REAL-TIME UPPER BODY POSE ESTIMATION FROM DEPTH IMAGES

Ming-Han Tsai, Kuan-Hua Chen, I-Chen Lin

National Chiao Tung University, Taiwan

ABSTRACT

Estimating upper body poses from a sequence of depth images is a challenging problem. Lately, the state-of-art work adopted a randomized forest method to label human parts in real time. However, it requires enormous training data to obtain favorable results. In this paper, we propose using a novel two-stage method to estimate the probability maps of upper body parts of users. These maps are then used to evaluate the region fitness of body parts for pose recovery. Experiments show that the proposed method can obtain satisfactory outcome in real time and it requires a moderate size of training data.

Index Terms—*Pose estimation, depth image, arm pose, randomized forest*

1. INTRODUCTION

Tracking human poses is one of the most important issues regarding depth image analysis. Former research treated depth images of a subject as a combination of 3D components, and employed silhouette regression [1], model fitting [4] or other methods to fit parts for a depth image. Other research introduced statistical or learning methods, such as associate Markov network [3] or randomized forest [6,7], to label the foreground pixels with respect to each body part. For each pixel, they usually took the part of maximum likelihood as its label and then identified the location of each body part by pixel clustering.

In this paper, we focus on the upper body pose estimation. That is because most of the commands and interactions through depth cameras are regarding the upper body postures. The positions of human hands, elbows, and shoulders are crucial to all of above applications. In our early trial, we found that existing classification algorithms usually calculated the per-pixel probability of each part and labeled the pixels in one step. However, to concurrently identify numerous body parts, the classifiers are prone to be sensitive to different or noise-disturbed data. For instance, when a user sit, the probabilities for arms or legs are usually close to each other and it results in ambiguity. To tackle this problem, the related methods required more than hundred thousands of training poses to obtain favorable results.

By contrast, we propose a two-stage classification models. After we obtain (segment) the full human body depth data, our system divides the target regions in a depth map into upper and lower body parts in advance, and then the extracted upper body region is further processed by the second randomized decision forest model. In the second classifier, the upper body part is divided into eight detailed parts. These parts are the head, torso, left shoulder, right shoulder, left upper arm, right upper arm, left forearm, and right forearm. In order to achieve real-time performance, the cascading two-stage classifiers are performed on graphics processing units (GPU).

With the two-stage approach, we turn a depth map into probability maps with respect to multiple upper body parts. However, when using only a moderate size of training data, the probability maps are still noisy. If we would like to retrieve the skeletons and joints of these parts by intuitive line fitting or region segmentation, the skeletons and joints will frequently jitter. Therefore, we further formulate objective functions with respect to multiple probability maps to more reliably estimate the joints and skeletons. A random sample consensus (RANSAC) method is used to efficiently approximate the optimal skeletons according to the objective functions. At last, the estimated skeleton poses are checked and rectified. The experiment demonstrates that the proposed approach outperforms the related method under a moderate size of training data. The flowchart of our system is shown in Fig. 1.

2. PROBABILITY MAP ESTIMATION BY TWO STAGES

The randomized forest [8] is an ensemble classifier containing multiple decision tree models. Component trees within a forest are usually randomly different from one another. This leads to no correlation between the trees. In recent years, this classifier has been proved to be effective for labeling various subjects from images. Shotton et al. [6,7] proposed an efficient framework for recognizing human parts from depth images. In their decision forest, they generated a large amount offset vectors for each pixel to find a set of vectors with distinct separation. However, when the classifier attempts to separate multiple small body parts, the offset vectors usually has to be small. In other words, they focus more on local variations, and it results in ambiguous situations between upper and lower limbs.

As mentioned above, we propose using two stages of randomized decision forests: one for rough upper/lower

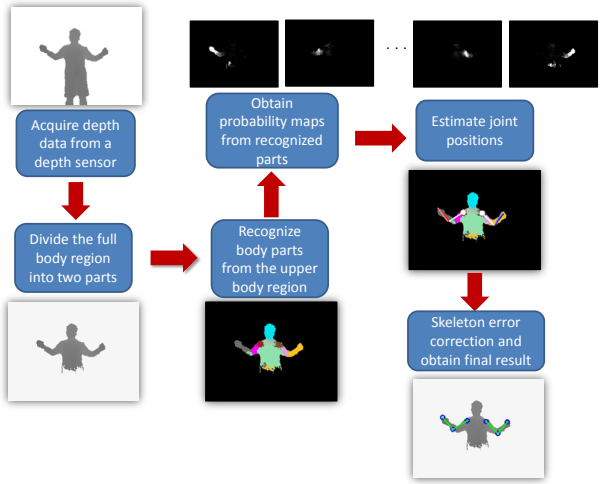


Fig.1 System overview.

body segmentation, and the other one for detailed upper body part segmentation. The first forest can employ offset vectors of a larger range to roughly separate upper and lower bodies, and the second forest then employs vectors of relatively smaller ranges for detail classification. Therefore, we can use fewer levels of decision trees to achieve higher recognition accuracy under identical training data. By introducing GPU acceleration, the total computation time of the two-stage approach is less than 1 millisecond for a single user, and it is extendable to multiple users.

Besides, different to method by Shotton et al. [6,7] where a pixel is directly assigned a label, in our framework, the outputs of random decision forest become several probability maps associated with body parts. The eight probability maps (shown in Fig. 2) are then used in the following skeleton extraction. For simplicity, we abbreviate the index of body parts *left shoulder*, *right shoulder*, *left upper arm*, *right upper arm*, *left forearm*, and *right forearm* to *ls*, *rs*, *lu*, *ru*, *lf* and *rf* in the following discussion.

3. SKELETON EXTRACTION FROM PROBABILITY MAPS

This section introduces the objective functions based on the aforementioned probability maps, and describes a random

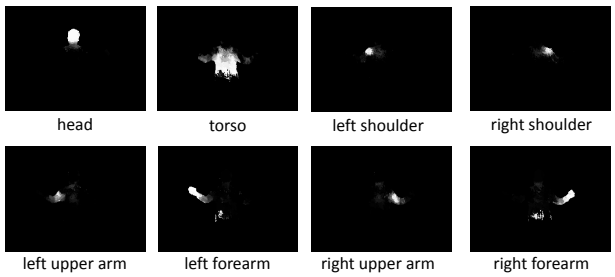


Fig.2 An example of probability maps of eight body parts.

sample consensus (RANSAC) process for posture approximation in real time.

3.1. Objective functions of body parts

Our idea is inspired by the pictorial structure [5,9,10,11]. Pictorial structure estimates the linkage of body parts as a concatenation of conditional probabilities. Our goal is to efficiently estimate the upper body postures, especially the critical forearm poses. Unlike the limb-based representation in conventional pictorial structure, we calculate the skeletons through joint representation. Given the probability map set M = {$\boldsymbol{M}_{head}$, $\boldsymbol{M}_{torso}$, $\boldsymbol{M}_{ls}$, $\boldsymbol{M}_{rs}$, $\boldsymbol{M}_{lu}$, $\boldsymbol{M}_{lf}$, $\boldsymbol{M}_{ru}$, $\boldsymbol{M}_{rf}$ } as in Fig. 2, an initial thought is that the correct skeleton position should be located at the center of the region with highest probabilities. For example, consider the right shoulder center position P_{rs} and its neighbor region R_{rs}, we define the objective function (response) S_{rs} of joint *rs* as the sum of the pixel probability values within region R_{rs} .

$$S_{rs} = \sum_{p \in R_{rs}} M_{rs}(p), \tag{1}$$

When a pixel is located at the correct position, it should have a high response value. The functions about the head, torso and right shoulders can be represented in a similar way; we retrieve the best P_{head}, P_{torso} P_{ls} and P_{rs} from the positions with maximum of S_{head}, S_{torso}, S_{ls} and S_{rs}.

By contrast, because an arm consists of two linked and movable parts, i.e. upper arm and forearm, we have to concurrently consider the two probability maps $\boldsymbol{M}_{ru}$ and $\boldsymbol{M}_{rf}$ or $\boldsymbol{M}_{lu}$ and $\boldsymbol{M}_{lf}$. In the following explanation, we take Fig. 3 as an example for right arm evaluation, and evaluation for the left arm is similar.

Given fixed shoulder position P_{rs} calculated in (1), the next target is to find the best locations of right elbow and hand, P_{re} and P_{rh}. These two variable points are drawn in red in Fig. 3. Based on a pair of P_{re} and P_{rs}, we can define an approximate upper arm region R_{ru} (the purple ellipse in Fig.3). Similarly, a pair of P_{rh} and P_{re} can define an approximate forearm region R_{rf} (the blue ellipse in Fig.3).

To find the upper arm region R_{ru} and forearm region R_{rf}, we define an objective function S_{rarm}.

$$S_{rarm} = \sum_{p \in R_{ru}} W_{ru} M_{ru}(p) + \sum_{p \in R_{rf}} W_{rf} M_{rf}(p), \tag{2}$$

where W_{ru} and W_{rf} are anisotropic Gaussian fields determined by region R_{ru} and R_{rf}. They can be regarded as variable templates of the upper arm and forearm. The two terms in (2) evaluate the correlations between the variable templates and the underlying probabilities of corresponding parts. In the first row of Fig. 3, the elbow P_{re} is not at a correct position and S_{rarm} is lower. In the second row, both the

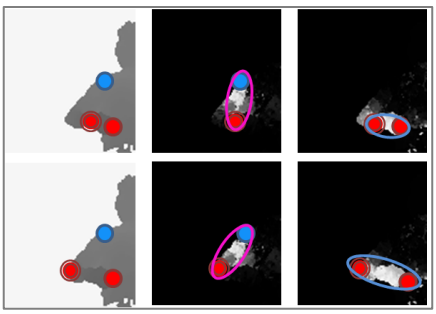
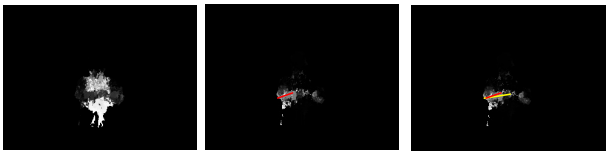


Fig. 3. Right arm fitting by an objective function. Blue point: the given shoulder point; red point: variable positions of the elbow and hand; purple ellipse: upper arm region according the current shoulder and elbow positions; blue ellipse: forearm region according to the elbow and hand positions.



(a) (b) (c)

Fig. 4. Effect of complementary term. (a) M_{torso}. (b) Estimated forearm (red) without the complementary term. (c) Estimated forearm (yellow) with the complementary term.

elbow and hand positions are correct and they generate adequate R_{ru} and R_{rf}. The S_{rarm} value is higher.

In practice, the per-pixel estimation of forearm probabilities by a random forest method is usually disturbed by other regions, especially the other hand's forearm. To address this problem, we included a complementary penalty term and rewrote the estimated function (2):

$$S'_{rarm} = \sum_{p \in R_{ru}} W_{ru} M_{ru}(p) + \sum_{p \in R_{rf}} w_{rf} M_{rf}(p) + \lambda \sum_{p \in R_{rf}} w_{rf} \left(1 - M_{\overline{rf}}(p)\right), \quad (3)$$

where $M_{\overline{rf}}(p)$ is the sum of all other parts' probabilities at position p except the left forearm and right forearm. λ controls the weight.

For example, in a hand-crossing case, the left and right forearm regions are mixed together, Eq. (2) generates shortened arms because the pixels of high forearm probabilities are separated. The complemental term helps the algorithm to find a dominating skeleton which covers the whole mixture region. Fig.4 shows an example with the complementary term.

3.2. Pose Approximation by RANSAC

Among the aforementioned objective functions, it is relatively straightforward to estimate individual joints, including head, torso and left/right shoulder joints. Their procedures are similar to the mean-shift algorithm [12], and the estimated points iteratively shift their positions toward centers of high probability regions.

By contrast, for the arm posture estimation, there are two variables Pre and Prh which can concurrently move. For efficiency and avoidance of trapping into the local minimum, we chose RANSAC algorithm to solve the corresponding objective functions. We randomly select two points from the upper-body region extracted in the first stage and calculate S'_{rarm} iteratively until an adequate result is retrieved. In practice, the possible position candidates are restricted only at positions with high probabilities with respect to the forearm. That is because when an upper arm and forearm overlap in a depth image, the forearm is mostly in front of the upper arm such that the upper arm region is partially occluded. Thus, elbow and hand joints should locate at the forearm region. The default iteration number of RANSAC is 1000.

4. POSE CORRECTION

Few failure cases may still occur with the above method and the estimated skeletons need to be rectified. Two additional steps are applied to keep the pose within reasonable scope. First, the estimated postures are projected into subspace generated by principal component analysis (PCA). Besides, our system amends unreasonable lengths of skeletons and invalid angles of human limbs.

It also checks whether the templates of the estimated body parts are able to fully cover the input depth regions, especially the arm parts. When low coverage situations occur, the proposed system automatically adjusts the elbow joints, and it increases the coverage between the synthetic and input regions.

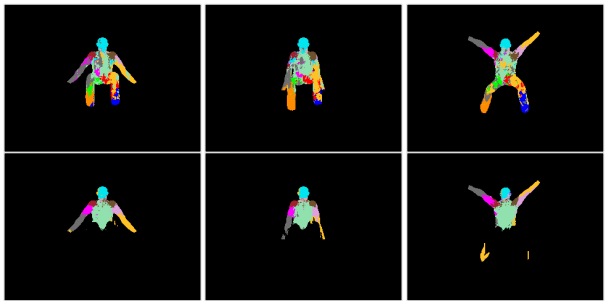


Fig.5 Classified result images of two methods at sitting poses. We illustrate the most probable part of each pixel. A color represents a certain body part. The upper row is estimated by a single-stage random forest. The lower row is estimated by the two-stage method.

Standing Poses	head	torso	Left shoulder	Left upper arm	Left fore-arm	Right shoulder	Right upper arm	Right fore-arm
Single stage	0.862249	0.901332	**0.819069**	0.583980	**0.892386**	**0.795690**	**0.681857**	**0.856755**
Two stage	**0.871207**	**0.928993**	0.737761	**0.612904**	0.869445	0.762663	0.564690	0.762886
Sitting Poses	head	torso	Left shoulder	Left upper arm	Left fore-arm	Right shoulder	Right upper arm	Right fore-arm
Single stage	0.662479	0.684488	0.639641	0.403708	0.302592	0.526079	0.493677	0.211280
Two Stage	**0.86800**	**0.906886**	**0.684056**	**0.670106**	**0.592193**	**0.535756**	**0.598233**	**0.506838**

Table I. The per-pixel labeling accuracy of each body part. A pixel is assigned to the body part of the maximum probability. The two-stage model generates more accurate results on the forearm regions in sitting poses.

5. EXPERIMENT RESULTS

The proposed method is designed for one- or multi-user skeleton estimation. Running on a desktop with Intel i7 CPU and Nvidia GTX770 GPU, our system can calculate more than two users' skeletons in real time.

We compared the performance of the single-stage random forests classifier and our two-stage method for upper body estimation. For the single-stage random forest method, we adopted approximately 30,000 synthesized depth images, of which the postures were acquired from the CMU mocap database [2] to train this classifier. For the two-stage estimation model, we firstly employed about 10,000 synthesized images to train the first upper and lower body classifier. For the second phase, we adopted about 20,000 images to train the detailed upper-body-part classifier.

We tested two pose datasets from CMU mocap database [2]: one contains standing actions, and the other contains sitting actions. Each dataset has around 1000 frames. We compared the classified results with the ground truth body part regions and calculated the accuracy. Fig. 5 shows several examples. Table I lists the per-pixel accuracy of labeling according to the maximum probability. In standing poses, both methods perform acceptably in accuracy, but in sitting poses, the two-stage model outperforms the single-stage one in arm regions, especially in forearm region which is important in gesture/motion tracking. We would like to stress that our skeleton extraction uses all the probability maps instead of the labels of maximum probability. Fig. 6 shows examples of the proposed upper body skeleton estimation from a live depth camera.

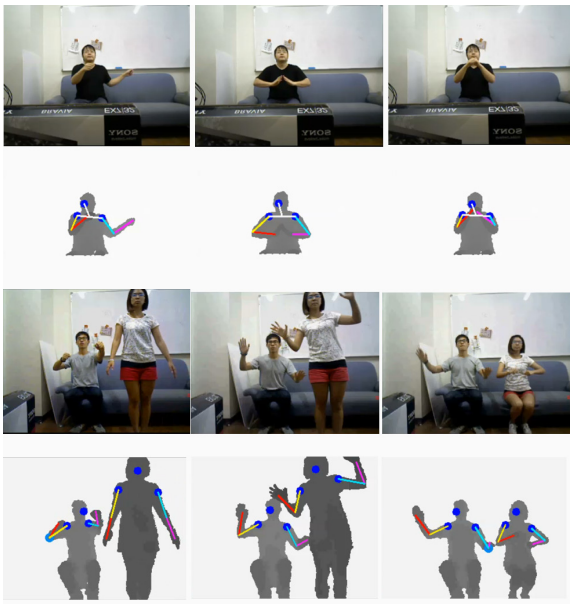


Fig.6 Upper body posture estimation from a live depth camera. In the first case, the lower body is even partially occluded.

6. CONCLUSION

This paper aims at estimating upper body postures from a sequence of depth images. In order to improve the per-pixel classification accuracy with only a moderate size of training data, we propose a two-stage classification model. The probability maps estimated by the classifiers are applied to objective functions for skeleton extraction, where a RANSAC method is used to estimate approximate results. The proposed framework is adapted to parallel computing with GPU acceleration, and therefore, it can estimate multi-user poses in real-time as Fig.6.

There are two main contributions in this paper. First, the two-stage method generates more stable results under the condition of compact training data. Second, the objective functions based on probability maps are presented to make the estimated pose more reliable, while per-pixel labeling methods usually fail due to self-occlusion and noise. One possible future work is to utilize temporal coherence or motion patterns [13] to further improve the estimation results.

ACKNOWLEDGEMENT

Authors would like to thank Josh Shu and Jing-Cheng Li for their help in data collection and GPU coding. This project was partially supported by Ministry of Science and Technology, Taiwan under grant no. NSC 102-2221-E-009-081-MY3 and MOST 104-2218-E-009-008.

8. REFERENCE

[1] A. Agarwal and B. Triggs , “Recovering 3D Human Pose from Monocular Images,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.28, no.1, pp. 44-58, 2006.
[2] CMU Mocap Database. http://mocap.cs.cmu.edu.
[3] D. Anguelov, B. Taskar, V. Chatalbashev, D. Koller, D. Gupta, and A. Ng, “Discriminative learning of markov random fields for segmentation of 3D scan data,” *IEEE Conference on Computer Vision and Pattern Recognition*, vol.2, pp. 169-176, 2005.
[4] D. Grest, J. Woetzel, and R. Koch, “Nonlinear body pose estimation from depth images,” *German Association for Pattern Recognition(DAGM)*, 2005.
[5] H. Lu, X. Shao, and Y. Xiao, “Pose Estimation With Segmentation Consistency,” *IEEE Transactions on Image Processing*, vol. 22, no. 10, Oct. 2013.
[6] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. “Real-time human pose recognition in parts from single depth images,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1297–1304, 2011.
[7] J. Shotton, R. Girshick, A. Fitzgibbon, 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. Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2821–2840, 2013.
[8] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001.
[9] M.A. Fischler and R.A. Elschlager. “The representation and matching of pictorial structures,” *IEEE Transactions on Computer*, 22(1), pp. 67–92, January 1973.
[10] M. Dantone, J. Gall, C. Leistner, and L. Van Gool, “Human Pose Estimation using Body Parts Dependent Joint Regressors,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3041-3048, 2013.
[11] P. Felzenszwalb and D. Huttenlocher, “Pictorial structures for object recognition,” *International Journal of Computer Vision*, vol. 61, no. 1, pp. 55–79, Jan. 2005.
[12] D. Comaniciu and P. Meer, "Mean Shift: A Robust Approach Toward Feature Space Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, 2002.
[13] I-C. Lin, J-Y. Peng, C-C. Lin, M-H Tsai, "Adaptive Motion Data Representation with Repeated Motion Analysis," *IEEE Trans. Visualization and Computer Graphics,* vol. 17, no. 4, pp. 527-538, April, 2011.