MULTI-VIEW HUMAN ACTIVITY RECOGNITION USING MOTION FREQUENCY

Neslihan Köse[⋆] *Mohammadreza Babaee*[⋆] *Gerhard Rigoll*[⋆]

[⋆] Institute for Human-Machine Communication, TU Munich, Germany

ABSTRACT

The problem of human activity recognition can be approached using spatio-temporal variations in successive video frames. In this paper, a new human activity 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/non-existence 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.

Index Terms— activity recognition, frame differencing, motion frequency

1. INTRODUCTION

Human activity recognition is important in fields such as video surveillance, video retrieval and human-computer interaction. This study aims to classify human actions from multi-view video sequences.

There are several approaches in human activity recognition, which use databases collected with a single camera. Since traditional cameras capture the 2D projection of a scene, the analysis of actions in the image plane is in fact a projection of the real actions. Thus, the projection of the actions depend on the viewpoint. In order to obtain full information about the actions, there are two common approaches. The first one is the use of 3D representations of reconstructed 3D data with the utilization of multi-cameras [1]. The second approach is based on feature extraction from 2D image views of a multi-view camera system. The proposed method in this paper is based on this second approach. Using multi-view camera system helps to accumulate complementary information from different views.

In the proposed approach, the difference images between adjacent frames of a video are used to find motion barcodes as in [2] for each pixel. Then for each view, a motion frequency of each pixel is determined by cumulative sum over all difference images of one view. For each camera view, we obtain one pixel motion frequency vector. Motion frequency vector of all camera views are concatenated together to get discriminative features. Pixel motion frequency vector provides information about frequency of presence/absence of motion in a particular pixel.

Our method is evaluated using two well-known publicly available i3DPost Multi-View Human Action Dataset [3] and INRIA Xmas Motion Acquisition Sequences (IXMAS) Multi-View Human Action Dataset [1]. We compare our results with other techniques applied on the same databases and observe that our approach is robust and provides better or comparable results to most other techniques.

The rest of this paper is organized as follows. Section 2 discusses the related work in activity recognition. Section 3 explains our proposed activity recognition technique. Section 4 presents the experimental results and analysis. Finally, Section 5 concludes the paper.

2. RELATED WORK

There are several activity recognition studies which use 2D multi-view image data.

In [4], the authors present a framework for learning a compact representation of primitive actions for simultaneous activity recognition and viewpoint estimation. In [5], the authors use the circular shift invariance property of the discrete Fourier transform (DFT) magnitudes to solve the view correspondence problem between train and test samples, and then use fuzzy vector quantization (FVQ) and linear discriminant analysis (LDA) for action recognition from multi-view silhouettes. In [6], binary body masks from frames of a multi-camera system are concatenated to produce the multi-view binary mask. These masks are rescaled and vectorized to create feature vectors. Next, FVQ is applied to represent movements and LDA is applied to map movements in a low dimensionality discriminant feature space. In [7], the authors analyze the motion from two cameras and infer rotation centers that best explain the observed motion. Then, they group pixel-level flow into dominant pointing vectors that each originate from a rotation center and merge across views to obtain 3D pointing vectors. The approach in [8] is based on data fusion of 2

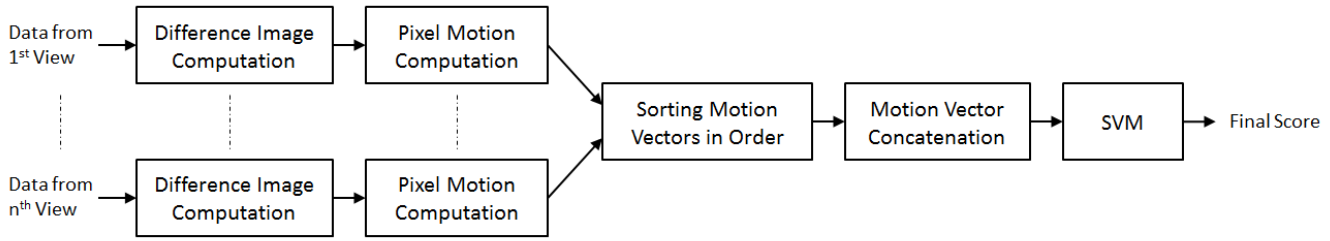

Fig. 1. Flowchart of the proposed human activity recognition technique.

orthogonal views. A low dimensional feature vector is used which consists of the projections of the width profile of the actor on to an action basis, which is built using eigen analysis of walking sequences of different people, and simple spatio-temporal features. In this approach, dynamic time warping (DTW) is applied for recognition in the final step. There are also some other techniques like metric learning [13] or action representation by feature-trees [14] which use 2D image data acquired by multiple cameras.

Furthermore, cross view activity recognition, which is based on recognizing actions by training on one view and testing on another different view, has been investigated in several studies such as [9], [10], [11] and [12].

3. PROPOSED ACTIVITY RECOGNITION TECHNIQUE

In this section, the proposed activity recognition approach is explained. Fig. 1 shows the flowchart of our approach.

In the first step, video frame is vectorized and then frame differencing is applied by computing the differences between successive frames in time t and $t+1$. The difference image at time t is given by:

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

$I_t(i,j)$ is the intensity of the pixel (i,j) in the t^{th} frame, w and h are the width and height of the image respectively. The absolute value of the difference image is compared with a predetermined threshold value. Motion information (T_k) of the difference image is calculated using:

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

where t is the threshold ($t = 30$ used in this paper). In this study, we tested the performances with several threshold values and observed the best performance with this threshold value. Note that this threshold value should be adapted according to the database used for the evaluations.

For an N-view camera system, there are N number of video sequences. For each video sequence, pixel wise sum of all difference images is computed to get cumulative pixel motion. If K is the total number of frames in one video sequence, then pixel motion feature vector is calculated as:

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

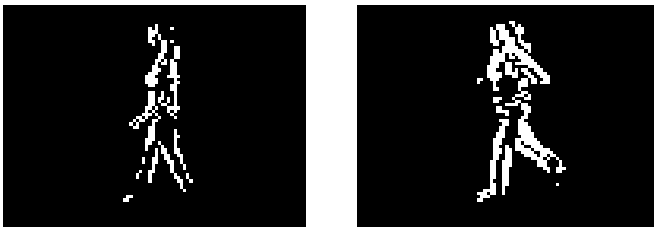

Fig. 2. Example difference images (T_k) for walking (1^{st} picture) and running (2^{nd} picture) activities from the i3DPost Multi-View Human Action Dataset.

Number of white pixels in a difference image give motion information. Difference image of an activity with higher speed contains more white pixels. For example, there are more white pixels in the difference image of running compared to the difference image of walking action as in Fig. 2. For any pixel, motion frequency will be less in running action than in walking action as pixel motion will be short lived in running due to high speed. The main advantage of using difference images to find pixel motion frequency in activity recognition is that it can be applied without requiring a static background.

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 [2] 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

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
[15]	89.6	97.5	87.5	80
[5]	NR	90	NR	NR
[16]	95.3	97.8	NR	NR
[17]	98.2	97.8	NR	NR
Proposed Method	**94.79**	**95**	**96.87**	**95.5**

Note: Here NR indicates result for this category is not reported.

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 activity classification using Support Vector Machines(SVMs).

In our approach, we combined all pixel motion frequency vectors to obtain feature vectors. For instance, using the i3DPost dataset, we first resized our images to 108x192 resolution hence obtained 20736 total number of pixels in our difference images. So we have a pixel motion frequency vector containing 20736 values per camera view (For 8 cameras, we have 165888 values per action). These feature vectors are then used for classification using linear multi-class SVM classifier. We used LIBSVM [18] library which applies the one-against-one approach for multi-class classification. If K is the number of classes, then $K(K-1)/2$ classifiers are constructed and each one trains data from two classes. In classification, a voting strategy is used.

4. EXPERIMENTAL RESULTS

The performance of our approach is tested on the publicly available i3DPost [3] and the IXMAS [1] multi-view human action datasets. The train-test set partitioning is important when the performances of different algorithms are tested on the same database. We report our results using leave-one-out technique.

4.1. Experiment 1: Tests Using the i3DPost Dataset

The i3DPost dataset[1] is a high-quality dataset, which was generated within the "Intelligent 3D Content Extraction and Manipulation for Film and Games" EU-funded research project. It consists of 8 actors performing 10 different actions,

[1]http://kahlan.eps.surrey.ac.uk/i3dpost_action, [3]

Fig. 3. The columns correspond to the 10 different actions performed by the 8 actors in the i3DPost Dataset, where the first 6 columns show the single actions and the last 4 columns show the combined actions. The first 8 rows show images from the 8 camera views. The 9^{th} row shows the corresponding 3D mesh models [19].

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). Fig. 3 shows the 10 different actions performed by the 8 actors, where the first 6 columns show the single actions and the last 4 columns show the combined actions.

Table 1 shows the comparison results with several existing techniques which use the same database with the same train-test set partitioning. In [15], for the classification of 6 single actions, 5 single actions, 4 combined actions and finally all 10 actions, the accuracies are reported 89.6%, 97.5%, 87.5% and 80%, respectively. In [15], the authors observe that in the classification of single actions: walk and run, and in the classification of combined actions: run-jump-walk and walk-sit actions are confused. Gkalelis et al. [5] reported an accuracy of 90.00% for the classification of 5 single actions. In [16], Iosifidis et al. reported an accuracy of 95.3% and 97.8% for the 6 and 5 single actions sets, respectively. In their recent study [17], the accuracies are reported 98.2% and 97.8% for the classification of the 6 and 5 single actions, respectively. For an exact comparison with these techniques, 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

Fig. 4. The view points from cameras 1-5 of the IXMAS dataset

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 [5] 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. In [20], the authors achieved 100% accuracy for the classification of 6 actions. However, an exact comparison with [20] is not possible since stand and sit actions are used instead of hand wave and jump in place actions in their classification. It should also be noted that our approach performs significantly better with the classification of combined actions than the classification of single actions.

4.2. Experiment 2: Tests Using the IXMAS Dataset

The IXMAS dataset (http://4drepository.inrialpes.fr/public/viewgroup/6, [1]) 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 dataset has been recorded by 5 calibrated and synchronized cameras resulting in a total of 2340 action instances. Fig. 4 shows an example for sitting action which shows the view points from cameras 1-5 of the IXMAS dataset. This dataset is a challenging one due to the fact that the subjects freely choose their position and orientation. Therefore, each camera has captured different viewing angles, which makes the recognition task harder. To be consistent with the experiments in [21], [22], and [23], we use the data of 10 subjects and 11 action categories excluding point and throw actions. Table 2 shows the comparison results of techniques which use IXMAS dataset. Number of actions and actors involved in the experiment is important for performance comparison. Our approach provides 94.07% accuracy for 11 class classification with 10 actors. 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. All methods in Table 2 except [21] use only 2D frames of video sequences. In [21], the authors apply local partitioning and hierarchical classification of 3D histogram of oriented gradients volumes using the 3D information provided in the dataset and still report the accuracy of 83.5%.

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

Method(%)	Actions	Actors	Train-Test Partitioning	Accuracy
[24]	12	12	5-fold	80.5
[23]	11	10	leave-one-out	76.5
[25]	11	12	leave-one-out	85.9
[21]	11	10	leave-one-out	83.5
[22]	11	10	leave-one-out	81.4
[26]	13	12	leave-one-out	78
Proposed Method	11	10	leave-one-out	**94.07**

5. CONCLUSION

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 multi-view 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.

6. REFERENCES

[1] 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.

[2] G. Ben-Artzi, M. Werman, and S. Peleg, "Event retrieval using motion barcodes," in *2015 IEEE International Conference on Image Processing (ICIP)*, Sept 2015, pp. 2621–2625.

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

[4] R. Souvenir and J. Babbs, "Learning the viewpoint manifold for action recognition," in *2008 IEEE Confer-*

ence on Computer Vision and Pattern Recognition, June 2008, pp. 1–7.

[5] N. Gkalelis, N. Nikolaidis, and I. Pitas, "View indepedent human movement recognition from multi-view video exploiting a circular invariant posture representation," in *2009 IEEE International Conference on Multimedia and Expo*, June 2009, pp. 394–397.

[6] A. Iosifidis, N. Nikolaidis, and I. Pitas, "Movement recognition exploiting multi-view information," in *2010 IEEE International Workshop on Multimedia Signal Processing*, Oct 2010, pp. 427–431.

[7] P. Matikainen, P. Pillai, L. Mummert, R. Sukthankar, and M. Hebert, "Prop-free pointing detection in dynamic cluttered environments," in *Face and Gesture 2011*, March 2011, pp. 374–381.

[8] S. Cherla, K. Kulkarni, A. Kale, and V. Ramasubramanian, "Towards fast, view-invariant human action recognition," in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, June 2008, pp. 1–8.

[9] Ali Farhadi and M. K. Tabrizi, *Learning to Recognize Activities from the Wrong View Point*, pp. 154–166, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[10] I. N Junejo, E. Dexter, I. Laptev, and P. Pérez, "Cross-view action recognition from temporal self-similarities," in *European Conference on Computer Vision*. Springer, 2008, pp. 293–306.

[11] J. Liu, S. Ali, and M. Shah, "Recognizing human actions using multiple features," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, June 2008, pp. 1–8.

[12] J. Liu and M. Shah, "Learning human actions via information maximization," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, June 2008, pp. 1–8.

[13] D. Tran and A. Sorokin, *Human Activity Recognition with Metric Learning*, pp. 548–561, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[14] K. K. Reddy, J. Liu, and M. Shah, "Incremental action recognition using feature-tree," in *2009 IEEE 12th International Conference on Computer Vision*, Sept 2009, pp. 1010–1017.

[15] M. B. Holte, T. B. Moeslund, N. Nikolaidis, and I. Pitas, "3d human action recognition for multi-view camera systems," in *2011 International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission*, May 2011, pp. 342–349.

[16] A. Iosifidis, A. Tefas, and I. Pitas, "View-invariant action recognition based on artificial neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 3, pp. 412–424, March 2012.

[17] 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.

[18] C. Chung Chang and Chih-Jen Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[19] M. B. Holte, C. Tran, M. M. Trivedi, and T. B. Moeslund, "Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, pp. 538–552, 2012.

[20] A. K. Singh Kushwaha and R. Srivastava, "Multiview human activity recognition system based on spatiotemporal template for video surveillance system," *Journal of Electronic Imaging*, vol. 24, no. 5, pp. 05–1004, 2015.

[21] Daniel Weinl, Mustafa zuysal, and Pascal Fua, "Making action recognition robust to occlusions and viewpoint changes," 2010.

[22] G. Srivastava, H. Iwaki, J. Park, and A. C. Kak, "Distributed and lightweight multi-camera human activity classification," in *2009 Third ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)*, Aug 2009, pp. 1–8.

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

[24] F. Baumann, J. Lao, A. Ehlers, and B. Rosenhahn, "Motion binary patterns for action recognition.," in *ICPRAM*, 2014, pp. 385–392.

[25] A. A. Chaaraoui, P. C. Pérez, and F. Flórez-Revuelta, "Silhouette-based human action recognition using sequences of key poses," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1799–1807, 2013.

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