Body Related Occupancy Maps for Human Action Recognition

Sanne Roegiers(✉), Gianni Allebosch, Peter Veelaert, and Wilfried Philips

Department of Telecommunications and Information Processing,
Image Processing and Interpretation, University of Ghent,
St-Pietersnieuwstraat 41, 9000 Ghent, Belgium
sanne.roegiers@ugent.be

Abstract. This paper introduces a novel spatial feature for human action recognition and analysis. The positions and orientations of body joints relative to a reference point are used to build an occupancy map of the 3D space that was occupied during the action execution. The joint data is acquired with the Microsoft Kinect v2 sensor and undergoes a pose invariant normalization process to eliminate body differences between different persons. The body related occupancy map (BROM) and its 2D views are used as feature input for a random forest classifier. The approach is tested on a self-captured database of 23 human actions for game-play. On this database a classification with an F1-score of 0.84 is achieved for the front view of the BROM from the complete skeleton.

Keywords: Human action recognition · Relative positioning · Body related · Occupancy map · Pose invariant normalization · Kinect · Skeleton

1 Introduction

Due to recent development of low-cost and dependable sensor technologies, significant research effort has been made into human action recognition. The strong interest into this research field is further enhanced by the many possible application areas such as intelligent visual surveillance [10], Human-Computer Interaction (HCI) [9], automatic annotation [2] and behavioral biometrics [7]. In particular, marker-less vision-based systems have great potential to deliver inexpensive, non-obtrusive solutions for human action recognition.

In this paper we introduce a novel way to recognize and analyze human actions using body related occupancy maps (BROMs). This method is illustrated on an application for physiotherapy that helps to motivate children in physical rehabilitation or a fitness program. It is not always easy for the children to sustain their efforts and keep up with their exercises. The application is therefore designed as a platform for exergaming which combines exercising with gaming. Not only does the platform present a fun way to exercise, it also offers the possibility of remote monitoring and coaching the subject in an e-environment.

J. Blanc-Talon et al. (Eds.): ACIVS 2017, LNCS 10617, pp. 15–27, 2017.
https://doi.org/10.1007/978-3-319-70353-4_2

The application framework consists of a gross motoric exergame where the child controls the game by performing the required exercises in front of a Kinect sensor. Automatic human action recognition is therefore a key part of the application.

In our approach to human action recognition and analysis, we measure how the subject uses the space around him during the action. The philosophy behind this idea is that different actions require the use of different zones in the personal space around the human body. For example, waving your arms takes primarily place next to the torso. In contrast, pushing something away with both arms happens in front of the torso. We look at the areas that are occupied by the person while the action is being performed. The occupied areas are indicated by an occupancy map, which is then utilized to classify the performed action with a general classifier. We will find that actions can already, for the most part, be recognized by only registering how a person uses the space around him.

In our approach, instead of considering the body as a whole, we also investigate the division of the human body into different body parts. In this paper we will research which particular division strategy is most useful. We will find that, when we classify simple actions, the separation into body parts is most fruitful when it is applied in a hierarchic way. Furthermore, to eliminate the differences of same actions performed in different postures, we only look at the positions of the body parts relative to a reference point. This means that the absolute position of the subject in the world coordinate system is replaced by the relative position in the reference coordinate system.

The BROMs look at a human action in a way that strongly differs from other action recognition methods. We will show that we can recognize a big set of actions by simply registering how a person utilizes the space around himself. Furthermore, some actions can even be recognized by considering only a few selected joints of the human body.

The rest of the paper is structured in the following manner. We start by introducing some related work that involves human posture or action recognition using the skeletal information of the Kinect sensor. Next we discuss some issues with the existing methods for action recognition. In Sect. 4 the skeletal data obtained with the Kinect sensor is explained. The section thereafter explains the different stages of our approach to human action recognition. The final section discusses the results of the experiments on the self-captured database for the game application. The paper is completed with our conclusions.

2 Related Work

Since the release of the Microsoft Kinect sensor, it has been extensively used for vision based human posture, gesture and action recognition as the user's skeletal information is accurately generated by the sensor from its depth images. The skeletal data acquired by the Kinect is usually transformed to extract important features. Common features are the locations, the angles and the velocities of the skeleton joints. The features are used to classify the human posture or action. General classifiers, such as Support Vector Machines (SVM), or template-matchers are commonly used for recognition of static gestures (i.e. postures,

sitting, lying, standing,...). Dynamic gestures however (i.e. running, walking, jumping,...), include a temporal dimension and are typically handled by Hidden Markov Models (HMM) or motion based models. Some existing methods for human posture and action recognition are here summarily explained.

The method by Zhang et al. [11] utilizes the Kinect skeletal information to generate 9 normalized vectors representing different body parts. With an optimized SVM they identify 22 pre-defined postures. Through principle component analysis (PCA) they found that in the reduced feature space of the three main orthogonal principle components, most body postures were well separated from each other. An accuracy of 99.14% was achieved.

Gahlot et al. used spherical angles and angular velocities for action recognition [4]. The pose of the subject is first estimated by three joints near the torso. The angles and velocities are then computed in reference to the torso joint. Horizontal symmetry is incorporated through a motion energy based method to account for actions performed by either the left or right side of the body. The classification is finally obtained with an individual HMM for every supported action. Standing, sitting and bending is classified with an accuracy of 90%.

Hussein et al. [5] use covariance matrices from 3D joint locations to classify human actions. To account for the temporal dimension, multiple covariance matrices over sub-sequences of the action are employed in a hierarchical manner. Normalization of the joint locations is used to make the method scale invariant. An off-the-shelf SVM classifier was used to validate the descriptor on three different datasets with accuracy ranging from 90.5% to 95.4%.

The sign base recognition method of Martinez et al. [6] is based on body parts relations and consists of two parallel phases: the recognition based on hand postures and based on hand gestures. The posture recognition works on basis of bag-of-words representation constructed from shape context descriptors. Classification is achieved with a multiclass SVM classifier. The gesture recognition is based on the relations between the hand and the rest of the joints. A visual dictionary is constructed and used to form the word sequence that represents a gesture. HMMs for each sign class are used for classification.

Lastly we discuss our earlier work that was performed for the exergaming project [3]. We used normalized skeleton joint locations to classify every body pose with a Random Forest Classifier. The idea behind this was that different actions show sufficient different body poses to be distinguishable from each other. A sliding window with the classification results of the last several consecutive frames decides on the final classification by a majority voting scheme. The strategy obtained an accuracy of 96.7% on the exergame dataset and an accuracy of 98.3% on the Microsoft Research Cambridge-12 Kinect dataset.

The classification methods all show promising results, but every method is focused on different features and works only on specific poses and actions. The space that is occupied during an action has not yet been analyzed. This research is based on the idea that this space contains important features that are not yet used to classify human action and poses. That is why we propose to use the occupied space as an addition for human action recognition.

3 Existing Issues with Human Action Recognition

There are a few issues that are not all handled by the existing solutions. The features that are extracted from the skeletal data should ideally be insensitive to small variations in a persons appearance, action execution and camera viewpoint. At the same time, the features must be sufficiently distinctive to allow for robust classification. Some methods are sensitive to the general location of the subject in relation to the sensor. A related problem is that generally the subject has to be facing the Kinect sensor, making it view-dependent. A last issue is the scale-invariance. The features have to be independent from the person's dimensions.

The classification method itself can also introduce some issues. By nature, posture recognition methods can only classify static actions. Dynamic actions however contain a temporal dimension that cannot be handled by the posture recognition systems. A much used solution for this problem is the HMM. However, the construction of a HMM is a complex and time consuming task. A last issue that we bring forward is real-time recognition. Many recognition approaches only work after the entire action is performed. Other designs need the starting and end point of the action. With online recognition, the actions can begin at any time and recognition has to be achieved as soon as possible.

The approach that is introduced in this paper tries to account for most of the problems brought forward with an additional feature that describes the space that is occupied during the human action performance.

4 Kinect Skeletal Data

We use the skeletal data that the Kinect acquires as input for our action recognition. The skeletal data is delivered at 30 fps and consists of 25 joint positions and orientations. These joints are head, neck, spine shoulder, spine mid, spine base and for both left and right: shoulder, elbow, wrist, hand, thumb, hand tip, hip, knee, ankle and foot. The joints are connected to each other through bones with each a parent joint and a child joint. The parent joint is the end joint of the bone that is closest to the reference joint, usually the spine base. For example, the upper arm bone has the shoulder joint as parent joint and the elbow joint as child joint. The elbow joint is in turn the parent joint of the lower arm bone.

Not every joint is considered for every BROM. We denote $SD = \{(\mathbf{p}_i, \mathbf{q}_i), i = 0, \ldots, n\}$ as the selected skeletal data for a body or body part in a frame with, for joint i, $\mathbf{p}_i = (x, y, z)$ the cartesian camera coordinates and $\mathbf{q}_i = (a, b, c, d)$ the orientation relative to the camera system and n the number of joints.

5 Human Action Recognition

5.1 Pose Invariant Normalization

Before the relative occupancy map can be built, the skeletal data has to be normalized. The variability across subjects in height and other dimensions has

to be eliminated to accurately classify the movements of the individual. A too large difference in body part sizes between different individuals will disturb the classifier performance. Therefore every (partial) skeleton will be transformed to a skeleton model with standard dimensions.

Since action recognition must be independent of absolute positions, a joint is chosen as reference point. The entire skeleton is re-positioned and rotated so the reference joint lies at the origin and its orientation coincides with the axes.

This normalization process is done in several phases. Given a $SD = \{(\mathbf{p}_i, \mathbf{q}_i)\}$ with joint $(\mathbf{p}_{\text{ref}}, \mathbf{q}_{\text{ref}})$ as the reference joint, the normalized skeletal data is then denoted by $N(SD) = \{(\mathbf{x}_i, \mathbf{r}_i);\ i = 1, \ldots, n\}$ with $\mathbf{x}_i$ the normalized cartesian coordinates and $\mathbf{r}_i$ the quaternion of the normalized orientation of joint i.

First, the joints are translated so the reference joint lies in the origin. The translated skeletal data is:

$$SD_{\text{trans}} = (\mathbf{p}_i - \mathbf{p}_{\text{ref}}, \mathbf{q}_i). \tag{1}$$

The next step is the rotation of the translated skeletal data with the conjugate of the orientation of the reference joint: ([1])

$$SD_{\text{rot}} = ((\mathbf{p}_i - \mathbf{p}_{\text{ref}})\mathbf{q}^*_{\text{ref}}, \mathbf{q}_i\mathbf{q}^*_{\text{ref}}) = (\mathbf{p}_{i,\text{rot}}, \mathbf{q}_{i,\text{rot}}). \tag{2}$$

In the last phase, the dimensions of the skeleton are normalized to the dimensions of a standard skeleton model. The standard model is based on the mid-sized male aviator [8].

$$N(SD) = \left(\frac{\mathbf{p}_{i,\text{rot}} - \mathbf{p}_{\text{parent}_i,\text{rot}}}{||\mathbf{p}_{i,\text{rot}} - \mathbf{p}_{\text{parent}_i,\text{rot}}||} l_i + \mathbf{x}_{\text{parent}_i}, \mathbf{q}_{i,\text{rot}} \right) = (\mathbf{x}_i, \mathbf{r}_i), \tag{3}$$

where l_i is the length of the limb that has $\mathbf{p}_i$ as child joint in the standard model skeleton. As can be seen in (3), the normalized position of the parent joint has to be known to calculate the normalized position of a joint. The order for transforming the joints is therefore crucial: parent joints before child joints. The very first joints to be transformed, are the child joints of the reference joint, since the normalized position of the reference joint is by definition the origin.

The skeleton is now normalized to a template that not only removes the individual body differences, but also the differences in the global positioning and orientation of the body. The normalization process makes it therefore possible to represent both adults and children with the same model despite the sizable differences in body shape. This is illustrated in Fig. 1 where the skeleton of a nine-year old child is compared to the skeleton of an adult before, during and after the pose invariant normalization process. Both subjects are standing in the T-pose while looking in different directions (Fig. 1a). In Fig. 1b the skeletons were translated and rotated in reference to their own spine base joint. The orientation

[1] The product of two quaternions is called the Hamilton product. The product of a vector and a quaternion is calculated by using the quaternion representation of the vector and then taking the Hamilton product of the two quaternions.

of the reference joint now coincides with the axes of the coordinate space. After remodeling the skeletons to the standard sized skeleton, the T-poses are clearly similar (Fig. 1c).

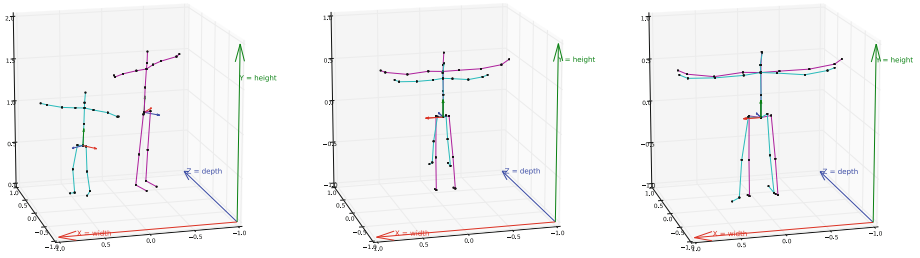

Fig. 1. Overview of the skeleton normalization process for a child (cyan) vs. an adult (magenta). The orientation of the reference joint, the spine base, is depicted by its x, y and z axes. (a) The original skeletons, (b) the skeletons in reference to the spine base joint, (c) the skeleton after the complete pose invariant normalization process.

5.2 Body Related Occupancy Map

The purpose of the body related occupancy map (BROM), is to keep track of which areas in the space around the skeleton were occupied during the action performance and how much of the time they were occupied.

The BROM can be constructed relative to a room, a specific object in space, a specific joint of a skeleton, or even relative to a body part that belongs to another person present in the room. The proposed method is broad enough to handle each of these cases. The skeletal data is re-oriented in reference to the chosen point. The process is essentially the same as the first two steps of the normalization of the skeletal data. If the reference point is the same as the reference joint in the normalization phase, the re-orientation can be skipped. If the entire skeleton is analyzed, the optimal reference is the spine base joint.

To build the BROM, we first divide the space around the skeleton into adjoining cells according to a rectangular grid, a spherical grid or a cylindrical grid depending on the application. For the test case in this paper we opted for dividing the space into cells by means of a rectangular grid. The reference point of the occupancy map always lies in the center cell of the grid. The cells form the foundation of the occupancy map that is filled frame by frame.

For every frame, the occupation of the 3D cells is considered. A cell is occupied if a joint from the skeletal data lies within its boundaries. Counters in the BROM for the current frame keep track of the number of joints that occupy every cell. The BROMs for every frame of the performed action are then added together and divided by the total number of frames for normalization:

$$BROM_f(u,v,w) = \sum_{x_i \in c_{u,v,w}} 1, \text{with } i = 0,\ldots,n. \tag{4}$$

$$BROM(u,v,w) = \frac{1}{F}\sum_{f=1}^{F} BROM_f(u,v,w), \tag{5}$$

where $BROM_f(u,v,w)$ is the occupancy counter for cell (u,v,w) for frame f, $c_{u,v,w}$ the 3D space corresponding to the grid cell (u,v,w) in the reference space, F the number of frames in the sequence and $BROM$ the occupancy map for the sequence. The values in the complete BROM thus represent the occupancy level of the cells by the human skeleton joints during the action.

The BROM can be partitioned into slices to make visualization as a 2D projection of the map possible. The front view (FV), side view (SV) and top view (TV) of the occupancy map are obtained by adding the 2D slices together and normalizing the resulting 2D grid. Only the formula for the FV is given as the SV and TV are similarly calculated.

$$FV(u,v) = \frac{1}{W}\sum_{w} BROM(u,v,w) \tag{6}$$

Examples of the FV projection of a BROM are shown in Fig. 2 for the action "flying with big arm moves". Due to the normalization of the skeletons, the FVs for a child and for an adult are indistinguishable from each other. It is also clear from the FVs that the action is primarily performed by the arms.

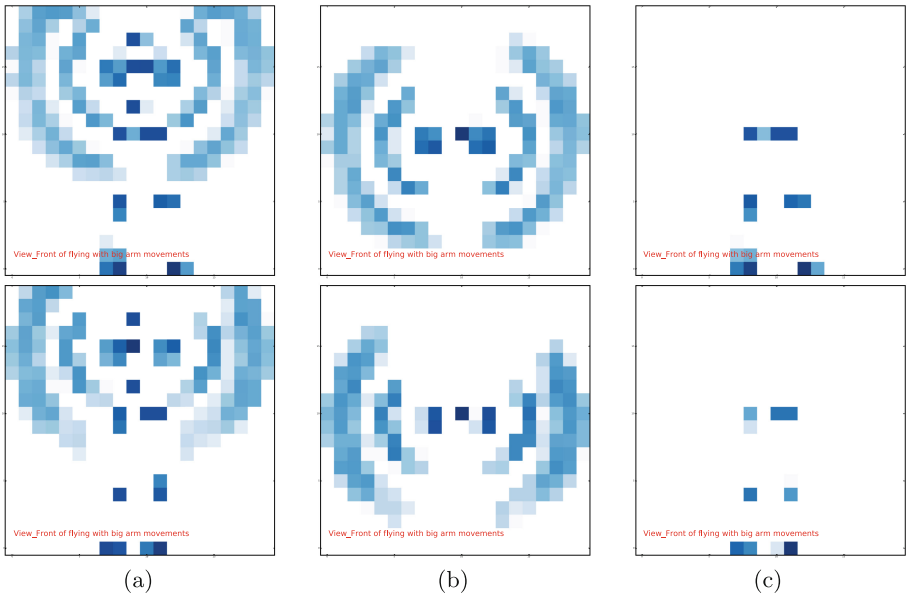

Fig. 2. Front view projections of the BROMs during the action "flying with big arm moves" for a child (top) and an adult (bottom). (a) FV for the entire skeleton, (b) for the arms, (c) for the legs.

6 Results

6.1 Evaluation Method

The proposed strategy is evaluated on the self-captured dataset for the exergame application. The actions are characterized by the sequence of frames from their start frame until their end frame. The BROM for the actions are built from the frame sequences.

The BROM discretizes the 3D cartesian coordinate space into $20 \times 20 \times 20$ bins with linear spacing. The entire map contains a volume of $2\,\mathrm{m} \times 2\,\mathrm{m} \times 2\,\mathrm{m}$ with the reference point in the center. Each bin thus has a resolution of 10 cm.

We have opted for random forest (RF) as the classification model for the BROMs. A RF, is a low-bias, low-variance ensemble classifier, trained with bagging and random feature selection. It has been proven that RF's are almost invariant to overfitting and are robust against noise.

The classifiers are trained on the leave-one-subject-out principle [1]. The data of all the subjects except one are used as training data for the classifier. The classifier is then evaluated on the data of the subject left out. This process is repeated for every subject in the database. Moreover, every experiment is repeated ten times to account for outliers that may result from the randomness in the RF classifiers. The overall evaluation of the classification method is the average of the results. The F1-score is used to evaluate each classifier.

6.2 Action Dataset

The action dataset was recorded with the Microsoft Kinect v2 sensor at the Sportlab of the department of Movement and Sport Sciences at Ghent University in Belgium. It consists of four male and one female subjects performing 22 specific actions three times with in between the neutral action standing still with arms along the body (24256 samples). The entire action list be found in Table 2.

6.3 Whole Body Classification Versus Body Part Classification

The classification performance of several BROM constructions are compared against each other. In the first experiment, the entire action set is classified by BROMs constructed from the complete skeleton body and BROMs constructed from different body parts. Table 1 gives an overview of the different body parts with their reference joint and included joints that are used in the BROMs.

The left side of Table 2 shows the resulting F1-scores per action. It is clear that the BROMs of the partial skeletons are not enough on their own for a reliable classification of the entire action set. But it gives a first indication of which actions are best classified by which body part. To make it more clear, the F1-scores that are greater than or equal to 0.75 were highlighted in green. The F1-scores between 0.5 and 0.75 were highlighted in orange. The classification of the BROM with the complete skeleton, results in a F1-score of an acceptable 0.80.

Table 1. Different body parts with their reference joint and included joints.

Body part	Reference joint	Included joints
All body	Spine_base	Spine_mid, Spine_shoulder, Neck, Head, Shoulder_left, Elbow_left, Wrist_left, Hand_left, Shoulder_right, Elbow_right, Wrist_right, Hand_right, Hip_left, Knee_left, Ankle_left, Foot_left, Hip_right, Knee_right, Ankle_right, Foot_right
Left arm	Shoulder_left	Elbow_left, Wrist_left, Hand_left
Right arm	Shoulder_right	Elbow_right, Wrist_right, Hand_right
Arms	Spine_shoulder	Shoulder_left, Elbow_left, Wrist_left, Hand_left, Shoulder_right, Elbow_right, Wrist_right, Hand_right
Left leg	Hip_left	Knee_left, Ankle_left, Foot_left
Right leg	Hip_right	Knee_right, Ankle_right, Foot_right
Legs	Spine_base	Hip_left, Knee_left, Ankle_left, Foot_left, Hip_right, Knee_right, Ankle_right, Foot_right

Table 2. F1-scores for the classification of the BROMs of different body parts for the complete action set (left) and for different action subsets (right).

	Complete action set							Subsets of actions					
	all body	left arm	right arm	arms	left leg	right leg	legs	left arm	right arm	arms	left leg	right leg	legs
neutral	**0.89**	0.74	0.70	0.74	0.72	0.68	**0.87**	**0.99**	**0.98**	**0.98**	**0.91**	**0.90**	**0.90**
walking	0.56	0.00	0.02	0.01	0.33	0.46	0.69				0.31	0.46	0.61
running	0.48	0.21	0.13	0.20	0.56	0.08	0.43				0.44	0.14	0.40
step left	**0.75**	0.03	0.05	0.07	0.62	0.41	0.65				0.62	0.53	0.68
step right	0.72	0.17	0.00	0.00	0.67	0.31	0.53				0.64	0.32	0.60
bowing	0.57	0.27	0.25	0.35	0.59	0.51	0.69				0.60	0.52	0.61
bow left	**0.81**	0.15	0.45	0.52	**0.78**	0.36	0.65				**0.77**	0.36	0.67
bow right	**0.81**	0.08	0.06	0.35	0.19	0.62	0.74				0.21	0.61	0.74
little jump	0.15	0.04	0.02	0.00	0.20	0.20	0.19				0.17	0.18	0.33
big jump	0.68	0.33	0.15	0.16	0.69	0.65	**0.75**				0.73	0.66	0.71
little jump hands up	0.39	0.47	0.41	0.35	0.13	0.22	0.31	0.49	0.41	0.40	0.19	0.28	0.37
big jump hands up	**0.82**	0.67	0.65	0.65	0.14	0.35	0.25	0.71	0.61	0.71	0.19	0.31	0.27
climbing	**0.79**	0.61	0.51	0.74	0.24	0.20	0.52	0.62	0.55	0.72			
hummingbird	**0.94**	0.67	**0.89**	**0.88**	0.02	0.00	0.01	0.72	**0.85**	**0.90**			
flying small moves	**0.90**	**0.76**	**0.86**	**0.92**	0.09	0.04	0.02	**0.81**	**0.86**	**0.89**			
flying big moves	**0.88**	**0.84**	**0.84**	**0.88**	0.00	0.00	0.01	**0.84**	**0.82**	**0.87**			
punch left	**0.84**	0.59	0.02	**0.80**	0.00	0.18	0.01	0.61		**0.79**			
punch right	**0.92**	0.23	0.73	**0.87**	0.00	0.00	0.03		0.72	**0.88**			
pushing forward	**0.85**	0.73	0.62	**0.84**	0.40	0.11	0.01	**0.77**	0.68	**0.90**			
high kick left	0.25	0.29	0.02	0.02	0.42	0.04	0.25				0.46		0.42
high kick right	0.48	0.00	0.34	0.04	0.55	0.68	0.65					0.71	**0.79**
low kick left	0.50	0.04	0.07	0.10	0.62	0.09	0.27				0.66		0.47
low kick right	0.71	0.01	0.15	0.01	0.11	**0.78**	0.63					0.77	0.74
total	**0.80**	0.56	0.51	0.62	0.55	0.49	0.62	**0.91**	**0.90**	**0.93**	**0.76**	0.73	**0.78**

Most of the mislabeled samples (57%) are classified as the neutral action. Especially the little jump shows this behavior as it is almost always misclassified as the neutral action. This is however not surprising as this action consists of only a slight bending of the knees as seen from the spine_center joint. 26% of the mislabeled samples are classified as an action that has a highly similar use of the space around the subject. Walking and running are two such actions that are frequently switched. Unfortunately, the biggest difference between these two actions is the speed of execution which is lost in the construction of the BROM.

This was an intentional choice in the design of the BROM because normalizing on the execution time, meant that, for example, slow flying and fast flying are both classified as flying.

The computational time for creating the BROMs of the entire body took 54 s for the complete dataset (in Python on a Intel Core i7 processor). This averages to 2.23 ms/sample or 83.17 ms/BROM. A notable 94% of this time was spent on the pose invariant normalization of the skeletons. The computational time was reduced to 0.42 ms/sample or 15.74 ms/BROM when the BROM was created for only the leg or the arm.

6.4 Action Subclasses

In the second experiment we investigate how the classification with the BROMs of the body parts performs if every body part has its own subset of actions assigned. The action set for each body part is derived from the previous experiment and is assigned according to the main body part that is used to perform the action. For example, "flying" is mainly performed by the arms. The action is therefore only classified by the BROMs of the left/right arm and both arms.

The right side of Table 2 gives an overview of the actions that were classified by the BROMs of each body part and the resulting F1-scores. A remarkable improvement in performance of the body part classifiers can be made if they are classified on their own action subset. The only actions that are still difficult to classify are running, little jump, little jump hands up and high kick left.

6.5 Combining Body Part Classifiers

The third experiment consists of combining the predictions of the body part BROM classifiers into one general prediction. There are two obvious ways to combine the classifiers. The first and simplest combination is by taking the majority vote of the separate classifiers. The second method takes the prediction probabilities for every action from the separate classifiers, adds them up and takes the action with the highest probability as the final classification.

An extra rule is implemented when the final prediction gives the neutral action. In the previous experiment, the body part classifiers were also tested on untrained actions. With very few exceptions, the untrained actions were classified as the neutral action. This was a welcome result and it means that the classifiers give either the correct classification or the neutral class back. The extra rule is based on this finding. If the final classification is the neutral action, the second most probable class is returned instead, unless all classifiers are unanimous.

The experiment proves that the classification performance decreases when the body is divided into separate body parts as shown in Table 3. The smaller the different body parts, the more the performance of the general classifier suffers. Also, the method of majority voting gains only significance if the number of classifiers in the combination increases.

In addition, this experiment points us in the direction that a hierarchic classification is necessary. The whole body BROM can be used to distinguish if the

Table 3. F1-scores for the combined classifiers of BROMs of different body parts.

Classifiers	Majority	Probability
Combination of 2 *Arms, Legs*	0.77	0.79
Combination of 4 *Left arm, Right arm, Left leg, Right leg*	0.68	0.67
Combination of 6 *Left arm, Right arm, Arms, Left leg, Right leg, Legs*	0.74	0.65

action is mainly performed by the arms, the legs or both. The second level in the hierarchy then only uses the BROM of the main body parts. This way the BROM of the arms is only used if there is sufficient indication of arm action.

6.6 2D View Classification

The BROM results in a large feature vector for the classifier, the last experiment studies if this can be reduced through only taking the 2D views of the BROM. Because the classifier for the whole skeleton proved to be the best classifier, only the views of this BROM is tested. Table 4 shows the results.

Table 4. Classification with the complete BROM versus classification with the 2D views of the BROM.

	Total F1-score
Complete BROM	0.80
Front view of BROM	0.81
Side view of BROM	0.74
Top view of BROM	0.74
Front, side and top view of BROM	0.84

Reducing the BROM to its 2D views, increases the performance of the classifier if the correct combination is chosen. Using the front view of the BROM raises the F1-score by 1%. However, using the side view or the top view of the BROM, reduces the performance by 6%. The best result is obtained when the three views together are used as the feature input of the classifier. The F1-score is then boosted by 4% to a value of 0.84.

7 Conclusion

In this paper a new feature for human action classification and analysis was presented. Skeletal data recorded with the Microsoft Kinect v2 sensor was used

as the initial input data. The skeletal data was first rotated and translated to a reference point to make the framework camera view-independent. To account for individual body differences between people, the skeleton was then transformed to a standard model. From the transformed skeleton, the entire skeleton or a body part was selected to build a body related occupancy map. The reference point formed the center of the occupancy map and all other joints were in relation to this point. The BROM was built by binning the relative positions from the joints in a 3D grid for every frame from the performed action and normalizing the resulting map. The BROM and the top, side and front views were each used to train a Random Forest classifier to predict the action that was performed.

Four experiments were done on the self-captured action dataset to test the approach. In the first experiment a classification result of an 0.80 F1-score was achieved for the BROM from the complete skeleton with the spine base joint as reference. The second experiment showed that better classification results could be achieved on the BROMs of different body parts if each body part was trained on their own specialized subset of the actions. The classification scores increased with 16% to 42%. The third experiment determined that the classifiers for the body parts trained on their own subset of actions, could be combined to form a general classifier that can predict the complete action set. The last experiment proves that decreasing the feature vector size can increase the F1-score to 0.84 if the top view, side view and front view is used as input for the classifier.

With the BROMs it is possible to largely classify a diversified set of human actions by only looking at the space that the human uses during the action performance. The BROMs can therefore be used as a great additional feature for human action classification and analyzation.

References

1. Arlot, S., Celisse, A., et al.: A survey of cross-validation procedures for model selection. Statistics surveys **4**, 40–79 (2010)
2. Ballan, L., Bertini, M., Del Bimbo, A., Seidenari, L., Serra, G.: Event detection and recognition for semantic annotation of video. Multimedia Tools Appl. **51**(1), 279–302 (2011). https://doi.org/10.1007/s11042-010-0643-7
3. Deboeverie, F., Roegiers, S., Allebosch, G., Veelaert, P., Philips, W.: Human gesture classi cation by brute-force machine learning for exergaming in physiotherapy. In: Computional Intelligence and Games 2016, CIG 2016 (2016)
4. Gahlot, A., Agarwal, P., Agarwal, A., Singh, V., Gautam, A.K.: Article: Skeleton based human action recognition using kinect. In: IJCA Proceedings on Recent Trends in Future Prospective in Engineering and Management Technology, RTFEM 2016, vol. 1, pp. 9–13, July 2016. Full text available
5. Hussein, M.E., Torki, M., Gowayyed, M.A., El-Saban, M.: Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations. IJCAI. **13**, 2466–2472 (2013)
6. Martínez-Camarena, M., Oramas M.J., Tuytelaars, T.: Towards sign language recognition based on body parts relations. In: ICIP (2015)
7. Muro-De-La-Herran, A., Garcia-Zapirain, B., Mendez-Zorrilla, A.: Gait analysis methods: an overview of wearable and non-wearable systems, highlighting clinical applications. Sensors **14**(2), 3362–3394 (2014)

8. Project, Y.S.O.A.R., Army, U.S., Force, U.S.A., Navy, U.S., Committee, T.S.A.R.P.T.S.: Anthropometry and Mass Distribution for Human Analogues: Military male aviators. Anthropology Research Project (1988). https://books.google.be/books?id=ZxRNXwAACAAJ
9. Rautaray, S.S., Agrawal, A.: Vision based hand gesture recognition for human computer interaction: a survey. Artif. Intell. Rev. **43**(1), 1–54 (2015). https://doi.org/10.1007/s10462-012-9356-9
10. Vishwakarma, S., Agrawal, A.: A survey on activity recognition and behavior understanding in video surveillance. Vis. Comput. **29**(10), 983–1009 (2013). https://doi.org/10.1007/s00371-012-0752-6
11. Zhang, Z., Liu, Y., Li, A., Wang, M.: A novel method for user-defined human posture recognition using kinect. In: 7th International Congress on Image and Signal Processing, CISP 2014, pp. 736–740. IEEE (2014)