Hand part classification using single depth images

Myoung-Kyu Sohn, Dong-Ju Kim and Hyunduk Kim

Department of IT Convergence, Daegu Gyeongbuk Institute of Science & Technology (DGIST), Daegu, South Korea

Abstract. Hand pose recognition has received increasing attention as an area of HCI. Recently with the spreading of many low cost 3D camera, researches for understanding more natural gestures have been studied. In this paper we present a method for hand part classification and joint estimation from a single depth image. We apply random decision forests(RDF) for hand part classification. Foreground pixels in the hand image are estimated by RDF, which is called per-pixel classification. Then hand joints are estimated based on the classified hand parts. We suggest robust feature extraction method for per-pixel classification, which enhances the accuracy of hand part classification. Depth images and label images synthesized by 3D hand mesh model are used for algorithm verification. Finally we apply our algorithm to the real depth image from conventional 3D camera and show the experiment result.

1 Introduction

Vision-based gesture recognition is one of the possible solutions for HCI (Human Computer Interface) as it provides natural interaction between people and all kind of devices. There have accordingly been many studies on gesture recognition techniques [1, 2]. Hand gesture is emerging topic, which enables interactions between human and computer more naturally. As only hand part of the human body is needed for analysis of gesture, hand gesture recognition is more efficient in some case in contrast to human activity recognition based on full body movement. It can simply be categorized three problems for recognizing hand gesture, which are hand detection, pose estimation and gesture classification.

Hand gesture recognition research has often focused on techniques to detect hand on the frame from regular RGB camera. Light condition, cluttered background, skin color et al. make it difficult to find the hand from the RGB image. With the commercial release of depth camera such as Kinect [3] and Xtion, the segmentation process are much more simplified through depth information. Thus, techniques for recognizing hand pose using segmented hand depth information have been developed [4, 6, 5].

Hand pose estimation generally can be categorized into two groups: shape-based, 3D model-based. Shape-based approaches generally match the shape features of the detected hand to a shape features from a predefined hand shape database. Doliotis et al. [7] extracts contour of the hand which implies the shape

and the boundary of the hand. After normalizing the features for translation invariance, the depth similarity between two images was measured to classify hand pose. Liu et al. [8] use depth images acquired by a time-of-flight camera for hand gesture recognition. The authors detect hand by depth difference between foreground and background and conduct Chamfer distance matching technique to measure shape similarity. Ren et al. [9] propose a Finger-Earth Mover's distance to measure the dissimilarities between hand shapes. Suryanarayan et al. [10] propose scale and rotation invariant hand pose recognition techniques. The authors suggest a volumetric shape descriptor which implies 2D image data with depth information.

In 3D model-based approaches, the pose estimation is an optimization problem that minimizes the difference between 3D hand models in database and a detected hand. Oikonomidis et al. [11] propose a generative single hypothesis model-based pose estimation method. They apply particle swarm optimization for 3D hand pose recovery. In [12], the authors present model-based tracking with an articulated hand model and recognize the pose using an unscented Kalman filter.

With the help of depth information, many researches have been carried out to classify hand pose with less effort in detecting hand. However, most researches focus on above approaches such as shape-base pose estimation or 3D model based matching algorithm. Skeleton-based approach can be used for pose estimation but has been still challenging task. Shotton et al. [13] present state of the art work in pose estimation by skeleton configuration. In their pioneering work, they use the random decision forests(RDF) [14] to classify human's body part for each pixel. Then, a local clustering method is applied to find a center of the each body part. In [15], the authors has applied the approach by Shotton et al. to hand image instead of body image and they show the feasibility of the skeleton-based hand pose estimation.

In this paper, we adopt the idea of human pose estimation by Shotton et al. in estimating hand skeleton. We suggest more robust feature extraction method to enhance the performance of the recognition system and compare the simulation results.

2 Data

For classifier to work well in practice, it is important that training data covers variety of hand pose encountered in real life. A solution is to collect a large number real data. Gathering large amount of real data is not easy problem, many research in this field has often focused on the techniques to overcome the lack of training data. However, Shotton et al. shows that synthesized large, varied dataset outperforms in their evaluation.

To classify each of hand part, 3D hand mesh models created by a modeling tool is used for synthesizing various hand pose data. The hand model has 21 different parts which have joints with their center. Depth image and label image are generated from the mesh model and the label image indicates the different

label as different color. Every finger has four parts except thumb with three part. Palm has relatively large area in hand so we divided it into two parts, upper palm and lower palm. We build five datasets with different hand gesture. Ten mesh models with different character are used for rendering depth image and label image. Size of 150K images with different character and different pose are synthesized by the 3D mesh model and automating script. Fig. 1 shows the rendered depth image and label image respectively.

a) depth image b) label image

Fig. 1. Synthesized image from 3D mesh model

3 Hand part classification and joint estimation

Random decision forests are used for classification of hand part. We employ per-pixel classification to recognize hand part from a image. Training is a procedure of deciding good parameters of each node for splitting data into child nodes with high information gain. Every node split the data with Eq. 1.

$$\begin{aligned} Q_L(\phi) &= \{(I,p)|f_\theta(I,p) > \tau\} \\ Q_R(\phi) &= \{(I,p)|f_\theta(I,p) \leq \tau\} \end{aligned} \tag{1}$$

Feature responses of input data are calculated using randomly selected candidate parameters $\theta = (u, v, \tau)$ at every node. From these feature response f_θ, information gain $G(\phi)$ is obtained. In training phase we choose splitting parameters which generate maximum information gain. The gain is computed by Eq. 2.

$$\begin{aligned} G(\phi) &= H(Q) - \sum_{S\in\{L,R\}} \frac{|Q_S(\phi)|}{Q} H(Q_S(\phi)) \\ \phi^* &= argmax G(\phi) \end{aligned} \tag{2}$$

Where H is Shannon entropy, information gain is estimated from distribution of the each labels in the all input data of each node. ϕ is set of candidate parameter and ϕ^* is the selected split parameter from tree learning which maximize the information gain at node.

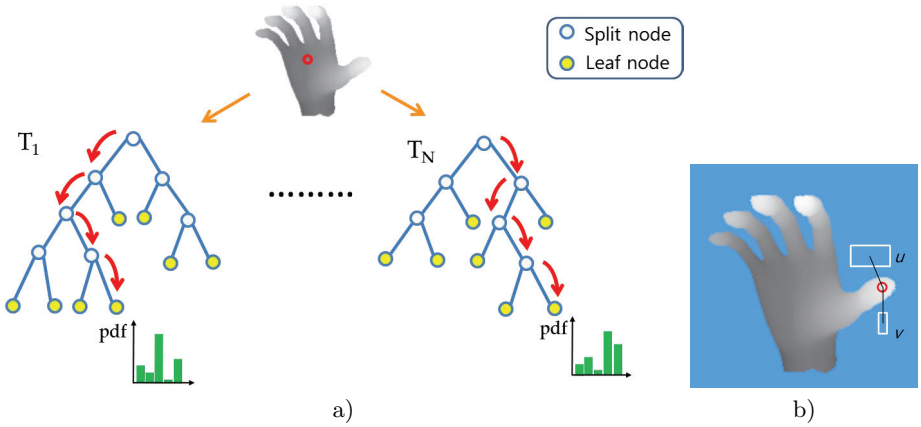


Fig. 2. a) Random decision forests. Every split node learns a split function in a training phase and test data traces the tree by split function. The leaf node has a probability density of class label. b) Feature extraction for a selected pixel position. Parameter u, v are random sized offset patch

In classification, input data goes to the leaf node by the split function at each node. Each input data has the learned posterior $p_t(c|I, p)$ at the leaf node in each tree. The posterior distributions of each tree in the forests are averaged for the final classification.

Feature of each pixel has a large impact on overall system and performance [16]. We suggest random offset patches from pixel x instead of using offset pixel. While the offset pixel was used as a feature extraction in the state-of-the-art algorithm in Shotton et. al.

$$f_\theta(I, p) = \frac{1}{k}\sum_{i=1}^{k} d_I(p + \frac{u(i)}{d_I^0}) - \frac{1}{l}\sum_{j=1}^{k} d_I(p + \frac{v(j)}{d_I^0}) \tag{3}$$

The element values of u and v are (x, y) values within the random sized offset patch. The center of the offset patch is a offset value from a position x. The size of patch is also normalized according to the depth d_I^0, which is averaged depth value of the hand. Feature value calculated by difference value between depth values by two patches, which are mean depth value of each patch. The normalized size of patch and the normalized depth value ensure 3D translation

invariant. Fig. 2(b) Shows two random offset patches for the feature extraction from a position.

Since each pixel is classified with its posterior probability, joints of hand part can be estimated using this information. The global 3D centers of mass for each part can be used as a simple method. In this method the outliers largely degrade the global estimate. Instead of using global center, we apply mean shift for finding local mode of the joint.

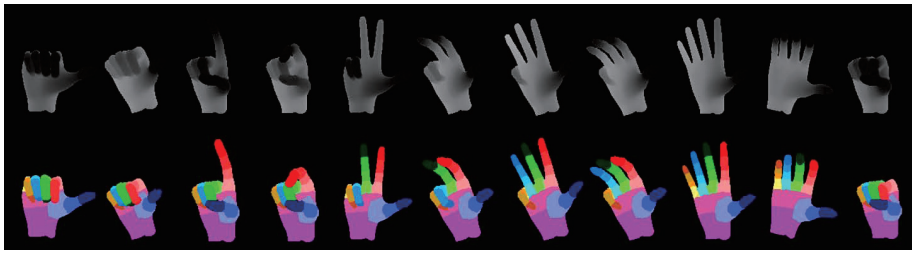
4 Experiments

To evaluate the proposed algorithm we have built five datasets and tried our algorithm to each datasets. Each dataset consists of consecutive hand pose frames from one pose to another pose. Each hand gestures are folding thumb, folding two fingers, folding tree fingers, fold four fingers and folding fiver fingers respectively. Training data has large effect on the performance of the recognition system. It is commonly known that containing good coverage of the variation such as scale, hand size and shape and camera pose in training data result in good performance in generalization issue. Depth and scale translation invariance are considered explicitly in the features of the random decision forests. For the hand size and shape invariance, we have used ten different hand characters including man, woman, child, etc. We have changed the camera pose using the automating script to handle the rotation invariance. Each dataset consists of approximately 30K images. Fig. 3 shows the example of different hand pose with different camera pose and the example of dataset-5 which has gesture images of folding five fingers.

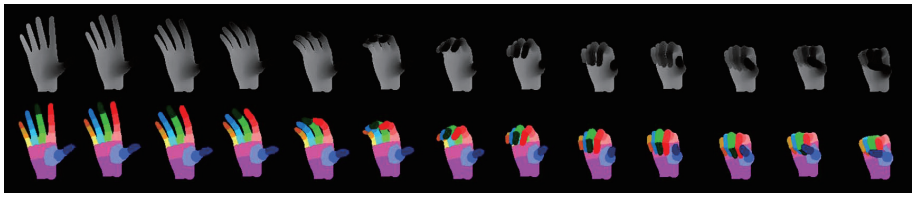
Random forests with ten trees have been used for training. Input data consists of 2500 random pixels from each images of the dataset. Pixels are selected randomly in the foreground of the image with 150x150 size. 2000 candidate features (θ) and 20 candidate threshold (τ) are used for training trees. We conduct 5x2-fold cross validation for each dataset. We report the average per-pixel accuracy for per-pixel classification and the average precision for joint estimation.

Feature. We have compared our feature to simple pixel depth comparison approach suggested by Shotton et al. In Fig. 4(a) we show the average per-pixel classification accuracy for each dataset and our approach outperforms the Shotton algorithm for all datasets. As mentioned, using the normalized average of the depth between two different patches instead of using depth between two pixels gives more reliable feature value for the split function in the forests. Feature by offset patch makes the classifier to be more robust to variations of hand pose and shape.

Maximum offset patch. We show how maximum offset patch size affects accuracy of the system. Width and height of the patch for feature is selected randomly from a range between 1 to maximum offset patch size. In this experiment we fixed the length of maximum offset u, v to 60 pixel meter. Fig. 4(b)

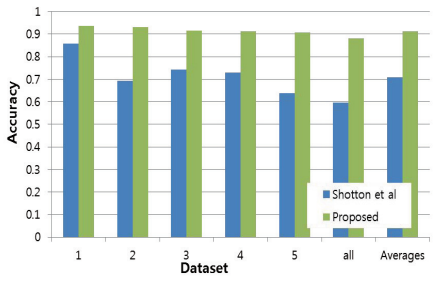


a) Examples from all the datasets

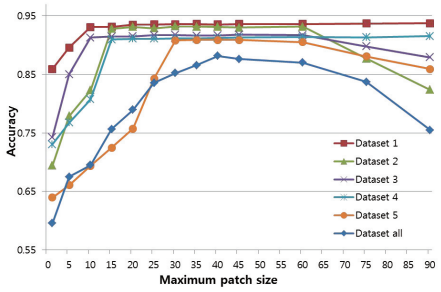


b) Examples from dataset-5 (images captured from five finger folding gesture)

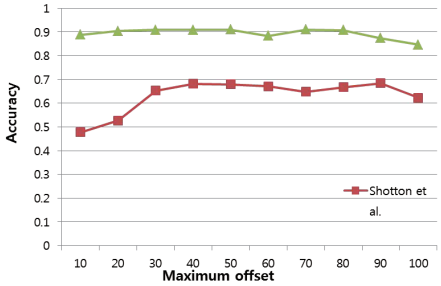
Fig. 3. Examples of test and training dataset



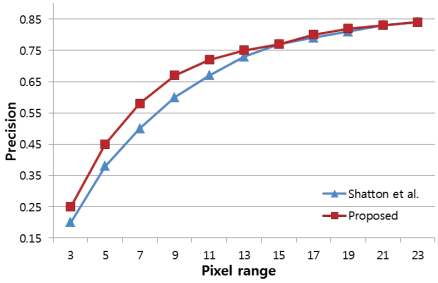
a) Average accuracy



b) Maximum offset patch size
(for maximum offset = 60)



c) Maximum offset
(for maximum offset patch = 40)

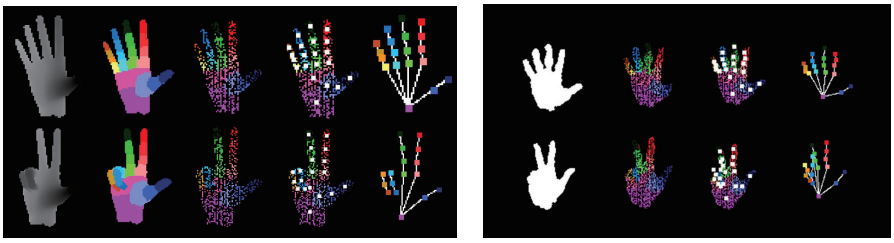


d) Skeleton estimation accuracy

Fig. 4. Accuracy results of the experiment and comparision

show the result for each dataset. Accuracy increases as maximum offset patch size increases in all dataset. But the accuracy tends to decrease from maximum patch size of 60. It is likely caused by that the offset patches tend to locate out of the image bound. The maximum offset patch having a size of 1 is same as offset pixel used in Shotton method. The dataset-5 is relatively harder gesture than other datasets and the dataset-1 is relatively easier gesture than other gesture. The result also shows that the harder gesture dataset has the lowest accuracy. We show maximum offset patch size affects more on the harder dataset. The best performance is found at 40 for maximum offset patch size.

Maximum offset. Offset is randomly selected from 0 to maximum offset value in Eq. 3. Shotton show the range of offset has also a large effect on accuracy, and as the maximum offset is increased, the classifier is able to use more spatial context to make its decision. Fig. 4(c) depicts the accuracy according to the offset value. This offset value has less effect on the accuracy in the proposed method, while the accuracy largely decreases in lower than 30 of maximum offset value in offset pixel method.



(a) result from synthetic data (Depth, Ground truth label, estimated label, joint proposal and skeleton respectively)

(b) result from real data (Depth, estimated label, joint proposal and skeleton respectively)

Fig. 5. Hand part classification and skeleton estimation result

Joint estimation. The ground truth of joint position is calculated using global 3D centers of mass for each hand part of the ground truth label image. We compare this value to the estimated joint position of the estimated label image. Joint occluded in the ground truth label image is excluded in precision calculation. Precision on joint estimation are shown in Fig. 4(d). Since our approach has higher accuracy in per-pixel classification, precision of joint estimation also higher accuracy as compared to Shotton et al. We applied the algorithm to the real depth data captured by conventional 3D depth camera which has 320x240 resolution. The final result images of synthetic and real data are shown in Fig. 5

5 Conclusion

In this paper we present a per-pixel classification for hand part and joint estimation. We synthesize depth image and ground truth label image to learn the random decision forests. Then we verify the algorithm using synthetic depth image and real depth image. Using the suggested depth comparison feature extraction method, we show the classification accuracy outperforms compared to the state-of-the-art method. To apply this algorithm to the various applications, various hand pose data should be trained. As a future work, we plan to construct more various hand pose dataset and improve random decision forests to classify the various hand pose.

Acknowledgement This work was supported by the DGIST R&D Program of the Ministry of Education, Science and Technology of Korea (14-IT-03). It was also supported by Ministry of Culture, Sports and Tourism (MCST) and Korea Creative Content Agency (KOCCA) in the Culture Technology (CT) Research & Development Program (Immersive Game Contents CT Co-Research Center).

References

1. Rautaray, Siddharth S., and Anupam Agrawal. "Vision based hand gesture recognition for human computer interaction: a survey." Artificial Intelligence Review (2012): 1-54.
2. Alon, Jonathan, et al. "A unified framework for gesture recognition and spatiotemporal gesture segmentation." Pattern Analysis and Machine Intelligence, IEEE Transactions on 31.9 (2009): 1685-1699.
3. Kinect camera [Online]. Available: http://www.xbox.com/en-US/kinect
4. Hackenberg, Georg, Rod McCall, and Wolfgang Broll. "Lightweight palm and finger tracking for real-time 3D gesture control." Virtual Reality Conference (VR), 2011 IEEE. IEEE, 2011.
5. Doliotis, Paul, et al. "Comparing gesture recognition accuracy using color and depth information." Proceedings of the 4th International Conference on Pervasive Technologies Related to Assistive Environments. ACM, 2011.
6. Tara, R., P. Santosa, and T. Adji. "Hand segmentation from depth image using anthropometric approach in natural interface development." Int. J. Sci. Eng. Res 3.5 (2012): 1-4.
7. Doliotis, Paul, et al. "Hand shape and 3d pose estimation using depth data from a single cluttered frame." Advances in Visual Computing. Springer Berlin Heidelberg, 2012. 148-158.
8. Liu, Xia, and Kikuo Fujimura. "Hand gesture recognition using depth data." Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on. IEEE, 2004.
9. Ren, Zhou, Junsong Yuan, and Zhengyou Zhang. "Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera." Proceedings of the 19th ACM international conference on Multimedia. ACM, 2011.
10. Suryanarayan, Poonam, Anbumani Subramanian, and Dinesh Mandalapu. "Dynamic hand pose recognition using depth data." Pattern Recognition (ICPR), 2010 20th International Conference on. IEEE, 2010.

11. Oikonomidis, Iason, Nikolaos Kyriazis, and Antonis A. Argyros. "Efficient model-based 3D tracking of hand articulations using Kinect." BMVC. Vol. 1. No. 2. 2011.
12. Stenger, Bjoern, Paulo RS Mendona, and Roberto Cipolla. "Model-Based Hand Tracking Using an Unscented Kalman Filter." BMVC. Vol. 1. 2001.
13. Shotton, Jamie, et al. "Real-time human pose recognition in parts from single depth images." Communications of the ACM 56.1 (2013): 116-124.
14. Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.
15. Keskin, Cem, et al. "Real time hand pose estimation using depth sensors." Consumer Depth Cameras for Computer Vision. Springer London, 2013. 119-137.
16. Lepetit, Vincent, Pascal Lagger, and Pascal Fua. "Randomized trees for real-time keypoint recognition." Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Vol. 2. IEEE, 2005.