CONTINUOUS DETECTION AND RECOGNITION OF ACTIONS OF INTEREST AMONG ACTIONS OF NON-INTEREST USING A DEPTH CAMERA

Neha Dawar, Nasser Kehtarnavaz

Department of Electrical and Computer Engineering, University of Texas at Dallas, USA

ABSTRACT

This paper presents a human action recognition approach using a depth camera for situations when actions of interest are performed in a continuous and random manner among actions of non-interest. The developed approach first performs detection of actions of interest by separating actions of interest from actions of non-interest in an on-the-fly manner and then classifies the detected actions of interest. Skeleton joint positions from depth images are used to achieve the detection of actions of interest. Recognition of detected actions of interest is then achieved by fusing the outcome of two classifiers, one classifier using skeleton joint positions and the other classifier using depth images. A continuous dataset consisting of actions of interest associated with the smart TV application is collected and made publicly available. The results obtained by applying the developed approach to this dataset indicate its effectiveness in detecting and recognizing actions of interest from continuous data streams.

Index Terms— Continuous action detection, continuous action recognition, continuous detection and recognition of actions of interest among actions of non-interest

1. INTRODUCTION

Human action recognition is an extensively researched topic in computer vision which has been utilized in many human-computer interaction applications. The literature includes a wide collection of papers involving the use of the Kinect depth camera for human action recognition, e.g. [1-5]. In these works, various features such as action graphs, random occupancy patterns (ROP), space-time occupancy patterns (STOP), depth motion maps (DMM), histogram of oriented gradients (HOG) have been extracted from depth images to achieve action recognition. 3D skeleton joint positions from depth images have also been used for action recognition, e.g. [6-7]. These joint positions are made available via the Microsoft Software Development Kit v2 [8]. Kinect v2 is capable of providing the 3D spatial locations of 25 skeleton body joints. In [9], both depth images and skeleton joint positions were used simultaneously for action recognition.

It is important to note that the considerable amount of research that has been conducted on human action or gesture recognition has focused primarily on recognizing actions which appear as single or isolated actions. It still remains a challenge to deal with continuous streams of activities composed of both actions of interest and actions of non-interest that appear in a random order. Continuous streams of activities constitute a more realistic scenario in many human-computer interaction applications such as smart TV and gaming.

A continuous action recognition approach using a depth camera was covered in [10]. However, the dataset examined only contained actions of interest. This paper deals with a more challenging situation where both actions of interest and actions of non-interest occur continuously and in a random order. As a result, both the problems of action detection and action recognition are addressed at the same time to allow recognizing actions of interest among actions of non-interest in an on-the-fly manner. The developed approach first detects the presence of actions of interest from continuous data streams and then classifies them. The major contribution of this paper is the development of a human action recognition approach which is capable of dealing with recognizing actions of interest among actions of non-interest in continuous data streams by simultaneously using depth and skeleton information captured by a Kinect depth camera.

The rest of the paper is organized as follows: The continuous dataset collected to analyze the developed approach is described in Section 2. Section 3 provides a detailed description of the approach. The experimental results are reported in section 4, and the paper is concluded in section 5.

2. CONTINUOUS DATASET

This work involves the detection and recognition of actions of interest from a continuous data stream consisting of actions of interest and actions of non-interest that appear in a random order with respect to each other. Apart from the video datasets provided in [11-12] that are captured by video cameras, there is no publicly available dataset that provides continuous data streams from a depth camera. Hence, as part of this work, a dataset for the wrist actions involved in smart TV gestures was collected and is made publicly available. This dataset can be downloaded from this link http://www.utdallas.edu/~kehtar/UTD-CAD.htm.

The actions of interest for the smart TV application consist of 'Waving a hand', 'Flip to Left', 'Flip to Right', 'Counterclockwise Rotation', and 'Clockwise Rotation'. For training, the subjects were asked to perform these actions of interest one action at a time. While for testing, the subjects were asked to perform these actions of interest continuously among various actions of non-interest in a random order. Subjects had the freedom to choose their own actions of non-interest. Example actions of non-interest included drinking water, eating snacks, stretching, walking around and reading a book.

Two scenarios were considered for data collection: subject-specific and subject-generic. For the subject-specific scenario, the training and testing were done by the same subject. For the training dataset, each of the actions of interest was done 15 times. For the testing dataset, the actions of interest were done randomly among actions of non-interest for 6 continuous data streams. For the subject-generic scenario, 5 different subjects were asked to repeat each of the actions of interest 5 times during the training. For the testing dataset, for each subject, a continuous data stream was collected which consisted of all the actions of interest randomly done among actions of non-interest.

3. DEVELOPED CONTINIUOUS ACTION DETECTION AND RECOGNITION

The action detection and recognition approach developed in this paper involves first segmenting and detecting actions of interest from a continuous data stream of skeleton joint positions and then classifying such detected actions of interest by utilizing both skeleton joint positions and depth images. The developed approach comprises three main steps: segmentation, detection and classification. The segmentation step involves identifying the presence of an action in a continuous data stream. These actions can be actions of interest or actions of non-interest. Next, in the detection step, a segmented action is labeled as an action of interest or an action of non-interest. This is achieved by using the method of support vector data description (SVDD) described in [13]. Then, in the classification step, the detected actions of interest are classified by using both the skeleton joint positions and depth images. A variable-length Maximum Entropy Markov Model (MEMM) classifier [14] is used for classification of skeleton information, while a Collaborative Representation Classifier (CRC) [15] is used for classification of depth information. Basically, in this paper, different existing techniques are integrated into a real-time approach to perform both detection and recognition of actions of interest among actions of non-interest performed in a continuous manner. A block diagram of the steps involved in the developed approach is shown in Fig. 1. In what follows, more details of these steps are mentioned.

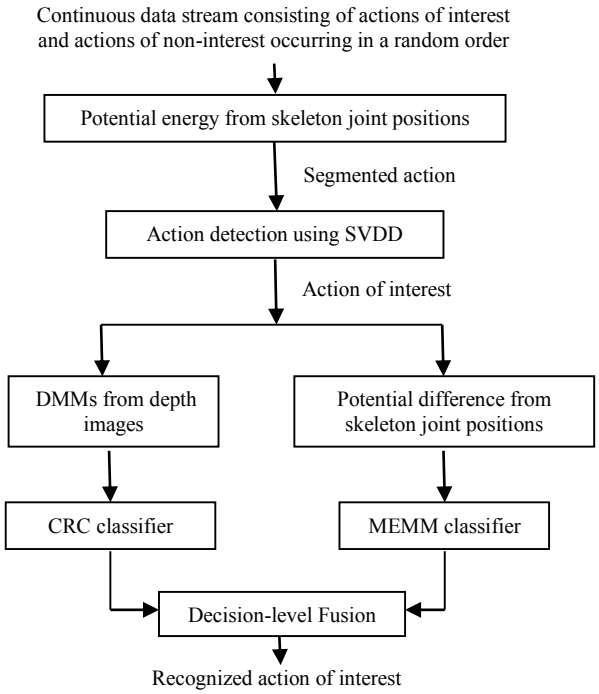

Fig. 1. Block diagram of the developed continuous action detection and recognition approach.

3.1. Segmentation Step

Segmentation of actions is achieved using skeleton joint positions via a technique similar to the one discussed in [14, 16]. The so called normalized relative orientations (NRO) of the joints are extracted and used as features for segmentation. The NRO of a joint i is computed with respect to its rotating joint j as follows:

$$F_{NRO}^{i} = \frac{L_i - L_j}{\|L_i - L_j\|} \quad (1)$$

where L_i and L_j denote the respective 3D locations of joints i and j and $\|\cdot\|$ represents the Euclidean distance. Let $F_t = (F_{NRO}^1, F_{NRO}^2, F_{NRO}^3, \dots)_t$ be the vector of all the joints NROs at frame t. Based on a sequence of NRO feature vectors $(F_1, F_2, \dots, F_t, \dots)$, a potential energy function at the t^{th} frame is obtained as follows:

$$PE(t) = \|F_t - F_r\|^2 \quad (2)$$

where F_r denotes a reference NRO feature vector, which is considered here to be the first frame in the sequence. This potential energy function is compared to a user specified threshold. If the potential energy of the frames appears below this threshold, it is set to zero. This threshold is set experimentally. For example, for the dataset collected, a

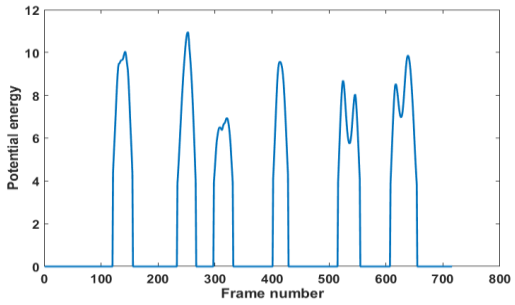

Fig. 2. Potential energy of a continuous data stream.

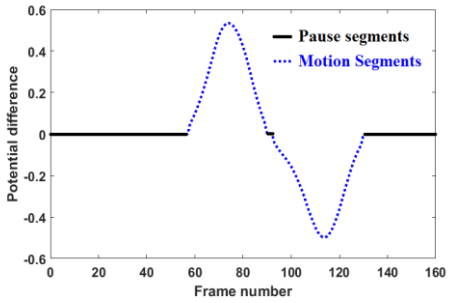

Fig. 3. Pause and motion segments in the action 'Flip to left'.

threshold in this range [1.05, 2.90] for the subject-specific scenario and a threshold in this range [3.70, 5.00] for the subject-generic scenario generated accurate detection. An example of the potential energy function for a continuous data stream is shown in Fig. 2. As can be seen from this figure, consecutive frames with positive values are marked as a segmented action. Segmenting frames in this manner provides the start and end of an action, which are then used to detect whether that action is an action of interest or not.

3.2. Detection Step

Detection of actions of interest from a segmented action is done based on a one-class SVDD classifier. The basic concept behind SVDD is to find a spherical boundary enclosing all the data of interest. Consider a training dataset X consisting of N data samples x_m, $X = \{x_m, m = 1, \dots, N\}$. If R is the radius and a is the center of the smallest sphere encircling all the data samples, the following quantity is minimized in SVDD [13, 17]

$$H(R, a, \xi_m) = R^2 + \gamma \sum_{m=1}^{N} \xi_m \tag{3}$$

subject to the constraints

$$\|\phi(x_m) - a\|^2 \leq R^2 + \xi_m, \text{ for all } m \tag{4}$$

and

$$\xi_m \geq 0 \tag{5}$$

where ξ_m is a slack variable that penalizes outliers, γ is a parameter which controls a trade-off between volume and error and the notation ϕ indicates a nonlinear transformation to a higher dimensional kernel space. The interested reader is referred to [13] for more details on SVDD.

Once an action is considered from a continuous data stream, the potential energy of that action is divided into three equal portions and the average NROs from these three portions are used as X in the above SVDD minimization. To examine a segmented action, the average NROs of its three equal portions are computed and mapped according to the nonlinear transformation. The distance of the feature vector from the center of the sphere is found. If this distance is less than the radius of the sphere, the corresponding action is considered to be an action of interest.

3.3. Classification Step

Next, the detected actions of interest are classified using a variable-length MEMM classifier for the skeleton information and a CRC classifier for the depth information. The operation of the MEMM classifier is similar to that of a Hidden Markov Model (HMM) classifier but it is computationally more efficient than HMM.

For classification of skeleton data, a potential difference at frame t is computed from the potential energy function as follows:

$$PD(t) = PE(t) - PE(t-1) \tag{6}$$

If this potential difference is less than a value close to zero, segments of an action are labeled as pause segments, otherwise they are labeled as motion segments. For the dataset examined in this work, the value 0.04 allowed separating pause and motion segments. An example of the pause and motion segments for a segmented action 'Flip to left' is shown in Fig. 3.

Based on pause and motion segments, a codebook is then setup which is used to perform recognition. The details associated with the training of the MEMM classifier is provided in [16] with the difference that in this work, the clustering as part of the training is applied to motion segments, not pause segments. As a result, instead of considering motion segments between every pair of pause clusters, the mean of pause segments between every pair of motion clusters is considered. Recognition of a segmented action is then carried out based on likelihood probabilities as discussed in [16].

Classification of the depth data is done by first extracting depth motion maps (DMM) as described in [18]. DMMs are derived from 2D projection maps corresponding to the front, side and top views of 3D depth data. For a depth sequence of n frames, a DMM is obtained as follows:

$$DMM = \sum_{k=1}^{n-1} |map^{k+1} - map^{k}| \tag{7}$$

Similar to [19], a l_2-regularized CRC is utilized here to classify actions of interest based on DMMs. Finally, the decision-level fusion approach discussed in [20] is adopted using uniformly distributed classifier weights. The label of

the segmented action is assigned to be the class with the largest probability.

4. RESULTS AND DISCUSSION

This section presents the results of the developed detection and recognition approach on the continuous dataset collected. Noting that there exists no approach in the literature that performs both detection and recognition of actions of interest when performed in a continuous and random manner among actions of non-interest, the evaluation of the approach developed in this paper is carried out via commonly used recognition measures. The average duration of a continuous data stream in the examined dataset is about 40s with the actions of interest occupying 10s of this duration on average and the actions of non-interest occupying the remaining time of 30s.

The outcome of the segmentation and detection steps is reported in Table 1. For the subject-specific scenario, there were a total of 30 actions of interest in the 6 continuous data streams. In this scenario, all the 30 actions of interest were correctly detected, while there was only one action of non-interest which was wrongly detected as an action of interest. Similarly, all the 25 actions of interest in the subject-generic scenario were correctly detected with only one action of non-interest wrongly detected as an action of interest.

Table 1. Outcome of the segmentation and detection steps

Scenario	Actions of interest correctly detected	Actions of non-interest detected as action of interest
Subject-specific	30/30	1
Subject-generic	25/25	1

For the overall detection and recognition results, since a sequence in a continuous data stream is unknown or cannot be matched to the ground truth, the evaluation of the overall approach was done using the widely used precision P, recall R and $F1$ score measures. These measures are defined as follows [21-22]:

$$P = {\#TP}/{(\#TP + \#FP)} \quad (8)$$

$$R = {\#TP}/{(\#TP + \#FN)} \quad (9)$$

$$F1 = {2P \cdot R}/{(P + R)} \quad (10)$$

where $\#TP$ denotes the number of true positives, $\#FP$ the number of false positives, and $\#FN$ the number of false negatives.

As mentioned in [23], whenever an action was found within a window of four frames, it was marked as a true positive, whereas the actions of non-interest wrongly detected as actions of interest or the actions of interest that were misclassified were marked as false positives. Actions of interest not detected or not correctly recognized were marked as false negatives.

The precision, recall and $F1$ score measures obtained for the subject-specific and subject-generic scenarios are reported in Tables 2 and 3, respectively. These tables also show the results of the situations when the classification was performed using the skeleton and depth information individually or separately. As can be seen from these tables, the values of the precision, recall and $F1$ score measures were increased when both the skeleton and depth information were used together due to correcting some of the misclassifications.

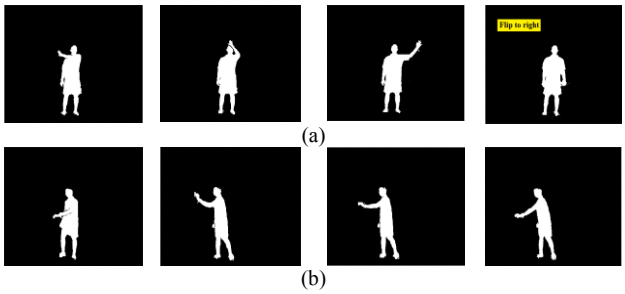

Fig. 4. (a) Snapshots of depth images from an action of interest 'Flip to Right' (b) an action of non-interest 'writing on a board'.

Table 2. Precision, recall and $F1$ score measures for subject-specific scenario

Modality used for classification	Precision	Recall	F1 score
Skeleton only	80.4%	83.1%	81.7%
Depth only	62.5%	64.5%	63.5%
Skeleton+Depth	85.6%	88.3%	86.9%

Table 3. Precision, recall and $F1$ score measures for subject-generic scenario

Modality used for classification	Precision	Recall	F1 score
Skeleton only	77.2%	80.2%	78.7%
Depth only	69.2%	72.0%	70.5%
Skeleton+Depth	86.8%	90.0%	88.3%

It is important to emphasize that detection was performed whenever an action was segmented in a continuous data stream and classification was performed whenever that action was labeled as an action of interest. An example of an action of interest and an action of non-interest from a test sequence is shown in Fig. 4.

5. CONCLUSION

In this paper, an action detection and recognition approach has been developed which is capable of dealing with continuous data streams captured by a depth camera. Such data streams for the smart TV application were collected which consisted of five actions of interest performed continuously and in a random order among various actions of non-interest. The results obtained indicate the effectiveness of the developed approach in separating actions of interest from actions of non-interest and classifying them in an on-the-fly manner. In our future work, it is planned to apply this approach to other applications involving different sets of actions of interest.

6. REFERENCES

[1] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pp. 9-14, June 2010.

[2] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu, "Robust 3d action recognition with random occupancy patterns," *Proceedings of Computer Vision–ECCV 2012*, Springer Berlin Heidelberg, pp. 872-885, 2012.

[3] A. Vieira, E. Nascimento, G. Oliveira, Z. Liu, and M. Compos, "Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences," *Iberoamerican Congress on Pattern Recognition*, Springer Berlin Heidelberg, pp. 252-259, September 2012.

[4] C. Chen, R. Jafari, and N. Kehtarnavaz, "Action recognition from depth sequences using depth motion maps-based local binary patterns," *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, Waikoloa, HI, pp. 1092-1099, January 2015.

[5] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," *Proceedings of the 20th ACM International Conference on Multimedia*, pp. 1057-1060, October 2012.

[6] L. Xia, C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3d joints," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 20-27, June 2012.

[7] X. Yang, and Y. Tian, "Eigenjoints-based action recognition using naive-bayes-nearest-neighbor," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 14-19, June 2012.

[8] http://www.microsoft.com/en-us/kinectforwindows/

[9] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," *Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR)*, pp. 1290-1297, June 2012.

[10] D. Huang, S. Yao, Y. Wang, and F. De La Torre, "Sequential max-margin event detectors," *Proceedings of European Conference on Computer Vision*, Springer International Publishing, pp. 410-424, September 2014.

[11] R. De Geest, E. Gavves, A. Ghodrati, Z. Li, C. Snoek, and T. Tuytelaars, "Online Action Detection", *Proceedings of European Conference on Computer Vision*, Springer International Publishing, pp. 269-284, October 2016.

[12] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," *Proceedings of IEEE Computer Vision and Pattern Recognition*, pp. 2929-2936, June 2009.

[13] D. Tax, and R. Duin, "Support vector data description," *Machine Learning*, vol. 54, no. 1, pp. 45-66, January 2004.

[14] G. Zhu, L. Zhang, P. Shen, and J. Song, "An Online Continuous Human Action Recognition Algorithm Based on the Kinect Sensor," *IEEE Sensors Journal*, vol. 16, no. 2, pp. 161, January 2016.

[15] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?," *Proceedings of IEEE International Conference on Computer Vision*, pp. 471-478, November 2011.

[16] G. Zhu, L. Zhang, P. Shen, J. Song, L. Zhi and K. Yi, "Human action recognition using key poses and atomic motions," *Proceedings of IEEE International Conference on Robotics and Biomimetics*, pp. 1209-1214, December 2015.

[17] F. Saki, and N. Kehtarnavaz. "Online frame-based clustering with unknown number of clusters," *Pattern Recognition*, vol. 57, pp. 70-83, September 2016.

[18] C. Chen, K. Liu, and N. Kehtarnavaz, "Real-time human action recognition based on depth motion maps," *Journal of Real-Time Image Processing*, pp. 1-9, August 2013.

[19] C. Chen, R. Jafari, and N. Kehtarnavaz, "A real-time human action recognition system using depth and inertial sensor fusion," *IEEE Sensors Journal*, vol. 16, no. 3, pp. 773-781, February 2016.

[20] C. Chen, R. Jafari, and N. Kehtarnavaz, "Fusion of depth, skeleton, and inertial data for human action recognition," *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2712-2716, March 2016.

[21] J. Davis, and M. Goadrich, "The relationship between Precision-Recall and ROC curves," *Proceedings of the 23rd International Conference on Machine Learning*, pp. 233-240, June 2006.

[22] C. Goutte, and E. Gaussier, "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation," *Proceedings of European Conference on Information Retrieval*, Springer Berlin Heidelberg, pp. 345-359, March 2005.

[23] V. Bloom, D. Makris, and V. Argyriou, "G3d: A gaming action dataset and real time action recognition evaluation framework," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 7-12, June 2012.