3D Hand Skeleton Model Estimation from a Depth Image

Chin-Yun Fan, Meng-Hsuan Lin
Inst. of ISA, National Tsing Hua University, Hsinchu, Taiwan
mayral1021@livemail.tw,ari12292000@gmail.com

Te-Feng Su, Shang-Hong Lai
Dept. of Computer Science, National Tsing Hua University, Hsinchu, Taiwan
{tfsu,lai}@cs.nthu.edu.tw

Chih-Hsiang Yu
Industrial Technology Research Institute, Chutung, Taiwan
SeanYu@itri.org.tw

Abstract

In this paper, we present an algorithm for estimating 3D hand skeleton model from a single depth image based on the Active Shape Model framework. We first collect a large amount of training depth images, representing all articulated hand shape variations, and a set of hand joint points are labeled on these depth images. To accommodate the wide variations of hand articulations, we represent the hand skeleton model with multiple PCA models that are learned from the training data. In the search stage, we iteratively compute the translation and rotation from the hand depth information and fit the 3D hand skeleton model with the multiple PCA models. In addition, we modify the model fitting procedure to handle the partial occlusion problem when only some fingers are visible. In our experiments, we demonstrate the proposed algorithm on our hand depth image datasets to show the effectiveness and robustness of the proposed algorithm.

1 Introduciton

Capturing human hand articulation from images or videos has received increasing attention for many computer vision applications, such as hand tracking, hand gesture recognition, human-computer interfaces, etc. Although there has been much work on hand articulation over the past decades, human hand motion exhibits high degrees of freedom with large viewpoint variations and partial occlusion, which still make the hand skeleton estimation problem very challenging.

In order to reduce the problem complexity, Wang and Popovi'c [3] used a single camera to track a hand wearing an ordinary cloth glove. However, such devices somehow constrain the field of applicability, since the hand requires wearing an additional device which is usually expensive. Several attempts have been made to overcome the problem by using markerless visual data [1, 2, 4, 6, 7]. Vision-based analysis of hand is probably the most natural way of achieving hand articulation estimation.

In recent years, the emergence of depth cameras, such as Kinect camera, has opened new possibilities for acquiring depth information. Following the popularity of depth sensor, depth information based hand articulation estimation gains considerable attention. Oikonomidis et al. [2] formulated 3D tracking of hand joints as an optimization problem that minimizes the discrepancy between the 3D structure and appearance of hypothesized 3D hand model instances. Qian et al. [5] modeled a hand simply by using a number of spheres. Then they proposed a hybrid method that combines gradient based and stochastic optimization methods to estimate the 3D hand model with fast convergence and good accuracy.

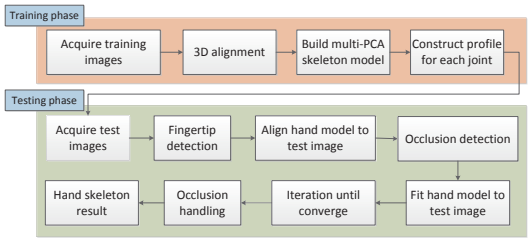
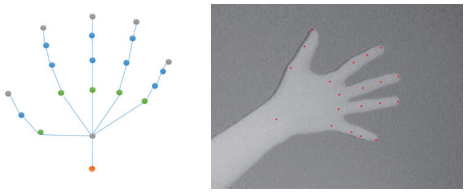


Figure 1. System flowchart of the proposed 3D hand skeleton model estimation.

In this paper, we propose a 3D hand skeleton model estimation algorithm from depth images by using Active Shape Model (ASM). Firstly, we collect a large amount of training depth images, representing all hand shape variations labeled with a set of landmark points. Here, hand joints are considered as landmarks. Then we model shape deformation as the statistical hand shape model by performing PCA which learns the deformation through a set of training examples. A hand shape is generated by adding a linear combination of some significant basis of variation to the mean shape. In the search procedure, we first give an initial estimation of the hand model. Then we compute the global translation and rotation of the hand model by using an iterative approach from the hand depth image. In addition, the proposed method is developed to handle the occlusion problem when some fingers are not visible in the image. Fig. 1 illustrates the flowchart of the proposed ASM-based algorithm for 3D hand articulation estimation from depth images.

2 Hand shape model with multiple PCA

In this paper, We present a hand articulation estimation method based on ASM. In order to build a 3D hand skeleton model by using PCA, we first collect a large amount of hand depth images of differ-



(a) (b)

Figure 2. (a)The example of hand articulations (b)labeled joints on the hand depth image.

ent pose variations along with labeled joint positions. The hand model is represented in 3D space by a set of 21 joint points, including 5 fingertips, 1 middle joint (on thumb) or 2 middle joints (on other fingers), 1 base joint between each finger and palm of each finger, wrist and palm. An example of hand skeleton model is shown in Fig. 2 (a). Here, we label the landmark points on color images and then transform them to depth images by homographic transforms estimated from the camera calibration process since these kinematic joints are difficult to identify in depth image. Fig. 2 (b) illustrates the labeled joints in the hand depth image.

2.1 Hand shape model

Consider a hand image I with n joint points and (x_q, y_q, z_q) represents 3D coordinate of the q-th joint point, where $1 \leq q \leq n$. Therefore, a hand skeleton can be represented as a vector $\mathbf{h} = (x_1, ..., x_n, y_1, ..., y_n, z_1, ..., z_n)$. Instead of aligning 2D sets of image features, a standard alignment between 3D point sets is required in this work. In our case where the 3D transformations are linear in the motion parameters, regular least squares can be used to align these skeleton joints over all training data.

The following energy E is minimized to compute the Euclidian motion between two hand point sets $\mathbf{h}_i$ and $\mathbf{h}_j$. The energy function is formulated as follows:

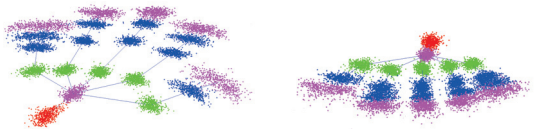
$$E = \sum_{\mathbf{x}_q \in \mathbf{h}_i, \mathbf{u}_q \in \mathbf{h}_j} \|\mathbf{x}_q - \mathbf{R}\mathbf{u}_q - \mathbf{t}\|^2 \qquad (1)$$

where $\mathbf{t}$ and $\mathbf{R}$ represent the translation and rotation between two point sets, respectively. The weighted centroids of the two point sets $\mathbf{c}_{\mathbf{h}_i}$ and $\mathbf{c}_{\mathbf{h}_j}$ can be used to estimate the translation $t = \mathbf{c}_{\mathbf{h}_j} - \mathbf{R}\mathbf{c}_{\mathbf{h}_i}$.

The rotation between two point sets $\hat{\mathbf{h}}_i = \mathbf{h}_i - \mathbf{c}_{\mathbf{h}_i}$ and $\hat{\mathbf{h}}_j = \mathbf{h}_j - \mathbf{c}_{\mathbf{h}_j}$, which are both centered at the origin, can be estimated by computing the SVD of 3×3 correlation matrix. Then the rotation matrix is obtained as $\mathbf{R} = \mathbf{U}\mathbf{V}^{\mathbf{T}}$. Thus a statistical hand shape model can be learned through a set of aligned training hand skeleton models. The hand mean shape and modes of variation are found by using PCA. Therefore, a hand shape is generated by adding a linear combination of some significant modes of variation to the mean shape. Thus, the hand articulation of joint positions can be represented as

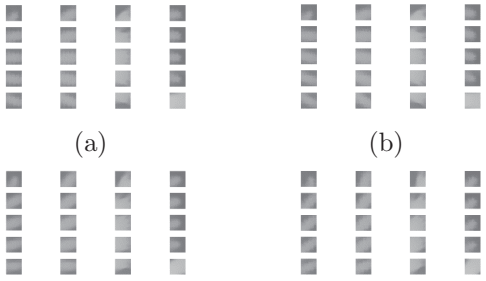
$$\mathbf{h}_{new} = \bar{\mathbf{h}} + \mathbf{P}\mathbf{b} \qquad (2)$$

where $\bar{\mathbf{h}}$ is mean shape, $\mathbf{P}$ is a matrix consisting of the eigenvectors computed from PCA and $\mathbf{b}$ is a vector



(a) (b)

Figure 3. The alignment result and mean shape (shown in line). Left: 2D view point. Right: 3D view point.



(a) (b) (c) (d)

Figure 4. Four different orientations of appearance template (exclude palm) (a) −20° (b) −30° (c) 0°(d) 20°

of the associated coefficients to be determined in the fitting process. Fig. 3 shows the aligned training data and mean shape of hand.

2.2 Appearance model

After the hand shape alignment and PCA process, we construct the appearance model for each joint. Take the rotation of the hand pose on an image plane into account, the appearance model used as the search for finding new positions for the joint points should be built with different orientations of hand models. Different from the traditional ASM, we use the same training images and rotate the appearance patches for four angles to compute the means and corresponding covariance matrices for each of the joint points as templates. Fig. 4 shows the appearance patches in four orientations of all joints.

2.3 Hand shape model with multiple PCA

To accommodate large variations in hand pose, we collect a large amount of images of different hand poses and cluster these labeled hand vectors in a high-dimensional space by k-means clustering. Based on the clustering result, we train PCA models individually for each of the multiple clusters and apply the multiple PCA models to hand tracking. Our experimental result proves the multi-PCA model outperforms the single PCA model for the case of large hand pose variations.

3 Hand articulation estimation

In the search procedure, for a new testing hand depth image, the fingertips and palm detection pro-

Algorithm 1 Occluded ASM based fitting algorithm

Input: depth image I, fingertips F, palm P position, and mean shape $h_{MeanShape}$
Output: hand joints position of remaining joints $\widetilde{h}_{new}$
Initialization: $T \leftarrow combination(F)$
Foreach $i \leftarrow 1, ..., M$ // M possible combinations
  $h_i \leftarrow Transform(h_{MeanShape}, F, P, T_i)$
  Compute $FittingError(h_i)$
End foreach
$\widetilde{h}_{new} \leftarrow h_i$
While
  Foreach remaining joint j in $\widetilde{h}_{new}$
    $\widetilde{h}_{new,j} \leftarrow ASMfitting(\widetilde{h}_{new,j}, I)$
  End foreach
    $\widetilde{h}_{new} \leftarrow \widetilde{h}_{MeanShape} + \widetilde{P}b$
Until converge

vides the position information of visible fingertips and palm. Then we can used it to transform the mean shape on the image to obtain an initial position of each joint. After that, in each iteration, new position of each joint in this image will be found through comparing with the appearance model built in the previous training stage. The different sizes of patches are chosen in a predefined search edge, e.g. 15×15, and we calculate the Mahalanobis distance for each of them with the template. The position as a center of a patch with minimum value of distance will be selected as the new position of this joint; this process is performed for local optimization and after all joints have updated their new positions, the trained PCA model will be used to adjust the hand shape. The coefficient b can be computed as follows:

$$b = P^T(h_i - h_{MeanShape}) \tag{3}$$

The whole fitting process iterates until the coefficient vector b converges, i.e. the change in b between consecutive iterations is insignificant. According to Eq. 2, the new hand joint positions are computed.

3.1 Occlusion handling

It is inevitable to suffer from the hand-articulated occlusion in practice. Intutively, extra occluded hand shape models are needed to estimate accurate hand joint positions. In the proposed method, instead of occluded hand shape models, we modify the ASM fitting procedure to solve the partial occlusion problem in hand articulation. For the occlusion case, we first detect the fingertips and palm. Then the fitting errors are evaluated to identify the occluded fingers by trying all combinations of visible fingers according to the depth value and locations of joints on the hand depth image. The new positions $\widetilde{h}_{new}$ of the remaining joints will be captured through the PCA model and ASM fitting method described previously. Consider the occluded joints as missing data in this work, the origin PCA model should be adjusted by excluding all joints lying on occluded fingers in mean shape and removing corresponding eigenvectors. Therefore, the coefficient vector b in Eq. 3 is used to adjust the whole hand shape through PCA model, which is given as follows:

Algorithm 2 Occluded joint detection procedure

Input: hand shape of remaining joints $\widetilde{h}_{new}$, depth image I
Output: Final hand shape h_{new}
Foreach $a \leftarrow 1, ..., M$ // occluded Fingers
  flag $\leftarrow 0$
  $Cp \leftarrow \widetilde{h}_{new,basejoint(a)}$ //position of start joint
  Foreach j in a // middle joints
    $h_{occl,j} \leftarrow Cp + FingerVector(h_{i,basejoint(a)}, h_{i,j})$
    while $h_{occl,j}$ in background **do**
      $h_{occl,j} \leftarrow h_{occl,j} -$
        $scale * FingerVector(h_{i,basejoint(a)}, h_{i,j})$
      $flag \leftarrow 1$ // next joint not exist
      $h_{occl,j} \leftarrow CheckDistance(h_{occl,j}, C_p,$
        $FingerLength_{basejoint(a)})$
      $Cp \leftarrow h_{occl,j}$
      If flag equal to 1
        break
      End if
    End foreach
End foreach

$$b = (\tilde{P}^T\tilde{P})^{-1}\tilde{P}^T(\tilde{h}_{new} - \tilde{h}_{MeanShape}) \tag{4}$$

Algorithm 1 presents the high-level ASM fitting procedure for partial joints with adjusted PCA model. For joint positions h_{occl} on the occlusion finger, we start from the base joint found in previous step and plus the finger vector providing the approximate position of the next point. If the depth value of that point lies on background, then search along opposite direction until it lies on foreground (hand). Moreover, checking the 3D distance from the searched point to base joint in a search range by the mean length of the finger derived in the training process to assure accuracy. Algorithm 2 summarizes the occluded finger joint detection procedure.

4 Experimental results

In order to construct a statistical hand shape model, we collected 558 training depth images from 62 subjects performing 2 totally open hands, 4 different opening shapes and 3 different fingers bendings. Both training and testing depth images are captured from Softkinetic camera. We can obtain both color and depth images; however, only depth images are utilized to detect hand joints. The test hand depth images contain different conditions, such as non-occlusion, finger bending and occlusions, and large variations of hand poses. In our work, a good initial guess for the hand articulation estimation is required. We firstly detect fingertips and palm by finding obvious peaks of the hand contour. Then the affine transformation is computed from the corresponding fingertips between learned hand shape model and those detected from the test image. Once the learned hand shape model is transformed to the coordinate of the test image, the following search procedure is performed for estimating the hand articulation.

Fig. 5 (a)$\sim$(d) shows the hand skeleton estimation results in the hand depth images without occlusions. It

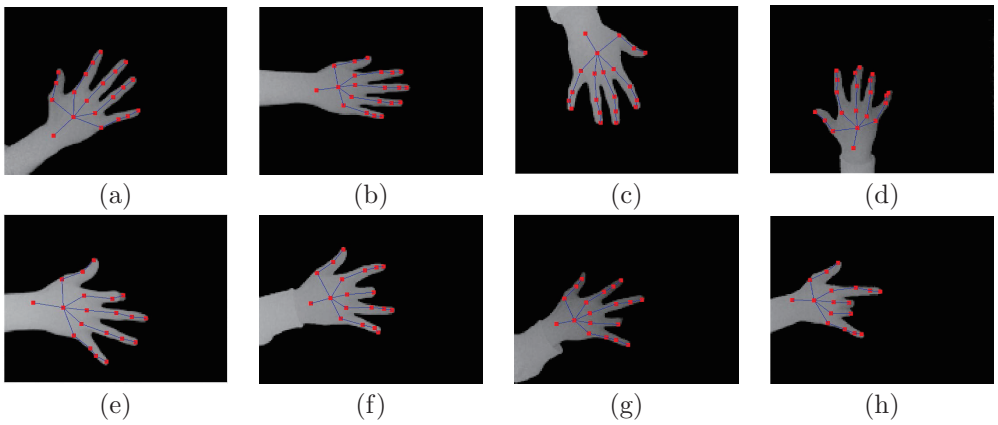


Figure 5. Examples of 3D hand skeleton model estimation for different cases (a)∼(d) without occlusion and (e)∼(f) with self-occlusion.

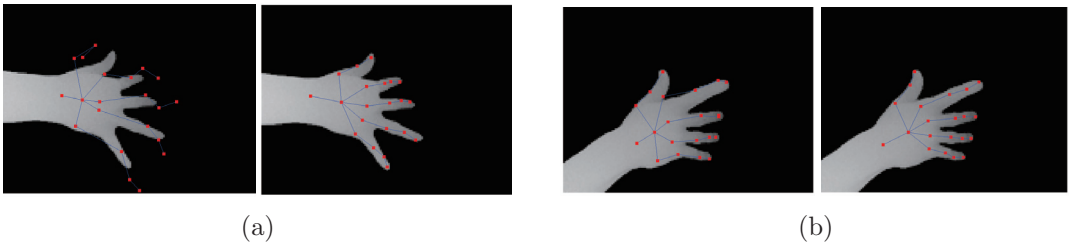


Figure 6. Two examples show the hand skeleton model estimation results by using the single-PCA hand shape model and multi-PCA hand shape models. Left: estimation by single-PCA hand shape model. Right: estimation by using the multi-PCA hand shape model.

is obvious that that the hand joints are captured accurately even though the hand pose of test images contain different variation of opening and bending of fingers. The proposed method can handle the rotation of hand as well. Furthermore, Fig. 5 (e)∼(h) demonstrates the hand skeleton estimation under self-occlusions which only partial articulations are visible in the hand depth image. We can see that accurate estimation of hand skeleton models can be obtained under different occlusion conditions of hand fingers by using the proposed algorithm. In addition, Fig. 6 shows the 3D hand skeleton model estimation results for large variations of hand poses by using the single-PCA and multi-PCA hand shape models, respectively. It is obvious that the multi-PCA hand shape model performs significantly better than the single-PCA hand shape model in our experiments. The results demonstrate the multi-PCA hand shape model is more suitable to handle the hand pose with large variations.

In addition, we also perform quantitative accuracy assessment of the skeleton model estimation with the ground truth data of the 3D finger joint points manually labeled. For 100 non-occluded hand depth images, the average fitting error is 12.0752 mm. For other 73 occluded hand depth images, the average fitting error is 14.628 mm.

5 Conclusions

In this paper, we proposed a new approach for 3D hand skeleton model estimation from a depth image by using an improved ASM algorithm. In our algorithm, a multi-PCA hand skeleton model and a modified ASM fitting process were developed to handle the occlusion problem when only partial fingers are visible in the image. In our experiments, we demonstrated the robustness of the proposed algorithm on different conditions, including partial occlusion, finger bending and large variations of hand pose.

References

[1] Y. Wu, J. lin, Thomas. S. Huang "Capturing natural hand articulation," *ICCV*, 2001.

[2] I. Oikonomidis, N. Kyriazis, and A. Argyros. "Efficient model-based 3D tracking of hand articulations using kinect," *BMVC*, 2011.

[3] R. Y. Wang and J. Popovic. "Real-time hand-tracking with a color glove," *ACM Transactions of Graphics (TOG)*, 2009.

[4] D. Tang, T.-H. Yu and T.-K. Kim. "Real-time articulated hand pose estimation using semi-supervised transductive regression forests," *ICCV*, 2013.

[5] C. Qian, X. Sun, Y. Wei, X. Tang and J. Sun. "Real-time and Robust Hand Tracking from Depth," *CVPR*, 2014.

[6] M. de La Gorce, D. J. Fleet, and N. Paragios. "Model-Based 3D Hand Pose Estimation from Monocular Video," *IEEE Trans. PAMI*, vol. 33, pp.1793-1805, 2011.

[7] D. Tang, H. J. Chang, A. Tejani and T.-K. Kim. "Latent Regression Forest: Structured Estimation of 3D Articulated Hand Posture," *CVPR*, 2014.