MULTI-VIEW HUMAN ACTIVITY RECOGNITION USING MOTION FREQUENCY

Neslihan Köse, Mohammadreza Babaee, Gerhard Rigoll

Institute for Human-Machine Communication, TU Munich, Germany

neslihan.koese@tum.de, reza.babaee@tum.de, rigoll@tum.de

INTRODUCTION

The problem of human activity recognition can be approached using spatio-temporal variations in successive video frames. In this paper, a new human action recognition technique is proposed using multi-view videos. Initially, a naive background subtraction using frame differencing between adjacent frames of a video is performed. Then, the motion information of each pixel is recorded in binary indicating existence/nonexistence of motion in the frame. A pixel wise sum over all the difference images in a view gives the frequency of motion in each pixel throughout the clip. The classification performances are evaluated using these motion frequency features. Our analysis shows that increasing number of views used for feature extraction improves the performance as different views of an activity provide complementary information. Experiments on the i3DPost and the INRIA Xmas Motion Acquisition Sequences (IXMAS) multi-view human action datasets provide significant classification accuracies.

PROPOSED ACTIVITY RECOGNITION TECHNIQUE

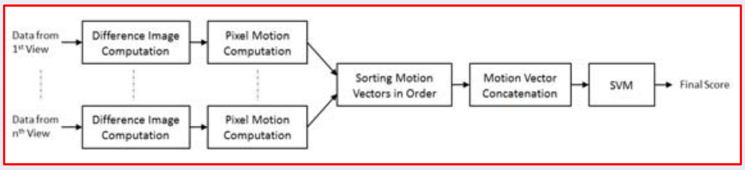
Data from 1st View
Data from nth View
Difference Image Computation
Difference Image Computation
Pixel Motion Computation
Pixel Motion Computation
Sorting Motion Vectors in Order
Motion Vector Concatenation
SVM
Final Score

Fig. Flowchart of the proposed human activity recognition technique.

Difference Image Computation:

$$D_t(i,j) = |I_t(i,j) - I_{t+1}(i,j)|, 1 \le i \le w, 1 \le j \le h$$

Thresholding:

$$T_k(i,j) = \begin{cases} 1 & \text{for}\ , D_k(i,j) > t \\ 0 & \text{for}\ , otherwise \end{cases}$$

Pixel Motion Computation:

$$PMI(i,j) = \sum_{k=1}^{K-1} T_k(i,j)$$

- Difference image of an activity with higher speed contains more white pixels.

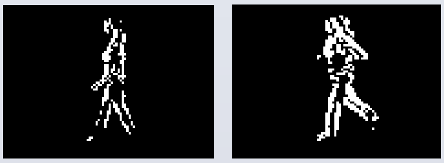

Fig. Example difference images (Tk) for walking (1st picture) and running (2nd picture) activities from the i3DPost Multi-View Human Action Dataset.

- Since subjects may enter the scene from different points, we have sorted the data in same order to have a placement as if all the subjects enter the scene from the same or nearly the same points before feature extraction.
- This is needed in case there is a significant variation in frames captured for subjects entering from different directions.
- We construct a matrix of pixel motion barcodes introduced in [11] from all difference images of a view. Taking the sum of each pixel barcode provides number of times significant motion is observed in that particular pixel.
- This process is repeated for all pixels and data is vectorized to obtain a vector form where each value represents number of times a motion existed in a pixel.
- We call this value as pixel motion frequency and all the vectors of different views combined together give us pixel motion frequency vector.
- This vector can be used for action classification using linear multi-class Support Vector Machines(SVMs) classifier.

EXPERIMENTAL RESULTS

- **Experiment 1: Tests using the i3DPost Dataset**

i3DPost dataset [12] consists of 8 actors performing 10 different actions, where 6 are single actions (walk, run, jump, bend, hand-wave and jump-in-place) and 4 are combined actions (sit-stand-up, run-fall, walk-sit and run-jump-walk). The database was recorded from 8 camera-views.

- For an exact comparison with the techniques in Table 1, we evaluate the performance of our approach for similar scenarios and obtained accuracies of 94.79%, 95%, 96.87% and 95.5% for the classification of 6 single actions, 5 single actions, 4 combined actions and all 10 actions, respectively.
- These results show that our approach outperforms the results of existing techniques for 4 combined actions and all 10 actions. For the 6 single actions, our approach gives slightly less accuracy (94.79%) with the best result being 98.2%. For the 5 single actions, our approach gives 95% classification accuracy which is better than [2] and slightly behind remaining methods. For the 4 combined actions and all 10 actions, we obtain the best performances with 96.87% and 95.5%, respectively.

Table 1. Comparison of Different Methods All Using the I3DPost Human Action Dataset with Leave-One-Out Partitioning

Method (%)	6 single actions	5 single actions	4 combined actions	10 actions
[1]	89.6	97.5	87.5	80
[2]	NR	90	NR	NR
[3]	95.3	97.8	NR	NR
[4]	98.2	97.8	NR	NR
Proposed Method	**94.79**	**95**	**96.87**	**95.5**

- **Experiment 2: Tests using the IXMAS Dataset**

The IXMAS dataset [13] has 12 subjects performing 13 daily-life actions 3 times each: check watch, cross arms, scratch head, sit down, get up, turn around, walk, wave, punch, kick, point, pick up and throw. The database was recorded from 5 camera-views.

- Table 2 shows that our technique obtains much better accuracy on the IXMAS dataset compared to other existing techniques with the advantage of being very simple.

Table 2. Comparison of Different Methods All Using the IXMAS Human Action Dataset

Method (%)	Actions	Actors	Train-Test Partioning	Accuracy
[5]	12	12	5-fold	80.5
[6]	11	10	Leave-one-out	76.5
[7]	11	12	Leave-one-out	85.9
[8]	11	10	Leave-one-out	83.5
[9]	11	10	Leave-one-out	81.4
[10]	13	12	Leave-one-out	78
Proposed Method	**11**	**10**	**Leave-one-out**	**94.07**

CONCLUSIONS

We proposed a multi-view activity recognition approach which applies the pixel based motion information for activity recognition. In this work, we only used multiple 2D views for feature extraction and not the 3D information provided by the datasets. Our pixel motion frequency vector provides information not only about the motion of an action but also detailed texture differences due to the involvement of frame differencing inside the proposed approach.

Experiments are conducted on two well-known Multiview action datasets. The classification rates are evaluated as 95.5% for all the 10 actions in the i3DPost dataset and 94.07% for the 11 actions in the IXMAS dataset. These results prove that our approach provides significant classification accuracies.

In future work, it would be interesting to see the performance of this approach for more complex datasets. We intend to improve activity recognition performance of this method by further incorporating temporal information.

ACKNOWLEDGEMENTS

This work has been performed by the project “Augmented Synopsis of Surveillance Videos in Adaptive Camera Networks” under Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) research grants programme. The authors would like to thank the DFG for the financial support.

REFERENCES

[1] M. B. Holte, T. B. Moeslund, et al., “3d human action recognition for multi-view camera systems,” Int. Conf. on 3D Imaging, Modeling, Processing, Visualization and Transmission, May 2011, pp. 342–349.

[2] N. Gkalelis, N. Nikolaidis, et al., “View independent human movement recognition from multi-view video exploiting a circular invariant posture representation,” IEEE Int. Conf. on Multimedia and Expo, 2009, pp. 394–397.

[3] A. Iosifidis, A. Tefas, et al., “View-invariant action recognition based on artificial neural networks,” IEEE Transactions on Neural Networks and Learning Systems, vol. 23, no. 3, pp. 412–424, 2012.

[4] A. Iosifidis, A. Tefas, and I. Pitas, “Multi-view action recognition based on action volumes, fuzzy distances and cluster discriminant analysis,” Signal Processing, vol. 93, no. 6, pp. 1445–1457, 2013.

[5] F. Baumann, J. Lao, et al., “Motion binary patterns for action recognition.,” ICPRAM, 2014, pp. 385–392.

[6] Z. Wang, J. Wang, J. Xiao, K. H. Lin, and T. Huang, “Substructure and boundary modeling for continuous action recognition,” IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 1330–1337.

[7] A. A. Chaaraoui, P. C. P´erez, and F. Fl´orez-Revuelta, “Silhouette-based human action recognition using sequences of key poses,” Pattern Recognition Letters, vol. 34, no. 15, pp. 1799–1807, 2013.

[8] DanielWeinl, Mustafa zuysal, et al., “Making action recognition robust to occlusions and viewpoint changes,” 2010.

[9] G. Srivastava, H. Iwaki, et al., “Distributed and lightweight multi-camera human activity classification,” Third ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC), 2009, pp. 1–8.

[10] P. Yan, S. M. Khan, and M. Shah, “Learning 4d action feature models for arbitrary view action recognition,” in IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–7.

[11] G. Ben-Artzi, M. Werman, and S. Peleg, “Event retrieval using motion barcodes,” IEEE Int. Conf. on Image Processing (ICIP), 2015, pp. 2621-2625.

[12] N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis, and I. Pitas, “The i3dpost multi-view and 3d human action/interaction database,” Conference for Visual Media Production, 2009, pp. 159–168.

[13] D.Weinland, R. Ronfard, and E. Boyer, “Free viewpoint action recognition using motion history volumes,” Computer Vision and Image Understanding, vol. 104, no. 23, pp. 249 – 257, 2006.

IEEE International Conference on Image Processing, ICIP, September 17-20 2017, Beijing, China