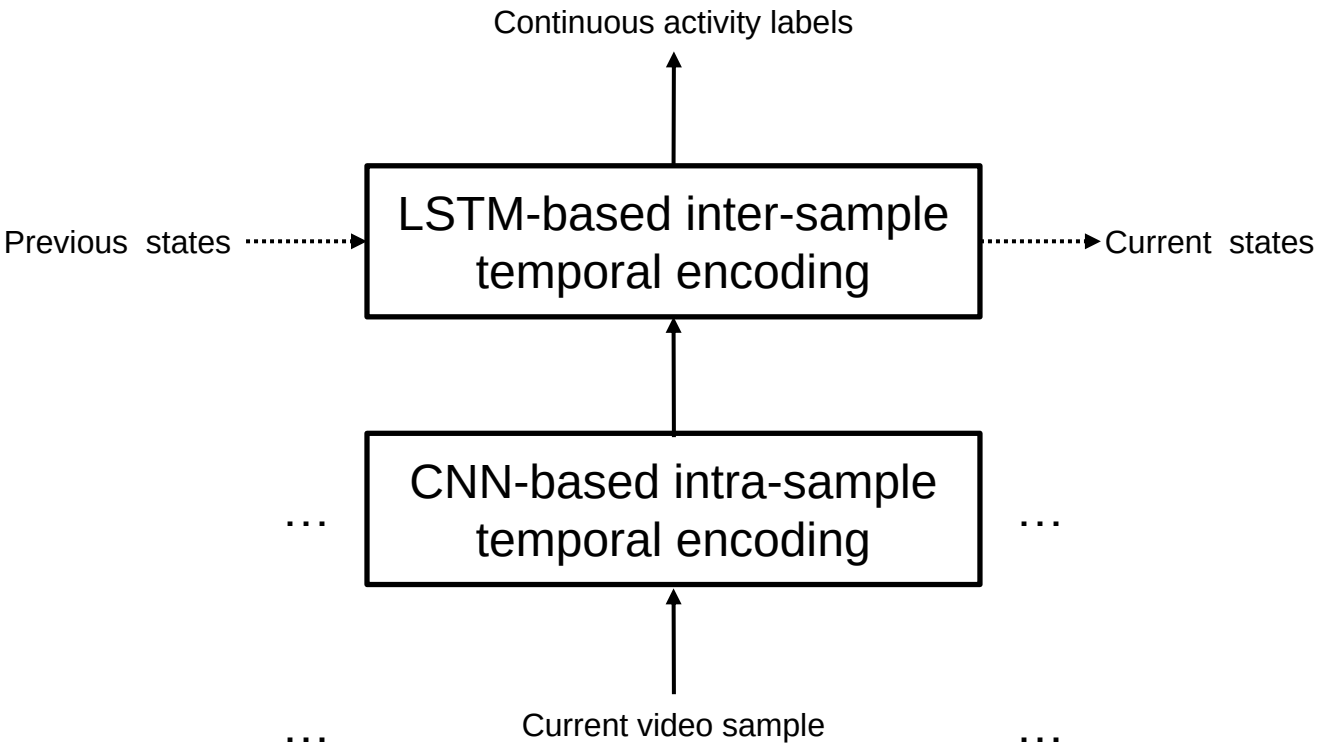
- Trajectory pooled deep descriptors (TDD) [wang2015CVPR]
  - Integrate hand-crafted improved trajectories with 2-stream convolution feature maps
  - Sum pooling is applied
  - High computational cost
  - Less effective to global-motion encoding in FPV

# Learning temporal representation

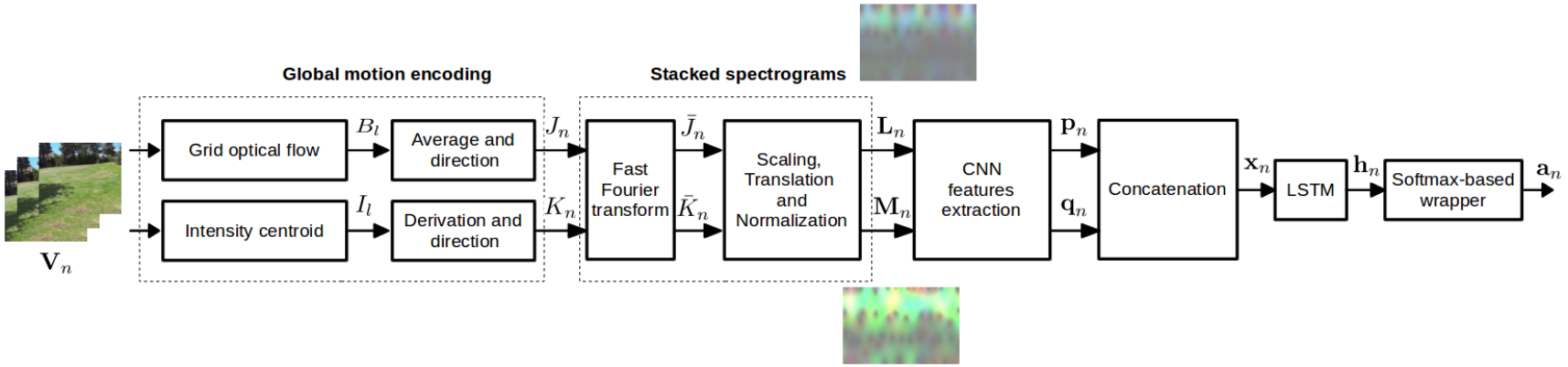
- Rank pooling - Video Darwin (VD) [fernando2015CVPR, 2016PAMI]
  - Learn to rank frame-level features chronologically
  - Simple and easy to implement
  - No exploitation of the deep knowledge (transfer learning)

# Proposed approach

- A CNN-LSTM framework for intra-sample and inter-sample temporal encoding in FPV activity recognition

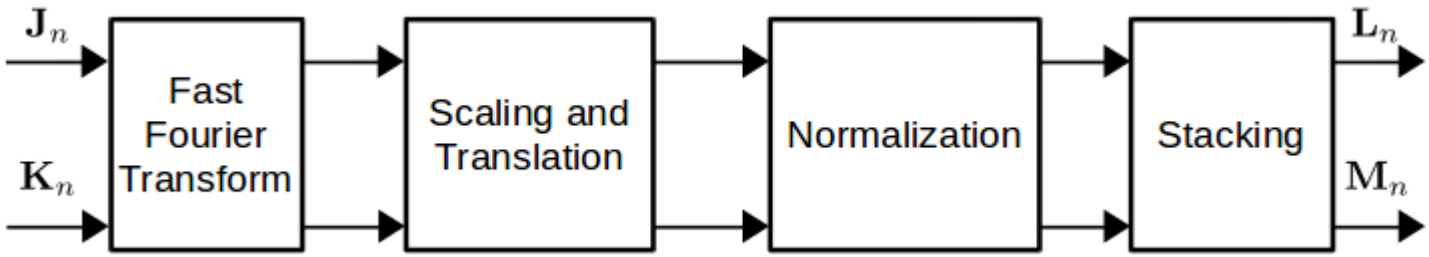


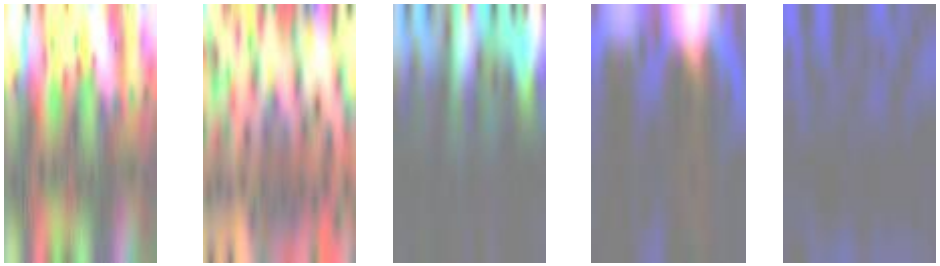
# Proposed approach in details




- Multi-stream global motion encoding
- Stacked spectrogram representation
- High-level intra-sample features extraction (CNN)
- Long-term temporal dependency encoding (LSTM)

# Stacked spectrogram representation



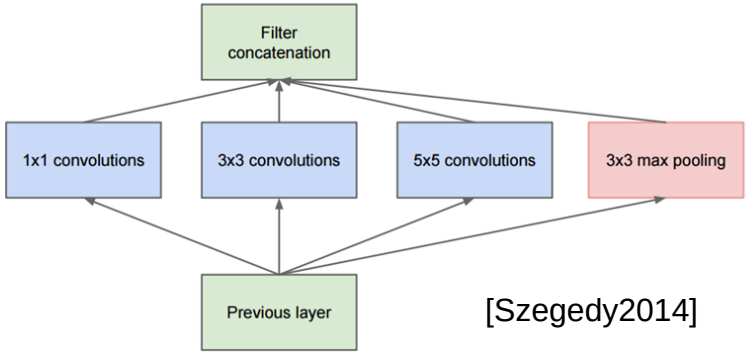



Go upstairs Run Walk Sit/Stand Static

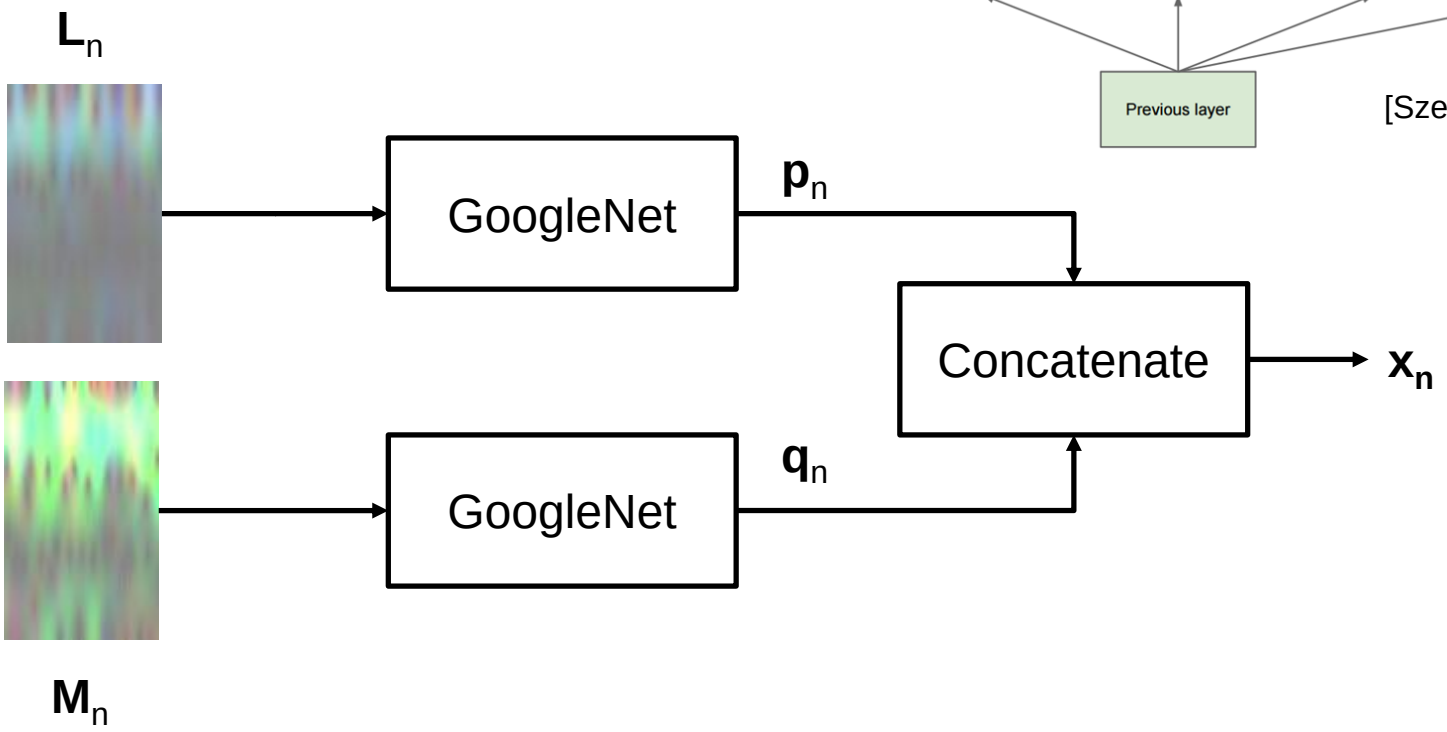
- Spectrogram of horizontal, vertical and direction components constitute a 3-channel representation
- Only requires 2D CNNs

# High-level intra-sample features extraction

- GoogleNet[Szegedy2014] is used on the spectrogram images
  - ImageNet[2014] winner
  - Inception module

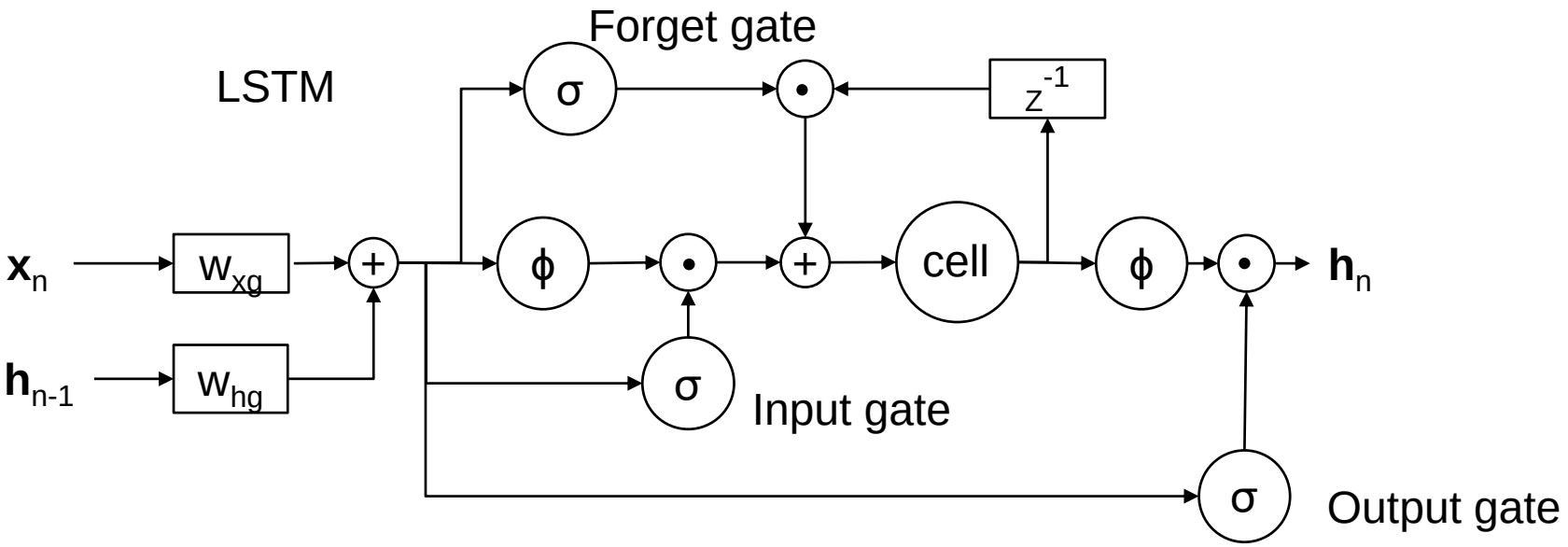



[Szegedy2014]



# Inter-sample temporal encoding

- LSTM-RNN is employed
  - effective for vanishing and exploding gradient




X: input, h: hidden state, W: weight

σ, ϕ : activation function

⊕ : sum, ⊙ : dot product

# Experiments: Datasets

- A proprioceptive subset (15 hrs) of HUJI [poleg2016WACV]
  - Going upstairs, Run, Walk, Sit/Stand, Static
- Train-test sets splits
  - HUJI:  50% for each train and test
- Pre-trained models used
  - Inception on ImageNet,
  - TDD on UCF101, 13K videos, 101 classes
  - C3D on Sport1M, 1M videos, 487 classes

# Results: Intra-sample encoding

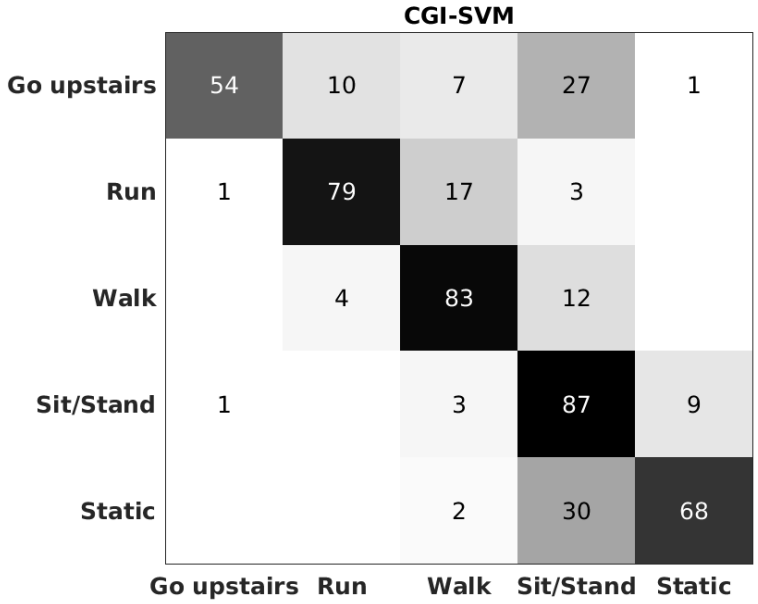
- The proposed inception features from stacked spectrograms outperform the existing representations

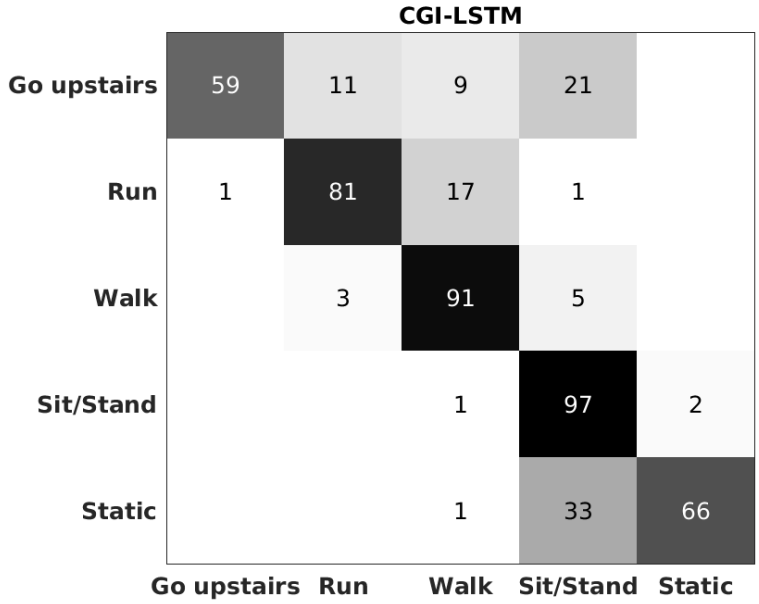
| | Precision (%) | Recall (%) | F-score (%) |
|---|---|---|---|
| C3D | 64 | 64 | 64 |
| VD | 59 | 62 | 60 |
| TDD | 63 | 73 | 68 |
| TGP | 57 | 61 | 59 |
| GI | 61 | 58 | 59 |
| CI | 64 | 69 | 66 |
| CGI (Proposed) | **70** | **74** | **72** |

*SVM is used for all the experiments

# Results: Inter-sample encoding

- The LSTM-based temporal dependency encoding improves the recognition performance



# Results: Inter-sample encoding (2)

- The LSTM network improves the state of the art

| | Without LSTM | | | | | With LSTM | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Up-stairs | Run | Walk | Sit/ Stand | Static | Up-stairs | Run | Walk | Sit/ Stand | Static |
| TGP | 52 | 34 | 82 | 57 | **81** | 52 | 34 | 83 | 60 | 84 |
| VD | 54 | 67 | 46 | 73 | 70 | 55 | 71 | 55 | 89 | 89 |
| C3D | 67 | 74 | 73 | 57 | 53 | 63 | 68 | 74 | 47 | **94** |
| TDD | **68** | 76 | **95** | 52 | 72 | **70** | 71 | 86 | 83 | 39 |
| GI | 51 | 69 | 49 | 66 | 55 | 52 | 76 | 53 | 88 | 51 |
| CI | 49 | 75 | 72 | 83 | 67 | 56 | 79 | 89 | **97** | 67 |
| CGI (Proposed) | 54 | **79** | 83 | **87** | 68 | 59 | **81** | **91** | **97** | 66 |

# Conclusion

- Proposed a CNN-LSTM framework to recognize proprioceptive activities continuously
  - Maintains the temporal information in the CNN block,
  - Provides less complex network (no 3D convolutions and pooling)
  - Exploits the temporal relationships among samples using the LSTM block
  - Applied on multi-modal (Ineritial-Vision) setting [Abebe2017ACVR]
- Limitation
  - Not end-to-end trainable

# Questions?

Source code:
http://www.eecs.qmul.ac.uk/~andrea/fpv-lstm.html.

# References


1. G. Abebe, A. Cavallaro, **A long short-term memory convolutional neural network for first-person vision activity recognition**, Proc. of ICCV workshop on Assistive Computer Vision and Robotics (ACVR), Venice, October 28, 2017
2. G. Abebe, A. Cavallaro and X. Parra, **Robust multi-dimensional motion features for first-person vision activity recognition,** Computer Vision and Image Understanding, Vol. 149, August 2016, pp. 229-248
3. M. S. Ryoo and L. Matthies. **First-person activity recognition:What are they doing to me?** In Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2730 – 2737, Portland, USA, June 2013.
4. B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars**. Rank pooling for action recognition**. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 39(4):773–787, 2017.
5. D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. **Learning spatiotemporal features with 3D convolutional networks**. In Proc. of IEEE International Conference on Computer Vision (ICCV), pages 4489–4497, Santiago, Chile, December 2015.
6. L. Wang, Y. Qiao, and X. Tang. **Action recognition with trajectory-pooled deep-convolutional descriptors**. In Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4305–4314, Boston, USA, June 2015.
7. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. **Going deeper with convolutions**. In Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1–9, Boston, USA, June 2015.
8. G. Abebe and A. Cavallaro. **Hierarchical modeling for first-person vision activity recognition**. Neurocomputing, 267:362–377, June 2017.
9. Y. Poleg, A. Ephrat, S. Peleg, and C. Arora. **Compact CNN for indexing egocentric videos**. In Proc. of IEEEWinter Conference on Applications of Computer Vision (WACV), pages 1–9, New York, USA, March 2016.
10. G. Abebe, A. Cavallaro, **Inertial-Vision: cross-domain knowledge transfer for wearable sensors**, Proc. of ICCV workshop on Assistive Computer Vision and Robotics (ACVR), Venice, October 28, 2017