Particle Filter-Based Fingertip Tracking with Circular Hough Transform Features

Martin Do, Tamim Asfour, and Rüdiger Dillmann
Karlsruhe Institute of Technology (KIT), Adenauerring 2, 76131 Karlsruhe, Germany.
{*martin.do,asfour,dillmann*}*@kit.edu*

Abstract

In this work, we present a fingertip tracking framework which allows observation of finger movements in task space. By applying a multi-scale edge extraction technique, an edge map is generated in which low contrast edges are preserved while noise is suppressed. Based on circular image features, determined from the map using Hough transform, the fingertips are accurately tracked by combining a particle filter and a subsequent mean-shift procedure. To increase the robustness of the proposed method, dynamical motion models are trained for the prediction of the finger displacements. Experiments were conducted on various image sequences from which statements on the performance of the framework can be derived.

1 Introduction

Towards an intuitive and natural interface to machines, markerless human observation has become a major research focus during the past years. Regarding coarse granular human tracking incorporating the torso, arms, head, and legs, considerable progress has been achieved whereas human hand tracking still remains an unresolved issue, although the hand is considered to be one of the most crucial body parts regarding the interaction with other humans and the environment.

Regarding markerless tracking and detection of human body parts, most systems have limited sensor capabilities which in common case are limited to a stereo camera setup. Full hand tracking approaches in joint angle space have been proposed in [1],[2],[3]. However, due to the highly complex structure of the hand whose motion involves 27 DoF, tracking can be only achieved at a low frame rate or on multiple views from different perspectives.

Using stereo vision, a reasonable solution lies in reducing dimensionality of the problem by shifting from joint angle space into task space. In [4], a finger tracking approach based on Active contours is presented for air-writing. The target to be tracked consists of a contour which is laid around the pointing finger. As a result, since no reliable statement can be made on the actual fingertip position, one has to assume that finger pose is not changing.

Hence, based on curvature properties, in [5] fingertips are detected within a contour which is extracted from skin blob tracking. A more elaborate approach is presented in [6] where particles are propagated from the center of the hand to positions close to the contour. Intersection of the contour with line segments at particles and examination of the transitions between non-skin and skin-area indicate whether a particle represents a fingertip. However, this method is specifically designed to detect tips of stretched fingers. Based on multi-scale color features [7] introduces a hierarchical representation of the hand consisting of blobs of different sizes with each blob representing a part of the hand. The blob features are matched with a number hierarchical 2D models each incorporating a specific finger pose. Therefore, tracking is accomplished under the assumption that the local finger poses regarding the hand remains fixed. In order to implement a continuous fingertip tracking method, we would like to rely on prominent features which can be extracted at any time of an image sequence. In [8], for detecting a guitarist's fingertips, circular features are proposed which are localized by performing a circular Hough transform. For the same application, [9] defined semicircular templates which are used to find the fingers' positions.

In our work, we adopt the concept of circular features to tackle the more complex problem of tracking fingertips of a freely moving hand, where overlap of finger and palm occur frequently leading to difficulties regarding the robust extraction of these features. For tracking, we combined particle filtering with a mean-shift algorithm. In addition, a dynamical motion model for predicting was trained to enhance the robustness of the proposed framework.

The paper is organized as follows. Section 2 describes the feature extraction consisting of an edge detection step and the Hough transform. Details on the tracking procedure performed on the resulting map are given in Section 3. Subsequently, first experiments with the framework are explained in Section 4. In Section 5, the work is summarized and notes to future works are given.

2 Feature Extraction

In order to generate the edge image, a skin color segmentation is performed for extracting the hand and finger regions. Morphological operators are applied on the segmented image to eliminate noise and to produce a uniform region. To detect the edges in this preprocessed image, image gradients are calculated on various scales.

2.1 Multi-Scale Edge Extraction

Considering the problem of fingertip tracking, due to small intensity variances between different parts of the hand, e.g. the fingernail and the skin, respectively, the finger regions and the palm, it is desired to detect edges where contrast can vary over a broad range. Depending on the parameters, applying standard algorithms, such as the Canny edge detectors on a wider scale, leads to an edge image where numerous, false edges occur. To preserve low contrast edges in certain areas while reducing noise close to high-contrast edges, based on the work of [10], we implemented a filter approach consisting of a steerable Gaussian derivative filter on multiple

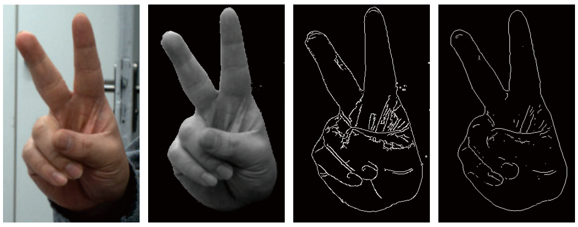

Figure 1. Left: Original input image. Center/Left: Color segmented image. Center/Right: Edge image using Canny detector. Right: Edge image using the method proposed in Section 2.1.

scales. The basis filter for x is defined as follows:

$$G_k^x(x,y;\sigma_k) = \frac{-x}{2\pi\sigma_k^4} e^{\frac{-(x^2+y^2)}{2\sigma_k^2}}. \tag{1}$$

$G_k^y(x,y;\sigma_k)$ is defined analogously. To determine the scale at which a gradient can be reliably estimated, the magnitude of the filter response $r_k^x(x,y;\sigma_k)$ and $r_k^y(x,y;\sigma_k)$, obtained by convolution of the image I with the filters from Eq. 1, is checked against a noise threshold. While the magnitude can be calculated according to:

$$r_k^m(x,y;\sigma_k) = \sqrt{r_k^x(x,y;\sigma_k)^2 + r_k^y(x,y;\sigma_k)^2}, \tag{2}$$

the threshold is set by following function:

$$c_k = \frac{\sqrt{-2ln(1-(1-\alpha)^R)}}{2\sigma_k^2\sqrt{2\pi}} s_l. \tag{3}$$

with s_l representing the standard deviation and α the significance level for an image with R pixels which defines an upper boundary for allowed misclassification of image pixels. To take into account local intensity and contrast conditions, we focus on local signal noise in a specific region rather than on global sensor noise. Therefore, Eq. 3 depends on the local standard deviation s_l calculated within a $2\sigma_k^{max} \times 2\sigma_k^{max}$-neighborhood where σ_k^{max} denotes the largest scale being examined. Hence, we calculate each gradient at the minimum reliable scale σ_k^{min} where the likelihood of error due local signal noise falls below a standard tolerance. This guarantees that a more accurate gradient map is estimated which is less sensitive to signal noise and errors caused by interference from nearby structures. The edge image obtained from the map is depicted in Fig. 1.

2.2 Hough Transformation for Circle Detection

The circle features representing the N fingertips are detected by applying a Hough transform with radius r. For each edge point (x,y) with known direction in the form of a rotation angle θ, a vote is assigned to possible circle feature positions (u,v) in two-dimensional Hough space I_H according to:

$$I_H(u,v) = I_H(u,v) + 1 \tag{4}$$

with $u = x \pm rcos(\theta)$ and $v = y \pm rsin(\theta)$. Unfortunately, curves around the fingertips do not always feature perfect circular arcs. To cope with noisy and slightly deformed curves, the voting is performed for a set of radii $R = \{-m \cdot 1.1 \cdot r, \ldots, m \cdot 1.1 \cdot r\}$ with $m \in \mathbb{N}$ whereby a range of pixels along the edge tangent is considered during the voting process. In order to increase the robustness of the tracking algorithm, a density distribution is formed in Hough space by convolving I_H with a Gaussian kernel $G(u,v;\frac{r}{2})$.

Since the hand motion occurs in 3D Euclidean space, a fixation of r is only valid if movement of the fingertip in direction of the z-axis of a camera is excluded during the tracking. Adaptation of r in each frame, allows to track fingers in all directions. Based on the generated density distribution in frame t a radius estimate $\hat{r}_t$ is determined by applying an Expectation Maximization algorithm. Further details are given in Section 3.3.

3 Tracking Fingertips

3.1 Prediction

Providing a prediction on the movement of the objects to be tracked increases the robustness of a statistical tracking framework. We train dynamical motion models in the form of a second-order auto-regresssive (AR) process as proposed in [4], which is described as follows:

$$q_t - \bar{q} = A_1(q_{t-1} - \bar{q}) + A_2(q_{t-2} - \bar{q}) + b_0\omega_k \tag{5}$$

where $q_t \in \mathbb{R}^D$ denotes the current configuration, $\bar{q}$ the mean configuration, and $\omega_k \in [0,1]$. To learn the AR parameters A_1, $A_2 \in \mathbb{R}^{D\times D}$ and $b_0 \in \mathbb{R}^D$, training data is provided in the form of a configuration sequence $Q = \{q'_0, \ldots, q'_M\}$ whereas the sequence is generated by manual labeling of fingertips in each frame of a recorded image sequence.

Two AR models are trained to provide predictions for the local fingertip movement concerning a static hand pose as well as the movement of the hand itself. Based on the assumption that the motion of each finger is influenced by the motion of the neighbored fingers, the first model is trained with training data whose instances $q'_i \in Q$ with $D = N$ consists of the length of the vector $v_t^{j,j+1} = q_t^{j+1 \bmod N} - q_t^j$ between the fingertip j and $j+1 \bmod N$:

$$q'_i(j) = \|v_t^{j,j+1}\| \quad j = 1, \ldots, N. \tag{6}$$

For finger j, this leads to following displacement vector:

$$\bar{v}_t^j = \frac{1}{2}\left(\sum_{i=j-1}^{j} A_1^N(i,i+1)v_{t-1}^{i,i+1} + A_2^N(i,i+1)v_{t-2}^{i,i+1}\right) + b_0^j\omega_k. \tag{7}$$

The second model which considers the global movement of the mean position of all fingertips p^m is trained with a data set formed of $q'_i = p^m$ with $D = 2$ resulting into an overall displacement:

$$\hat{v}_t^j = A_1^2(p_t^m - p_{t-1}^m) + A_2^2(p_{t-1}^m - p_{t-2}^m) + b_0^m\omega_k + \bar{v}_t^j. \tag{8}$$

Due to the coupled fingertip movements, the models behave well resulting in reasonable prediction of the finger displacements which supports the state estimation in the ensuing tracking procedure.

3.2 Particle Filter Tracking

For the proposed fingertip tracking framework, a state hypothesis s of a particle (s,w) consists of the

N fingertip positions of the hand with each position being denoted by the coordinates (x,y) within the image. Particle filtering is an iterative algorithm where, first, at time t M samples are drawn from a set of previous particles $X_{t-1} = \{(s^i_{t-1}, w^i_{t-1})\}$ proportionally to their likelihood w^i_{t-1}. Subsequently, from each drawn sample a new state hypothesis s^i_t is generated. Adding a Gaussian random variable ω and the displacement vector $\hat{v}_j$ from Eq. 8, s^i_t can be written as:

$$s^i_t = s^i_{t-1} + \hat{v}_t + \omega. \tag{9}$$

To determine a particle set X_t, for each s^i_t the likelihood w^i_t is computed. In order to compute the weights for the new set of particles, one has to approximate the likelihood function $p(z_t|s_t)$ with z_t representing the current observation. Our approximation of $p(z_t|s_t)$ is based on two cues: a contour and a distance cue. The contour cue is derived by exploiting the external energy functional E_{img} of a contour C^i_t obtained by connecting the single points in s^i_t according the finger order. The E_{img} is determined in terms of an edge image Z^E_t which is constructed by drawing lines between the set of maximum bins Z^V_t that can be found in I_H. As a result, the likelihood function can be written as:

$$p_c(z_t|s_t) \propto w_c(s_t) = exp\left\{\frac{-E_{img}(Z^E_t, C_t)}{\sigma^2_c}\right\}. \tag{10}$$

The distance cue is calculated from the Euclidean distance between s^i_t and Z^V_t which consists of the sum of minimal distances between $s^i_t(j)$ and Z^V_t. Based on this cue, the likelihood function can be defined as

$$p_d(z_t|s_t) \propto w_d(s_t) = exp\left\{\frac{-\sum_{j=1}^N \min(\|s_t(j) - Z^V_t\|)}{N\sigma^2_d}\right\}. \tag{11}$$

The final likelihood function is constructed from Eq. 10 and Eq. 11, hence, we define the computation of weights as follows:

$$w^i_t = \frac{\sqrt{w_c(s^i_t)w_d(s^i_t)}}{\sum_{k=1}^M \sqrt{w_c(s^k_t)w_d(s^k_t)}}. \tag{12}$$

One obtains a current state estimate of the fingertip configuration by evaluating the following sum:

$$s_t = \sum_{i=1}^M w^i_t s^i_t. \tag{13}$$

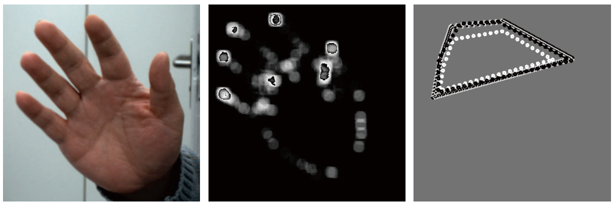

Figure 2. Left: Original input image. Center: Visualization of the Hough space. Right: Generated contour for the particle filter tracking. The particle with the highest likelihood (black dotted line) and the particle with the lowest (white dotted line) are depicted.

Further details concerning the particle filter algorithm can be found in [11].

3.3 Mean-Shift

To obtain more accurate position estimates, a mean-shift algorithm is applied to move the estimated fingertip position $p_j = s_t(j)$ towards the peak of local density distribution. We adopted the EM-like mean-shift algorithm proposed in [12] which in addition provides the possibility to estimate the covariance of the local density distribution. The covariance estimation allows us to adapt the radius r corresponding to the current circular image features. Hence, taking into account movement in the depth of the camera, for tracking circular features in Hough space one has to incorporate an adaptation of radius r_t. Under the assumption that the distribution can be modeled as a Gaussian, we want to find parameters $\bar{p}_j$ and $\bar{V}_j$ representing center and covariance matrix of the distribution that maximize following function for M independent samples:

$$f(\bar{p}_j, \bar{V}_j) = \sum_{i=1}^M G(p_i; \bar{p}_j; \bar{V}_j) I_H(p_i), \tag{14}$$

which can be solved iteratively by, first, calculating λ_i according to:

$$\lambda_i = \frac{G(p_i; \bar{p}_j; \bar{V}_j)) I_H(p_i)}{\sum_{i=1}^M G(p_i; \bar{p}_j; \bar{V}_j) I_H(p_i)}. \tag{15}$$

A new estimation for the center can be derived from followin equation:

$$\hat{p}_j = \sum_{i=1}^M \lambda_i p_i \tag{16}$$

whereas a covariance matrix estimation is obtained by evaluating following term:

$$\hat{V}_j = c \sum_{i=1}^M \lambda_i (p_i - \bar{p}_j)(p_i - \bar{p}_j)^T \tag{17}$$

with c being a constant. If convergence is achieved, the radius is determined from the covariance matrix.

4 Results

The proposed fingertip tracking framework was applied on several image sequences which were captured with a static stereo camera setup and a resolution of $R = 640 \times 480$ pixels. For edge extraction, the method presented in Section 2.1 is applied with $\sigma_k = 4, 2, 1, 0.5$ and $\alpha = 0.5$. Currently, initialization of the tracking is done manually by defining a region I^n_H where finger n is to be found. Using the Hough transform with different radii constructed with $m = 3$, The maximum bin in I^i_H is labeled as finger n according to the finger order $n = \{Thumb = 0, Index = 1, Middle = 2, Ring = 3, Pinkie = 4\}$.

Taking into account the predicted displacements of the fingers, the particle filter tracking algorithm with minimum 600 particles shows good performance. Around 3 mean-shift iterations needed to achieve convergence, The number of iterations for the subsequent mean-shift algorithm depends on the numbers of particles meaning less particles result in more mean-shift

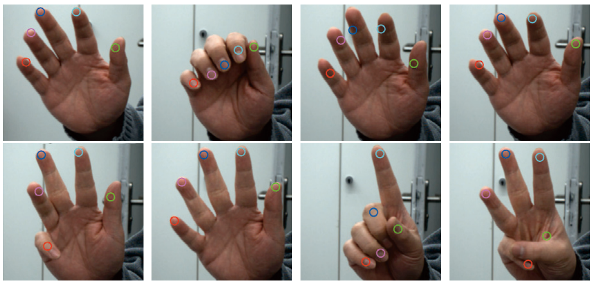

Figure 3. Images of the tracking results. Upper row: Simultaneous closing of the fingers. Lower row: Sequential flexing of the fingers. Labeling of fingertips: Thumb (green), index (light blue), middle (dark blue), ring (pink), and pinkie (red).

iterations and vice versa. Exploiting the combination of both, reasonable accuracy of ≈ 7 pixels mean deviation is achieved for translation movements. For a rotation, opening and closing movement, the error increases up to 20 pixels. These measurements are depicted in Fig. 4. The tracking procedure fails if a finger is lost, which is the case if the movements are too fast.

In case of failure, currently, a very rudimentary reinitialization is performed consisting of a search for maximum bins in the vicinity of the last known estimation and arranging of the fingers according accord position polar space, assuming that fingers are arranged clockwise, respectively counter clockwise, depending on the hand that is observed. Since this assumption is not valid for several finger poses, hence, this might lead to mislabeling.

Since this algorithm operates on monocular images, for each view a tracking instance is created whereas the 3D finger positions are calculated by exploiting epipolar geometry. The presented framework is capable of online tracking of fingertip motion with a frame rate of 15 Hz on a 2.40 GHz dual core CPU. Sample images during the tracking process are depicted in Fig. 3.

5 Conclusion

In this work, we presented a fingertip tracking which allows observation of fine granular human actions such as grasping in an efficient manner. Using Hough transform and a combination of particle filter and meanshift tracking, circular features representing the fingertips could be localized and tracked. Currently, the proposed framework is applied for capturing human grasping movements for online imitation learning using the on-board stereo camera pair of a robot.

However, in the experiments we conducted, we were able to observe that the error on the fingertip localization increases, when the hand performs movements which go beyond translation. These can be led back to the use of a single dynamical motion model for the prediction. In the near future, the fingertip prediction module will be implemented in the form of multiple intertwined motion models to provide better predictions. Concerning the motion model of the hand, we realized that it needs to be extended by an angular dimension to cover the hand rotation. To enable full online observation of the human upper body the fingertip tracking will be integrated into an upper body tracker and its implementation will be improved to raise its efficiency.

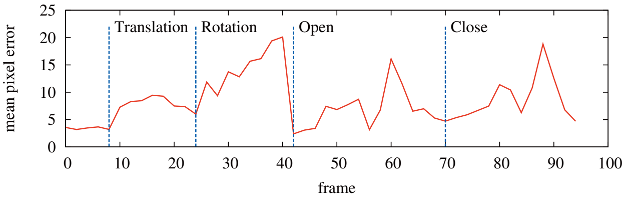

Figure 4. Error plot for a sequence of four hand and finger movements: Translation and rotation of the hand, close and open movement of fingers.

6 Acknowledgments

The work described in this paper was conducted within the EU Cognitive Systems project GRASP (FP7- 215821) funded by the European Commission.

References

[1] J. Rehg and T. Kanade, "Digiteyes: Vision-based hand tracking for human-computer interaction," in *Proc. Workshop Motion of Non-Rigid and Articulated Bodies*, November 1994, pp. 16–22.

[2] B. Stenger, P. R. S. Mendonca, and R. Cipolla, "Model-based 3d tracking of an articulated hand," in *Proc. Int. Conf. Computer Vision and Pattern Recognition*, December 2001, pp. 310–315.

[3] I. Oikonomidis, N. Kyriazis, and A. A. Argyros, "Markerless and efficient 26-dof hand pose recovery," in *Proc. 10th Asian Conf. Computer Vision*, November 2010.

[4] A. Blake and M. Isard, *Active Contours: The Application of Techniques from Graphics, Vision, Control Theory and Statistics to Visual Tracking of Shapes in Motion*, 2000.

[5] A. A. Argyros and M. Lourakis, "Vision-based interpretation of hand gestures for remote control of a computer mouse," in *Proc. HCI06 Workshop*, May 2006, pp. 40–51.

[6] K. J. Hsiao, T. W. Chen, and S. Y. Chien, "Fast fingertip positioning by combining particle filtering with particle random diffusion," in *Proc. Int. Conf. Multimedia and Expo*, June 2008, pp. 977–980.

[7] L. Bretzner, I. Laptev, and T. Lindeberg, "Hand gesture recognition using multi-scale colour features, hierarchical models and particle filtering," in *Proc. Int. Conf. Aut. Face and Gesture Recognition*, May 2002, pp. 423–428.

[8] A. M. Burns and M. M. Wanderley, "Visual methods for the retrieval of guitarist fingering," in *Proc. Int Conf. New Interfaces for Musical Expression*, Paris, France, June 2006, pp. 196–199.

[9] C. Kerdvibulvech and H. Saito, "Markerless guitarist fingertip detection using a bayesian classifier and a template matching for supporting guitarists," in *Proc. 10th Virtual Reality Int. Conf.*, April 2008.

[10] J. Elder and S. Zucker, "Scale space localization, blur, and contour-based image coding," in *Proc. Int. Conf. Computer Vision and Pattern Recognition*, June 1996.

[11] P. Azad, *Visual Perception for Manipulation and Imitation in Humanoid Robots*, 2009.

[12] Z. Zivkovic and B. Kroese, "An em-like algorithm for color-histogram-based object tracking," in *Proc. Int Conf. Computer Vision and Pattern Recognition*, Washington D.C., USA, June 2004, pp. 798–803.