High-Speed Estimation of Multi-finger Position and Pose for Input Interface of the Mobile Devices

Yoshihiro Watanabe∗, Atsushi Matsutani*, Takehiro Niikura†, Takashi Komuro†, and Masatoshi Ishikawa∗

∗Graduate School of Information Science and Technology, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan.
Email: Yoshihiro_Watanabe@ipc.i.u-tokyo.ac.jp
†Graduate School of Science and Engineering, Saitama University
255 Shimo-Okubo, Sakura-ku, Saitama City, 338-8570, Japan.

Abstract—Mobile devices are too small to operate freely using their input surfaces. To solve this problem, non-contact and natural gesture interfaces have been the focus of recent research. In this paper we propose a method of estimating multi-finger position and pose for operating such devices at high speed using a single camera. Our method achieves the finger tracking based on the appearance and shape deformation model by estimating the translational movements and the degree of bent finger. The experimental results show that our method can obtain the position of the hand and the pose of the each finger within 9.7 ms.

I. INTRODUCTION

Handheld mobile devices are now widespread in everyday life and are becoming more and more sophisticated in functions and smaller in size. However, the downsizing causes the narrow operation space. To make full use of the available features and to enhance usability, there is a great deal of research exploring alternative interaction systems for operating mobile devices, not only traditional desktop GUIs.

For example, Harrison et al. [1] developed a system that allows users to move their finger in space for input operations on mobile devices. In this system, a magnetic device is placed on the finger. Misty et al. [2] proposed a wearable interface that offers natural hand gestures. By using a camera attached to a pendant, this system recognizes the gestures as the movements and arrangements of colored markers on the fingers. Niikura et al. [3] developed a vision-based interface with a display and a single high-speed camera. In this system, the user holds the device and types on a virtual keyboard by moving his or her finger quickly and freely. However, this system is limited in only one finger operation.

We focus on the strong potential in Niikura’s system and propose the extended operations based on multi-fingered input such as typing as shown in Fig. 1. The key technology is the vision-based multi-finger recognition.

Erol et al. [4] reviewed the region of finger pose estimation. Two general approaches have been explored to estimate finger pose. The first one is the 3D model-based approach, which takes advantage of 3D hand models, and the second one is the appearance-based approach, which directly connects 2D image features with hand configurations. It has been difficult for both of these approaches to achieve the high-speed operation and the high-accuracy estimation. Also the two approaches commonly rely on high quality segmentation and feature extraction of the hand.

In order to solve the conventional problems, our method introduces a multi-finger deformation model based on Active Appearance Model (AAM) [8]. The proposed method estimates the position and the rotation of the hand and the pose of the four fingers. Fig. 1 illustrates the estimated parameters. In this paper, thumb is not included in order to make the model compact, although it is possible to be added in theory. Also, in the focused task, the estimation speed is important in order to observe the high-speed finger motion and provide a quick response to the user operation.

II. REQUIREMENTS FOR MULTI-FINGER TRACKING

This section describes the required functions of the vision system for the in-air input in the mobile devices. As the new styles of input, we focus on providing the typing and 3D operation in the wide air area. In order to realize this type of input, the motions shown in Fig. 1 are required to be observed. The motions include translations \((X_f, Y_f)\), rotation about the optical axis \(\phi_f\), bending posture of the 1st and 2nd finger joint \((\psi_{f1}, \psi_{f2})\) and the depth position \(Z_0\). \(f\) indicates the label of each finger. In this paper, four fingers shown in Fig. 1 are the targets to be observed. Therefore, the label is \(f = 1, 2, 3, 4\).

We assume the finger posture changes from straighten state to the bend state where the fingertip direction is perpendicular to the palm. Therefore, when we define the 1st and 2nd joint as \(\psi_{f1}\) and \(\psi_{f2}\) (when straightened, \(\psi_{f1} = 0\)), the range of the bend is \(\psi_{f1} + \psi_{f2} < \pi/2\). Based on the anatomical insight, those two parameters \((\psi_{f1}, \psi_{f2})\) can be assumed to work with...
each other. This assumption is modeled as $\psi_{f1} = \frac{2}{5} \psi_{f2} [4]$. Using the bend parameter $\beta_f \in [0, 1]$, those joint angles are expressed as $\psi_{f1} = \beta_f \cdot \pi/5, \psi_{f2} = \beta_f \cdot 3\pi/10$.

These hand and finger motions are observed in images as translation $(t_{f1}, t_{f2})$, rotation $\theta_f$, hand size $k$ and finger shape $s_f$. Using these observed image features, the method is required to estimate the original motion state. In this paper, it is assumed that the user’s palm and the camera image plane is parallel and the occlusions of the fingers are not occurred. Also, we assume that the maximum speeds of the translation and the wrist rotation are 1 m/s and 300 deg/s respectively.

Next, we describe the camera specification. In this paper, we assume to use a VGA camera (resolution: 640 × 480) and a wide-angle lens (angle of view: 120 × 90 deg). The hand is assumed to be moved around 20 cm from the camera. The frame rate is important for high-accuracy estimation of the high-speed moving hand because the high-frame-rate imaging can reduce the difference between frames. This point has great impact on the tracking. In this paper, we assume to use 100-fps camera. Such camera has been introduced in general commercial devices [6], [7] and it is considered to be acceptable. Under these conditions, the difference between frames can be calculated. Considering the motion at maximum speed, the observed change in the successive frames is around 10 pixel/frame. Therefore, our method focuses on meeting this requirement.

III. MULTI-FINGER TRACKING AND POSE ESTIMATION

A. Overview

As advance preparation, various sample images capturing the finger motion are collected and the finger deformation model is generated. Using this model, the original states of the fingers captured in input image are estimated. The overview of the online operation is shown in Fig. 2. First, we estimate the parameters to fit the deformation model to the fingers in the captured image. Next, using the estimated parameters, the finger shapes are obtained. Finally, the bend posture of each finger is estimated using this shape.

In the fitting step, we use two types of deformation models. One is the four-finger model and the other is one-finger model. Four-finger model is the deformation of the region covering all fingers. One-finger model expresses the each finger deformation independently. Four-finger model is good to find the finger region from the input image roughly but difficult to reduce the small variation. Therefore, we first apply the fitting by using the four-finger model and secondly apply the fitting by using the one-finger model. For precise alignment, as shown in Fig. 3, in the one-finger-model fitting, we use the edge image. This is effective to fit to the contour of the finger which expresses the finger bend.

![Fig. 2. Overview of the proposed method](image)

![Fig. 3. staged estimation using two types of deformation models](image)

B. Building finger deformation model

Our method builds two sets of deformation model including four-finger model and four one-finger models. The deformation is described in two types of models including shape and appearance. Using the training dataset, these models are constructed as follows.

The main feature points $m_k(k = 1, \cdots 7)$ is set to each finger as shown in Fig. 4. There is a point on the top of the finger and are six points on the wrinkle of the three joints. In addition with this main feature points, $n$ optional feature points $i^k_i (k = 1, \cdots 6, \quad i = 1, \cdots n)$ are set as interpolating between $m_k$ and $m_{k+1}$. In this paper, we set $n = 4$ and the total number of points is 31. The appearance of the finger is expressed by using a texture within the patch which is segmented by Delaunay triangulation using these points. This patch segmentation has advantages because the relative location of the patch is not changed with the deformation caused by the finger bending. These settings are applied to each finger. We call each one-finger model.

![Fig. 4. One feature points and interpolation points](image)
where \( s_0 \) is the average shape and the coefficient \( p_i \) is the weight of the \( i \)-th eigenvector. The shape variation is represented by changing the shape parameter \( \lambda = [\lambda_1 \ \lambda_2 \ \cdots \ \lambda_m]^T \).

The appearance of an image is an image vector \( A(x) \) (brightness value) defined over the pixels \( \forall x = (x, y)^T \in s_0 \). Like the shape model, the appearance model can be expressed as a base appearance \( A_0(x) \) plus a linear combination of \( m \) appearances:

\[
A(x) = A_0(x) + \sum_{i=1}^{m} \lambda_i A_i(x) \quad \forall x \in s_0
\]

We express the appearance parameter as \( \lambda = [\lambda_1 \ \lambda_2 \ \cdots \ \lambda_m]^T \).

In our method, we generate shape and appearance model for the four-finger model. On the other hand, only the shape model is generated in the one-finger model. These models are learned in advance using large datasets of finger motion.

2D similarity transformation (translation, rotation, and scaling) can be represented as affine warping \( q \), which is a different parameter set from that of deformation, by building the model after normalization.

C. Staged estimation using multiple models

Our method employs staged estimation using the multiple models for robust performance. First, the method achieves the rough estimation using the four-finger model. This step is good to find the finger region from the input image but difficult to reduce the small variation. Therefore, secondly, the fine estimation is achieved by using the one-finger model. In this step, we utilize the finger contour in order to recognize the precise finger bending.

1) Rough Estimation using four-finger model: Based on the method of the Inverse Compositional Image Alignment [8], this step can be formulated as the minimization problem:

\[
\min_{\Delta p, \Delta q} \sum_{x \in s_0} \left[ A_0(N(W(x; \Delta p); \Delta q)) - I(N(W(x; p); q)) \right]^2
\]

\[
I(x) \text{ is the input image, } W(x; p) \text{ stands for a piecewise warping appearance } A \text{ from } s \text{ to } s_0 \text{ by } p. \text{ Similarly, } N(x; q) \text{ stands for a 2D similarity transformation. This transformation is described by scale change } k, \text{ rotation } \theta, \text{ and translation } (t_x, t_y). \text{ The problem in equation (3) is solved based on quasi-Newton’s method. Using the estimated } \Delta p \text{ and } \Delta q, \text{ } p, \text{ } q \text{ are updated as follows.}
\]

\[
N(W(x; p); q)) \leftarrow N(W(x; p); q) \circ N(W(x; \Delta p); \Delta q)^{-1}
\]

2) Estimation using contour one-finger model: In order to improve the accuracy, this step re-estimate the finger shape in which the feature points are located on the contour of the finger in an image accurately. In this re-estimation step, the shape parameter \( s \) in one-finger model is estimated by maximizing the following equation:

\[
\max_s \sum_{x \in s} I'(x)
\]

\( I'(x) \) is the gradient input image. All the pixel value is positive. This maximization problem is solved by moving the each feature point iteratively with the limited deformation. First, at each feature point \( x \), the normal direction is calculated, the pixel values on the positive and negative direction within a pixel (4 pixel in the experiment) are obtained, and the feature point is moved to the location of the maximum value. After this operation, the model parameter \( p \) is calculated. If the condition \( p_i > m \sigma_i \) ( \( m \) is constant value and \( \sigma_i \) is the corresponding eigen value in PCA) is meet, \( p_i \) is replaced with \( m \sigma_i \). This technique is introduced to maintain the finger shape described in the model. This search is repeated until its convergence.

D. Multi-Finger pose estimation

Since the motion of a bending finger at the first joint and the second joint has correlation between joints, the bending pose could be expressed as the degree of bending with a small number of statistical parameters. Therefore, we introduce a bending parameter \( \beta \in [0, 1] \) to represent the degree of bending. In particular, we assume the following equation based on the anatomical insight: \( \psi_{f1} = \beta_f \cdot \pi/5, \psi_{f2} = \beta_f \cdot 3\pi/10. \)

\( f : f = 1, \ldots, 4 \) indicates the label of each finger.

Moreover, we linearly approximate the relationship between this bending parameter \( \beta_f \) and the shape parameter vector \( p_f \) based on one-finger model, given by

\[
\beta_f = R_f \left[ p_f, 1 \right]^T
\]
IV. EXPERIMENTS

A. Experimental condition

This section describes the experimental results. Fig. 6 shows the experimental setup. The used camera was Point Grey Research Inc. Firefly MV. The resolution was 640 × 272 and the frame rate was 100 fps. As the master data, we used the successive images capturing the finger bending motion, hand translation and the motion spreading the fingers. Each motion consists of 92, 71 and 100 images respectively. Four-finger model and one-finger model are 36 and 9 dimensions after applying the PCA. The contribution rate of the four-finger and one-finger model was 99.6 % and 97 % respectively.

B. Fitting accuracy

We evaluated the fitting accuracy using the finger bending motion. In this experiment, the initial parameter for the estimation was generated by changing the parameters of the deformation, translation, rotation such that the average displacement between the true and the changed feature points was within μ pixel. The example input setting is shown in Fig. 7. In order to evaluate the performance effected by the displacements between successive frames, we changed the value μ and calculated the fitting error. The fitting error was calculated as the root mean square of the difference in the feature points between the true data and the results.

Also in this experiment, we compared the methods applying only the four-finger model fitting and the proposed method. The result is shown in Fig. 8. The horizontal axis shows the displacement μ and the vertical axis shows the fitting error. The red solid line shows the results using only four-finger model. Blue dash line shows the results using the proposed method introducing the staged estimation. Fig. 9 show the improvement example.

C. Estimation of finger bending

This section describes the experimental results of the estimation of finger bending. In this experiment, both of the repeat counts in the four-finger-model fitting and one-finger-model fitting were fixed to 10. In the test movie, at first, all four fingers were bent. Second, each finger was bent in turn. At the last, all four fingers were bent again.

D. Processing time

This section describes the processing time required for the proposed method. In this experiment, the repeat counts in the four-finger-model fitting and the one-finger-model fitting were fixed to 10 and 25 respectively. Also if the changes of the position per a single feature point by the update in the iterative calculation were within 1.0 pixel, we stop the operation at the frame.

The total processing time for a single image was 9.7 ms. The used CPU was Intel Corei7 2.9GHz. This performance can realize the 100-fps operation. The four-finger-model fitting took 3.4 ms, the one-finger-model fitting took 6.2ms, and the estimation of the bending parameter took 0.1 ms. Other example of the tracking results based on this setting are shown
Fig. 10. Example tracking results. In this tracking, fingers were bent in turn.

Fig. 11. Example tracking results for the images observing the moving hand.

Fig. 9. Improvement example. Left: the result using only four-finger model. Right: the result using the proposed method.

Fig. 12. Change of the bending parameters in Fig. 11.

V. CONCLUSION

We propose a new method of the single-camera-based multi-finger pose estimation for input interface in mobile devices. The method can be achieved at high speed and realize high-accuracy bending estimation.

REFERENCES