Learning Hierarchical Feature Representation in Depth Image

Yazhou Liu[1], Pongsak Lasang[2], Quansen Sun[1] and Mel Siegel[3]

[1]Nanjing University of Science and Technology
[2]Panasonic R&D Center Singapore
[3]Carnegie Mellon University
{yazhouliu, sunquansen}@njust.edu.cn
Pongsak.Lasang@sg.panasonic.com
mws@cmu.edu

Abstract. This paper presents a novel descriptor, geodesic invariant feature (GIF), for representing objects in depth images. Especially in the context of parts classification of articulated objects, it is capable of encoding the invariance of local structures effectively and efficiently. The contributions of this paper lie in our multi-level feature extraction hierarchy. (1) Low-level feature encodes the invariance to articulation. Geodesic gradient is introduced, which is covariant with the non-rigid deformation of objects and is utilized to rectify the feature extraction process. (2) Mid-level feature reduces the noise and improves the efficiency. With unsupervised clustering, the primitives of objects are changed from pixels to superpixels. The benefit is two-fold: firstly, superpixel reduces the effect of the noise introduced by depth sensors; secondly, the processing speed can be improved by a big margin. (3) High-level feature captures nonlinear dependencies between the dimensions. Deep network is utilized to discover the high-level feature representation. As the feature propagates towards the deeper layers of the network, the ability of the feature capturing the data's underlying regularities is improved. Comparisons with the state-of-the-art methods reveal the superiority of the proposed method.

1 Introduction

Photometric local descriptor [1] is one of the most powerful tools for image and video analysis. It has attracted extensive research efforts in recent years, and remarkable progress has been achieved [2–7]. Local descriptor encodes the micro-structure or statistical information of a region and generates a new description of it. Ideally, this description should be at least partially invariant to photometric (changes in brightness, contrast, saturation or color balance) and geometric (mainly affine transformations, like translations, rotations and scale changes) transforms [8]. Therefore, it has a wide variety of applications in the fields of object detection and classification [9], texture analysis [10], image retrieval [11, 12], object tracking [13, 14], and face recognition [15, 16].

Recently, with the rapid development of range sensors, such as Kinect and SwissRanger, 3D depth information can be readily obtained with low cost. These devices use either the structured light or time-of-flight (ToF) to measure the distance between objects and cameras. The images that captured by these devices are known as range/depth images. Since depth image resolves the distance ambiguities which exist in the photometric image, a large number of recent applications have emerged based on it. For instance, 3D scanning and reconstruction [17, 18], pose and action recognition[19–22].

Comparing depth images with photometric images, the differences exist in the following aspects.

1) *Geometrical structure.* Pixels in a depth image indicate calibrated depth in the scene, rather than a measure of intensity or color [23]. Therefore, depth images capture the geometrical structure information and resolve the depth ambiguities, which can greatly simplify some processing step such as background subtraction.

2) *Weak texture.* In the depth image, the color and texture variations induced by clothing, hair, and skin are not observable.

3) *High noise.* Comparing with the advanced photometric image sensors, the noise rates of depth sensors are relatively higher, especially in the environments of strong ambient light.

4) *Low resolution.* The lateral resolution of time-of-flight cameras is generally low compared to the standard 2D video cameras, with most commercially available devices at 320×240 pixels or less. Kinect claims its lateral resolution as 640×480.

With these essential differences, the well-developed local descriptors for the photometric images cannot be readily applied for the depth images. For example, scale invariant feature transform (SIFT) [7] descriptor and its variants [2] encode the gradient distribution with respect to the orientations within a local region. However, without photometric texture, the interest points with distinct local structure and statistics cannot be identified in depth images. Local binary pattern (LBP) descriptor [24] and its variants represent the statistics of micro structures of a region. For depth images, the meaningful structures only exist at the boundary regions of objects. Therefore, different parts within the objects cannot be differentiated successfully.

Finding a descriptor that can encode the local information of depth images effectively is the motivation and target of this paper. In order to achieve this target, the properties of depth images and their special targeting applications must be considered during the feature design process. The desirable properties of the depth descriptors are summarized as follows: Firstly, the descriptors in the depth should have some invariant properties which are preferable for the specific applications. Secondly, because the depth images are noisy and textureless, the units that being processed should be changed from pixels to some middle level representations, which might reduce the processing time and improve the robustness of the descriptor. Thirdly, in order to capture the nonlinear nature of the feature, some high level representations should be discovered to improve the discriminant of the feature.

The contribution of this paper lies in the multi-level feature extraction hierarchy for depth images. The above targeting properties have been addressed in the different levels of the hierarchy.

Low-level representation encodes the invariance to the articulate motion. An active research topic based on depth image is pose and action recognition [20–22]. In this context, most of the interested objects are non-rigid and consist of multiple parts, for instance, human/animal body contains torso and limbs, and human hands contain fingers. Parts of the objects are connected by joints and have multiple degree of freedoms (DOFs). The overall motion patterns of objects are articulate motion. It is desirable that the descriptors can provide consistent representations of the parts in different poses and gestures. The object is model as a map whose nodes correspond to the pixels on the object and whose edges represent the neighborhood relationship between the pixels. Based on this map model, the geodetic gradient is introduced to rectify the feature extraction which is supposed to be invariant to the articulate motion as long as the local connection relationship does not change during the motion.

Mid-level representation reduces the computation cost. By unsupervised clustering, the pixels of the depth image are grouped into clusters according to their 3D positions in the scene. And the clustering result is referred to as superpixel representation, which is used to replace the rigid structure of the pixel grid. On the one hand, this representation provides a convenient primitive from which to compute image features and greatly reduce the complexity of subsequent image processing tasks [25]. Especially for the texture-less depth image, in which only the pixels on the boundaries of objects have the distinct structural information, the superpixel based representation reduces the image redundancy dramatically. On the other hand, since the noise level of depth sensors is higher than photometric sensors, superpixels can suppress the noise effect of the individual pixels and yield a more robust representation.

High-level representation captures the nonlinear information. Because of the complexity of data distribution, nonlinear mapping which maps the data from their original space to some latent feature space is used to achieve better discriminant. Deep learning [26–29] is utilized to find high order dependency between the dimensions of the feature, which has been successfully applied in many fields including computer vision [30–32], natural language processing and speech recognition [33–35]. In this work, the employed deep network is based on stacked denoising autoencoders (SdA), as the feature evolved towards the deeper layers of SdA, two desirable properties for classification have been observed: sparsity and better discrimination.

The overall processing flow of the method is shown in Fig.1. The rest of the paper is structured as follows. Section 2 introduces the proposed geodesic invariance feature (GIF) in detail and Section 3 extends GIF to the superpixel based mid-level representation. Section 4 presents a deep learning based method for feature mining. Experimental results and comparisons are provided in Section 5. Finally, we conclude this work in Section 6.

2 Geodesic Invariant Feature (GIF)

The proposed method is inspired by the early works [3, 23, 21] which use the binary strings generated by pairwise comparisons to represent a local patch. In [3], the comparisons are based on the intensity of pixels and in [23, 21] the comparisons are based on the depth values of pixels. The difference of the proposed method is that the geodesic gradient is used to rectify the pairwise comparisons process, which endows the feature with the robustness to the articulate motion.

Given a depth image I, $I(p)$ represents the pixels depth value at position $p = (x, y)^T$. $D_{C_{p_c,r},F}$ denotes a local descriptor which determined by two parameters: coverage $C_{p_c,r}$ and feature list F. $C_{p_c,r}$ represents the coverage region of a local descriptor within the image, where p_c is the center and r is the radius of the coverage region. $F = \{P_i, \ldots, P_n\}$ denotes the list of random feature pairs where $P_i = (p^i_u, p^i_v)$ is a pair of random positions and n is the number of position pairs.

The simple comparison function is defined as follows:

$$\tau(p_u, p_v) = \begin{cases} 1, & \text{if } |I(p_u) - I(p_v)| > t \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where (p_u, p_v) is a random pair from the list F and t is the comparison threshold.

By applying the comparison function $\tau(\cdot)$ to the feature list F, a binary string $f \in \{0, 1\}^n$ is obtained and serves as the feature vector of the descriptor $D_{C_{p_c,r},F}$ as shown in Fig.2(a)∼(b). In [23, 21], f is used to train the randomized decision forests and regression forests.

Distance Invariance. In order to make a descriptor be invariant to distance variation, the coverage of the descriptor should be constant in the *real world* space. Based on the knowledge of projection geometry, the radius r of feature coverage $C_{p_c,r}$ on the image should be defined as:

$$r = \frac{\alpha}{I(p_c)} \tag{2}$$

where $I(p_c)$ is the depth value of center pixel p_c, and α is a constant determined by the size of the coverage in the *real world* space and imaging focus. Intuitively, this equation tells that if the object is closer to the camera, the size of the descriptor on the image should become larger and vice versa.

Articulation Invariance. In order to endow the descriptor with the articulation invariance, we introduce the geodesic gradient to rectify the feature extraction, which is referred to as canonical direction $\mathit{\Gamma}$. Now we are going to introduce how to calculate the feature with the geodesic gradient. The properties of geodesic gradient will be presented in the following parts.

By assigning a consistent orientation to each descriptor based on local properties, the descriptor can be represented relative to this orientation and therefore

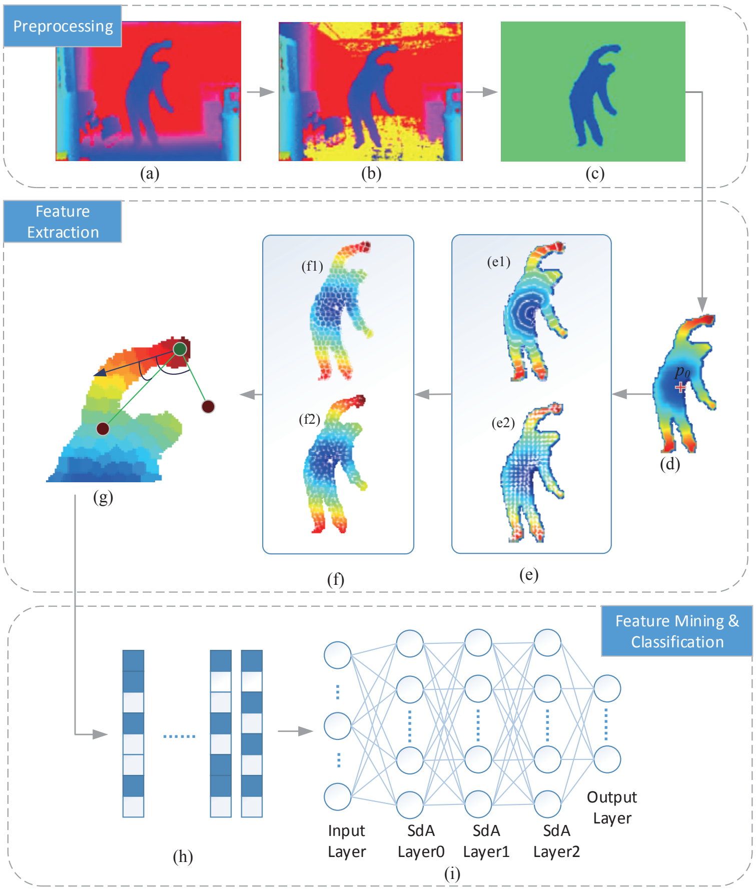


Fig. 1. The processing flow of the proposed method (Best viewed in color). (a) The input depth image. (b) Floor and ceiling detection (c) The foreground segmentation result. (d) Geodesic distance map of the foreground object. (e) Low-level feature: geodesic invariant feature extraction: (e1) is the isoline map and (e2) is the geodesic gradient map. (f) Mid-level feature: superpixel constrained geodesic invariant feature extraction: (f1) is the superpixel segmentation result and (f2) is the geodesic gradient map of the superpixels. (g) Orientation regularized binary feature calculation. (h) Binary feature strings. (i) High-level feature: depth network for feature mining.

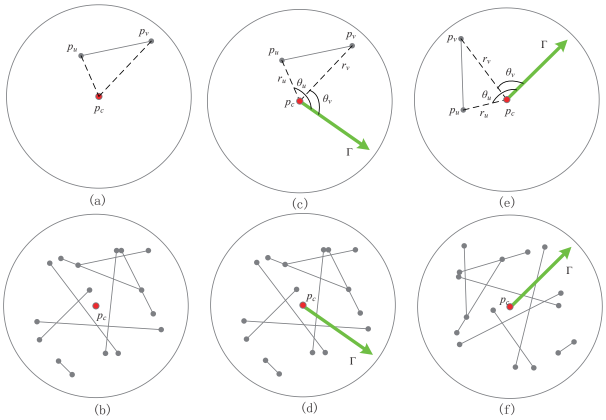


Fig. 2. Using the geodesic gradient to rectify the feature extraction. (a) Comparison between the two points generates one bit binary feature value. (b) A region is represented by the binary string produced by the multiple random comparisons which have been used in [23, 21]. (c) Random points generated with respect to the canonical direction $\mathit{\Gamma}$. (d) Descriptor contains multiple pair comparisons. (e)∼(f) the random point-pairs are covariant with the canonical direction $\mathit{\Gamma}$.

achieve invariance to the rotation [7]. For a geodesic invariance descriptor, its coverage is denoted by $C_{p_c,r,\mathit{\Gamma}}$, where $\mathit{\Gamma}$ is used to represents the canonical direction of the descriptor. The random point-pairs are generated in the polar coordinates where p_c is the origin and $\mathit{\Gamma}$ is the polar axis, as shown in Fig.2(c)∼(d). Take random point p_u for instance, it determined by two parameters, the angle $\theta_u \in [0, 2\pi)$ and the distance $r_u \in [0, r)$. Since θ_u represents the relative angle between the point p_u and $\mathit{\Gamma}$, all the point-pairs are covariant with the canonical direction $\mathit{\Gamma}$, as shown in Fig.2(e)∼(f).

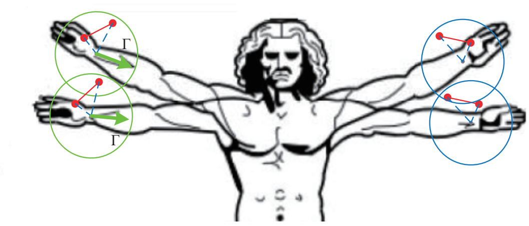


Fig. 3. An intuitive example of GIF descriptor: the GIF is represented by the circles (green) on the right hand of Vitruvian man; the feature of [23, 21] is represented by the circles (blue) on the left hand.

Fig.3 gives an intuitive example which shows the contribution of the canonical direction $\mathit{\Gamma}$. The feature which covers the right hand of the Vitruvian man is the GIF descriptor and the one covers the left hand represents the feature of [23, 21]. This example shows that the variation introduced by the articulation is canceled out by the canonical direction Γ, therefore the positions of the given point pair are relatively stable with respect to the local body parts. But for the feature of [23, 21], this invariance cannot be readily maintained.

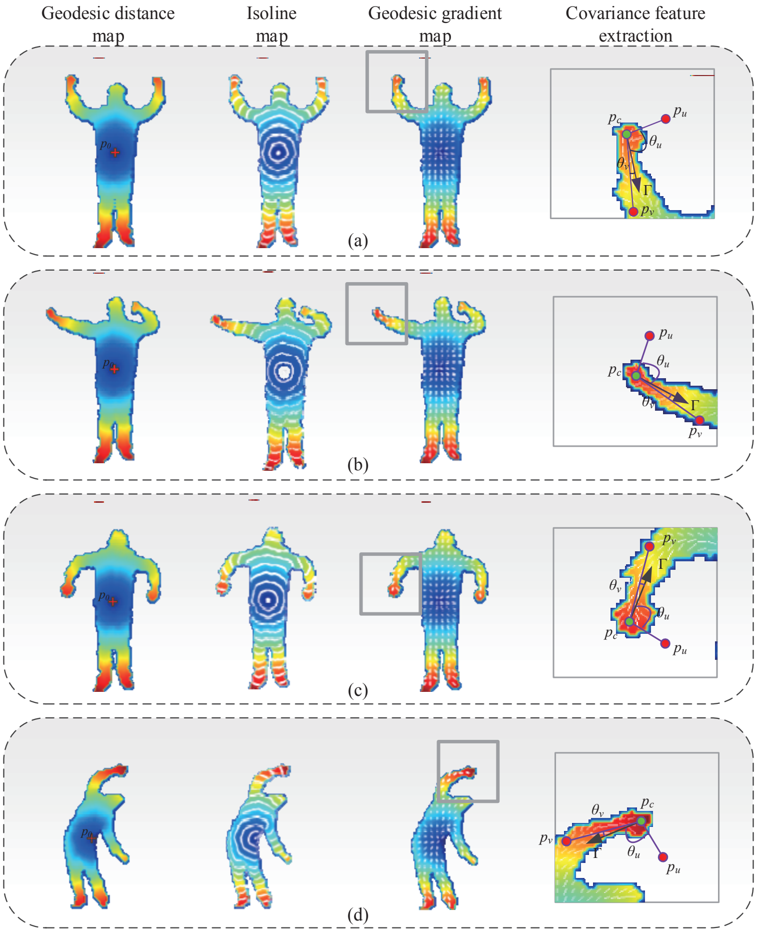


Fig. 4. Geodesic invariance feature. (a)∼(d) show the invariance of a GIF descriptor under four different poses. The geodesic distance map, isoline map, geodesic gradient map and feature extraction are shown from left to right.

Calculation of canonical direction Γ is inspired by the insight that geodesic distances on a surface mesh are largely invariant to mesh deformations and rigid transformations [36]. More intuitively, the distance from the left hand of a person to the right hand along the body surface is relatively unaffected by her/his posture. For the given object, f_g represents the foreground object segmentation results as shown in Fig.1(c). p_0 represents the centroid pixel of f_g which marked as the red cross in Fig.1(d). Calculation of canonical direction $\mathit{\Gamma}$ contains following steps:

1. The geodesic distance map I_d is generated by calculating geodesic distances between the pixels on the f_g and p_0 using Dijkstra's algorithm, as shown in the first column (from left) of Fig.4.
2. The pixels with equal geodesic distances to p_0 are marked in the isoline map as shown in the second column (from left) of Fig.4.
3. For each pixel, canonical direction can be either calculated as

$$\mathit{\Gamma} = \arctan\left(\frac{\partial I_d}{\partial x}, \frac{\partial I_d}{\partial y}\right) \tag{3}$$

or as the direction that pointing along the shortest path obtained from step 1. In this paper, we adopt the second approach and the examples of canonical direction are shown in the third column (from left) of Fig.4.

The nature of canonical direction $\mathit{\Gamma}$ is presented in the right most column of Fig.4, in which the hand patches of four different poses are shown. Rectified by $\mathit{\Gamma}$, the positions of the point-pairs are stable with respect to the body parts in different poses. Therefore, the invariance to the articulation can be obtained.

3 Superpixel Constrained Geodesic Invariant Feature (ScGIF)

In this section, we develop the superpixel constrained geodesic invariant feature (ScGIF) and use it as the mid-level representation of the depth data. The benefit of this mid-level representation is twofold: firstly, the processing speed can be improved by a big margin; secondly, better robustness to the noisy depth data can be achieved. The motivation of ScGIF is based on the following observations:

1. The time complexity of the Dijkstra's algorithm used in Section 2 for building the geodesic distance map is $O(|E|+|V|log|V|)$, where $|E|$ is the number of edges and $|V|$ is the number of vertices in the map. The processing speed is directly related to the number of pixels on the foreground object f_g. Therefore, if the number of pixels can be reduced, the processing speed can be improved.
2. The depth data obtained by the range sensor are noisy. The noise may come from the shadows of the objects, the strong ambient light that overwhelm the IR light or object materials that scatter the IR light. Therefore, the per-pixel feature extraction/classification are prone to be affected by the noise.

Based on the above observations, we replace the rigid structure of pixel grid by perceptually meaningful atomic regions, superpixels. The superpixel method

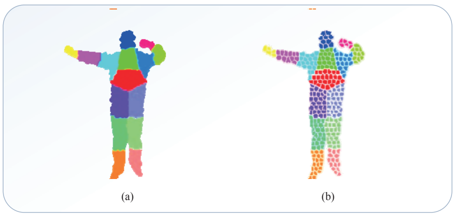


Fig. 5. Superpixel segmentation results. (a) Ground truth label of the depth data. (b) SLIC superpixel segmentation of the data

we used is based on SLIC [25], and a segmentation example is illustrated in Fig.5. Original SLIC method clusters the pixels based on their $[l, a, b, x, y]$ components where l, a and b are the color components in Lab space and x and y are the coordinates of the pixel. In our case, the clustering is performed based on $[x, y, z, L]$ where x, y and z are the coordinates in the real world space and L is the label of pixel. L is optional for the clustering and only used for offline training and evaluation. Using L, we can make sure that the pixels within the same superpixel have consistent labels, as shown in Fig5(a)$\sim$(b). During online classification, only real world coordinate $[x, y, z]$ is used for superpixel segmentation.

For each superpixel, we record the mean depth value of all the pixels that belongs to it. The random point pair comparison is replaced by the superpixel pair comparison. As illustrated in Fig. 6, the point pair (p_u, p_v) is mapped to their corresponding superpixel (p'_u, p'_v), and comparison is carried out between their mean depth values. The canonical direction Γ pointing along to the shortest path towards the object centroid p_0.

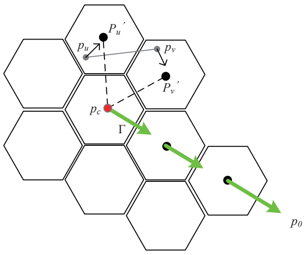


Fig. 6. Feature calculation for the superpixels

By SLIC clustering, the unit being processed is changed from pixel to superpixel. For depth frame with VGA size, the foreground objects may contains tens of thousands pixels, but this lead to only a few hundreds of superpixels. Therefore, the processing load is reduced dramatically. In additional, using the mean value to replace the individual pixel depth can further improve the robustness to the noise. These improvements are going to be verified in the experimental section.

4 Feature Mining through Deep Network

Generally, high dimensional data have complex distribution in their feature space. If the nonlinearity of the data is handled properly, the performance can be improved dramatically. There are two possible approaches to exploit nonlinearity of the data. The first approach is to leave this task to the classifiers, such as support vector machines (SVM) which use kernel functions to build nonlinear classifiers to model the complex data distribution. The second approach is to use an intermediate stage, between the original feature and the classifier, to map the feature from its original space to some latent feature space where the data may have more compact distribution. The second approach also called representation learning or feature learning which is the essence of deep learning. The sequential training manner makes it especially suitable for learning task under the big data environment.

In this section, we attempt to exploit the high order nonlinearity of the data by deep networks. Specifically, the employed deep network is based on stacked denoising autoencoders (SdA) [37–39]. Through SdA, the data are projected nonlinearly from its original feature space to some latent representations. We refer to these representations as SdA-layer$x$ feature spaces. SdA can eliminate the irrelevant variation of the input data while preserving the discriminant information that can be used for classification and recognition. Meanwhile, the propagation process of the data from the top layers to the deep layers of the SdA generates a series of latent representations with different abstraction ability. The deeper the layer, the higher level of abstraction.

The structure of our SdA deep network is illustrated in Fig.7(a). It contains five layers: 1 input layer, 3 hidden/SdA layers and 1 output layer. Each layer contains a set of nodes and the nodes between the adjacent layers are fully connected. The number of nodes in the input layer equals to n, which is the number of random pairs. The binary strings of ScGIF are feed directly to the network as the input layer. The number of nodes in the output layer is d which equals to the number of labels. The rationale and training details of SdA is beyond the scope of this paper, please refer to [37–39] for more details.

Since we have claimed that the representations learnt by SdA can eliminate the irrelevant variation of the input data while preserving the information that is useful for the final classification task, it is important to investigate what actually have been learnt through this deep feature hierarchies. We plot the features of different layers of SdA in Fig.7(b)∼(e), which provide us with some insight of SdA learning. Fig.7(b) is the binary string of ScGIF descriptor which feed directly to the input layer. Fig.7(c)∼(e) are the features that have been learnt by SdA layer $0 \sim 2$. A very interesting observation is that as the data propagate towards the deeper layers of the network, a trend of sparsity can be clearly observed. Until final layer of SdA, the number of no-zero entries reduced to 331, which accounts for only 16.6% of the 2000-dimentional feature space.

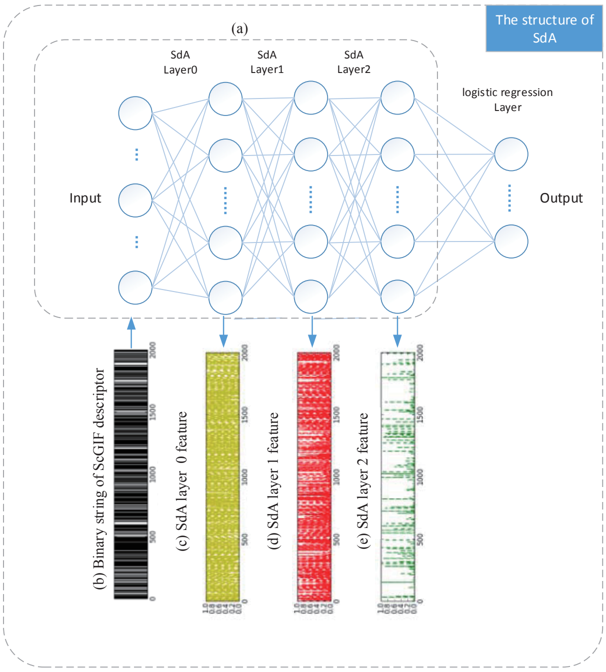


Fig. 7. The structure of SdA and its layer outputs. As the feature propagate from the SdA-layer0 feature space to the SdA-layer2, a trend of sparsity can be clearly observed.

5 Experiments

In this section, we present the details of parameter setting, dataset, and the comparison results with the state-of-the-art methods.

Dataset. We collect a depth image dataset using Kinect sensor. Four actors perform different poses and actions. There are 6930 depth frames of VGA size. For each human body, 15 body parts are labeled. An example of the ground truth label is illustrated in Fig.5(a). The total number of labeled pixels is around 80 million. We randomly select 50% frames for training, 10% frames for validation, and the rest 40% frames for testing.

Baselines. The first baseline descriptor that we used is the accumulative geodesic extrema descriptor proposed by Christian et al.[36], and it is referred as to AGEX. The second baseline descriptor is presented by Shotton et al. [23, 21] which is variant of the BRIEF[3] descriptor in the depth image. Therefore, this descriptor is referred to as BRIEFd. The geodesic invariant feature presented in

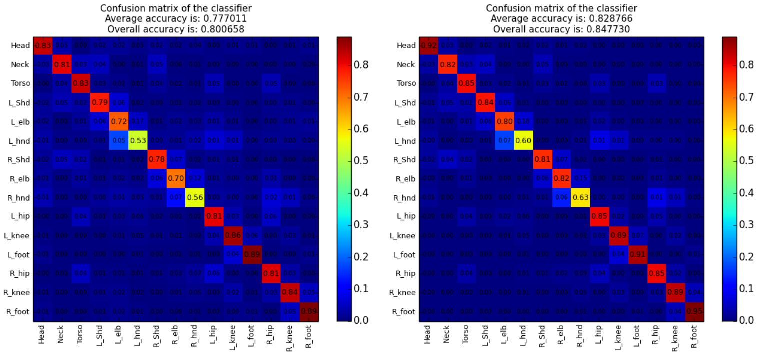
Section 2 is denoted as GIF, the superpixel constrained geodesic invariant feature in Section 3 is denoted as ScGIF, and the feature obtained by deep learning in Section 4 is denoted as ScGIF+SdA.

5.1 Contribution of canonical direction rectification

To highlight performance improvement obtained by using the canonical direction, we compare the BRIEFd and GIF in more details. Both of these two methods are pixel-wise descriptors and the only difference between them is with or without canonical direction rectification. Random forests are learnt as the classifiers which contain 3 random trees and the maximum level of each tree is 20.

The confusion matrices are illustrated in Fig.8, in which (a) is the confusion matrix of BRIEFd and (b) is the results of GIF. Average accuracy is increased from 77.7% to 82.9% and overall accuracy is improved from 80.0% to 84.8%. From these results, two observation can be obtained.

1. With the help of canonical direction, about 5% accuracy improvement can be achieved for both overall accuracy and average accuracy.
2. The improvements for parts on the limbs are higher than the parts on the torso. A possible explanation is that the articulate variation are more prominent on the limbs, and canonical direction can counteract these variations effectively.



Confusion matrix of the classifier
Average accuracy is: 0.777011
Overall accuracy is: 0.800658

	Head	Neck	Torso	L_Shd	L_elb	L_hnd	R_Shd	R_elb	R_hnd	L_hip	L_knee	L_foot	R_hip	R_knee	R_foot
Head	0.83	0.03	0.00	0.02	0.02	0.03	0.02	0.02	0.04	0.00	0.01	0.01	0.00	0.01	0.01
Neck	0.03	0.81	0.03	0.04	0.00	0.01	0.05	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
Torso	0.00	0.04	0.83	0.03	0.02	0.01	0.04	0.02	0.01	0.05	0.00	0.00	0.05	0.00	0.00
L_Shd	0.02	0.05	0.02	0.79	0.06	0.02	0.00	0.01	0.02	0.00	0.00	0.00	0.00	0.01	0.00
L_elb	0.02	0.00	0.01	0.06	0.72	0.17	0.01	0.02	0.03	0.01	0.01	0.00	0.00	0.01	0.01
L_hnd	0.01	0.00	0.00	0.01	0.05	0.53	0.00	0.01	0.02	0.01	0.01	0.00	0.00	0.00	0.00
R_Shd	0.02	0.05	0.02	0.01	0.01	0.02	0.78	0.07	0.02	0.01	0.01	0.00	0.00	0.00	0.00
R_elb	0.01	0.00	0.01	0.01	0.02	0.02	0.06	0.70	0.12	0.01	0.00	0.00	0.01	0.00	0.00
R_hnd	0.01	0.00	0.00	0.00	0.01	0.03	0.01	0.07	0.56	0.00	0.00	0.00	0.02	0.01	0.00
L_hip	0.00	0.00	0.04	0.01	0.03	0.06	0.02	0.02	0.02	0.81	0.03	0.00	0.06	0.00	0.00
L_knee	0.01	0.00	0.00	0.00	0.01	0.05	0.01	0.01	0.02	0.04	0.86	0.06	0.00	0.03	0.01
L_foot	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.01	0.00	0.04	0.89	0.00	0.00	0.01
R_hip	0.00	0.00	0.04	0.01	0.01	0.01	0.01	0.03	0.07	0.06	0.00	0.00	0.81	0.03	0.00
R_knee	0.01	0.00	0.00	0.01	0.02	0.02	0.00	0.01	0.05	0.00	0.02	0.01	0.03	0.84	0.05
R_foot	0.01	0.00	0.00	0.01	0.02	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.05	0.89

(a)

Confusion matrix of the classifier
Average accuracy is: 0.828766
Overall accuracy is: 0.847730

	Head	Neck	Torso	L_Shd	L_elb	L_hnd	R_Shd	R_elb	R_hnd	L_hip	L_knee	L_foot	R_hip	R_knee	R_foot
Head	0.92	0.03	0.00	0.02	0.01	0.02	0.02	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.00
Neck	0.03	0.82	0.03	0.04	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Torso	0.00	0.04	0.85	0.03	0.02	0.01	0.03	0.02	0.00	0.03	0.00	0.00	0.03	0.00	0.00
L_Shd	0.01	0.05	0.01	0.84	0.06	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
L_elb	0.01	0.00	0.01	0.06	0.80	0.18	0.00	0.00	0.01	0.02	0.00	0.00	0.00	0.00	0.00
L_hnd	0.00	0.00	0.00	0.01	0.07	0.60	0.00	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.00
R_Shd	0.02	0.04	0.02	0.00	0.00	0.00	0.81	0.07	0.02	0.00	0.00	0.00	0.01	0.00	0.00
R_elb	0.00	0.00	0.01	0.00	0.00	0.00	0.06	0.82	0.15	0.00	0.00	0.00	0.01	0.00	0.00
R_hnd	0.01	0.00	0.00	0.00	0.00	0.01	0.01	0.06	0.63	0.00	0.00	0.00	0.01	0.01	0.00
L_hip	0.00	0.00	0.04	0.00	0.02	0.08	0.00	0.00	0.01	0.85	0.02	0.00	0.05	0.00	0.00
L_knee	0.00	0.00	0.00	0.00	0.01	0.05	0.00	0.00	0.01	0.03	0.89	0.07	0.00	0.02	0.00
L_foot	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.04	0.91	0.00	0.00	0.01
R_hip	0.00	0.00	0.04	0.00	0.00	0.01	0.01	0.02	0.09	0.04	0.00	0.00	0.85	0.02	0.00
R_knee	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.05	0.00	0.03	0.00	0.03	0.89	0.04
R_foot	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.04	0.95

(b)

Fig. 8. Confusion matrix of the classification results (random forest): (a) Confusion matrix obtained using the BRIEFd feature. (b) Confusion matrix of the GIF feature.

5.2 Comparison with the state-of-the-art methods

The detailed comparison of the per-class classification accuracy are presented in Fig.9. Regarding the AGEX method [36], since we are working on the classification task, we use their patch based descriptor without the geodesic EXtrema

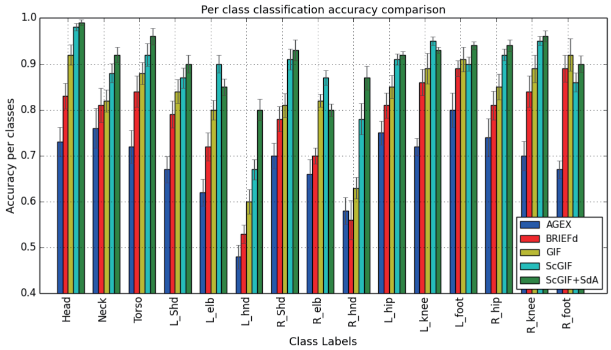
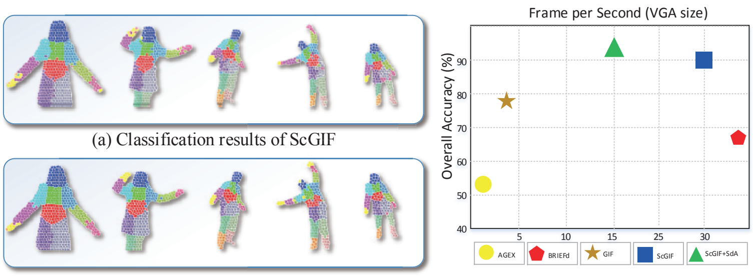


Fig. 9. Comparison results of each body parts.

detector. The random forest is used as the classifier and the training process is very similar to BRIEFd and GIF. The only difference is that we replace the random pair comparison with the random dictionary patch filtering. Following the setting in [40], 2000 random sub-patches for each joint are randomly generated and used as the dictionary. The error bars of all the methods are obtained by 5-round of cross validation. The visualization of the classification results of ScGIF and ScGIF+SdA are illustrated in Fig.10(a) and (b). From this result, we have following observations:

1. Among all the five methods, the performance of the dictionary patch matching based descriptor (AGEX) is lower with the others. One possible explanation is that the dictionary patches for the depth data are lack of textures and cannot provide enough discrimination information[40].
2. The proposed GIF descriptor and its two variants can outperform the other baseline methods. Especially for the hands and feet, accuracies have been improved by a big margin. This further verify the contribution of the canonical direction rectification.
3. Comparing the pixel based methods (BRIEFd and GIF) with the superpixel based methods (ScGIF, ScGIF+SdA), improvements for both of the mean classification accuracy and the standard deviation have been observed, which indicate superpixel is helpful for resenting the noise and texture less depth data.

In addition, we analyze the efficiency of the proposed methods. The accuracy versus rum time figure is illustrated in Fig.10(c). Here, only classification time is considered and the foreground segmentation time is not took into account. The testing platform consist of Intel i7 3.7G processor and 32G RAM. BRIEFd is the fastest one since there are only simple pixel comparison involved in the evaluation. Superpixel is critical for the speed improvement. It increase the speed of GIF from 3.7 fps to 30 fps (ScGIF). This verifies another assumption about the superpixel: the superpixel based representation can reduce the image redundancy and improve the processing speed dramatically. The accuracy winner



(a) Classification results of ScGIF

(b) Classification results of ScGIF+ SdA

(c) Overall accuracy versus run time

Fig. 10. Classification results of superpixel and superpixel+SdA.

ScGIF+SdA runs at 15 fps on a GTX 780i GPU using python lib theano (only the SdA part is running on GPU, superpixel is still on CPU). Since the evaluation process of the deep networks has high parallelism, the GPU time (including copy the data from host to device) for each frame is only 28.7 ms.

6 Conclusion

In this work, we presented a geodesic invariant feature and its two variants to encoding the local structure of the depth data. These new descriptors were applied in the context of human body parts recognition. Specially, the proposed descriptors form a multi-level feature extraction hierarchy: pixel based low-level representation addressed the articulation variation of human motion by introducing the canonical direction to rectify the feature extraction process; superpixel based mid-level representation replaced the rigid structure of the pixel grid by perceptually meaningful atomic regions which reduced the computation cost and improve the robustness to the noise; high-level feature exploit the nonlinearity of the data by deep networks which further improve the performance of the descriptor. We compare the proposed method with the state-of-the-art methods. Encouraging results have been observed. The proposed method can achieve superior classification accuracy and visual quality.

7 Acknowledgment

This work is supported by NSFC (Grant No 61300161, 61371168 and 61273251), Doctoral Fund of Ministry of Education of China (Grant No 20133219120033), Open Project Program of Jiangsu Key Laboratory of Image and Video Understanding for Social Safety (Grant No JSKL201306) and Programme of Introducing Talents of Discipline to Universities (Grant NoB13022).

References

1. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence **27** (2005) 1615–1630

2. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: Surf: Speeded up robust features. Computer Vision and Image Understanding **110** (2008) 346–359
3. Calonder, M., Lepetit, V., Ozuysal, M., Trzcinski, T., Strecha, C., Fua, P.: Brief: Computing a local binary descriptor very fast. IEEE Transactions on Pattern Analysis and Machine Intelligence **34** (2012) 1281 – 1298
4. Calonder, M., Lepetit, V., Strecha, C., Fua, P.: Brief: Binary robust independent elementary features (2010)
5. Chen, J., Shan, S., He, C., Zhao, G., Pietikinen, M., Chen, X., Gao, W.: Wld: A robust local image descriptor. IEEE Transactions on Pattern Analysis and Machine Intelligence **32** (2009) 1705–1720
6. Ke, Y., Sukthankar, R.: Pca-sift: A more distinctive representation for local image descriptors (2004)
7. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60** (2004) 91–110
8. Valle, E.: Local-Descriptor Matching for Image Identification Systems. Thesis (2008)
9. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence **32** (2010) 1627–1644
10. Chen, J., Zhao, G., Salo, M., Rahtu, E., Pietikinen, M.: Automatic dynamic texture segmentation using local descriptors and optical flow. IEEE Transactions on Image Processing **22** (2013) 326–339
11. Rahmani, R., Goldman, S.A., Zhang, H., Cholleti, S.R., Fritts, J.E.: Localized content based image retrieval. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (2008)
12. Shen, X., Lin, Z., Brandt, J., Wu, Y.: Detecting and aligning faces by image retrieval (2013)
13. Subrahmanyam, M., Maheshwari, R., Balasubramanian, R.: Local maximum edge binary patterns: A new descriptor for image retrieval and object tracking. Signal Processing **92** (2012) 14671479
14. Ta, D.N., Chen, W.C., Gelfand, N., Pulli, K.: Surftrac: Efcient tracking and continuous object recognition using local feature descriptors (2009)
15. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: Application to face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **28** (2006) 2037–2041
16. Zhang, W., Shan, S., Gao, W., Chen, X., Zhang, H.: Local gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face representation and recognition (2005)
17. Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., Fitzgibbon, A.: Kinectfusion: Real-time 3d reconstruction and interaction using a moving depth camera (2011)
18. Newcombe, R.A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A.J., Kohli, P., Shotton, J., Hodges, S., Fitzgibbon, A.: Kinectfusion: Real-time dense surface mapping and tracking (2011)
19. Helten, T., Baak, A., Bharaj, G., Mller, M., Seidel, H.P., Theobalt, C.: Personalization and evaluation of a real-time depth-based full body tracker (2013)
20. Lallemand, J., Pauly, O., Schwarz, L.: Multi-task forest for human pose estimation in depth images (2013)
21. Shotton, J., Girshick, R., Fitzgibbon, A., Sharp, T., Cook, M., Finocchio, M., Moore, R., Kohli, P., Criminisi, A., Kipman, A., Blake, A.: Efcient human pose

estimation from single depth images. IEEE Transactions on Pattern Analysis and Machine Intellingence **35** (2013) 2821 – 2840

22. Ye, M., Yang, R.: Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera (2014)
23. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images (2011)
24. Ojala, T., Pietikinen, M., Menp, T.: Multiresolution gray scale and rotation invariant texture analysis with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence **24** (2002) 971–987
25. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Ssstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence **34** (2012) 2274–2282
26. Arel, I., Rose, D.C., Karnowski, T.P.: Deep machine learning a new frontier in artificial intelligence research. IEEE Computational Intelligence Magazine **5** (2010) 13–18
27. Bengio, Y.: Learning deep architectures for ai. Foundations and Trends in Machine Learning (2009)
28. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence **35** (2013) 1798–1828
29. Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., Bengio, Y.: Theano: a cpu and gpu math expression compiler (2010)
30. Farabet, C., Couprie, C., Najman, L., LeCun, Y.: Learning hierarchical features for scene labeling. IEEE Transactions on Pattern Analysis and Machine Intelligence **35** (2013) 1915–1929
31. Kavukcuoglu, K., Sermanet, P., Boureau, Y.L., Gregor, K., Mathieu, M., LeCun, Y.: Learning convolutional feature hierachies for visual recognition (2010)
32. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks (2012)
33. Bordes, A., Glorot, X., Weston, J., Bengio, Y.: Joint learning of words and meaning representations for open-text semantic parsing (2012)
34. Socher, R., Huang, E.H., Pennington, J., Ng, A.Y., Manning, C.D.: Dynamic pooling and unfolding recursive autoencoders for paraphrase detection (2011)
35. Socher, R., Pennington, J., Huang, E.H., Ng, A.Y., Manning, C.D.: Semi-supervised recursive autoencoders for predicting sentiment distributions (2011)
36. Plagemann, C., Ganapathi, V., Koller, D., Thrun, S.: Real-time identification and localization of body parts from depth images (2010)
37. Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.: Greedy layer-wise training of deep networks (2007)
38. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders (2008)
39. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A.: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of Machine Learning Research **11** (2010) 3371–3408
40. Torralba, A., Murphy, K.P., Freeman, W.T.: Sharing features: efficient boosting procedures for multiclass object detection (2004)