09-22

15th IAPR International Conference on Machine Vision Applications (MVA)
Nagoya University, Nagoya, Japan, May 8-12, 2017.

Human Bodypart Classification Using Geodesic Descriptors and Random Forests

Sebastian Handrich, Ayoub Al-Hamadi, Erik Lilienblum and Zuofeng Liu
Otto-von-Guericke-University
Magdeburg, Germany
sebastian.handrich@ovgu.de

Abstract

A new approach to classify human body parts in depth images is proposed. The approach is based on geodesic descriptors. Such a descriptor randomly samples the local geodesic neighborhood of each depth pixel. During a training phase, a random forest classifier learns the correct body part from these descriptors. The experimental evaluation shows that we can robustly classify 19 body parts for several different poses and body proportions. We further compare our approach and the classification based on geodesic distance features to those that were used in previous works.

1 Introduction

The robust estimation of human poses has a wide range of applications like human-computer interaction, gesture and action recognition, but is still a challenging task as the human body is capable of an enormous range of poses. Furthermore, different body sizes and -proportions have to be taken into account. Over the last two decades, many different approaches were proposed. They can be distinguished e.g. by whether 3d information is available or the approach is pure image based. Another criterion is whether the approach employs any kind of a priori knowledge, for example in the form of pretrained classifiers or pose databases to which an observed pose is tried to be matched. Further distinctions can be made by whether the approach attempts to estimate the complete pose in form of joint positions or whether the different body parts are to be detected. Image based methods use features like silhouettes [3], skin colour or contours [2], but often lack the ability to resolve ambiguities, e.g. like self-occlusions. Approaches using databases are often restricted to previously trained poses. In contrast, methods without any prior knowledge [1] can estimate arbitrary human poses but rely on an exact feature point extraction. In [4], the authors used geodesic distances to segment human body parts and optical flow was used to solve for ambiguities during occlusions. The authors in [6] matched the positions of detected geodesic feature points to a predefined pose database. In his break-through paper [5], Shotton utilizes randomly sampled depth differences and random forests to classify the body part of a depth pixel and also to determine the joint positions of a skeleton model. A similar approach was made by the authors in [8] who used regression forests to learn the direction towards a skeleton joints based on depth difference features. A graph based method was presented in [7] which determines the pose from volumetric 3d data. And a recent approach was presented in [9], where deep convolutional networks were used to determine human poses in 2d images. In this work, we propose a new approach based on depth images that utilizes local geodesic distance based features to classify the body part to which a depth pixel belongs. We thus follow the suggestions made in [5] and replace the simple depth based features by more complex ones which contain a much higher information value per pixel. This enables us to robustly classify the body parts of a person in several poses.

2 Training Data

For training the classifiers, the correct body part of each depth pixel has to be known. Generating this ground truth data from real depth data recordings is very subjective to the labeling person and practically not possible. Following the approach made in [5], we use synthetic data. Next to the existence of reliable ground truth data, a major benefit is that several parameters like body proportions, camera parameters and noise can be varied and their influence on the pose estimation results can be examined without the need of multiple recordings. We use animated 3d character models taken from MakeHuman to create synthetic depth images. To obtain pose data, i.e. sequences of skeleton joint positions, a human actor performed several poses in front of a Kinect sensor. In our rendering tool, the 3d mesh is then animated using linear skinning and rendered into a depth buffer. This takes into account that a camera only *sees* the frontal face of the human body and that are hidden body parts due to self occlusions. Different 3d characters were used to simulate different body proportions (Fig. 1a). For each pose, there is a also map $\mathbf{B_k}$ that stores the body part of each depth pixel. The 19 used body parts are $\mathcal{B}$ = {*shoulders(2), feet(2), hands(2), upper arms(2), lower arms(2), upper legs(2), lower legs(2), head(1), neck(1), hip(1), torso(1), belly(1)*} (Fig. 1b). Some exemplary poses are shown in Fig. 2.

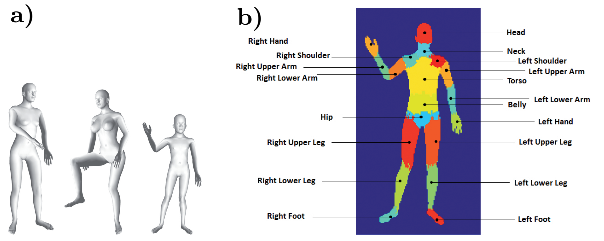

Figure 1: **a)** Different 3d models used to create a set of articulated human poses. **b)** Ground truth body parts (19). For each depth pixel the body part is known.

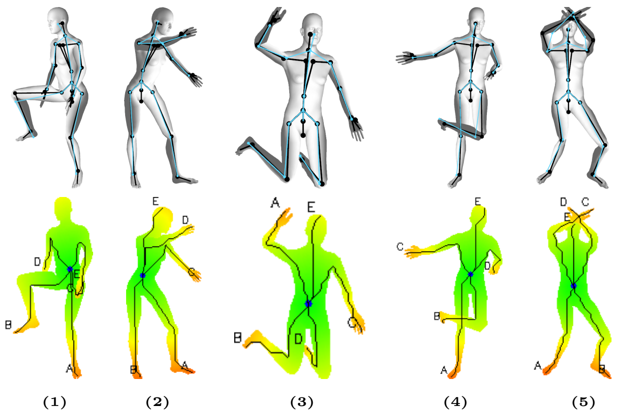

Figure 2: **Top:** Exemplary poses of the training data set. **Bottom:** Results of the geodesic distance calculation.

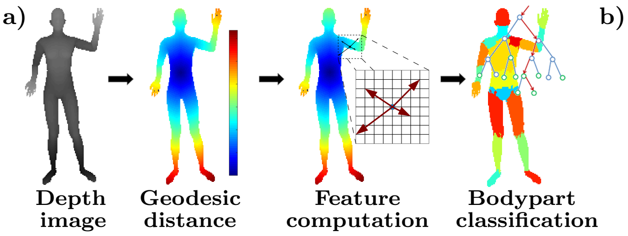

Figure 3: **a)** Overview of the proposed approach.

3 Bodypart Classification

A brief overview of the proposed approach is provided in Fig. 3. Our goal is to determine for each pixel of a depth image $\mathbf{D_k}$ the body part $b_j \in \mathcal{B}$. The first step is to compute the geodesic distances to the torso center. This results in a geodesic distance map $\mathbf{G_k}$. From $\mathbf{G_k}$, a set of feature vectors $\mathbf{\Gamma}$ (*geodesic descriptors*) describing the local geodesic neighborhood of a depth pixel is extracted. Based on those descriptors and a pre-trained Random Forest, the body part is classified in a final step.

Geodesic distances: The first step is to transform the depth image $\mathbf{D_k}$ into a 3d point cloud $\mathbf{W_k}$, based on a pinhole camera model $\boldsymbol{C} = \{c_x, c_y, c_k\}$, in which $[c_x, c_y]^T$ denotes the camera's principal point and c_k is the camera constant. $\mathbf{W_k}$ is an organized point cloud, i.e. for each 3d point $\mathbf{w}_i$ there is a corresponding 2d point $\mathbf{w}'_i$ in the depth image. The point cloud is then described as a weighted graph $G = (\mathbf{W_k}, E)$. Two 3d points are connected by an edge, if either an edge criterion $\mathcal{C}_1$ or $\mathcal{C}_2$ is fulfilled. The former connects two nodes if their euclidean distance is below a threshold ϵ_{c1} and if they correspond to adjacent pixels in the depth image (Eq. 1). The second one (Eq. 2) connects two nodes, if all depth image pixels between them have a lower depth value, i.e. are closer to the camera:

$$\mathcal{C}_1(i,j) = ||\mathbf{w_i} - \mathbf{w_j}||_2 \leq \epsilon_{c1} \wedge s(\mathbf{w'_i}, \mathbf{w'_j}) \leq 1, \quad (1)$$

$$\mathcal{C}_2(i,j) = d_m < \min(d_i, d_j), \forall \mathbf{w_m} \in \overline{\mathbf{w'_i w'_j}}, \quad (2)$$

where $s(\,\cdot\,)$ is the spatial 2d distance and $d_m()$ is the depth value of point $\mathbf{w'_m}$. To each edge, a weight $\omega(E)$ is assigned, which is the Euclidean distance between the connected nodes. Using Dijkstra's algorithm, the shortest possible connection (path $\mathcal{P}$) between two arbitrary 3d points $\mathbf{a}, \mathbf{b} \in \mathbf{W_k}$ is determined. The geodesic distance (Eq. 3) between $\mathbf{a}$ and $\mathbf{b}$ is then given by the cumulative weight of all edges that are element of the path.

$$g(\mathbf{a}, \mathbf{b}) = \sum_{\mathcal{E} \in \mathcal{P}(\mathbf{a}, \mathbf{b})} w(\mathcal{E}) \quad (3)$$

In Fig. 2 (bottom) some results of the geodesic distance calculation are shown. The depicted color represents the geodesic distance from the point cloud center, ranging from green (low distances) to red (high distances). It can be seen that the geodesic distance are also computed correctly when the body is occluded by a limb. The second criterion is crucial for a correct distance calculation as it prevents the fragmentation of the graph into several isolated graph segments in the case of self occlusion. Despite its simplicity, the used method is superior to other approaches, in which the fragmentation problem is either completely neglected [4] or separated graph segments were connected based on their Euclidean distance.

Geodesic descriptors: We now describe the features that are used as input to the body part classification. The main assumption is that each body part can be classified by analyzing its local geodesic neighborhood, i.e. that a pixel in a shoulder *sees* another set of geodesic distances in its 2d neighborhood as a pixel located within the torso. Let $\mathbf{x_i}$ denote the 2d position of a depth pixel. Its feature vector $\mathbf{F}_i$ is composed of a subset of feature vectors that we call geodesic descriptors (Eq. 4). Each descriptor contains a set of geodesic distances that were randomly sampled from the geodesic distance map $\mathbf{G_k}$ (Eq.5).

$$\mathbf{F}_i = \{\Gamma_j\}_{j=1 \dots N_t} \quad (4)$$

$$\Gamma_j = \frac{\{g^j_m\}_{m=1 \dots N_f}}{g_{\max}}, \text{ with} \quad (5)$$

$$g^j_m = \mathbf{G_k}(\mathbf{x_i} + \vec{\mathbf{u}}^j_m), \forall m = 1 \dots N_f \quad (6)$$

Here, $\vec{\mathbf{u}}^j_m$ is a 2d offset vector that is randomly sampled from the 3d interval $I_w = [(-N_w, -N_w, 0) \dots (+N_w, +N_w, 0)]$ and mapped to 2d using the pinhole camera model and the actual depth value at $\mathbf{x_i}$. For the sake of clarification, we want to emphasize that the offset vectors $\vec{\mathbf{u}}^j_m$ are randomly sampled for each geodesic descriptor - and not for each pixel. So, each descriptor is based on its own set of offset vectors.

There are three free parameters in our approach: (1)N_t is the number of geodesic descriptors Γ that are evaluated per pixel; (ii) N_f is the number of offset vectors $\vec{\mathbf{u}}^j$ in each geodesic descriptor and (iii) N_w determines the maximum 3d size of a geodesic neighborhood. In this work, we are particularly interested in the analysis of these parameters and their effect on the classification performance (see section 4).

Furthermore, we want to note that the features used in our approach are similar to those in [5]. Instead of geodesic distances, their features are composed of a set of depth differences at randomly sampled offset positions (Eq.7). Whereas the use of simple depth differences was strongly motivated by their computational efficiency, the use of geodesic distances is motivated by the fact that they contain much more information per pixel that can be used to classify the body part. The authors of [5] were already aware of this and suggested to use more complex features like depth integrals or local descriptors.

$$f_m^j = \mathbf{D_k}(\mathbf{x_i} + \frac{\vec{\mathbf{u}}_m^j}{\mathbf{D_k}(\mathbf{x_i})}) - \mathbf{D_k}(\mathbf{x_i} + \frac{\vec{\mathbf{v}}_m^j}{\mathbf{D_k}(\mathbf{x_i})}) \quad (7)$$

Random Forest classification: A random forest was used as a classifier. There is one tree for each geodesic descriptor Γ_j, so that each tree learns the correct body part from a different sampled geodesic neighborhood. N_t is therefore also equivalent to the numbers of trees. The tree was trained using the CART algorithm ([10]). As we want our classifier to generalize, the maximum tree depth is limited to 9. The class voting is based on the posteriori class probability distributions $P_j(\mathcal{B}|\Gamma_j)$. These define for a given depth pixel and its observation Γ_j the probability to be in class $b_i \in \mathcal{B}$. We combine the probability distributions of all trees and assign to a given depth pixel the body part class that maximizes the combined probability distribution (Eq. 8).

$$b_i = \underset{b \,\in\, \mathcal{B}}{\arg\max} \sum_{j=1}^{N_f} P_j(\mathcal{B}|\Gamma_j) \quad (8)$$

4 Experimental Results

Several experiments were carried out, in which we evaluate the classification performance of the proposed approach and examine the impact of the three free parameters (section 3) on the classification results. The experiments were run on the training data (section 2) which contains approx. 1000 different poses and three different character models. For each depth image all steps described in section 3 were performed. All results shown here are 10-fold cross-validated. In each iteration, 90% of the complete training data were used to train the classifier and the remaining 10% were used for testing. This was repeated 10 times and in each iteration a different subset was used for testing.

As a performance measurement of the classifier, the precision $\mathbf{p_i}$ and recall values $\mathbf{r_i}$ per class were computed. In the confusion matrices shown, rows denote the actual classes and columns refer to the predicted ones. The numbers in the confusion matrices are normalized to the number of actual pixels in each class and thus are true positive rates. The last row in each confusion matrix is the precision per class.

In a first experiment, the number of trees N_t of the random forest classifier were varied from 1 to 40 and the other parameters were kept constant ($N_f = 5$, $N_w = 10$ cm). A subset of the confusion matrices C are shown in Fig. 5(a-d) and the mean precision $\overline{\mathbf{p}}$ and recall $\overline{\mathbf{r}}$ values per tree number are depicted in Fig. 4a. One can see that when only a small number of trees (1 or 2) are used, the classification performance is quite low (both recall and precision are between 0.61 and 0.67). The main reason for this is that the classifier is not able to model the subtle differences in the geodesic neighborhood and therefore can not distinguish between body parts that have a similar geodesic distance from the torso center. Using one tree (Fig. 5a), for example only 30% of the pixel that actually belong to the right lower arm (rlA) were correctly classified, but the same amount was determined to be part of the left lower arm (llA). The same is true for the right and left upper arm. As the tree number increases from 1 to 10, both the precision and recall significantly increase (Fig. 4a). One can see that there is however a clear cutoff number. When the tree number exceeds 10, the further improvement is very low (2% - 3%).

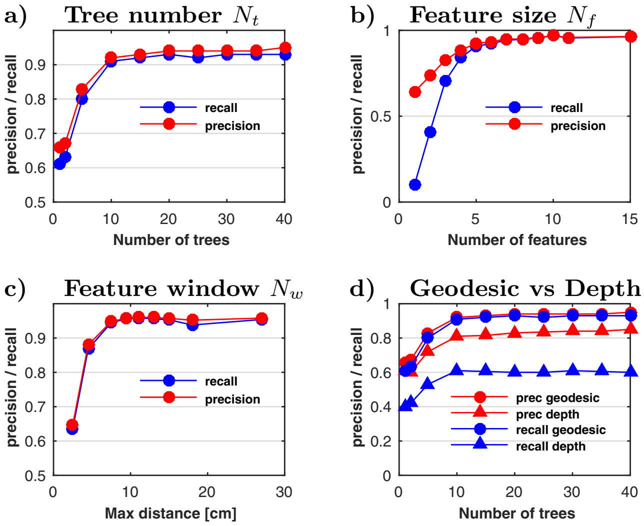

Figure 4: Performance of the body part classification (mean precision and recall values) for (a) different numbers of decision trees, (b) different geodesic descriptor sizes, (c) varied sizes of the local geodesic neighborhood. In (d), precision and recall compared to approach proposed in [5] are shown.

In a second experiment, we varied the size of each geodesic descriptor, N_f, i.e. the number of offset vectors (Fig. 4b and Fig. 5e-f) from 1 to 15. When only one offset vector is used, the classification performance is extremely low (mean recall $\overline{\mathbf{r}} = 0.1$). This is an expected result, as it means that body parts must be distinguished based on only one geodesic distance. There is, however, more than one body part with the same geodesic distance from the torso center. Again, it can be seen that there is a clear cutoff number of features ($N_f = 5$) at which any further improvement in both precision and recall is only very low. Similar experiments were performed for the maximum size the geodesic neighborhood N_w (Fig. 4c). The number of trees were set to 10, and also 10 features per geodesic descriptor were used. Using a distance that is too small results in a poor classification (approx. 0.63 in both recall and precision for $N_w = 2$ cm). We assume the reason for this is that such a small maximum distance is below the error of the geodesic distance computation. Based on those results, we set N_w to 10 cm in the following experiments. Finally, we compared our approach to the one described in [5]. Instead of geodesic based features (see Eq. 5 in section 3), the authors in [5] have used features that are based on local depth differences (Eq. 7). The results are shown in Fig. 4d. Using the depth based feature approach, the approximated mean recall on our data set was $\bar{\mathbf{r}} \approx 0.6$ and the mean precision value was $\bar{\mathbf{p}} \approx 0.8$. These results match the ones reported by Shotton et.al. Using the proposed approach both the recall and precision were approximately $\bar{\mathbf{p}} \approx \bar{\mathbf{r}} \approx 0.9$. Considering the low recall value of 0.6, the geodesic distance based approach clearly outperforms the one using depth differences. As already mentioned in the method section, this result is expected by both us and the authors of [5], since these features contain more information than simple depth differences.

5 Conclusion

We proposed a new approach to detect and classify human body parts in depth images. The approach determines for each depth image pixel to which body part it belongs. Following the suggestions made by Shotton et. al. we replace their depth difference based features by more complex ones that describe the local geodesic neighborhood. For this, the geodesic distances to the torso center are computed and randomly sampled in the 2d neighborhood of each depth pixel. We call these features a geodesic descriptor. A random forest classifier is then used to learn the correct body part from these descriptors. The presented experimental evaluation shows that we can robustly identify 19 human body parts in different poses as long as the the person is facing the camera ($< 60°$) and there is no other object between the user and the camera. We have further evaluated the classification performance based on both geodesic and depth difference based features. As expected, the geodesic based features contain much more information per pixel and led to an increased classification performance of $\approx 50\%$ (recall value). As such, the proposed approach can build the basis for a subsequent human pose estimation approach.

References

[1] G. Pons-Moll, A. Baak, T. Helten, M. Muller, H.-P. Seidel, and B. Rosenhahn: Multisensor-fusion for 3d full-body human motion capture. CVPR. 663-670, 2010

[2] L. Qinghua and M. Zhenjiang: Markerless human pose estimation using image features and extremal contour. ISPACS, 1–4, 2010

[3] D. Chen and C.B. Fookes: Labelled silhouettes for human pose estimation. ISSPA, 2010.

[4] L.A. Schwarz, A. Mkhitaryan, D. Mateus and N. Navab: Human skeleton tracking from depth data using geodesic distances and optical flow. Image Vision Comput., vol. 30, no. 3, pp. 217–226, 2012

[5] J. Shotton, A.F. Girshick, T.Sharp, M.Cook, M.Finocchio, R.Moore, P.Kohli, A.Criminisi, A.Kipman and A.Blake: Efficient human pose estimation from single depth images. IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 35, no. 12, pp. 2821–2840, 2013.

[6] A.Baak, M. Mller, G. Bharaj, H.-P. Seidel, C. Theobalt: A Data-Driven Approach for Real-Time Full Body Pose Reconstruction from a Depth Camera. Consumer Depth Cameras for Comp. Vis., Springer London, 71-98, 2013

[7] M. Straka, M, Rther, S. Hauswiesner and H. Bischof: Skeletal graph based human pose estimation in realtime. 69.1–69.12, 2011

[8] H.Y. Jung, S. Lee, Y.S. Heo and I.D.Y: Random Tree Walk toward Instantaneous 3D Human Pose Estimation. CVPR, 2467–2474, 2015

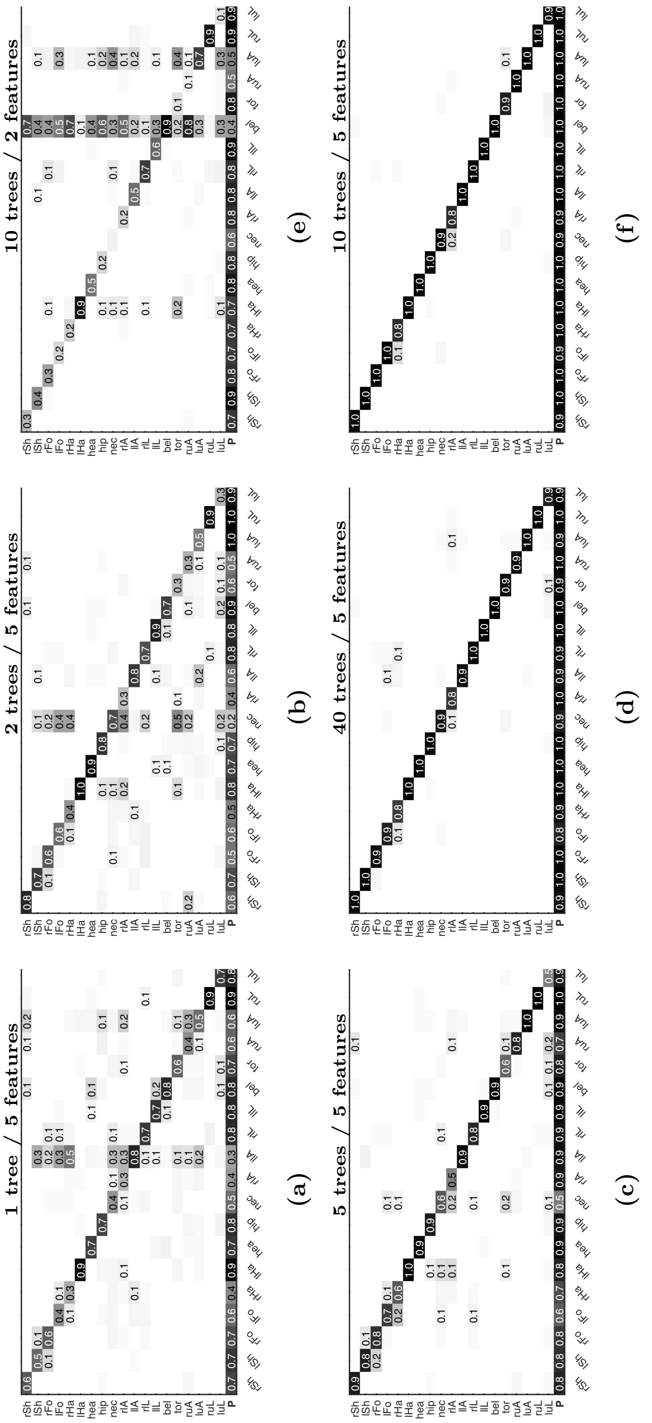

Figure 5: Confusion matrices of the body part classification (ten-fold cross-validated) with a random forest of 1(a), 2(b), 5(c) and 40(d) trees and five features per geodesic descriptor. (e) Result using 10 trees and 2 features. (f) Result using 10 trees and 5 features. Columns refer to the predicted classes. The numbers denote the true positive rates. Last row are precision values.

[9] V. Belagiannis and A. Zisserman: Recurrent Human Pose Estimation: CoRR abs/1605.02914, 2016.

[10] D. Coppersmith, S. J. Hong and J. R. M. Hosking: Partitioning Nominal Attributes in Decision Trees. Data Mining and Knowledge Discovery, vol. 3, 1999, 197-217.