An Approach Towards Action Recognition using Part Based Hierarchical Fusion

Aditya Agarwal and Bipasha Sen

{adiagar, bise}@microsoft.com

Microsoft

ISVC 2020
Virtual
October 5-7, 2020

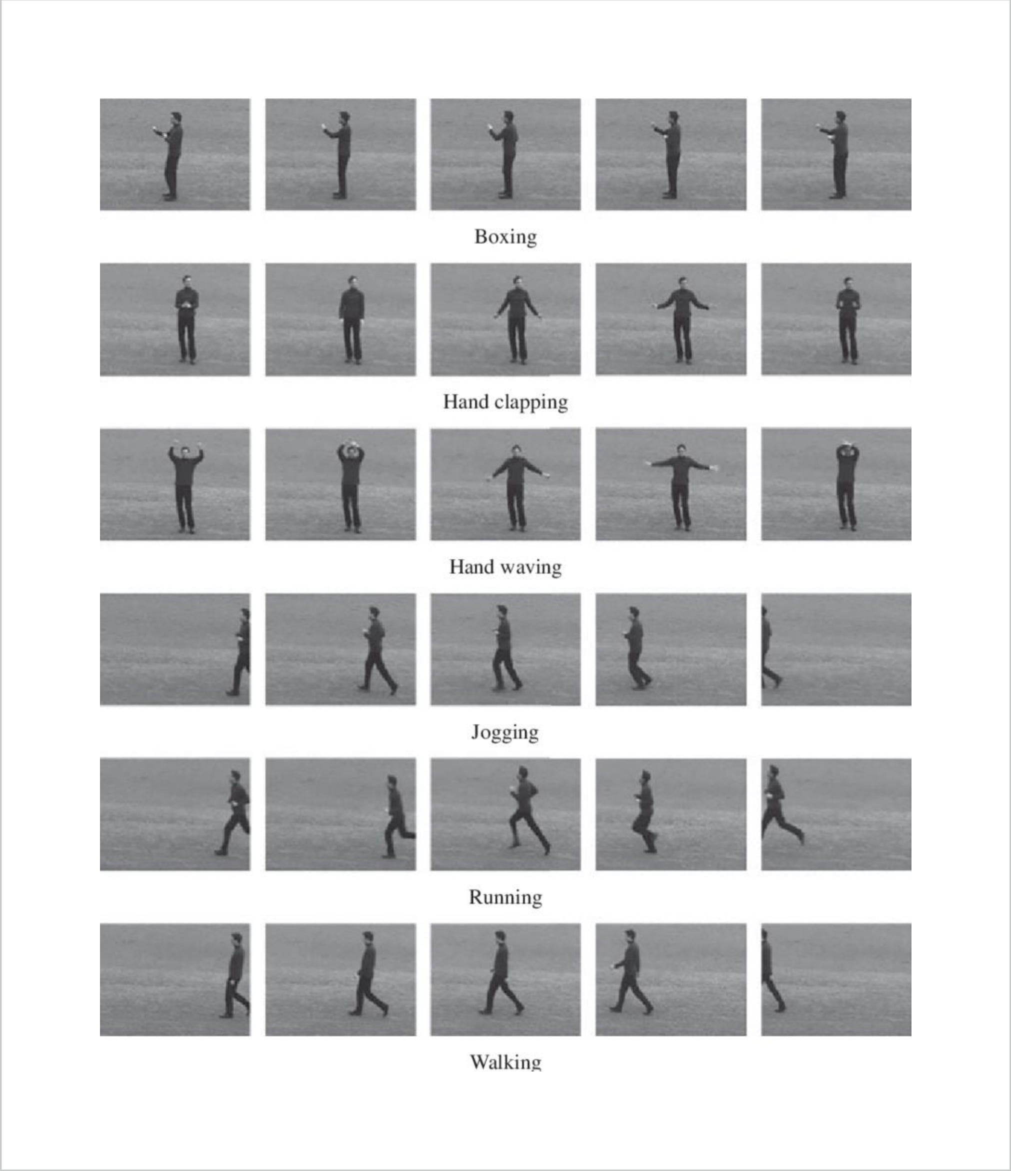
15th International Symposium on Visual Computing

Overview

- Motivation
- Problem Definition
- Human body parts
- Proposed Approach
- Human Action Recognition Pipeline
- Experimental Dataset
- Enhancements
    - Class Imbalance
    - Origin Shift of Pose Coordinates
- Comparative Architectures
- Conclusion
- References

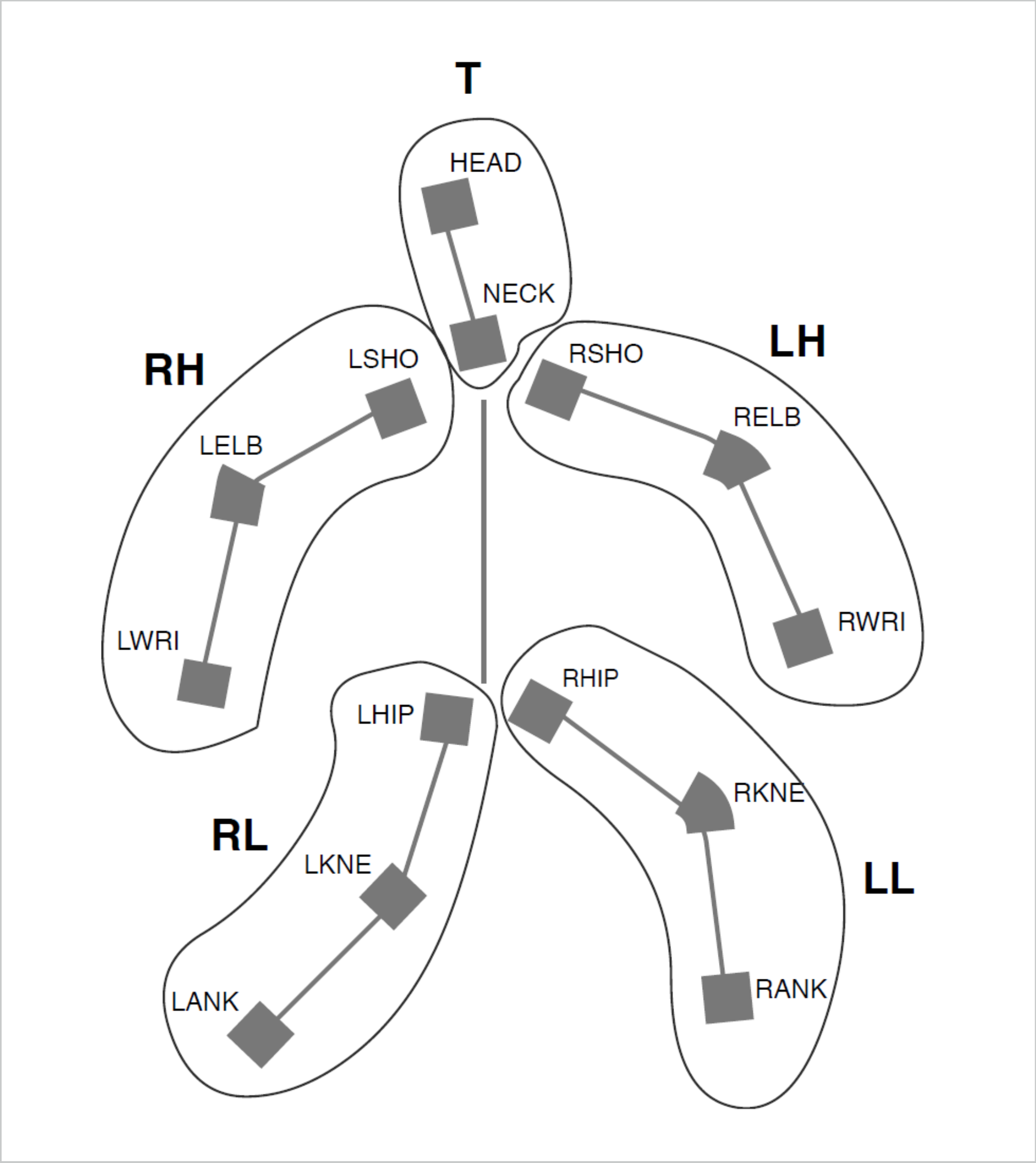
Motivation

- Human action recognition is a challenging problem and has applications in a wide variety of areas like –
  - Video based search and retrieval
  - Intelligent surveillance systems
  - Automated driving
  - Human Computer Interaction
  - Robotics



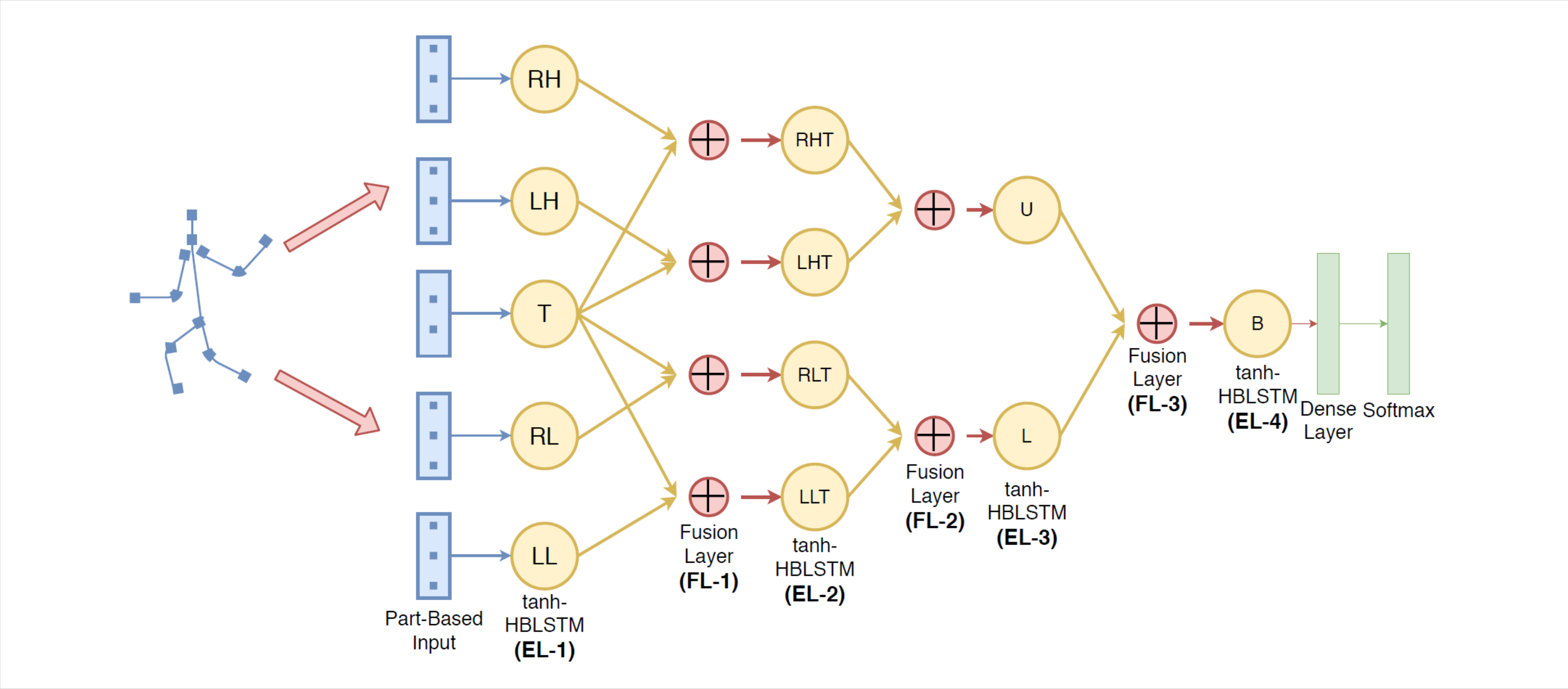
Problem Definition

Given a video with a human as the subject of interest, identify the corresponding action.



Human body parts

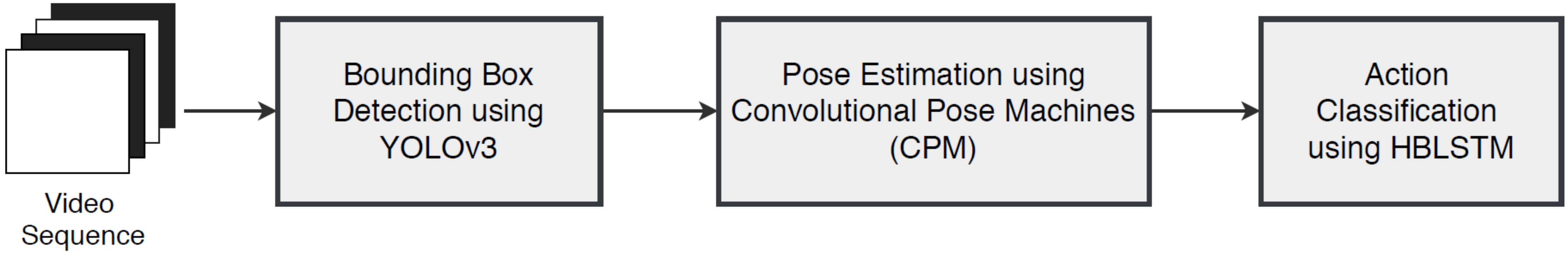
- Human body can be articulated as a system of rigid and hinged joints. These joints can be combined to form the limbs and the trunk.
- Human body can be decomposed into five parts – two arms, two legs and a trunk. The global action can be modeled as the collective action of these five parts.

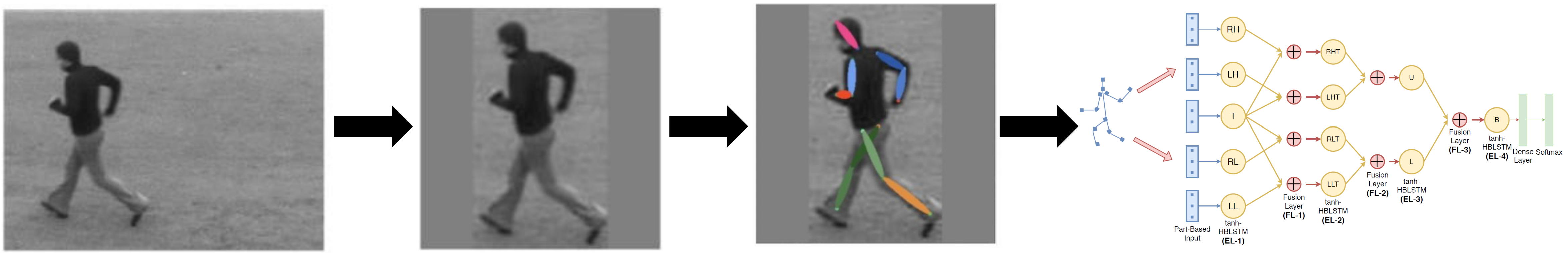


Proposed Approach

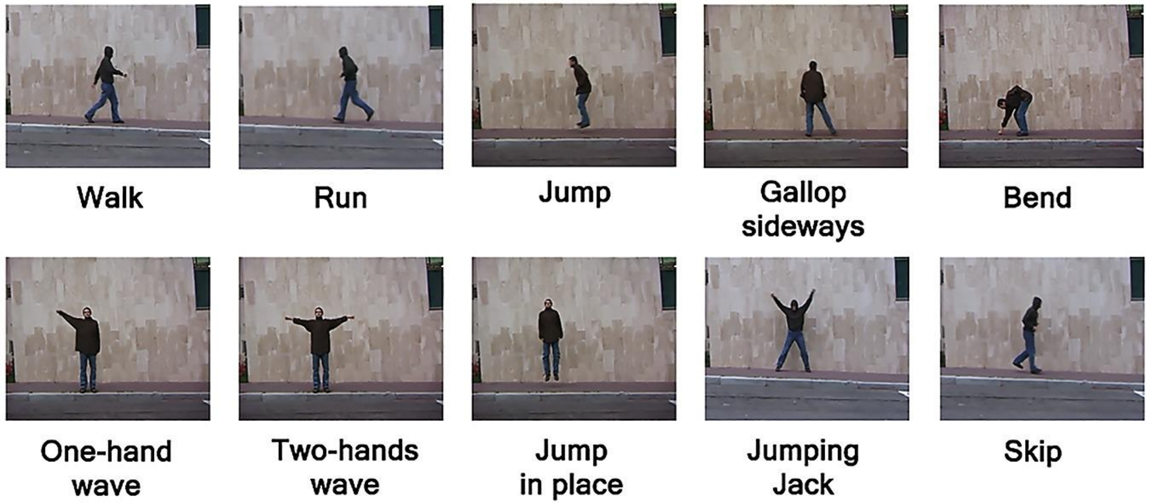
- $h_{i,j}^{t} = \overrightarrow{h_{i,j}^{t}} \oplus \overleftarrow{h_{i,j}^{t}}$ (the encoded part representation of part j at i^{th} layer for time t)
- $I_{i+1,p}^{t} = h_{i,j}^{t} \oplus h_{i,k}^{t}$ (the newly fused p^{th} representation for the fusion layer at time t)
- $O = v_{h_{4,body}^{t}} . h_{4,body}^{T} + b_{h_{4,body}^{t}}$ (the output of the dense layer)
- $(c_k) = \frac{e^{O_k}}{\sum_{j=1}^{C} e^{O_j}}$ (the output class probabilities)

HAR Pipeline

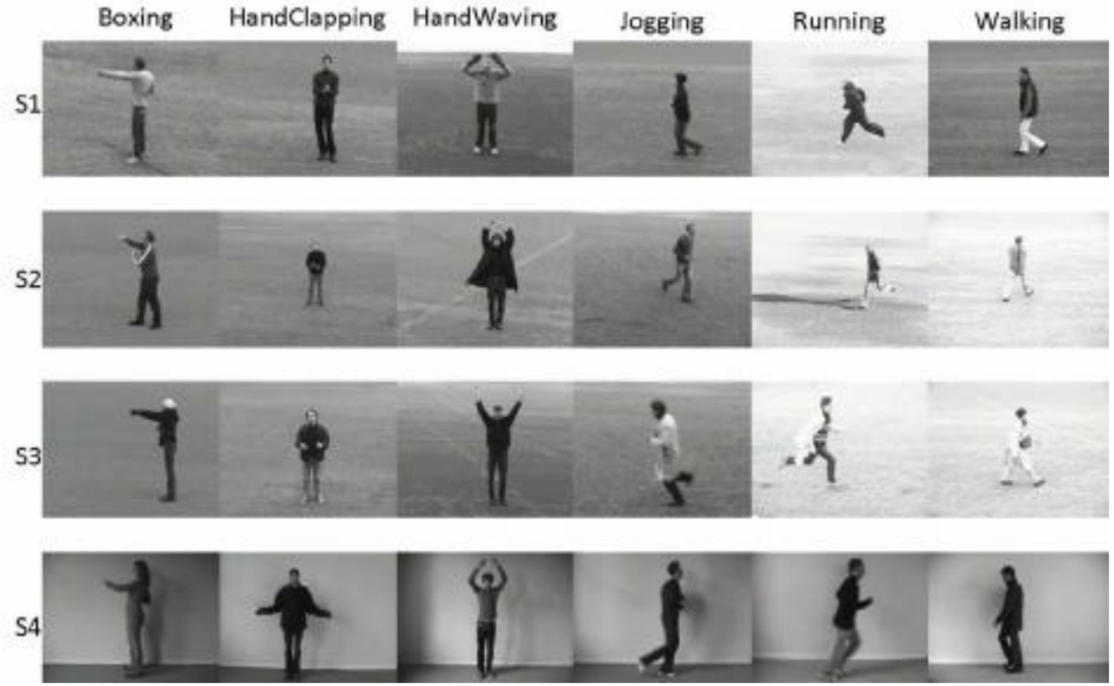




Experimental Dataset

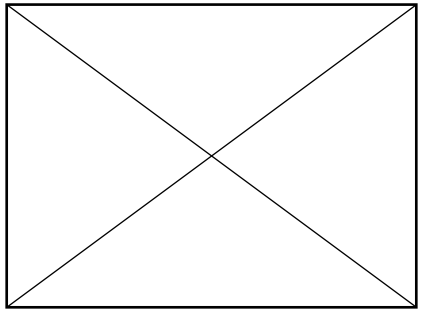


Weizmann dataset



KTH dataset

Enhancements – Class Imbalance



Frame sampled from action class – running　　*Corresponding bounding box*

- Issues:
  - In KTH, for some actions, the subject performing the action only appears for short duration.
  - Class imbalance data with actions like walking, jogging, running having relatively fewer frames.
- Solution:
  - Frames not containing the human as the object of interest are discarded.
  - Dataset is augmented by adding a moving window of size 10 to handle class imbalance.
- Gains:
  - The accuracy of action recognition improved on an average by ~3 points.

Enhancements – Origin Shift of Coordinates

- Human actions are independent of their absolute spatial positions.
- The pose coordinates are shifted w.r.t the coordinates of neck and center of body.
- The new origin is computed as:
  - $O = \frac{(P_{head}+P_{lhip}+P_{rhip})}{3}$
- The joint coordinates are shifted w.r.t to the new origin as
  - $P'_{N,x}, P'_{N,y} = (P_{N,x}, P_{N,y}) - (O_x, O_y)$
- The average recognition rates improved by an average of ~5 points.

Comparative Architectures

- 6 comparative architectures
- Architectures that operate directly on the trajectory of pose coordinates:
  - Deep Bidirectional RNN (DBRNN).
  - Deep Unidirectional LSTM (DULSTM).
  - Deep Bidirectional LSTM (DBLSTM).
- Models with hierarchical connections:
  - Point based Hierarchical BLSTM (PointHBLSTM).
  - Part based Hierarchical BLSTM (PartHBLSTM)
  - Proposed Approach

Methods	KTH	Weizman
DBRNN	82.4%	81.2%
DULSTM	89.8%	91.7%
DBLSTM	92.7%	94.8%
PointHBLSTM	94.1%	96.6%
PartHBLSTM_1	98.9%	99.9%
PartHBLSTM_2	98.4%	99.7%
Proposed Approach	**99.3%**	**100%**

Recognition rates with different experiments

Conclusion

- We proposed a technique based on part based hierarchical fusion for action recognition.
- We designed a pipeline composed of several independently trained modules.
- Further, we propose and experiment with different enhancements like class imbalance, origin shift. These techniques can be applied universally to action recognition.
- Overall, we achieve 99.3% and 100% recognition rates on the KTH and Weizmann dataset, respectively.
- Future Work: Extending this approach to more challenging dataset.

ISVC 2020

Virtual

October 5-7, 2020

Thanks for watching

Microsoft

References

1. Laptev, Lindeberg: Space-time interest points. In: Proceedings Ninth IEEE International Conference on Computer Vision. (2003) 432–439 vol.1

2. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing 28 (2010) 976 – 990

3. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In Leonardis, A., Bischof, H., Pinz, A., eds.: Computer Vision – ECCV 2006, Berlin, Heidelberg, Springer Berlin Heidelberg (2006) 428–441

4. Perˇs, J., Suli´c, V., Kristan, M., Perˇse, M., Polanec, K., Kovaˇciˇc, S.: Histograms of optical flow for efficient representation of body motion. Pattern Recognition Letters 31 (2010) 1369 – 1376

5. Wang, H., Kl¨aser, A., Schmid, C., Liu, C.: Action recognition by dense trajectories. In: CVPR 2011. (2011) 3169–3176

6. Zhou, Z., shi, F., Wu, W.: Learning spatial and temporal extents of human actions for action detection. IEEE Transactions on Multimedia 17 (2015) 1–1

7. Wang, L., Qiao, Y., Tang, X.: Action recognition with trajectory-pooled deepconvolutional descriptors. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)

8. Yemin Shi, Wei Zeng, Tiejun Huang, Yaowei Wang: Learning deep trajectory descriptor for action recognition in videos using deep neural networks. In: 2015 IEEE International Conference on Multimedia and Expo (ICME). (2015) 1–6

9. Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. 2014 IEEE Conference on Computer Vision and Pattern Recognition (2014)

10. Brau, E., Jiang, H.: 3d human pose estimation via deep learning from 2d annotations. In: 2016 Fourth International Conference on 3D Vision (3DV). (2016) 582–591

11. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation (2016)

12. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines (2016)

13. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: Realtime multiperson 2d pose estimation using part affinity fields (2018)

References

14. G¨uler, R.A., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild (2018)

15. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: CVPR 2011. (2011) 1297–1304

16. Ye, M., Zhang, Q., Wang, L., Zhu, J., Yang, R., Gall, J. In: A Survey on Human Motion Analysis from Depth Data. (2013) 149–187

17. Siddiqui, M., Medioni, G.: Human pose estimation from a single view point, realtime range sensor. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops. (2010) 1–8

18. Plagemann, C., Ganapathi, V., Koller, D., Thrun, S.: Real-time identification and localization of body parts from depth images. In: 2010 IEEE International Conference on Robotics and Automation. (2010) 3108–3113

19. Vemulapalli, R., Arrate, F., Chellappa, R.: Human action recognition by representing 3d skeletons as points in a lie group. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. (2014) 588–595

20. Gong, D., Medioni, G., Zhao, X.: Structured time series analysis for human action segmentation and recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (2014) 1414–1427

21. Zhang, Y., Zhang, Y., Zhang, Z., Bao, J., Song, Y.: Human activity recognition based on time series analysis using u-net (2018)

22. Kim, H., Kim, I.: Human activity recognition as time-series analysis. Mathematical Problems in Engineering 2015 (2015) 1–9

23. Yong Du, Wang, W., Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015) 1110–1118

24. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement (2018)

25. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A LargeScale Hierarchical Image Database. In: CVPR09. (2009)

26. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Doll´ar, P.: Microsoft coco: Common objects in context (2014)

References

27. Ramakrishna, V., Munoz, D., Hebert, M., Bagnell, J., Sheikh, Y.: Pose machines: Articulated pose estimation via inference machines. (2014) 33–47

28. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing 45 (1997) 2673–2681

29. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. Volume 3. (2004) 32–36 Vol.3

30. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as spacetime shapes. In: The Tenth IEEE International Conference on Computer Vision (ICCV'05). (2005) 1395–1402

31. Gao, Z., Chen, M.y., Hauptmann, A.G., Cai, A.: Comparing evaluation protocols on the kth dataset. In Salah, A.A., Gevers, T., Sebe, N., Vinciarelli, A., eds.: Human Behavior Understanding, Berlin, Heidelberg, Springer Berlin Heidelberg (2010) 88–100

32. Fathi, A., Mori, G.: Action recognition by learning mid-level motion features. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. (2008) 1–8

33. Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., Baskurt, A.: Sequential deep learning for human action recognition. In Salah, A.A., Lepri, B., eds.: Human Behavior Understanding, Berlin, Heidelberg, Springer Berlin Heidelberg (2011) 29–39

34. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as spacetime shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (2007) 2247–2253

35. Gilbert, A., Illingworth, J., Bowden, R.: Fast realistic multi-action recognition using mined dense spatio-temporal features. (2009) 925 – 931

36. Moussa, M.M., Hamayed, E., Fayek, M.B., El Nemr, H.A.: An enhanced method for human action recognition. Journal of Advanced Research 6 (2015) 163 – 169

37. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild (2012)

38. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: A large video database for human motion recognition. In: 2011 International Conference on Computer Vision. (2011) 2556–2563