Robust Geodesic Skeleton Estimation from Body Single Depth

Jaehwan Kim(✉) and Howon Kim

Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea
jh.kim@etri.re.kr

Abstract. In this paper, we introduce a novel and robust body pose estimation method with single depth image, whereby it is possible to provide the skeletal configuration of the body with significant accuracy even in the condition of severe body deformations. In order for the precise identification, we propose a novel feature descriptor based on a geodesic path over the body surface by accumulating sequence of characters correspond to the path vectors along body deformations, which is referred to as GPS (Geodesic Path Sequence). We also incorporate the length of each GPS into a joint entropy-based objective function representing both class and structural information, instead of the typical objective considering only class labels in training the random forest classifier. Furthermore, we exploit a skeleton matching method based on the geodesic extrema of the body, which enhances more robustness to joints misidentification. The proposed solutions yield more spatially accurate predictions for the body parts and skeletal joints. Numerical and visual experiments with our generated data confirm the usefulness of the method.

Keywords: 3D body parts classification · Joints identification
Skeleton estimation · Random forest · Geodesic descriptor
Joint entropy · Dynamic time warping

1 Introduction

3D body pose estimation, whose goal is to recover the poses of body parts and joints with naturally articulated movements, plays a key role and is a well investigated problem in variety of areas such as computational vision, human-computer interface, and computer animations, and so on. Especially, in the works of Shotton et al. [1,2], random forest algorithm proposed by [3] is employed to predict body poses from single depth image. The random forest is an ensemble learning method, which has proven fast and effective multi-class classifiers for various works such as image classification, object tracking, facial expression recognition, pose estimation and so on [4–6]. The solution proposed by Shotton et al. [1] is embedded within the commercial product 'Microsoft Kinect sensor^{TM}', which is readily available off-the-shelf gaming system. Moreover, the depth comparison proposed by [1] is popularly used in many works [2,7] as learning features

J. Blanc-Talon et al. (Eds.): ACIVS 2018, LNCS 11182, pp. 342–353, 2018.
https://doi.org/10.1007/978-3-030-01449-0_29

for the random forest classifier. Within the framework of body pose estimation based on the classified body parts [1], the accuracy and reliability of the body parts classification are important because they might influence the consequent learning process to infer the positions of 3D body joints. Furthermore, although the depth comparison features proposed by [1] are easy to compute and efficient in characterizing the change in body parts, the features themselves encode the only local information for the body parts not a global information such as the deformed whole body or the skeletal structure of the body joints. The depth comparison features are insufficient to empower the discriminative ability of the classifier.

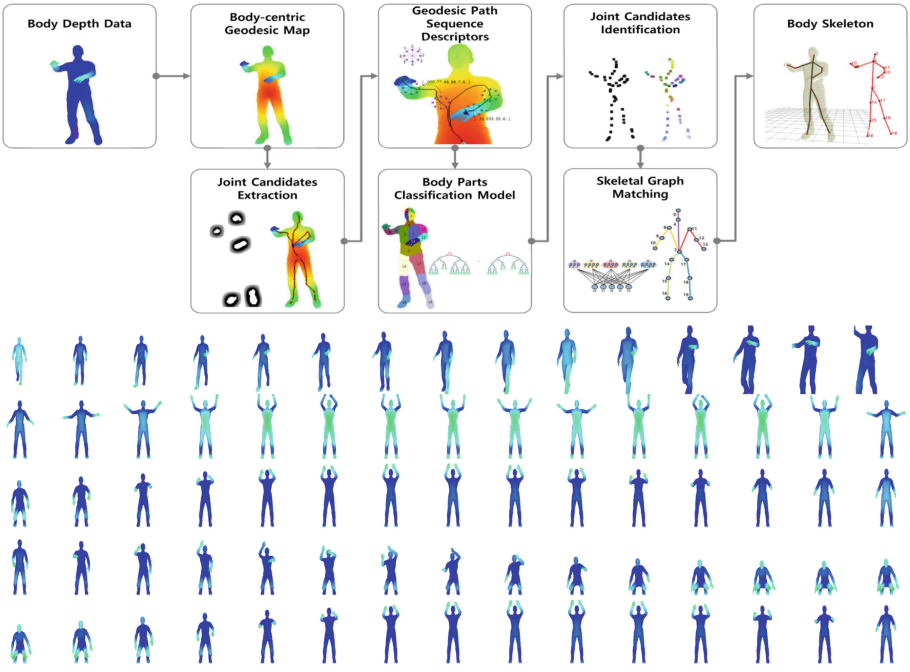


Fig. 1. Systematic overview of our system and our ground-truth samples (normalized to the depth [0,1]): from the top row, forward walking (T2), hand waving1 (T3.a), hand waving2 (T3.b), sitting (T4), and upstanding (T5) motions.

Our approach for 3D body pose estimation from a single depth-map is related to the previous works from [8,9] as they exploit a geodesic distance graph of the body depth image to localize the skeletal joints of the body.

In works [10], a variety of objective functions with the geodesic distance transforms based features for identifying interest objects in the semantic image segmentation with random forest. Moreover, in the context of a decision forest, a joint objective function for pixel classification and shape regression is

introduced in [11], which yields more spatially consistent predictions than results from the typical objective function only considering the data labels.

Motivated by existing works [1,8,11], we propose a new feature descriptor based on a geodesic path over the body surface, referred to as GPS (Geodesic Path Sequence), which is derived by concatenating sequence of characters correspond to the vectors along deformation paths. In order for the body parts classification, we also incorporate the length of each GPS descriptor into the joint entropy-based objective function involving both the body parts labels themselves and their geodesic structural information, leading to more accurate predictions. The geodesic descriptors reflect a geometry of body surface well, which is expected to improve our body parts classification performance. In addition, we exploit a skeleton matching method based on the geodesic extrema of the body, thereby reducing the misidentification problems for the joints and their bones in the skeletal configuration. As with the step in [1], we develop a ground-truth generator and cheaply create varied realistic data by synthesizing an avatar 3D body model with some interesting poses sampled from a large motion capture data set, which consists of five different motions: standing (T1), walking (T2), hand waving1 (T3.a), hand waving2 (T3.b), sitting (T4), and upstanding (T5) (see Fig. 1, samples similar to standing (T1) set are included in the other sets). In this paper, our final goal is to predict an accurate skeletal configuration of the body pose rather than the standard anatomic positions of the body joints.

2 Geodesic Path Sequence Descriptor

We show how well our GPS provides significant patterns with discriminative information across anatomically different body parts, through the empirical comparison of affinity matrices derived from two different types of features (i.e., our GPS and depth values). Then, we describe how to incorporate our GPS descriptors into the joint entropy-based objective in learning the random forest classifier.

In order to take the human body manifold structure into account, we exploit the geodesic distances and their paths among all points over the body surface and a their barycenter point as feature descriptors for the random forest classifier. At first, we construct an undirected weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ from the body points set $\{\boldsymbol{p}_{\boldsymbol{x}}\} \subseteq \mathcal{V}$, where $\mathcal{V}$ and $\mathcal{E}$ denote a set of vertices and a set of edges with pairwise distances being assigned as edge weights, and each $\boldsymbol{p}_{\boldsymbol{x}_i}$ is a 3D position vector consisting of a 2D coordinate $\boldsymbol{x}_i$ and its depth $d_D(\boldsymbol{x}_i)$ in the body depth image. The set of edges are defined as:

$$\mathcal{E} = \{d_E(\boldsymbol{p}_{\boldsymbol{x}_i}, \boldsymbol{p}_{\boldsymbol{x}_j}) \in \mathcal{V} \times \mathcal{V} \mid (||\boldsymbol{p}_{\boldsymbol{x}_i} - \boldsymbol{p}_{\boldsymbol{x}_j}||_2 < \delta) \wedge (||\boldsymbol{x}_i - \boldsymbol{x}_j||_\infty \leq 1)\}, \quad (1)$$

Each edge $d_E(\boldsymbol{p}_{\boldsymbol{x}_i}, \boldsymbol{p}_{\boldsymbol{x}_j}) \in \mathcal{E}$ is stored as a weight $w(d_E)$, where a 3D Euclidean distance of less than δ. The Dijkstra geodesic distance d_G is then

computed along the shortest path $\mathcal{P}$ between $\boldsymbol{p}_{\boldsymbol{x}_p}$ and $\boldsymbol{p}_{\boldsymbol{x}_q}$, which is defined as:

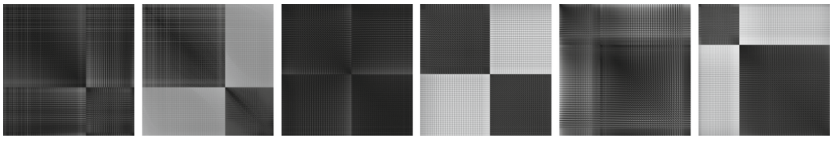
$$d_G(\boldsymbol{p}_{\boldsymbol{x}_p}, \boldsymbol{p}_{\boldsymbol{x}_q}) = \sum_{d_E \in \mathcal{P}(\boldsymbol{p}_{\boldsymbol{x}_p}, \boldsymbol{p}_{\boldsymbol{x}_q})} w(d_E) \tag{2}$$

The graph based geodesic descriptors are invariant to large motion deformations and geometric transforms as long as the local connection relationships remain, which well reflect the local body structure [8,12]. We then generate a body-centric geodesic map by measuring the Dijkstra geodesic distances for all N points on the body, $\{d_G(\boldsymbol{p}_{\boldsymbol{x}_i}, \boldsymbol{p}_{\boldsymbol{x}_0})\}_{i=1}^{N}$. Each Dijkstra geodesic distance, $d_G(\boldsymbol{p}_{\boldsymbol{x}_i}, \boldsymbol{p}_{\boldsymbol{x}_0})$, is associated with the sum of edge weights along a shortest path between a point, $\boldsymbol{x}_i$, over the body surface and a barycenter of the body, $\boldsymbol{x}_0$, under an assumption that points on anatomically similar body parts maintain a nearly constant geodesic distance. From the body-centric geodesic map, we finally define a descriptor based on the geodesic path which is represented by accumulating sequences of characters correspond to the vectors along the body's deformation path. The GPS for a point $\boldsymbol{x}_i$ is defined as:

$$\boldsymbol{d}_g(\boldsymbol{x}_i) = [c_1, c_2, \cdots, c_i], \tag{3}$$

where c_i is a character indicating the direction of between $\boldsymbol{p}_{\boldsymbol{x}_{i-1}}$ and $\boldsymbol{p}_{\boldsymbol{x}_i}$.

Dynamic time warping (DTW) is a powerful algorithm for measuring similarity between two time series by finding an optimal alignment. In here, we employ the fast DTW algorithm [13] in order to compute the similarity between two GPS descriptors in the binary test function of random forest within linear time. Fig. 2 shows that affinities between the inter- and intra- body parts for data aligned in the parts. All distance values for the affinities are normalized between 0 and 1. The more the affinity matrix has well-formed block diagonal structure, the better the partitioning of different parts. As shown in Fig. 2, the simple depth comparison features empirically do not provide enough discriminative power in learning the classifier. In case of two points having similar depth values, but located at different parts of the body, the features likely lead to erroneous predictions in the classification problem. Meanwhile, our proposed GPS is robust to large motion deformation, and it is effectively discriminative for different body parts.



(a)l.hand(13)-r.hand(10) (b)r.leg(15)-r.chest(5) (c)l.foot(19)-l.thigh(17)

Fig. 2. From the left, each pair of affinity matrices are depth-based and GPS-based similarities between two different body parts for data in Fig. 1's overview.

3 Joint Entropy-Based Body Parts Classification

For formulation of the body parts classification from single depth image, we assume that a set of N training samples $\mathcal{Q} = \{(\boldsymbol{f}_{\theta i}, \boldsymbol{l}_i)\}_{i=1}^{N}$ is given. The input variable $\boldsymbol{f}_{\theta i}$ corresponds to a feature for an individual pixel $\boldsymbol{x}_i$. The output variable is a discrete label $\boldsymbol{l}_i \in \mathcal{C}$, where $\mathcal{C}$ is a finite set of body labels.

In a given pixel $\boldsymbol{x}$ in depth image D, we propose a GPS comparison feature similar to the existing depth comparison feature [1], which is defined as:

$$\boldsymbol{f}_\theta(D, \boldsymbol{x}) = d_W\left(\boldsymbol{d}_g\left(\boldsymbol{x} + \frac{\boldsymbol{i}}{|\boldsymbol{d}_g(\boldsymbol{x})|}\right), \boldsymbol{d}_g\left(\boldsymbol{x} + \frac{\boldsymbol{j}}{|\boldsymbol{d}_g(\boldsymbol{x})|}\right)\right), \tag{4}$$

where $d_W(\boldsymbol{d}_g(\boldsymbol{x}_{\theta i}), \boldsymbol{d}_g(\boldsymbol{x}_{\theta j}))$ is a warp path distance between $\boldsymbol{d}_g(\boldsymbol{x}_{\theta i})$ and $\boldsymbol{d}_g(\boldsymbol{x}_{\theta j})$ descriptors. $\theta = (\boldsymbol{i}, \boldsymbol{j})$ is a pair of offsets to the pixel $\boldsymbol{x}$, and the scale invariance of depth is considered through the normalized by the length of $\boldsymbol{d}_g(\boldsymbol{x})$. Each node in tree is trained over a set of splitting candidates $\phi = \{(\theta, \tau)\}$, where feature parameter θ and partition threshold τ. The split candidates ϕ are randomly sampled from uniform distribution. For each ϕ ($m = |\phi|$), the subsets $\mathcal{Q}_L$ and $\mathcal{Q}_R$ partitioned from the original set of data $\mathcal{Q}$ are evaluated with our various energy functions at the current node. The partitioning is performed as follows:

$$\begin{aligned} \mathcal{Q}_L(\phi) &= \{(D, \boldsymbol{x}) \mid \boldsymbol{f}_\theta(D, \boldsymbol{x}) < \tau\} \\ \mathcal{Q}_R(\phi) &= \mathcal{Q} \setminus \mathcal{Q}_L(\phi) \end{aligned} \tag{5}$$

For the forest training procedure, the goal is to find optimal splitting parameters of each node and build partitioning binary tree which minimizes the objective function $\mathcal{J}$ defined as follows:

$$\phi^* = \arg\min_\phi \mathcal{J}(\mathcal{Q}, \phi). \tag{6}$$

An optimal criteria $\phi^* = \{\theta^*, \tau^*\}$ is defined as the split parameters of the node, and later used for prediction of new input data. The entropy is the expected value of the information contained in each message. The Shannon's entropy is generally used for training forests. Our goal is now to learn the joint probability $p_t(\boldsymbol{l}, \boldsymbol{g}|\boldsymbol{f}_\theta)$, where new variable $\boldsymbol{g} \in \mathbb{R}^3$ is a continuous regression variable for describing the relative 2D offsets between the depth pixel $\boldsymbol{x}$ and a barycenter of the body $\boldsymbol{x}_0$, and the geodesic distance $d_G(\boldsymbol{p}_{\boldsymbol{x}}, \boldsymbol{p}_{\boldsymbol{x}_0})$. By using the chain rule, we rewrite the joint distribution as $p_t(\boldsymbol{l}, \boldsymbol{g}|\boldsymbol{f}_\theta) = p_t(\boldsymbol{l}|\boldsymbol{f}_\theta)p_t(\boldsymbol{g}|\boldsymbol{f}_\theta, \boldsymbol{l})$, where we assume that $p_t(\boldsymbol{g}|\boldsymbol{f}_\theta, \boldsymbol{l})$ is a multivariate normal distributions. That is, $p_t(\boldsymbol{g}|\boldsymbol{f}_\theta, \boldsymbol{l}) \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{g}|\boldsymbol{l}}, \boldsymbol{\Sigma}_{\boldsymbol{g}|\boldsymbol{l}}|\boldsymbol{g}, \boldsymbol{f}_\theta, \boldsymbol{l})$ is one distribution per class label $\boldsymbol{l}$. We

actually define the joint objective function $\mathcal{J}$ as follows:

$$\mathcal{J}(\mathcal{Q},\phi) = \sum_{p\in\{L,R\}}\sum_{\boldsymbol{x}\in\mathcal{Q}_p}\frac{|\mathcal{Q}_p|}{|\mathcal{Q}|}\psi_E(\boldsymbol{l},\boldsymbol{g};\mathcal{Q}_p), \tag{7}$$

$$\begin{aligned}\psi_E(\boldsymbol{l},\boldsymbol{g};\mathcal{Q}_p) &= -\sum_{\boldsymbol{l}\in\mathcal{C}}\int_{\boldsymbol{g}\in\mathbb{R}^3} p_t(\boldsymbol{l},\boldsymbol{g}|\boldsymbol{f}_\theta)\ \log(p_t(\boldsymbol{l},\boldsymbol{g}|\boldsymbol{f}_\theta))d\boldsymbol{g},\\ &= \underbrace{-\sum_{\boldsymbol{l}\in\mathcal{C}} p_t(\boldsymbol{l}|\boldsymbol{f}_\theta)\log(p_t(\boldsymbol{l}|\boldsymbol{f}_\theta))}_{\psi_E(\boldsymbol{l};\mathcal{Q}_p)}\\ &\quad \underbrace{+\sum_{\boldsymbol{l}\in\mathcal{C}} p_t(\boldsymbol{l}|\boldsymbol{f}_\theta)\Big(\frac{1}{2}\log((2\pi e)^3|\boldsymbol{\Sigma}_{\boldsymbol{g}|\boldsymbol{l}}|)\Big)}_{\psi_E(\boldsymbol{g};\mathcal{Q}_p|\boldsymbol{l})},\end{aligned} \tag{8}$$

where $|\boldsymbol{\Sigma}|$ denotes the determinant of a matrix.

Finally, the models of random forest are achieved by optimizing the joint objective function Eq. (8), including the conventional objective $\psi_E(\boldsymbol{l};\mathcal{Q}_p)$ for a discrete label $\boldsymbol{l}$ as well as the objective $\psi_E(\boldsymbol{g};\mathcal{Q}_p|\boldsymbol{l})$ for a continuous variable $\boldsymbol{g}$ given $\boldsymbol{f}_\theta$ and $\boldsymbol{l}$ variables. In here, the posterior that we are interested in is about the body parts classification. The overall prediction of the forest with T trees is estimated by averaging the individual predictions together and the output is predicted by inferring:

$$\boldsymbol{l}^* = \arg\max_{\boldsymbol{l}\in\mathcal{C}} p(\boldsymbol{l}|\boldsymbol{f}_\theta) = \arg\max_{\boldsymbol{l}\in\mathcal{C}}\frac{1}{T}\sum_{t=1}^{T} p_t(\boldsymbol{l}|\boldsymbol{f}_\theta). \tag{9}$$

4 Body Joints and Skeleton Identification

Given a body-centric geodesic map, as with the way in [8,9,12], the extreme points are computed by incrementally maximizing geodesic distances on the body surface. Based on the classified body parts and the geodesic paths between the body's barycenter and its geodesic extrema (i.e., end-nodes of the human skeletal graph), we localize and identify the joint candidates lying on the paths. The joint candidates are selected with $\angle(\overrightarrow{\boldsymbol{x}_{k-1}\boldsymbol{x}_k}, \overrightarrow{\boldsymbol{x}_k\boldsymbol{x}_{k+1}}) > \epsilon$). Here, $\angle(\overrightarrow{\boldsymbol{x}_{k-1}\boldsymbol{x}_k}, \overrightarrow{\boldsymbol{x}_k\boldsymbol{x}_{k+1}})$ is an angle between two vectors $\overrightarrow{\boldsymbol{x}_{k-1}\boldsymbol{x}_k}$ and $\overrightarrow{\boldsymbol{x}_k\boldsymbol{x}_{k+1}}$, where the three points $(\boldsymbol{x}_{k-1},\boldsymbol{x}_k,\boldsymbol{x}_{k+1})$ being around the point $\boldsymbol{x}_k$, are on the same GPS $\boldsymbol{d}_g(\boldsymbol{x}_k)$. ϵ is a threshold depending on the body type, and it is empirically set to about 30 in our experiments. After obtaining the joint candidates set $\{\boldsymbol{x}_i\}$, the representative label $\boldsymbol{l}^{'}$ is evaluated as Eq. (10) from the local window at each joint candidate, where the local patches are based on the already classified body parts. Fig. 3 describes the meta-examples generated at each step.

$$\boldsymbol{l}^{'}_{\boldsymbol{x}_i} = \arg\max_{\boldsymbol{l}^*_{\boldsymbol{u}}}\sum_{\boldsymbol{u}\in\mathcal{W}_{\boldsymbol{x}_i}}\sum_{\boldsymbol{l}\in\mathcal{C}}\delta(\boldsymbol{l}^*_{\boldsymbol{u}},\boldsymbol{l}), \tag{10}$$

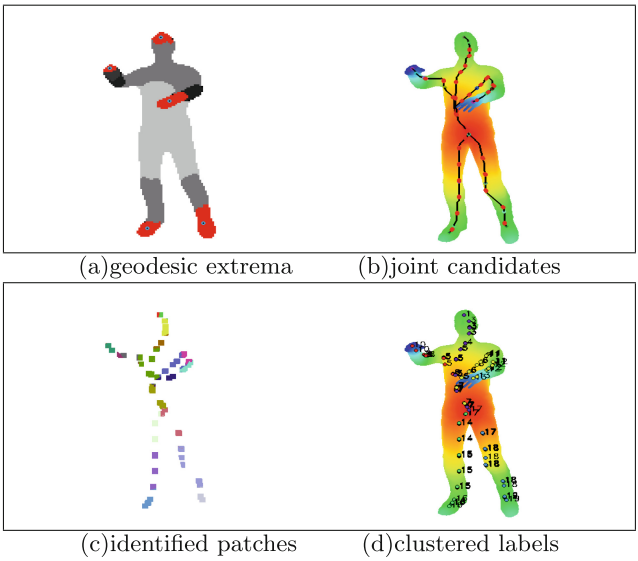
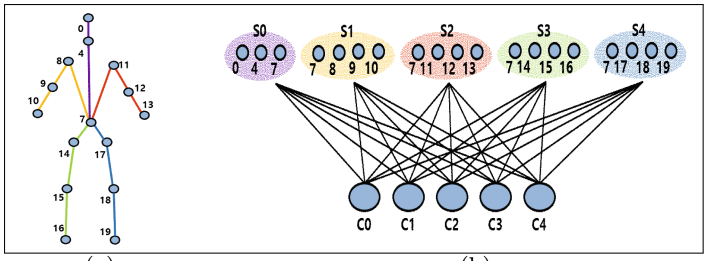


Fig. 3. (a) Geodesic extrema (blue points in red regions); (b) joint candidates (red points) on five GPSs (black lines); (c) color-labeled patches on joint candidates; (d) labels set $\{l^{'}\}$ classified into five sub-skeletons (i.e., each skeleton for four limbs and one trunk). (Color figure online)

where $\boldsymbol{l}_{\boldsymbol{u}}^{*}$ is the label for the position $\boldsymbol{u} \in \mathbb{R}^2$ being in the local window $\mathcal{W}_{\boldsymbol{x}_i}$ centered at $\boldsymbol{x}_i$. δ is a kronecker delta function. Our main idea is to match two graphs by comparing the labeled sets of ordered points on the paths between the body center and the geodesic extrema of the skeletal configuration under the assumption that there are meaningful joints for the human skeletal structure in the set of joint candidates. In here, the body center and the geodesic extrema labels are defined as 7 (center), 0 (head), 10 (right hand), 13 (left hand), 16 (right foot), and 19 (left foot), respectively. All joint candidates are identified and clustered as in Fig. 3(d) by matching with a given template graph as Fig. 4(a). In order to match the sub-skeletons (i.e., each skeleton for limbs and trunk) with the template graph, we consider a weighted bipartite graph such as illustrated in Fig. 4(b), which is with two vertex sets, a set of sub-skeleton labels and a set of joint labels, and the weight of each edge is defined as a DTW distance between two consecutive joint labels. Given the bipartite graph, the matching is performed by using the Hungarian method [14]. Finally, the skeletal graph with 15 labeled nodes is extracted, which correspond to the whole body skeleton (see Fig. 5(b)).

5 Numerical Experiments

We show the usefulness of our method, through the empirical comparison to different objective functions based on different types of features (i.e., our GPS



(a) (b)

Fig. 4. (a) Template skeleton model consisting of four limb sub-skeletons and one trunk sub-skeleton; (b) bipartite graph with two vertex sets (s#: set of joint labels for each sub-skeleton in the template; c#: set of candidate joint labels for each geodesic path).

and depth comparison feature [1]). We applied our method to samples from our ground-truth data sets, consisting of five types of motions: forward/backward walking, hand waving1, hand waving2, sitting, and standing; each motion group has approximately 100 frames. As in a conventional leave-one-out training scheme, the sequences for each model is evaluated with the trained model from other models. For quantitative evaluation of estimated joint positions and skeleton accuracy, we present three different types of measurements: (a) we estimate the mean absolute error (MAE) Eq. (11) in order for the training error evaluation of the classified body parts; (b) the mean average precision (mAP) is evaluated by averaging the precision of the estimated 15 joints on each frame, which is to determine whether the position of the estimated joint is within a given threshold relative to the ground-truth (in here, the threshold is fixed to $max(\{|\boldsymbol{s}_i|\}_{i=0}^{4})/10$); (c) the other is a new measurement of similarity between the estimated skeletons and the ground-truth skeleton by comparing their DTW score, which is referred to as mean average matching (mAM) and defined as Eq. (12).

$$MAE = \frac{1}{N} \sum_{i} \sum_{\boldsymbol{l} \in \mathcal{C}} |\boldsymbol{l}_i^* - \boldsymbol{l}_i^G| \in [0, 1], \tag{11}$$

where $\boldsymbol{l}_i^*$ and $\boldsymbol{l}_i^G$ represent the predicted label of data $\boldsymbol{x}_i$ and the corresponding ground-truth label, respectively.

$$AM(i) = \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} \frac{d_W(\boldsymbol{d}_g(\boldsymbol{s}_{fi}^*), \boldsymbol{d}_g(\boldsymbol{s}_{fi}^G))}{max(\{|\boldsymbol{s}_i|\}_{i=0}^{4})} \in [0, 1], \tag{12}$$

where $AM(i)$ is an average matching score between the estimated sub-skeleton $\boldsymbol{s}_i^*$ and the corresponding ground-truth sub-skeleton $\boldsymbol{s}_i^G$, one body skeleton has five sub-skeletons, $\{\boldsymbol{s}_i^*\}_{i=0}^{4}$, and the mean matching score is normalized by the maximum length of the five sub-skeletons. $\mathcal{F}$ is a set of target frames.

In [1,2], they apply a local mode-finding approach based on mean-shift with a weighted Gaussian kernel for each classified body part to infer the final positions of 3D skeletal joints. However, as shown in Fig. 5(a) and 6, the local modes

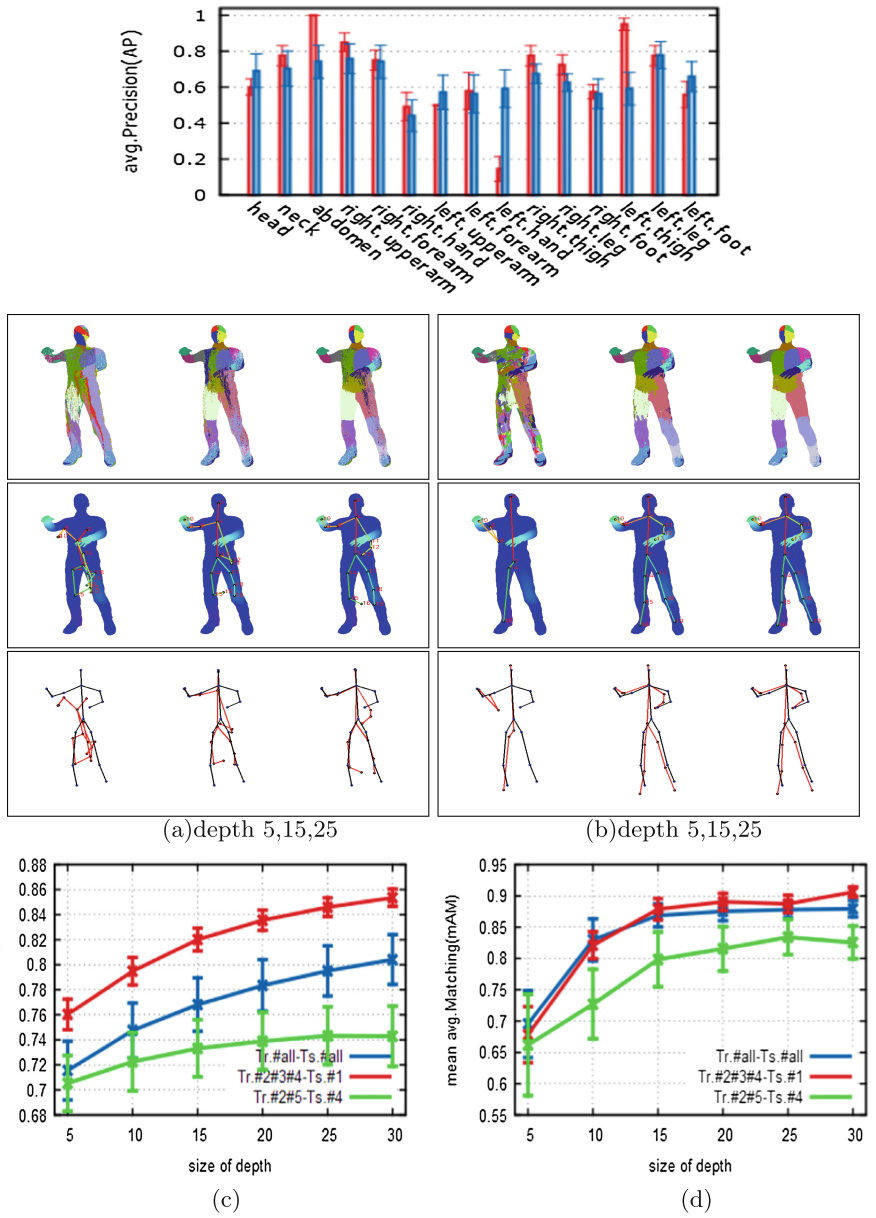


Fig. 5. Performance comparison: (top) APs for each joint at depth level 30, based on ours (blue) and [1] (red) with Training (T4)+Testing (T2, T5) sets; (mid) the predicted body parts for data in Fig. 1 and its skeletons with ground-truth (black lines); (bot) mAM values with different data sets, using [1] (a, c) and ours (b, d) methods, repectively. (Color figure online)

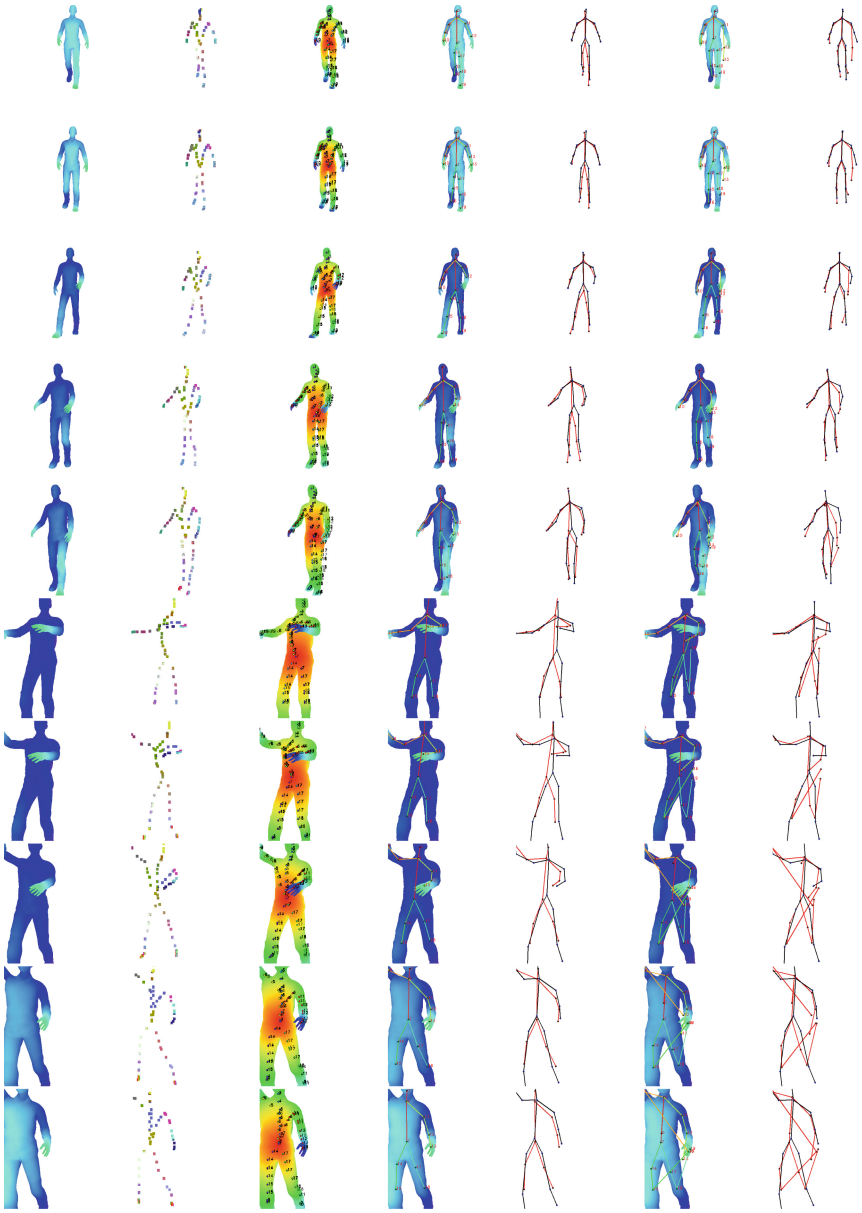


Fig. 6. Experimental results of estimated skeletons for forward-walking samples in Fig. 1: from the left, body depth, color-labeled patches on the joint candidates, identified & clustered joint candidates, body skeleton overlapping with the depth, skeletons from our method, and (the last two) skeletons from Shotton2011cvpr with ground-truth (black lines). (Color figure online)

Table 1. MAE, mAP, and mAM results for data sets used in Fig. 5(c, d) at depth level 30.

Method	$\psi_E(\boldsymbol{l}, \boldsymbol{g}; \mathcal{Q}_p)$+GPS	$\psi_E(\boldsymbol{l}; \mathcal{Q}_p)$+Depth
MAE	**0.0338**($\pm$0.0033)	0.0788($\pm$0.0047)
mAP	0.6463($\pm$0.0855)	**0.6700**($\pm$0.0513)
mAM	**0.8560**($\pm$0.0206)	0.8011($\pm$0.0218)

obtained from the misclassified outlying parts such as the left hand (label 13) and the right foot (label 16) cause failing skeleton results. Meanwhile, as shown in Fig. 5(b) and the MAE in Table 1, our method provides well-classified body parts as well as well-matched body skeletons through our GPS based-joint entropy and skeletal matching methods. As shown in Fig. 5 and Table 1, although our method is slightly less accurate than [1] in the mAP, our method offers the advantage of more accurately matching the body skeleton. In our method, the position of a target joint to be predicted depends on its GPS, geodesic distance, and the inclination angle between its neighboring joint candidate vectors on the GPS. Because of this assumption, it can be seen that the measured mAP value at the anatomical joint position reference is slightly lower. On the other hand, the mAM value makes us confirm that our proposed method well reflects not only the local features in the body depth data, but also the global structures in the skeletal configuration. Figure 6 shows a visual comparison of the predicted results.

6 Conclusions

We have presented a novel geodesic path sequence (GPS) descriptor, joint entropy-based objective with the GPS, and skeleton matching method for 3D body pose estimation based on the body parts classification, whereby it is possible to robustly predict the skeleton's position under severe body deformations. We also incorporate the GPS descriptor into a joint entropy-based objective function for learning both class and structural information about the body parts. Useful aspects of our proposed method could be summarized as follows: (a) The GPS descriptors can be widely used in variety of fields as a descriptor for deformable object representation; (b) The joint entropy objective function based on our GPS comparison features well reflects geodesic structural information over the body surface, leading to more accurate predictions in the random forest classifier; (c) The skeleton matching & identificaton based on the geodesic extrema of the body, which enhance more robustness to joints mis-identification. Empirical comparison with the conventional solution, single entropy-based objective with depth comparison features, confirmed the high performance of our method.

Acknowledgments. This research is supported by Ministry of Culture, Sports and Tourism (MCST) and Korea Creative Content Agency (KOCCA) in the Culture Technology (CT) Research & Development Program R2016030043.

References

1. Shotton, J., et al.: Real-time human pose recognition in parts from a single depth image. In: Cipolla, R., Battiato, S., Farinella, G. (eds.) Proceedings of International Conference on Computer Vision and Pattern Recognition, pp. 1297–1304. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-28661-2_5
2. Shotton, J., et al.: Efficient human pose estimation from single depth images. IEEE Trans. Pattern Anal. Mach. Intell. **35**, 2821–2840 (2013)
3. Breiman, L.: Random forests. J. Mach. Learn. **45**, 5–32 (2001)
4. Gall, J., Lempitsky, V.: Class-specific hough forests for object detection. In: Criminisi, A., Shotton, J. (eds.) Proceedings of International Conference on Computer Vision and Pattern Recognition, pp. 1022–1029. Springer, London (2009). https://doi.org/10.1007/978-1-4471-4929-3_11
5. Tan, D.J., Ilic, S.: Multi-forest tracker: a chameleon in tracking. In: Proceedings of International Conference on Computer Vision and Pattern Recognition, pp. 1202–1209 (2014)
6. Dapogny, A., Bailly, K., Dubuisson, S.: Pairwise conditional random forests for facial expression recognition. In: Proceedings of International Conference on Computer Vision and Pattern Recognition, pp. 3783–3791 (2015)
7. Girshick, R., Shotton, J., Kohli, P., Criminisi, A., Fitzgibbon, A.: Efficient regression of general-activity human poses from depth images. In: Proceedings of International Conference on Computer Vision, pp. 415–422 (2011)
8. Schwarz, L., Mkhitaryan, A., Mateus, D., Navab, N.: Estimating human 3d pose from time-of-flight images based on geodesic distances and optical flow. In: Proceedings of International Conference on Automatic Face and Gesture Recognition, Santa Barbara, CA, pp. 700–706 (2011)
9. Baak, A., Müller, M., Bharaj, G., Seidel, H., Theobalt, C.: A data-driven approach for real-time full body pose reconstruction from a depth camera. In: Proceedings of International Conference on Computer Vision, pp. 1092–1099 (2011)
10. Kontschieder, P., Kohli, P., Shotton, J., Criminisi, A.: GeoF: geodesic forests for learning coupled predictors. In: Proceedings of International Conference on Computer Vision and Pattern Recognition, pp. 65–72 (2013)
11. Glocker, B., Pauly, O., Konukoglu, E., Criminisi, A.: Joint classification-regression forests for spatially structured multi-object segmentation. In: Proceedings of European Conference on Computer Vision, Florence, Italy, pp. 870–881 (2012)
12. Plagemann, C., Ganapathi, V., Koller, D., Thrun, S.: Real-time identification and localization of body parts from depth images. In: Proceedings of International Conference on Robotics and Automation, pp. 3108–3113 (2010)
13. Salvador, S., Chan, P.: FastDTW: toward accurate dynamic time warping in linear time and space. In: KDD Workshop on Mining Temporal and Sequential Data, pp. 70–80 (2004)
14. Kuhn, H.: The hungarian method for the assignment problem. Nav. Res. Logist. Q. **2**, 83–97 (1955)