Tracking of Fingers in Dynamic Image of Omnidirection Camera by CONDENSATION Algorithm

Masahide Sakata, Norikazu Ikoma and Motonori Doi*
Faculty of Engineering, Kyushu Institute of Technology
1-1 Sensui-cho, Tobata-ku, Kita-Kyusyu, Fukuoka 804-8550, Japan
* Department of Telecommunications and Computer Networks,
Osaka Electro-Communication University
18-8 Hatsu-cho, Neyagawa, Osaka 572-8530, Japan
E-mail: sakata@sys2.comp.kyutech.ac.jp

Abstract

Omnidirection camera is a special camera having 360 degrees view. Gripping the camera with fingers, and using the camera image with finger motion for human machine interface is proposed in [1]. In this paper, we track the motion of fingers in the image of the same setting in [1] by using a state space model, which represents state and image observation of the fingers. State estimation is conducted by CONDENSATION algorithm[3], which is one realization of particle filters[2] maintaining posterior distribution of the state by many samples (particles) according to the distribution and is approximated by degree of dense of the samples. Experiment of finger tracking has been conducted, and usability of the algorithm is demonstrated.

1. Introduction

Omnidirection camera is a special camera having 360 degrees view and it is usually used for the purpose such as security, surveillance and observations, and robot eyes. There is a research using this camera with unusual way, gripping the camera with fingers(Figure 1), and use the camera image of finger motion for human machine interface[1]. The advantage of this interface is to be a more human-friendly one since fingers' motion, which is natural operation of human, is used as input to the interface and it can serve functions in the target system that are invoked by operation of fingers intuitively matching to each function. To develop the interface using fingers' image of omnidirection camera, it is necessary to know state of the fingers (position, angle, their velocities, and the number of fingers) from the dynamic image. In this research, we track the motion of fingers in the image by estimating the state of a state space model, which represents the state and the image observation of fingers. The state estimation of the model is conducted by CONDENSATION algorithm[3]. The CONDENSATION algorithm is one realization of particle filters[2], which maintain posterior distribution of the state as many samples (particles) according to the distribution, and the distribution is approximated by degree of dense of the samples. CONDENSATION algorithm keeps typical shape of tracking object as a template, and compares affine transformation of the template with actual image. Where, each sample has parameters of the affine transformation. The samples are weighted according to their fitness (likelihood) to the image obtained as a result of the comparison, and are resampled with probability proportional to the weights. Here, the samples are supposed to evolve according to given dynamics. The state estimation is achieved by proceeding these procedures at each time when observation (image frame) is obtained.

Structure hereafter is as follows; Chapter 2 defines a model for the motion tracking of fingers in omnidirection camera. Chapter 3 describes the state estimation by the CONDENSATION algorithm. Chapter 4 shows result of experiment to track the fingers in dynamic image of omnidirection camera by the proposed method. Conclusion and future works are described in chapter 5.

Figure 1. Gripping omnidirection camera.
2. Model

The finger image interface[1] uses omnidirection camera turning sideway (to the right, if right-handed) and grasping it by fingers as shown in figure 1, and it controls the target system synchronized with the motion of fingers reflecting in the camera image. Where, it is possible to move the fingers to direction of depth of gripping, wrist rotation, and tapping the cylinder of the omnidirection camera by finger. Figure 2 shows typical dynamic image obtained in this situation.

State of fingers and process to obtain image of fingers are modeled as follows. In this paper, the number of fingers, \( F \), is supposed to be known and fixed. First, one finger that reflects in omnidirection camera is illustrated in figure 3. \( O \) is the center of image of omnidirection camera, \( h \) is length of the finger that reflects in omnidirection camera, \( w \) is width of the finger, \( \theta \) is central angle of finger position, and \( \phi \) is inclination of the finger. Together with these parameters, we define the state of one finger by

\[
x = (h, w, \theta, \phi).
\]

To represent multi-fingers, subscript \( f \in \{1, 2, \ldots, F\} \) is given for each element of expression (1). They also have (discrete) time \( t \) in the subscript since the state of fingers may vary with time. That is, the state of \( f \)-th finger at time \( t \) is written as \( x_{f,t} \), and the elements of the vector are written in the same manner. Let multi-finger state by collecting each finger state be

\[
X = (x_1, \ldots, x_F).
\]

Next, a model concerning the motion of fingers is defined. It is modeled for each finger as follows. Although the motion of fingers is unknown, it is thought as smooth in motion. Thus the motion can be defined by random walk model

\[
x_{f,t} = x_{f,t-1} + v_{f,t}, \quad v_{f,t} \sim N(0, \tau),
\]

where variance matrix is defined as \( \tau = \text{diag}(\tau_1, \tau_2, \ldots, \tau_F) \).

From (3), system model of one finger can be represented in a conditional probability density from

\[
x_{f,t} \sim f(x_{f,t} | x_{f,t-1}),
\]

and we represent multi-finger case by

\[
X \sim f(X | X_{-1}), f(X_f | X_{-f}), \ldots, f(X_F | X_{-F}).
\]

Let us model how fingers with state above reflect in image. In this paper, we employ a simplified version of template, although the template originally proposed in the CONDENSATION algorithm is more complicated. Template is given beforehand for tracking. Template of finger is provided by a set of control points \( \{s_m\}_{m=1}^M \) with \( M=13 \) as shown in figure 4. The image observation on the template is obtained based on intensity of the image along with line segment assigned to each control point. We assign line segments as shown in figure 4: oblique for \( m=6 \) and 8, parallel to the y axis for \( m=7 \) and parallel to the x axis for other \( m \)'s.

At time \( t \), intensity \( I_{x,z}(z; x_{f,t}) \) along with line segment of the \( m \)-th control point is obtained from the image, where \( z \) shows position in the line segment, and \( x_{f,t} \) is the state of target finger. From the intensity, we extract contour position at the line segment by detecting the position where its 1st derivative is greater than a certain threshold. We choose the nearest contour position to the
origin among the extracted positions when more than one positions are obtained. Where the origin of the line segment is taken at the position of the control point. Position of contour in line segment is denoted by \( z_m(x_t) \). This is the observed value from the image. We suppress the notation \( x_t \) when it is not necessary and simply write the position as \( z_m \). Let observation of all control points on the template by collecting each observation value \( z_m \), be

\[
z_t = (z_{m_1}, \ldots, z_{m_T})
\]

(6)

Uncertainty of observation (observation error such as image processing) is represented by Gaussian distribution. Then fitness (likelihood) of the \( m \)-th control point to the image is

\[
h_m(z_m|x_t) \propto \sigma^{-1} \exp\left(-1/2\sigma^2 \min([z_m]_j, (t/2)^2)\right).
\]

(7)

Fitness (likelihood) of the template to the image is supposed to be product of fitnesses of control points then it is represented as

\[
h(z_t|x_t) = \prod_{m=1}^{M} h_m(z_m|x_t)
\]

(8)

Discussion so far was the situation of one finger. Subscript \( f \) is written for multi-finger case. That is, the observation of \( f \)-th finger at time \( t \) is denoted by \( z_{tf} \). Let multi-finger observation by collecting each finger observation be

\[
z_t = (z_{tf_1}, \ldots, z_{tf_T})
\]

(9)

Additionally, fitness (likelihood) of template for multi-finger case is supposed to be a product of the fitness of each finger such that

\[
H(Z_t|X_t) = \prod_{f=1}^{F} h(z_{tf}|x_{tf})
\]

(10)

\[3. \text{ CONDENSATION Algorithm}\]

The CONDENSATION algorithm[3] approximates posterior distribution of the state of tracking object by many pairs of sample and weight \( \{X^{(n)}_t, \pi^{(n)}_t\}_{n=1}^{N} \). Where \( X^{(n)}_t \) is \( n \)-th realization of multi-finger state, and \( \pi^{(n)}_t \) is normalized weight such that \( \pi^{(n)}_t \geq 0 \) and \( \sum_{n=1}^{N} \pi^{(n)}_t = 1 \) hold. The state estimation is achieved by applying three steps, namely, prediction, observation, and selection steps, to these samples.

In the prediction step, one-step-ahead prediction sample is generated by using sample set of the uniform weight \( \{X_{t-1\rightarrow 1}^{(n)}\}_{n=1}^{N} \), which was calculated at previous time. New sample is generated by using the system model as

\[
X^{(n)}_t = F(X^{(n)}_{t-1}),
\]

(11)

The set of generated samples \( \{X_{t-1\rightarrow 1}^{(n)}\}_{n=1}^{N} \) approximates one-step-ahead prediction distribution \( p(X_t|Z_{t-1}) \). Where \( Z_{t-1} = (Z_{t-1,1}, \ldots, Z_{t-1,T}) \), it is a notation to represent the history of observations from time \( 1 \) to \( t-1 \).

Next, in the observation step, calculate weights of the samples generated in the prediction step. Then weight is obtained from the fitness (likelihood) of the template, which is generated from the sample, to the image as

\[
\pi^{(n)}_t = H(Z_t|\tilde{X}^{(n)}_t)
\]

(12)
The set of weight and sample pairs, $\{x_{t_1}^{(a)}, \pi_t^{(a)}\}_{a=1}^N$, obtained in the prediction and the observation steps approximates the filter distribution $p(x_t | z_{1:t})$.

In the selection step, resample $X$ from $\{x_{t_1}^{(a)}, \pi_t^{(a)}\}_{a=1}^N$ according to the probability proportional to $\pi_t^{(a)}$, then obtain sample set $\{x_{t_1}^{(a)}\}_{a=1}^N$. $\{x_{t_1}^{(a)}\}_{a=1}^N$ is also an approximation of the filter distribution $p(x_t | z_{1:t})$. After}

the sample set $\{x_{t_1}^{(a)}\}_{a=1}^N$ of the filter distribution is obtained, sample average

$$\overline{x}_t = \frac{1}{N} \sum_{a=1}^{N} x_{t}^{(a)}$$

is employed as the estimated value of the present state.

Figure 5 shows these three steps of CONDENSATION algorithm.

![Figure 5. CONDENSATION algorithm.](image)

4. Experiment

By letting the number of fingers be $F=2$ and $3$, tracking experiments of the fingers have been conducted using the dynamic scene of kind of figure 2. As we can see in the image that various background objects other than the fingers reflect in the image, so it is thought that the tracking of the fingers is difficult by simple method.

First as $F=2$, we use thumb and forefinger as tracking objects in dynamic scene. The dynamic image is shown in figure 6. In the image, depth of gripping is almost the same degree throughout the image sequence, two fingers rotate anti-clockwise first, then rotate clockwise. After that, the fingers start to rotate against clockwise again, with alternatively moving one finger and other. As the conditions of the state estimation, we let the number of samples be $N=1000$, and we have given initial distribution as follows; For $\theta$, manually extracted value of $\theta$ of each finger by referring the first frame is used as the mean of the initial distribution of Gaussian with variance 5.0. For $h$
and \( w \), uniform distributions with given maximum and minimum values are used, and \( \phi \) is assumed to be Gaussian distribution with mean 0.0 and variance 3.0. \( r_i = 1.0 \), \( r_o = 1.0 \), \( r_i = 5.0 \), \( r_o = 2.0 \) are given as system noise variances. We have used template of figure 4 with lengths of line segments be the same as \( t = 30 \) for all control points. Estimation result is shown in Figure 6 with rectangle symbols showing the control points of the template obtained by the estimation, and lines connecting the symbols showing estimated contour of the fingers.

Next as \( F=3 \), we use thumb, forefinger and middle finger as tracking objects in dynamic scene of figure 7. In the image, three fingers rotate anti-clockwise while they are changing the depth of gripping of omnidirection camera, then rotate again clockwise. Then, we let the number of sample be \( N = 1500 \). \( r_i = 0.1 \), \( r_o = 0.1 \), \( r_i = 5.0 \), \( r_o = 3.0 \) are given as system noise variances. Other parameters are the same as in \( F=2 \) case. Estimation result is shown in Figure 7 by the same manner in figure 6.

We can recognize from figure 6 and 7 that the estimated contours are almost coincident to the actual of the fingers in the dynamic image. Even the algorithm lost the tracking as shown 51-st frame in figure 6 and 1st frame in figure 7, after some frames it can recover to correct position by robustness of the algorithm.

5. Conclusion

Supposing an interface that uses dynamic image of multiple fingers in omnidirection camera, we have proposed a tracking method of the fingers by the CONDENSATION algorithm to know the state of the fingers (position, angle, etc.). After modeling the state of fingers and the image observation by representing the dynamics and the observation process into a state space model, we have tracked the fingers by estimating the state with the CONDENSATION algorithm. As the result of the tracking experiment of multi-finger, we have successfully tracked contour of the fingers. From this result, we have confirmed that performance of the tracking multi-finger by the CONDENSATION algorithm is sufficient even for the dynamic image with clutter backgrounds.

For future works, use of color information in image could make the tracking more robust instead of using the intensity only. It is also better to use template having different color (or intensity) for inside and outside of the template. These are ideas to improve the algorithm within fixed number of fingers situation.

Another problem is online and real-time processing while it has been conducted with offline dynamic image and by batch process in the experiment of this paper.

Treating the case where the number of fingers varies with time is one another problem. By achieving this, it becomes possible to deal with operation of tapping the cylinder of omnidirection camera by fingers.

Final destination of this research is to develop an interface directly reflecting the result of fingers’ state estimation by CONDENSATION algorithm. Thus it is necessary to design and develop an interface system to invoke some actions from the tracking result are also interesting theme.

Acknowledgement

This study was carried out under the ISM Cooperative Research Program (2004-ISM-CPR-2020).

References

Figure 6. Estimation result for $F=2$ tracking: Control points and templates.

Figure 7. Estimation result for $F=3$ tracking: Control points and templates.