3D ACTIVITY RECOGNITION USING MOTION HISTORY AND BINARY SHAPE TEMPLATES

Saumya Jetley, Fabio Cuzzolin

Artificial Intelligence and Vision group

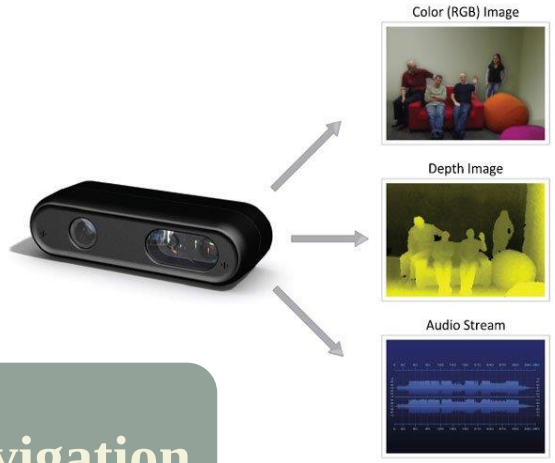
Department of Computing and Communication Technologies

Oxford Brookes University (UK)

http://cms.brookes.ac.uk/staff/FabioCuzzolin/

- **unconstrained human activity recognition** is crucial to manifold real-world applications

Video Surveillance



Video Indexing/ Retrieval

Robot Navigation

Surgical Assistance

Gaming Consoles

- **Affordable depth camera technology** has provided alternative ways of approaching the problem
- Techniques using depth images can broadly be categorized into **skeleton-based** and **non-skeleton** based

Skeleton Tracking based Techniques

- [Wang et.al. 2012[1], Yun et.al. 2013, Shuzi et.al. 2013, Yu et.al. 2011]
- Received much impetus after Shotton et.al. 2011 work on quick and accurate 3D joint estimation from a single depth image
- **Provide good accuracies**
- **Highly prone to noise and occlusions**

Non-Skeleton Tracking based Techniques

- [Li et.al. 2010, Vieira et.al. 2012, Wang et.al. 2012[2], Yang et.al. 2012[2], **Our Present Work]**
- **Comparatively lower accuracies**
- But **robust to noise and occlusions (and thus much better suited to an unconstrained setting)**

OUR APPROACH

- Is **simple**, **intuitive** and **efficient**
- Shows that an **elegant and smart combination of 2 primary features – Motion and Change in Shape** can **yet yield very competitive results**
- Is **better than/ or at par** with top competitors on **3-out-of-4 datasets**

- **Highlights** of our approach:

Preserves motion sequence information

by using **Motion History Templates**

Performs efficient temporal analysis

by using small **overlapping frame-sets** to ensure accurate motion information and maintain continuity

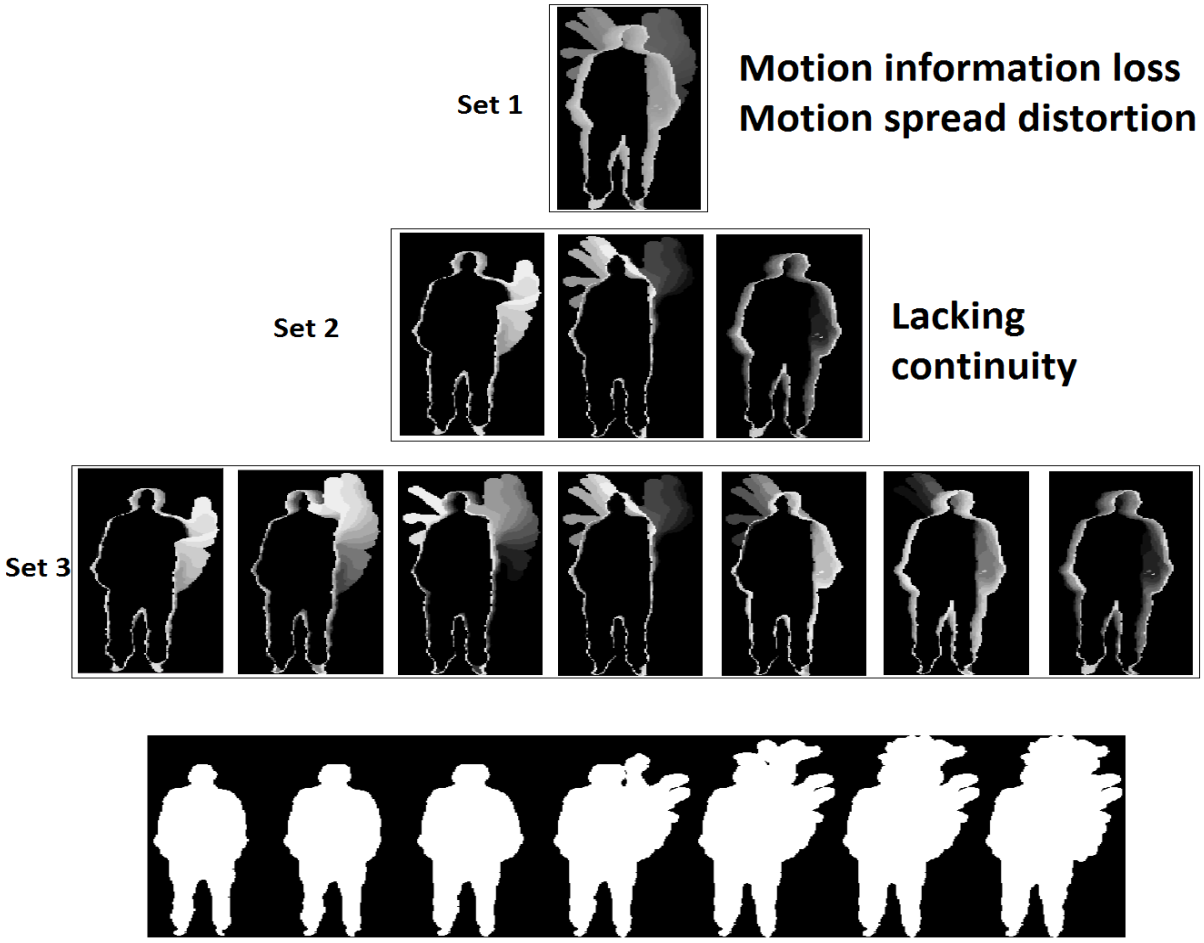
Captures shape information

by using **Binary Shape Templates** which track boundary growth across frames

Performs the above analysis on all 3 orthogonal views

thus leveraging three-dimensional shape and motion information

- **Motion and Shape Information combined in a novel Temporal Analysis**

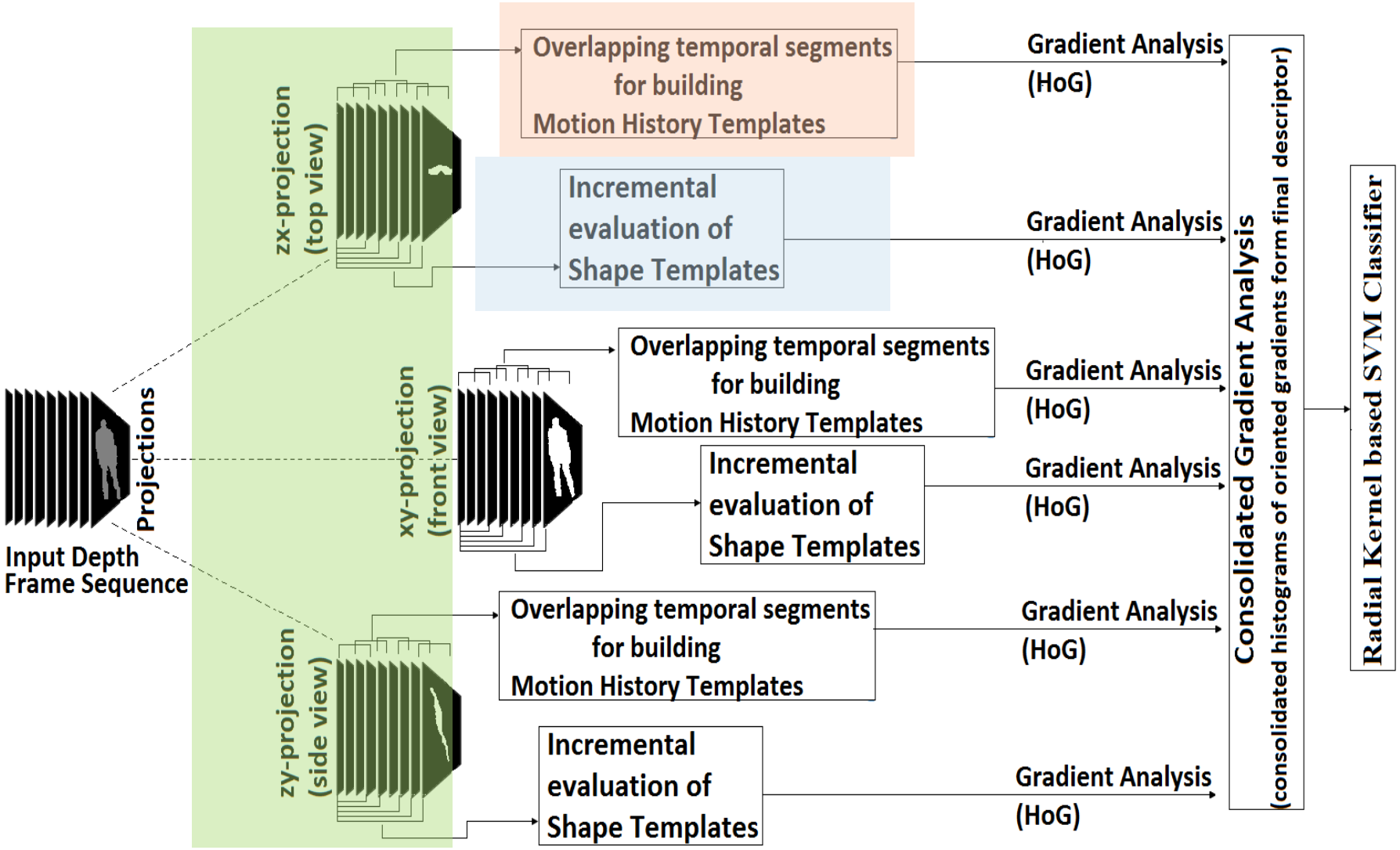


Motion history templates for:

- Set 1: the complete sequence;
- Set2: non-overlapping temporal blocks
- **Set 3: overlapping temporal blocks;**

Binary Shape Templates for a 'draw tick' action focused in the top-right region

- **Block Diagram**



- **Comparison with some related approaches**

[Yang et.al. 2012[2]]

Motion **Energy** Templates are used (for 3 orthogonal views)

Hence, temporal motion information is lost

[Megavannan et. al. 2012]

Frontal MHIs leave out 3D shape information

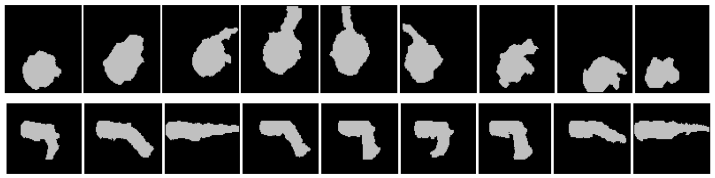
Average Depth Image does not consider motion and hence is susceptible to background clutter

[Weinland et. al. 2006]

Fourier Analysis along concentric circles of Motion History Volume is not efficiently discriminative causing **confusion between well distinct actions such as sit-down & pick-up, turn-around & majority-of-other-actions, walk & pick-up, walk & kick**.

Also, the used **1D-FT overlooks motion symmetry around z axis** and may fail to distinguish between single & two arm waves, forward & side punch, and forward & side kick.

- We have experimented with **4 standard datasets**

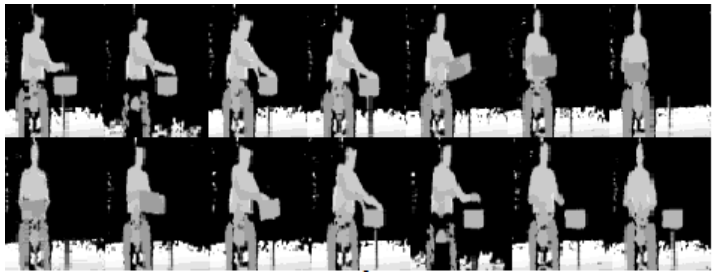


MSR 3D Gesture Dataset

12 dynamic American Sign Language gestures

MSR 3D Action Dataset

20 full body action categories



3D Action Pair Database

6 Action pairs- similar motion and shape but their co-occurrence is in different spatiotemporal order

UT Kinect Action Database

10 Action types – ridden with frequent occlusions and shifting background clutter

- Our approach overcomes such issues, yielding **high recognition rates**

3D Gesture Recognition

Approach	Accuracy
Proposed (MHI + BST based Gradient Analysis)	96.6%
Oreifej & Liu 2013	92.45%
Wang et al. 2012[2]	88.5%
Kurakin et al. 2012	87.77%

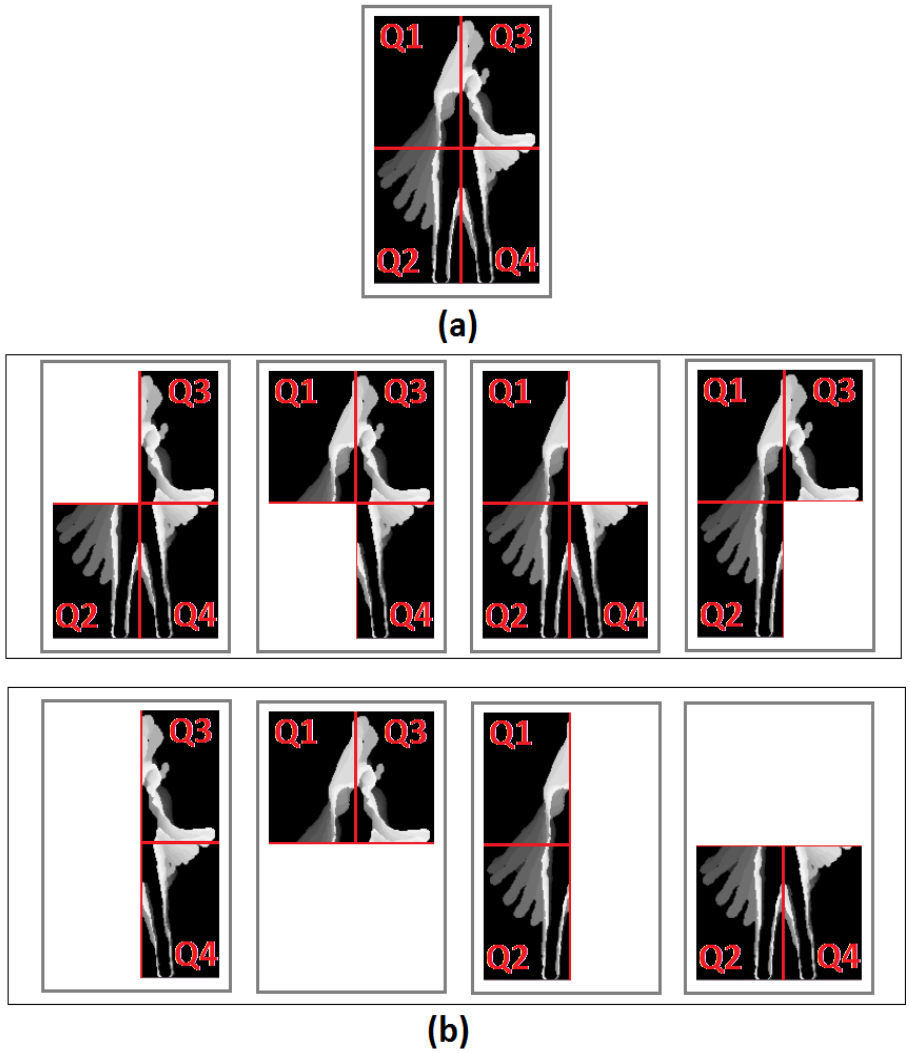
3D Action-Pair Recognition

Approach	Accuracy
Proposed (MHI + BST based Gradient Analysis)	97.22%
Oreifej & Liu 2013	96.67%
Wang et al. 2012[1]	82.22%
Yang et al. 2012	66.11%

3D Action Recognition

Approach	Accuracy
Proposed (MHI + BST based Gradient Analysis)	83.8%
Oreifej & Liu 2013	88.89%
Wang et al. 2012[1]	88.2%
Wang et al. 2012[2]	86.5%
Yang et al. 2012[2] /[1]	<84.1/ 85.52%
Vieira et al. 2012	<81.30%
Li et al. 2010	<71.90%

- Has **robust performance under occlusion** as well



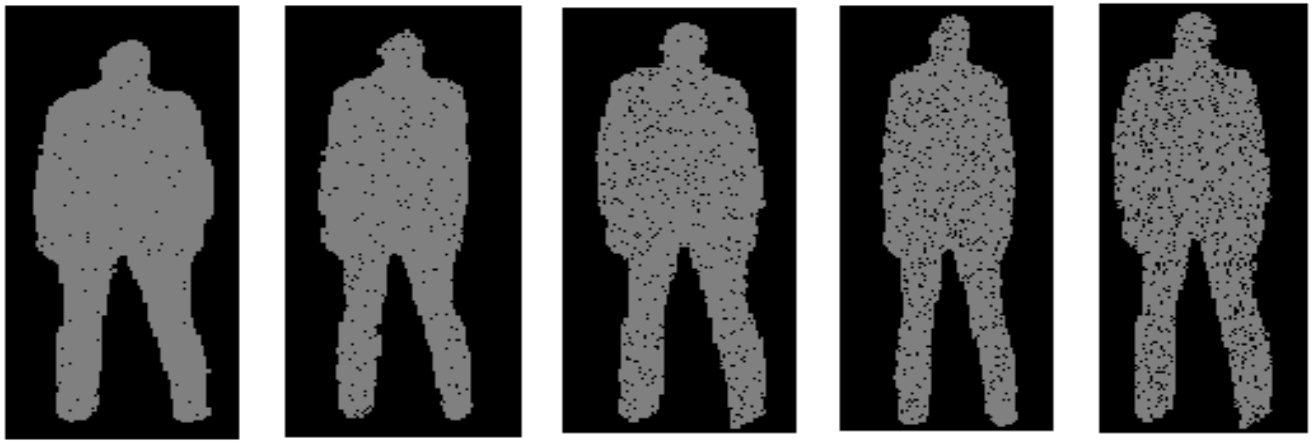
3D Action Recognition

Occluded Quadrant(s)	Relative Accuracy(%)
none	100
Q1	91.57
Q2	95.99
Q3	64.66
Q4	96.39
Q1+Q3	43.38
Q3+Q4	62.25
Q4+Q2	8.95
Q2+Q1	65.87

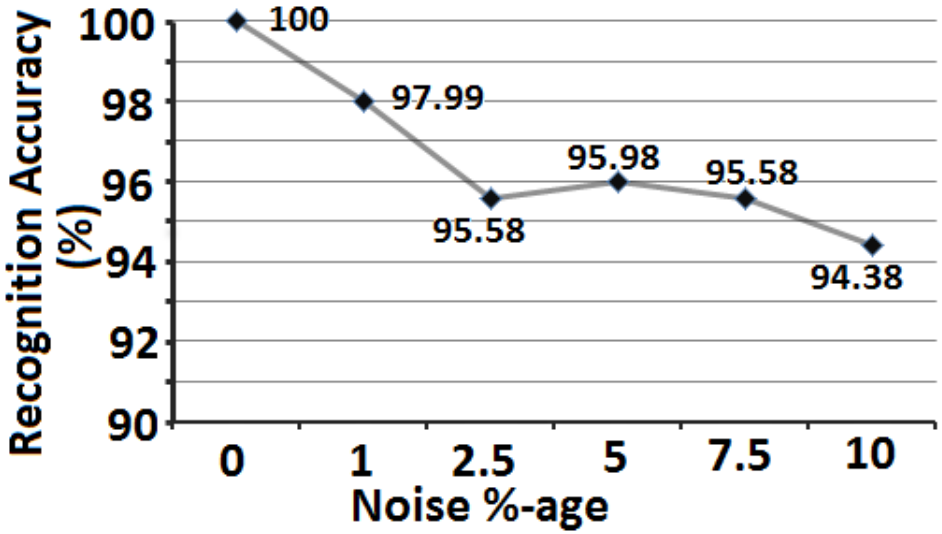
Reason of robust **performance under occlusion**

- **We perform an independent region-wise analysis**
- **Even if a particular body part is occluded, the visible regions are unaffected and can be accurately analysed**
- **On the other hand, methods using combined interplay of regional information (like relative 3D joint position variations) are highly likely to underperform**

- Performance under noise **(Discontinuities in Depth – Pepper Noise)**



3D Action Recognition



Reason of robust **performance under noise**

- **In each projection plane, we are interested in capturing the general motion spread (MHTs) and closed object boundary variations (MSTs)**
- **Thus, pixel-level discontinuities can be conveniently removed by basic morphological operations w/o affecting the overall features that our approach extracts!**

Questions ?

References

- Oreifej, O., Liu, Z.: Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. (2013)
- Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. (2012)

- 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: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. (2011)
- Yun, H., Sheng-Luen, C., Jeng-Sheng, Y., Qi-Jun, C.: Real-time skeleton-based indoor activity recognition. In: Control Conference (CCC), 2013 32nd Chinese (2013)
- Shuzi, H., Jing, Y., Huan, C.: Human actions segmentation and matching based on 3d skeleton model. In: Control Conference (CCC), 2013 32nd Chinese. (2013)
- Yu, X., Wu, L., Liu, Q., Zhou, H.: Children tantrum behaviour analysis based on kinect sensor. In: Intelligent Visual Surveillance (IVS), 2011 Third Chinese Conference on. (2011)

References

- Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3d points. In: Computer Vision and Pattern RecognitionWorkshops (CVPRW), 2010 IEEE Computer Society Conference on. (2010)
- Vieira, A., Nascimento, E., Oliveira, G., Liu, Z., Campos, M.: Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences. In Alvarez, L., Mejail, M., Gomez, L., Jacobo, J., eds.: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. (2012)

- Wang, J., Liu, Z., Chorowski, J., Chen, Z., Wu, Y.: Robust 3d action recognition with random occupancy patterns. In: Proceedings of the 12th European Conference on Computer Vision - Volume Part II. ECCV'12, Berlin, Heidelberg, Springer-Verlag (2012)
- Yang, X., Zhang, C., Tian, Y.: Recognizing actions using depth motion maps-based histograms of oriented gradients. In: Proceedings of the 20th ACM International Conference on Multimedia. MM '12, New York, NY, USA, ACM (2012)
- Kurakin, A., Zhang, Z., Liu, Z.: A real time system for dynamic hand gesture recognition with a depth sensor. In: Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European. (2012) 1975