Human-Object Interaction Recognition by Learning the distances between the Object and the Skeleton Joints

Meng Meng, Hassen Drira, Mohamed Daoudi, Jacques Boonaert

Human-Object Interaction Recognition by Learning the distances between the Object and the Skeleton Joints

Meng Meng [1], Hassen Drira[1], Mohamed Daoudi[1], and Jacques Boonaert [2]
[1] Institut Mines-Télécom/Télécom Lille, CRIStAL (UMR CNRS 9189) Lille, France
[2] Ecole des Mines de Douai, France

***Abstract*— In this paper we present a fully automatic approach for human-object interaction recognition from depth sensors. Towards that goal, we extract relevant frame-level features such as inter-joint distances and joint-object distances that are suitable for real time action recognition. These features are insensitive to position and pose variation. Experiments conducted on ORGBD dataset following state-of-the-art settings show the effectiveness of the proposed approach.**

I. INTRODUCTION AND RELATED WORK

A. Introduction

Imaging technologies have recently shown a rapid advancement with the introduction of low cost depth cameras with real-time capabilities, like Microsoft Kinect. These new acquisition devices have stimulated the development of various promising applications, including human pose estimation, scene flow estimation, hand gesture recognition, smart surveillance, web-video search and retrieval, quality-of-life devices for elderly people, and human-computer interfaces. Given the initial success of bag-of-words methods for action classification, the field is gradually moving towards more structured interpretation of complex human activities involving multiple people and objects as well as interactions among them in various realistic scenarios.

Most of the previous related works have been focused on simple human action recognition such as boxing, kicking, walking, etc. Several datasets were collected to serve as benchmark for researchers algorithms like MSR dataset [11].

Less effort have been spent on dynamic human object interaction. To the best of our knowledge, the only work reporting results on human object interaction is [1]. Actually, when people talk over phone for example, the hand that holds the phone is usually close to the ear no matter whether the person is sitting, bending, standing, or walking.

B. Related Work

Researchers have explored different compact representations of human actions recognition and detection in the past few decades. There is a large amount of work on static and 2D video. The release of the low-cost RGBD sensor Kinect has brought excitement to the research in computer vision, gaming, gesture-based control, and virtual reality. Shotton et al. [6] proposed a real-time method to predict 3D positions of body joints in individual depth map without using any temporal information from Kinect. Relying on the joints location provided by Kinect, in Xia et al. [5] an approach for human action recognition with histograms of 3D joint locations (HOJ3D) as a compact representation of postures is proposed. The HOJ3D computed from the action depth sequences are reprojected using LDA and then clustered into several posture visual words, which represent the prototypical poses of actions. The temporal evolutions of those visual words are modeled by discrete hidden Markov models (HMMs). Raviteja et al. [4] represented a human skeleton as a point in the Lie group which is curved manifold, by explicitly modeling the 3D geometric relationships between various body parts using rotations and translations. Using the proposed skeletal representation, it modeled human actions as curves in this Lie group then mapped all the curves to its Lie algebra, which is a vector space, and performed temporal modeling and classification in the Lie algebra. There are several works [10][11][12][13][14] relyed on skeleton information and developed features based on depth images for human object interaction recognition.

There are tremendous works on human action recognition but a few works on recognition of human-object interaction in the past few decade. Most of them were based on static images. Yao et al [3] proposed a grouplet feature for recognizing human-object interactions. Grouplets encode detailed and structured information in the image data. A data mining method incorporated with a parameter estimation step is applied to mine the discriminative grouplets. In another work of Yao et al [9], it treated object and human pose as the context of each other in human-object interaction activities. This approach developed a conditional random field model that learns cooccurrence context and spatial context between objects and human poses. Chaitanya et al [7] develop a discriminative model of human-object interactions. It based on modeling contextual interactions between postured human bodies and nearby objects.

All of the works of human object interaction introduced above based on static images. To the best of our knowledge, there is only one work on recognizing human-object interactions from dynamic depth sequences. Yu et al [1] proposed a novel middle level representation called orderlet [8] for recognizing human object interactions. It presented an orderlet mining algorithm to discover the discrimitive orderlets from an large pool of candidates. In addition, they collected a new dataset adopted in our paper for recognizing human-object interactions by a RGB-D camera.

We propose in this paper an approach for dynamic human

This work was supported by Institut Mines-Télécom

object interaction recognition using frame-based feature and a memory of $k = 0$ previous frames. For $k = 0$, we achieve the recognition without any memory of previous frames. There is no temporal variation between features which is the different part from [1], that means the proposed feature can be suitable for on line human-object interaction recognition.

The rest of the paper is organized as follows. Section 2 presents the spatio-temporal modeling of dynamic skeleton data. The features used for encoding the data, the classification method and an overview of the proposed approach are detailed in this section. Section 3 presents experimental results, then the conclusion and future works are given in the Section 4.

II. SPATIO-TEMPORAL MODELLING

The 3-D humanoid skeleton can be extracted from depth images (via RGB-D cameras, such the Microsoft kinect) in real-time thanks to the work of Shotton et al. [6]. This skeleton contains the 3-D position of a certain number of joints representing different parts of the human body and provides strong cues to recognize human-object interaction. We propose in this paper to recognize human object interaction based on the skeleton joints and the object position in each frame.

A. OVERVIEW OF OUR METHOD

An overview of the proposed approach is given in Fig.1. The object detection step is done based on adopted LOP algorithm [1]. Then, a feature vector based on the object and the joints is calculated for each frame. Next the feature vectors of n successive frames are concatenated together to build the feature vector for the sliding window. Finally, the human object interaction recognition is performed by Random Forest classifier.

B. INTRINSIC FEATURES

In order to have real time approach, one needs to use frame-based features. These features have to be invariant to pose variation of the skeleton but discriminative for human object interaction. We propose to use the inter-joints distances. Morever, we consider the object as an additional joint, the object position was detected by the LOP algorithm [1].

Then we obtained our new skeleton information that is donated as S which contains 20 joints from the original data and object joint represented by j_o.

$$S = \{j_1, j_2, ..., j_{20}, j_o\}$$

V refers to the set of the pairwise distances between the joint a and joint b from S.

$$V = \{d(a, b)\}, a \in S, b \in S$$

Thus the feature vector is composed by the all pairwise distances between the joints and the distances between the object and the joints. The size of this vector is equal to $m \times (m-1)/2$, with $m = 21$: 20 joints and the object joint.

We report in Fig. 2 examples of different human action interactions. The object is reported in red and the joints are reported in green. The proposed features are the pairwise distances between all these points in 3D.

C. DYNAMIC SHAPE DEFORMATION ANALYSIS

To capture the dynamic of object and skeleton deformations across sequences, we consider the inter-joint distances and object-joints distances computed at n successive frames. In order to make possible to come to the recognition system at any time and make the recognition process possible from any frame of a given video, we consider sub-sequences of n frames as sliding window across the video.

Thus, we chose the first n frames as the first sub-sequence. Then, we chose n-consecutive frames starting from the second frame as the second sub-sequence. The process is repeated by shifting the starting index of the sequence every one frame till the end of the sequence.

The feature vector for each sub-sequence is built based on the concatenation of individual features of the n frames of the sub-sequence.

Thus, each sub-sequence is represented by a feature vector of size the number of distances for one frame times the size of the window n. For the sliding window of size $n \in [1, L]$ that begins at frame i, the feature vector is:

$$x = [V_i, V_{i+1}, ..., V_{i+n-1}],$$

with L the length of the sequence.

For $n = 1$, our system is equivalent to recognition frame by frame without any memory of previous frames. If $n = L$, the length of the video, our system will provide only one decision at the end of the video. The effect of the size of the window on the performance is studied later in experimental part.

D. RANDOM FOREST-BASED REAL TIME HUMAN ACTION INTERACTION

For the classification task we used the Multi-class version of Random Forest algorithm. The Random Forest algorithm was proposed by Leo Breiman in [2] and defined as a meta-learner comprised of many individual trees. It was designed to operate quickly over large datasets and more importantly to be diverse by using random samples to build each tree in the forest. Diversity is obtained by randomly choosing attributes at each node of the tree and then using the attribute that provides the highest level of learning. Once trained, Random Forest classify a new action from an input feature vector by putting it down each of the trees in the forest. Each tree gives a classification decision by voting for that class. Then, the forest chooses the classification having the most votes (over all the trees in the forest). In our experiments we used Weka Multi-class implementation of Random Forest algorithm by considering 50 trees. A study of the effect of the number of the trees is reported later in the experimental part.

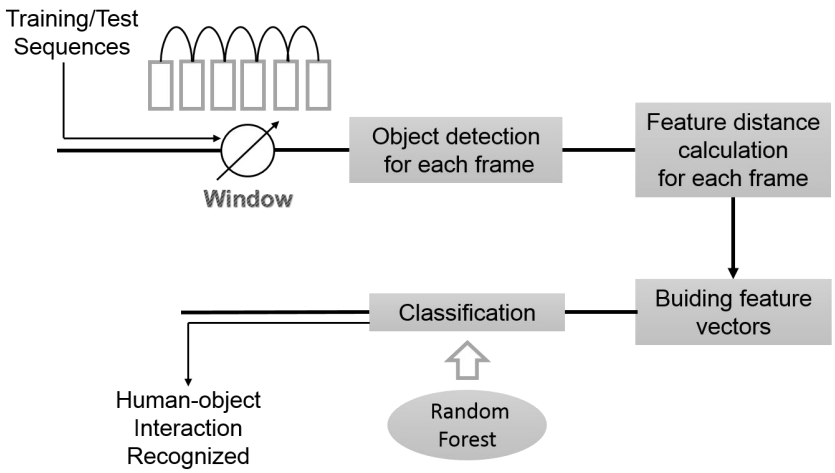

Fig. 1. Overview of our method. Three main steps are shown: Object detection and feature extraction for each frame; Buiding the feature vector for sliding window and Random Forest-based classification. Note that the both train and test sequences can go through the upper and lower path in the block-diagram.

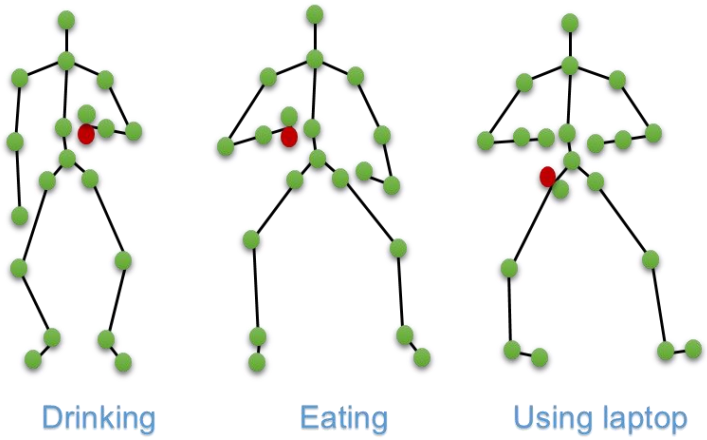

Fig. 2. Examples of our features based on pairwise joint distance. The red one refers to object joint for each action.

III. Experiments

This section summarizes our empirical results and provides an analysis of the performance of our proposed approach on ORGBD dataset [1] compared to the state-of-the-art approaches.

A. Dataset and experimental protocol

We evaluate the performance of our approach for action recognition based on the dataset "Online RGBD Action dataset(ORGBD)" [1]. This dataset contains seven types of actions which all those actions are human-object interactions: drinking, eating, using laptop, picking up phone, reading phone (sending SMS), reading book, and using remote. The bounding box of the object in each frames is manually labelled. In our approach, we use the object labels to locate our object feature.

All of the videos are captured by Kinect. Each action was performed by 16 subjects for two times. We compare our approach with the state-of-the-art methods on the cross-subject test setting, where half of the subjects are used as training data and the rest of the subjects are used as test data.

B. Experimental results and comparison to state-of-the-art

Up to now, most of the work related to human recognition based on the MSR-Action3D dataset or MSRDaliyActivity 3D dataset. Our work is the second one based on the ORGBD dataset. The first work based on the ORGBD dataset is [1]. The performance presented in this paper uses temporal variation of a joint location. We compare our approach with the state-of-the-art method on the 2-fold cross-validation and the comparison of the performance is shown in Table I.

The recognition rate of the discriminative Orderlet Mining is 71.4%, we note that the feature used in their method is with memory of previous frames. Fig.4 shows the confusion matrix without memory of previous frames of the proposed method. Over all seven action categories, 'eating', 'using laptop' and 'reading book' have the best recognition rate. 'drinking' is the most confused action in all cases; it is mostly confused with 'reading phone'.

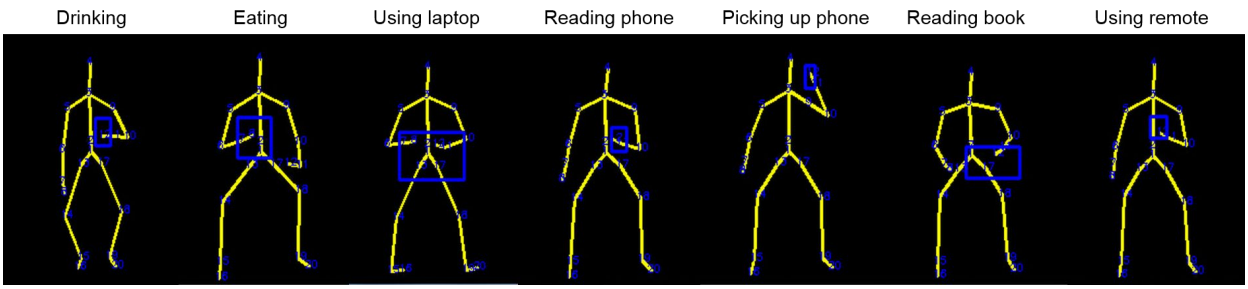

Fig. 3. Skeleton frames from Online RGBD Action dataset

TABLE I
REPORTED RESULTS COMPARISON TO STATE OF THE ART

Method	Accuracy
Discriminative Orderlet Mining (with memory of previous frames)	71.4%
Proposed approach (without memory of previous frames)	72.1%
Proposed approach with memory of 10 frames	75.8%

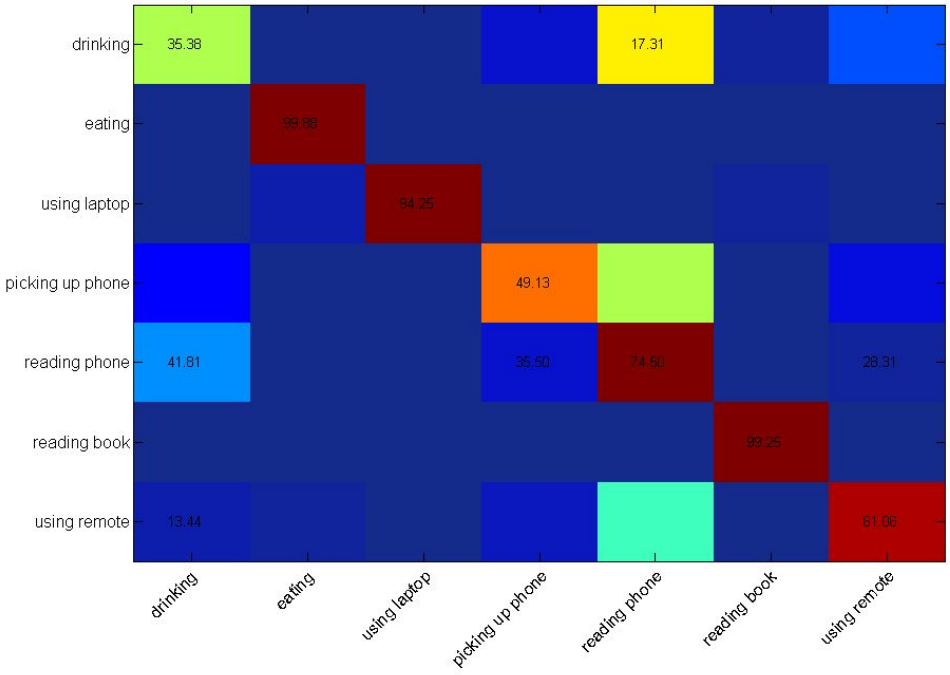

Fig. 4. Confusion matrix using one frame for the proposed approach on ORGBD dataset.

Actually, they used temporal variation cross frames to build their feature vector. In our method, when $n = 1$, the feature has no memory of previous frames, the mean value of the recognition rate by the Random Forest is 72.05%. For fair comparison, we need to compare our recognition rate with $n > 1$ to the result in [1]. We achieve 75.8% recognition rate for $n = 10$, which represents better performance than [1]. This result can be due to the use of the distance between the object and the joints that is more relevant than the object position used in [1].

C. *Effect of the number of trees in Random Forest algorithm*

The performance of Random Forest classifier varies with the number of trees. Thus, we perform the experiments with different numbers of trees; the results of this experimentation is shown in Fig. 5. As illustrated in this figure, the recognition rate raises with the increasing number of trees until 60, when the recognition rate reaches 72.5%, and then becomes quite stable. Thus, in the following we consider 50 trees and we report detailed results with this number of trees in Fig.5.

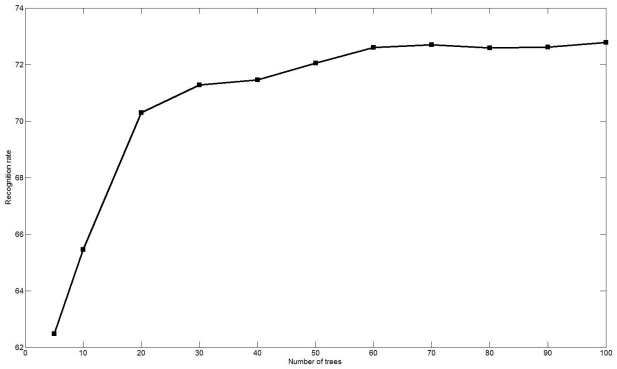

Fig. 5. Human-Object interaction recognition results using a Random Forest classifier when varying the number of trees.

D. Effect of the temporal size of the sliding window

We have conducted additional experiments when varying the temporal size of the sliding window used to define the sub-sequences. In Fig. 6, we report results for a window size equal to 2, 5 and 10 frames. The recognition rates are respectively 72.8%, 74.8% and 75.8%. Finally, we use the whole length of the sequence (on average this is about 100 frames). From the figure, it clearly emerges that the action recognition rate increases when increasing the temporal length of the window. This reveals the importance of the temporal dynamics and shows that the spatio-temporal analysis outperforms a spatial analysis of the frames. By considering the whole sequences for the classification, the result reaches 82%.

E. Relevant features

We reveal the relevant features for human object interaction recognition. The distances between the object and the joints are selected ones in general. In order to better understand the behavior of the proposed approach, we perform binary classification of each interaction. For action 1 (drinking), we label the data from action 1 as the first class, the second class includes all the remaining actions. The best features to classify action 1 (drinking) are revealed. We repeat this experiments for all the remaining actions separately. Fig. 7 shows the results of this experiment. The pairwise distances between the the yellow and red joints are the best features to recognize each human object interaction.

For example, the best features for drinking (action 1) are the pairwise distances between the object joint and the skeleton joints which are on the right hands, on both sides of the crotch, on the left hand and on the left feet. Another example, for eating (action 2), the best features are the pairwise distances between the object joint and the skeleton joints which are on the left hand and on the right hand. There is another situation, for using laptop (action 3), the best features are the pairwise distances between the object joint and the skeleton joints which are on the crotch and on the spinal part. Based on the attributed distances, we know which joint on the skeleton data for each action is more meaningful for recognizing different human object interactions.

IV. CONCLUSION AND FUTURE WORK

This paper proposed relevant features which accurately describes human object interaction across human action sequences acquired with depth sensor.

A human was represented by the set of joints located in the skeleton, the inter-joints distances and the object-joints distances were used to classify the human object interaction. These features have the advantage to be real time extracted and pose, position invariant. The classification is based on Random Forest algorithm using 50 trees. Experiments conducted on RGB-D dataset, following state-of-the-art setting demonstrate the effectiveness of the proposed approach.

In future, we will focus on more complicated scenarios, where one video contains several actions, and/or interactions. A segmentation part should be done before action recognition.

REFERENCES

[1] G. Yu, Z. Liu, J. Yuan, *Discriminative Orderlet Mining For Real-time Recognition of Human-Object Interaction*, Asian Conference on Computer Vision (ACCV) 2014.

[2] L. Breiman, Random forests. *Machine Learning*, vol. 45, pp 5-32; 2001.

[3] B.Yao, F. Li, *Grouplet: A Structured Image Representation for Recognizing Human and Object Interactions*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010.

[4] R. Vemulapalli, F. Arrate, R. Chellapan, *Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group*,IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[5] L. Xia, C. Chen, J.K. Aggarwal, *View Invariant Human Action Recognition Using Histograms of 3D Joints*, International Workshop on Human Activity Understanding from 3D Data in conjunction with 23th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, June 2012.

[6] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, A. Blake, *Real-time human pose recognition in parts from single depth images*, Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, Colorado Springs, Colorado, USA (June 2011) 1-8.

[7] C. Desai, D. Ramanan,C. Fowlkes, *Discriminative models for static human-object interactions*, International Journal of Computer Vision 95.1 (2011): 1-12.

[8] J. Wang, Z. Liu, Y. Wu, J. Yuan, *Mining actionlet ensemble for action recognition with depth cameras*,IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

[9] B.Yao, F. Li, *Recognizing Human-Object Interactions in Still Images by Modeling the Mutual Context of Objects and Human Poses*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 34.9 (2012): 1691-1703.

[10] G.Dian, G.Medioni,*Dynamic Manifold Warping for View Invariant Action Recognition*, IEEE International Conference on Computer Vision (ICCV), 2011.

[11] W.Li, Z.Zhang, Z. Liu, *Action Recognition Based on A Bag of 3D Points*, IEEE Computer Vision and Pattern Recognition Workshops (CVPRW), 2010.

[12] O. Omar, Z. Liu, *HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013.

[13] R. Slama, H. Wannous, M. Daoudi, A. Srivastava, *Accurate 3D action recognition using learning on the Grassmann manifold*, Pattern Recognition 48.2 (2015): 556-567.

[14] M. Devanne, H. Wannous, Stefano Berretti, Mohamed Daoudi, Alberto Del Bimbo, *Space-Time Pose Representation for 3D Human Action Recognition*, New Trends in Image Analysis and Processing–ICIAP 2013. Springer Berlin Heidelberg, 2013. 456-464.

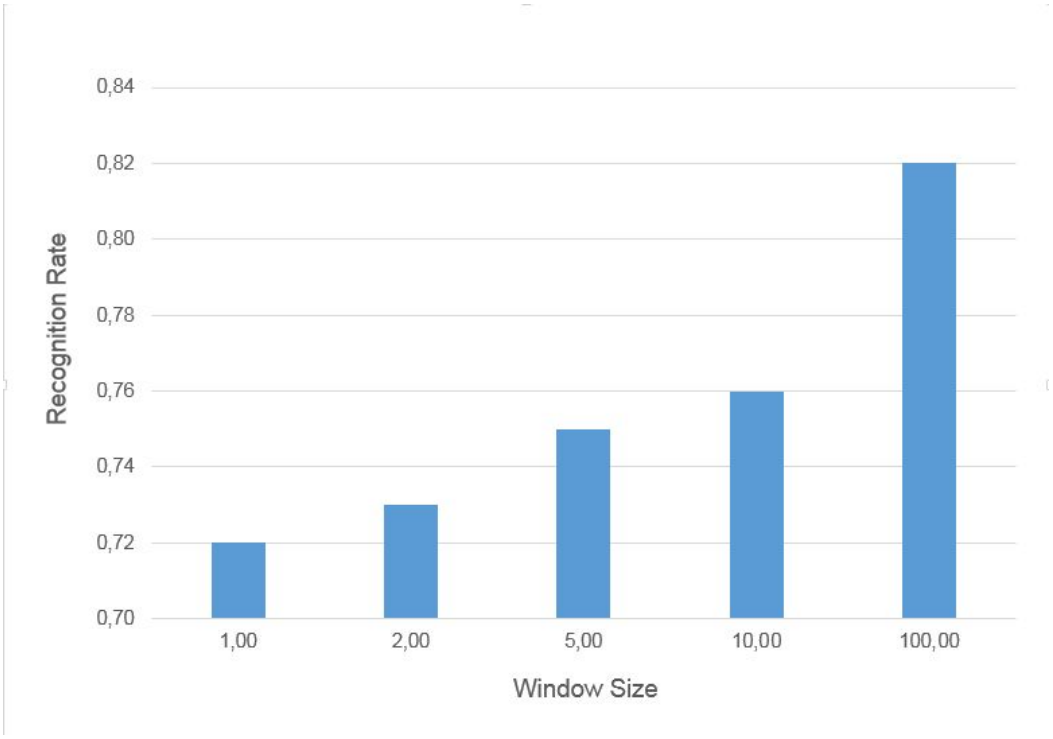

Fig. 6. Effect of the temporal size of the sliding window on the results. The classification rates increase when increasing the length of the temporal window.

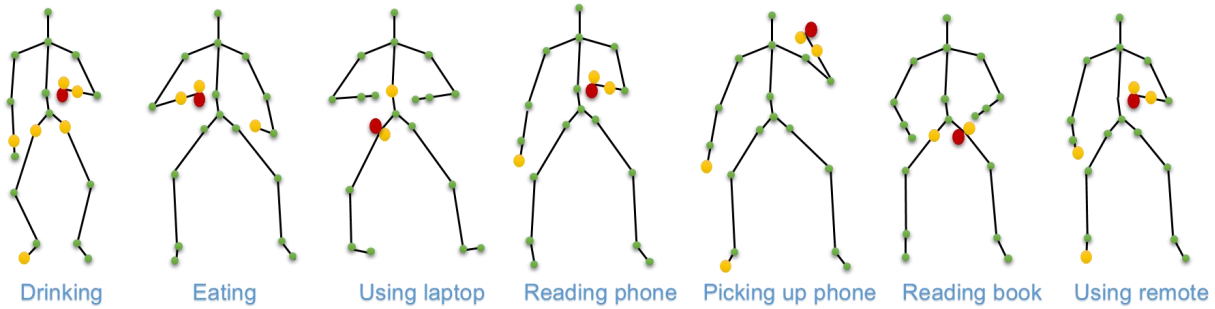

Fig. 7. Selected features for each interaction, the best features are the distances between the yellow and red joints.