Depth imaging system for human posture recognition

Dorra Riahi Wassim Bouachir Youssef Ouakrim Neila Mezghani

LICEF research center
TÉLUQ University
Montréal (Québec), Canada
dorra.riahi@teluq.ca, wassim.bouachir@teluq.ca, youssef.ouakrim@teluq.ca, neila.mezghani@teluq.ca

***Abstract*—This paper presents a human posture recognition system based on depth imaging. The proposed system is able to efficiently model human postures by exploiting the depth information captured by an RGB-D camera. Firstly, a skeleton model is used to represent the current pose. Human skeleton configuration is then analyzed in the 3D space to compute joint-based features. Our feature set characterizes the spatial configuration of the body through the 3D joint pairwise distances and the geometrical angles defined by the body segments. Posture recognition is then performed through a supervised classification method. To evaluate the proposed system we created a new challenging dataset with a significant variability regarding the participants and the acquisition conditions. The experimental results demonstrated the high precision of our method in recognizing human postures, while being invariant to several perturbation factors, such as scale and orientation change. Moreover, our system is able to operate efficiently, regardless illumination conditions in an indoor environment, as it is based depth imaging using the infrared sensor of an RGB-D camera.**

***Index Terms*—RGB-D camera, depth imaging, posture recognition, SVM.**

I. Introduction

Human posture recognition is an attractive topic in computer vision with several applications, including content-based video annotation, human-machine interaction, and especially ambient assisted living and health care applications, where human behaviour analysis is required. Due to this wide range of applications, numerous research works tried to address this open problem. In general, posture recognition works can be divided into two categories depending on the captured visual data: (1) RGB-based systems using standard acquisition devices to build 3D models, and (2) 3D inherent acquisition systems based on depth and infrared sensing.

Traditional posture recognition approaches focused on exploiting RGB data captured by visible light cameras. RGB-based methods generally infer a human body model through background subtraction or geometric transformation. In [1], a background subtraction method is used to extract the human silhouette. A 3D model of posture is then built using geometric features covering the position and the orientation. Authors in [4] proposed an alternative representation to build the silhouette 3D model of the body, where human posture estimation is based on kinematic constraints. A later work [5] is proposed to build a 3D human body from monocular silhouettes using direct sparse regression of joint angles. The silhouettes are extracted using background segmentation and encoded by an histogram-of-shape-context. Following a different approach, Fossati et al. [6] proposed a complex system combining detection and tracking algorithms to build a posture model using a moving camera. Human postures are detected using both motion and generative model. A non-linear and a circular temporary encoding methods are finally applied to find the best trajectories view match.

The use of conventional visible light cameras presents inherent limitations. For example, most RGB-based methods depend on background subtraction, and thus they lack the flexibility to handle challenging cases, such as dynamic backgrounds, camera motion, and silhouette deformation. Moreover, using RGB data implies analyzing the projection of a 3D world on bi-dimensional images, which is a source of computational complexity.

With the emergence of infrared sensors and RGB-D cameras, the posture recognition problem has been explored in a different approach. Since these types of sensors provide the 3D spatial information of the scene, they significantly alleviate the difficulties of RGB imaging, by offering a new way for human body modeling. In this direction, works in [2], [3] describe a technique for inferring the human body posture using a 3D visual-hull constructed essentially from a set of invisible cast shadow silhouettes. The authors used a multi-view setup based on one infrared camera and multiple infrared light sources, turning on and off cyclically using an electronic system. Although this setup can overcome some problems of multi-camera systems (calibration, synchronization, etc.), it uses cumbersome equipment and requires a complex setting, which limits its use to laboratory environments.

With the emergence of low-cost and easy-operation depth sensors such as Microsoft Kinect, it has become possible to design less complex systems that are able to operate in unconstrained environments. Using this new technology, a straightforward way to represent human posture is to gen-

*This research was supported by the Natural Sciences and Engineering Research Council Grant (RGPIN-2015-03853) and the Canada Research Chair on Biomedical Data Mining (950-231214).

eralize the visual descriptors designed for RGB images. In this sense, many recent methods for human body modeling relied on Space Time Interest Points (STIPs) [9], an extension of the spatial interest points into the spatio-temporal domain. The STIP frameworks were thus applied on depth images in different ways. The authors in [10] applied the Harris3D detector [9] to extract STIP features from depth images. The HOGHOF [11] descriptor is then used to describe the regions around the STIPs. Hernandez et al. [12] also used the Harris3D detector for detecting STIPs on both RGB and depth images. The extracted features are finally fused to construct a single bag of words model. Still with the aim of combining depth and RGB features, the authors in [13] extract STIPs from both RGB and depth channels. The interest points are then described using the 3D gradient information (along the 2D space and time).

But applying RGB-based descriptors on depth images is not the optimal choice, as leads to representing a 3D problem in the 2D space. Moreover, RGB images and depth images have different properties. For example, depth images generally contain a large amount of noise (corresponding to false depth values) compared to RGB data. As a consequence, visual descriptors based on intensity features such as gradient and interest points are not suitable for human body modeling from depth images.

To deal with this limitation, recent approaches (e.g. [8], [14] and [15]) used the skeleton detector [7] to build high-level features characterizing the 3D configuration of the human body. This representation is an efficient alternative to RGB-D based features, and it conforms to the early Johansson's biological studies on how human understand actions [16].

Our method proceeds along this direction for body modeling using joint-based features as a high-level descriptor. The proposed posture recognition system uses a single depth image to perform pose recognition based on the skeleton spatial configuration. Due to our practical setup including a single RGB-D camera, our system can be flexibly used in unconstrained indoor environments. Its high precision and low computational complexity makes it suitable for real-time applications such as health care applications using video surveillance techniques. Moreover, due to the only use of depth images formed by the infrared device of the RGB-D camera, pose recognition can be efficiently achieved regardless the illumination conditions (even in total darkness).

The rest of the paper is organized as follows. In the next section, we present the proposed method. Section 3 provides experimental results. Finally, section 4 concludes the paper.

II. The Proposed Method

Our system uses a single RGB-D camera (Microsoft Kinect v2) to capture depth maps and build a body skeleton. We then use the 3D joint positions to compute posture features including (1) the 3D pairwise joint distances and (2) the geometrical angles defined by adjacent body segments. A supervised classification method is finally applied to recognize statics human postures. In summary, the system consists of three main modules (see figure 1): skeleton estimation, feature computation, and postures classification.

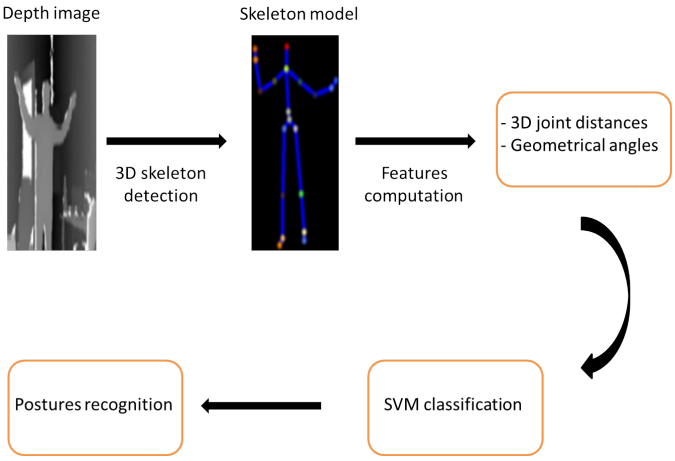

Fig. 1. System architecture

A. 3D skeleton estimation

To build the 3D model of human posture, we use the skeleton detection method [7] to localize different body parts. This method is quite accurate in modeling the human body as an articulated structure of connected segments. This can be achieved in real-time by the only use of depth images. The method of Shotton et al. uses [7] a random forest classifier to segment the different human body parts through a pixel-level classification from single depth images. A mean shift mode detection is then applied to find local centroids of the body part probability mass and generate the 3D locations of body joints. The obtained 3D skeleton formed by 25 joints represents the human posture as illustrated in the figure 2.

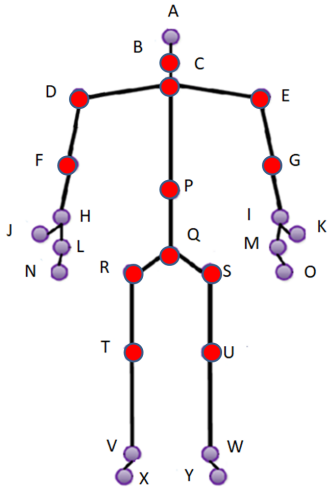

Fig. 2. The 3D joints defining the skeleton. Joints corresponding to the considered geometrical angles are represented in red.

B. Computing features

Constructing a robust feature set is a crucial step in posture recognition systems. In our work, we used 2 types of features to represent human postures: the 3D pairwise joint distances and the geometrical angles defined by adjacent segments.

Considering the set of N = 25 joints (shown in figure 2), the 3D joint coordinates are extracted as:

$$X = \{J_i = (x_i, y_i, z_i), i = 1, 2, ..., N\} \quad (1)$$

where, J_i is the joint number i obtained from the body skeleton and the triplet (x_i, y_i, z_i) represents the 3D coordinates of J_i. Once joint positions are extracted, we compute 2 types of features representing human body posture: 1) the 3D pairwise distances between joints, and 2) geometrical angles of adjacent segments.

1) 3D joint distances: Given the body skeleton obtained from depth images, we calculate the relative 3D distances between pairs of joints. Such features were successfully used to model human body for action recognition [8]. We calculate the feature vector as:

$$D = \{dist(J_i, J_j)|i, j = 1, 2, ..., N; i \neq j\} \quad (2)$$

where $dist(J_i, J_j)$ is the Euclidean distance between two joints J_i and J_j. To ensure scale invariance and to remove the impact of body height variation, the 3D distances are normalized with respect to the person's height. The normalization process is performed by considering the distance $\left\|\vec{BP}\right\|$ (the 3D distance between the spine shoulder joint and the spine middle joint). The choice of this segment is based on experimental observations concluding that the distance $\left\|\vec{BP}\right\|$ is sufficiently stable with respect to several body deformations, in addition to its proportionality to the person's height.

2) Geometrical Angles: In addition to 3D joint distances, our joint-based features include geometrical angles corresponding to relevant joints. Based on our dataset, we selected experimentally a subset of joints according to their importance in characterizing the studied human postures. In the figure 2, the selected joints are represented in red. We considered $m = 12$ angles as follows: $(\vec{CB}, \vec{CP})$, $(\vec{CP}, \vec{CD})$, $(\vec{CE}, \vec{CP})$, $(\vec{DC}, \vec{DF})$,$(\vec{EC}, \vec{EG})$, $(\vec{DR}, \vec{DF})$, $(\vec{ES}, \vec{EG})$, $(\vec{PC}, \vec{PQ})$, $(\vec{QT}, \vec{QU})$, $(\vec{RT}, \vec{RQ})$, $(\vec{SQ}, \vec{SU})$ and $(\vec{QR}, \vec{QS})$.
Geometric angles defined by adjacent segments are directly estimated from joint positions. We define the geometrical angle feature vector as:

$$A = \{\theta_k(\vec{u}, \vec{v})|k = 1, 2, ..., K\} \quad (3)$$

The angle θ_k between the two adjacent segments $\vec{u}$ and $\vec{v}$ is calculated using the 3D coordinates as:

$$\theta_k(\vec{u}, \vec{v}) = \arccos \frac{\vec{u}.\vec{v}}{\|\vec{u}\| \, . \, \|\vec{v}\|} \quad (4)$$

The two feature sets D and A are then concatenated in a single feature vector to form the final representation of human posture $F = [D, A]$.

C. Posture Classification

Once the body pose is modeled, posture recognition is performed in a supervised classification approach using SVM classifier. With its different kind of kernels, SVM has the ability to generate non-linearly as well as high-dimensional classification issue. The SVM classifier also supports multiple-class problems by computing the hyperplane between each class and the rest. In our work, we used an SVM with a linear kernel, since it outperformed other classifiers as discussed in the next section.

III. Experiments

A. Dataset

We created a new dataset that we named KSRD (Kinect Posture Recognition Dataset) using the Microsoft Kinect sensor v2 [1]. To the best of our knowledge, this is the first dataset created with the second version of the Kinect sensor (using Time of flight technology for depth sensing). Note that a recent study demonstrated that the second version of the Kinect sensor surpasses Kinect v1 significantly in terms of accuracy and precision [18]. In our experimental setup, the camera is mounted on a tripod placed in a corner of our laboratory. We consider 5 classes of postures: standing, bending, sitting, walking, and crouching. Each posture was performed by 10 participants of different ages, gender, and morphological characteristics. In order to increase the variability of our dataset and evaluate our system in handling scale and orientation change, each posture is captured at 4 different orientations and 4 different distances with respect to the camera. The distance varies from 1 meter to 4 meters while the orientation angle is changing between 0 and 360 degrees. Our dataset includes a total number of 800 observations that we used for training and testing the proposed system. The dataset is randomly divided into training and testing set using the K-Fold Cross-Validation method (with $K = 10$).

Figure 3 shows examples of human body skeletons corresponding to the 5 classes.

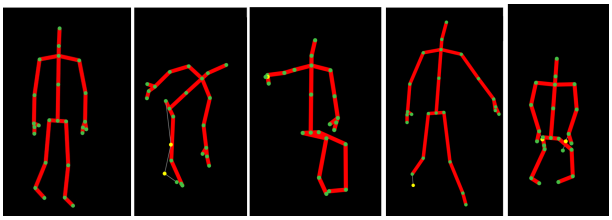

Fig. 3. Examples of body skeletons. From left to right : standing, bending, sitting, walking, and crouching.

B. Results and discussion

We designed our experiments in order to analyze the performance of our method according to several aspects.

[1]The dataset will be published online upon acceptance

Classifier	*Geometrical angles*	*3D joints distances*	*Combined features*
Linear Discriminant	64.4%	50.9 %	52.8%
Quadratic Discriminant	62.3%	51.9%	53.0%
Linear SVM	65.2%	87.9 %	**88.3%**
Quadratic SVM	70.6%	87.9 %	87.7%
Cubic SVM	70.5%	87.9 %	87.9%

TABLE I
PERFORMANCE RESULTS OF SEVERAL CLASSIFIERS USING DIFFERENT TYPES OF FEATURES.

Our first experiment set aimed to identify the optimal choices regarding the two main system components: 1) the feature set and 2) the classification method. We thus investigated several choices implying 5 classifiers (linear discriminant, quadratic discriminant, linear SVM, quadratic SVM, and cubic SVM) and 3 feature sets (geometrical angles, pairwise joint distances, and the combination of the 2 types of features).

Table I summarizes accuracy results for the tested classifiers applied to different feature sets. We can remark that the best recognition rates of linear discriminant classifiers were obtained by using the geometrical angles as features. However, SVM classifiers clearly outperformed linear discriminant methods regardless the feature set. These investigation shows that the linear SVM applied to the entire feature set achieves the best recognition rate (88.3%). We therefore base our recognition system on 1) linear SVM classification, and 2) a combination of the 2 types of features.

To further analyze the performance of the proposed pose recognition system, we used the confusion matrix shown in table II. This evaluation tool provides a comparison of reference postures to the predicted postures. We can see that the diagonal elements show high classification rates for all the classes. The classification accuracy is higher than 86% for almost all the classes, except for the crouching class (79%). The highest classification accuracy was obtained for the bending class, since this posture is characterized by distinctive joint-based features, especially for the upper body segments. Although a recognition rate of 86%, 8% of sitting poses were confused with the crouching pose. This can be explained by the similarity of joint configurations of these 2 postures. On the other hand, the relatively low recognition rate for the crouching posture can be explained by the inaccuracy of skeleton construction. This is mainly due to the large body deformation causing the occlusion of several body joints when the person is crouching. Even if the skeleton estimation algorithm of Shotton et al. [7] predicts the most likely position of the occluded joint, we noticed that this prediction is often inaccurate in the case of a large body deformation.

Postures	*Standing*	*Bending*	*Sitting*	*Walking*	*crouching*
Standing	**86%**	7 %	0%	7%	0%
Bending	3%	**88%**	3%	3%	3%
Sitting	0%	6 %	**86%**	0%	8%
Walking	3%	3 %	3%	**87%**	4%
crouching	0%	4 %	7%	10%	**79%**

TABLE II
CONFUSION MATRIX OF THE PROPOSED POSTURE RECOGNITION METHOD.

We then investigated the impact of feature reduction on recognition accuracy. For this purpose, we applied principal component analysis (PCA) on the entire feature set (formed by 3D pairwise distances and geometrical angles). As presented in table III, feature reduction resulted in the decrease of the recognition rate for the 5 classes, while the overall accuracy decreased to 75%. We thus concluded that dimensionality reduction caused a loss of relevant information for SVM classification.

Postures	*Standing*	*Bending*	*Sitting*	*Walking*	*crouching*
Standing	**82%**	7 %	0%	11%	0%
Bending	3%	**74%**	6%	9%	8%
Sitting	0%	3 %	**75%**	0%	22%
Walking	6%	6 %	3%	**83%**	2%
crouching	0%	12 %	14%	10%	**64%**

TABLE III
CONFUSION MATRIX OF THE DEVELOPED POSTURE RECOGNITION METHOD USING PCA FEATURE REDUCTION.

One of the major difficulties of pose recognition is the scale change. Our method is designed to handle this problem through the normalization of 3D distance features. On the other hand, our dataset includes participants of different heights, whose poses are captured at 4 different distances ranging from 1 to 4 meters from the camera. In order to evaluate the scale invariance of the proposed system, we carried out another sequence of tests where we trained 4 linear SVM classifiers. Each classifier is trained using poses captured at 3 distances (by excluding 1 subset corresponding to 1 distance in each training procedure). The tests are then performed on the excluded subsets. Table IV shows that the proposed method is scale invariant, as it is able to achieve high recognition rates regardless the distance to the camera. In these experiments, the system performance was stable for the 4 tested distances, with a recognition rate greater than (81%). Note that the highest recognition rate was achieved at about 2 meters. This is consistent with the experimental evaluation of the Kinect v2 [19], showing that depth accuracy is optimal at this distance. To sum up, our method allowed to achieve high accuracy, while being flexible and invariant to several perturbation factors.

Classifier	*Distance1*	*Distance2*	*Distance3*	*Distance4*
Linear SVM	81.2%	85.6 %	83.7%	83.4%

TABLE IV
RECOGNITION ACCURACY RESULTS FOR THE 4 DISTANCES.

IV. Conclusion

We presented a novel method for human posture recognition using an RGB-D camera. Our method includes 3 main steps: skeleton estimation, feature computation, and posture classification. Firstly, a 3D skeleton is estimated from depth images. Joint-based features are then computed for modeling the observed human pose. Posture classification is finally carried out using a linear SVM. Our method was validated on a challenging dataset released in our laboratory. The proposed system demonstrated high recognition rates and a significant invariance to important perturbation factors, including scale, orientation, and illumination changes. Our future work will focus on developing appropriate mechanisms addressing other difficulties, such as partial occlusion and fast body movements. This also includes extending the dataset by considering these challenging conditions.

References

[1] B.Boulay, F.Bremond, and M.Thonnat. Human posture recognition in video sequence. *IEEE International Workshop on VS-PETS, Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2003.

[2] R.Gouiaa and J.Meunier Human posture recognition by combining silhouette and infrared cast shadows. In *Image Processing Theory, Tools and Applications (IPTA), 2015 International Conference on*, proceedings, 2015.

[3] R.Gouiaa and J.Meunier Human posture classification based on 3D body shape recovered using silhouette and infrared cast shadows. In *Image Processing Theory, Tools and Applications (IPTA), 2015 International Conference on*, proceedings, 2015.

[4] T.Moeslund and E.Granum 3D human pose estimation using 2D-data and an alternative phase space representation. *Procedure Humans*, 2000.

[5] A.Agarwal and B.Triggs Recovering 3D human pose from monocular images. *IEEE transactions on pattern analysis and machine intelligence*, 28(1): 44–58, 2006.

[6] A.Fossati, M.Dimitrijevic, V.Lepetit, and P.Fua From canonical poses to 3D motion capture using a single camera. *IEEE transactions on pattern analysis and machine intelligence*, 32(7): 1165–1181, 2010.

[7] J.Shotton, T.Sharp, A.Kipman, A.Fitzgibbon, M.Finocchio, A.Blake, M.Cook, and R.Moore. Real-time human pose recognition in parts from single depth images. In *Communications of the ACM*, 56(1): 116–124, 2013.

[8] W.Bouachir and R.Noumeir Automated video surveillance for preventing suicide attempts. In *IET, 2016 International Conference on*, 2016

[9] I.Laptev On space-time interest points. In *International journal of computer vision (2005), 107-123, Springer.*, 2005

[10] B.NI, and al. A Colour-Depth Video Database for Human Daily Activity Recognition. In *Proceedings of IEEE International Conference on Computer Vision Workshops (ICCV Workshops), 6-13.*, 2011

[11] I.Laptev, M.Marszalek, C.Schmid, and B.Rozenfeld Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on (2008), 1-8.*, 2008

[12] A.Hernández-Vela, and al. BoVDW: Bag-of-Visual-and-Depth-Words for gesture recognition. In *Pattern Recognition (ICPR), 2012 21st International Conference on (2012), 449-452.*, 2012

[13] H.Zhanf, and L.Parker 4-dimensional local spatio-temporal features for human activity recognition. In *Intelligent robots and systems (IROS), 2011 IEEE/RSJ international conference on (2011), 2044-2049.*, 2011

[14] M.Zanfir, M.Leordeanu, and C.Sminchisescu The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection. In *Proceedings of the IEEE International Conference on Computer Vision (2013), 2752-2759.*, 2013

[15] L. Xia, C.Chen, and J.Aggarwal View invariant human action recognition using histograms of 3d joints. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on(2012), 20-27*, 2012

[16] G.Johansson Visual perception of biological motion and a model for its analysis. In *Perception & psychophysics (1973), 201-211.*, 1973

[17] J.Wang, Z.Liu, Y.Wu, and J.Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, 1290-1297*, 2012

[18] H.Gonzalez-Jorge, and al. Metrological comparison between Kinect I and Kinect II sensors. In *Measurement 2015 Elsevier, 21-26*, 2015

[19] L.Yang, and al. Evaluating and improving the depth accuracy of Kinect for Windows v2. In *IEEE Sensors Journal (2015), 4275-4285*, 2015