We propose a method for extracting human limb regions by the combination of optical flow-based motion segmentation and nonlinear optimization-based image registration. First, rotating limb regions with rough boundaries are extracted and motion parameters are estimated for an approximated model. Then the extracted region and estimated parameters are used as initial values for nonlinear optimization that minimizes residuals of two successive frames and estimates motion parameters. Combining the two steps reduces computational cost and avoids the initial state problem of optimization. According to estimated parameters, the limb region is extracted by a Bayesian classifier to obtain accurate region boundaries. Experimental results on real images are shown.

Keywords: extraction, motion estimation, optical flow, EM algorithm, gesture recognition

1. Introduction

It is important to extract human regions from a movie as a part of a human activity recognition system, such as gesture recognition for human interfaces and motion reconstruction in virtual reality. For such applications, detecting and extracting human arms in a scene plays a key role [1] because it is a crucial cue to know where a subject is and what he/she acts, especially when recognizing gestures in which arm movements mainly determine meaning.

In the last decade, many human activity recognition studies often used parameterized volumetric human body models to reconstruct actual human posture [2, 3]. However, a common problem is that these methods require that a background is known or at least is uniform in color to make subtraction easy; otherwise there must be no moving object except the subject. The assumptions about background are hurdles to developing methods so that a recognition system adapts to various real environments.

To overcome these problems, we have proposed a method [4] for finding and recognizing human arms in a general scene. The method extracts regions of rotating human limbs represented by a stick model and estimates their motion parameters. The extraction method proposed is an indirect method; that is, with optical flow of two successive images calculated in advance, it segments an image into several moving regions and estimates motion parameters of each region. It can extract arm regions from optical flow of a real image sequence contaminated by much noise. It is essentially impossible, however, to compute optical flow where the motion correspondence cannot be found, especially at the edge of motion, and the indirect method would fail to extract the exact arm region boundary.

In contrast, a method of motion segmentation by comparing intensities of two successive frames directly [5, 6] has been proposed. This direct method can deal with the motion discontinuity at motion edges and estimate more precise motion parameters because it does not use optical flow that causes failure with the indirect method. A problem of the direct method is its high computational cost because it uses nonlinear optimization to minimize intensity residuals of two frames all over the image with certain initial parameters that may sometimes deviate greatly from true values.

In this paper, we propose a method to extract regions of rotating human limbs with an accurate boundary by combined use of indirect and direct methods. First, limb regions are extracted by the indirect method using optical flow and motion parameters are estimated. Then, accurate boundaries of regions are obtained by the direct method based on the extracted region and estimates produced by the indirect method; the results of the indirect method are used as initial values for the direct method. This combination is expected to decrease computational cost and improve extraction and estimation results of the indirect method. We describe the indirect method of extraction and estimation based on optical flow in section 2, and the direct method using nonlinear optimization in section 3. Then an extraction with MAP estimation is discussed in section 4. Finally, we provide experimental results of real images in section 5, and discuss the performance of the proposed method in section 6.
2. Indirect method using optical flow

2.1. Motion Model

In this section, we describe the indirect method of extracting limb regions using optical flow. To achieve rough extraction and estimation, we use a rather simplified motion model because optical flows in real images usually involve so much noise that parameters of a precise motion model can not be recovered. Therefore, by using the simple model, we divide an image into regions with coherent movement and determine which region moves or rotates.

Here a limb is assumed to rotate on a plane that is not parallel to the image plane. As shown in Fig. 1, there are two cases. In general, a rotation in three dimensions is represented as a revolution about an axis, however, both cases can be used as an approximated 3D motion model. Then the limb is projected onto the image plane by orthographic projection. In both cases, a motion of point \( p_j = (x_j, y_j) \) on a rotating limb and its velocity \( \dot{p}_j = (u_j, v_j) \) in the image plane are modeled as follows [4],

\[
\dot{p}_j = A_j q,
\]

where

\[
A_j = \begin{pmatrix} y_j & 1 & 0 & 0 \\ 0 & 0 & x_j & 1 \end{pmatrix}, \quad q = (\alpha, \beta, \gamma, \delta)^T.
\]

Here, motion parameters (angular velocity \( \omega \) and rotation center \((c_x, c_y)\)) of the rotating limb are calculated from \( q \) as follows:

\[
c_x = -\delta / \gamma, \quad c_y = -\beta / \alpha, \quad \omega = -\text{sign}(\alpha) \sqrt{-\alpha \gamma}.
\]

Similarly, \( \phi \) and \( \psi \) are retrieved from \( q \). But it is important that we discriminate rotating regions by \( \omega \); a rotating arm region has a large angular velocity, while in a moving region but without rotation \( \omega \) becomes quite small.

2.2. Motion Clustering

Although motion of a point on a limb is modeled as explained above, optical flow computed from real images involves inevitable noise. Therefore, we find groups of flows corresponding to each moving region by clustering optical flow.

We assume that the distribution of \( \dot{p}_j \) is subject to a two-dimensional Gaussian,

\[
p(\dot{p}_j | p_j, q, \Sigma) = \frac{1}{2\pi|\Sigma|^2} \exp \left\{ -\frac{1}{2} \left( \dot{p}_j - A_j q \right)^T \Sigma^{-1} \left( \dot{p}_j - A_j q \right) \right\},
\]

where \( \Sigma = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix} \) is a covariance matrix that assumes that the errors for \( u \) and \( v \) are mutually independent because \( \alpha, \beta \) and \( \gamma, \delta \) are estimated separately.

Then, posterior probability of the optical flow is modeled as a mixture density comprised of \( M \) densities of clusters \( R_i \),

\[
p(\dot{p}_j | p_j) = \sum_{i=1}^M \xi_i p(\dot{p}_j | p_j, q_i, \Sigma_i),
\]

where \( \xi_i \) are weights for each density.

The problem becomes an estimation of \( q_i \) for each cluster \( R_i \) and a segmentation of points \( p \) based on the weights of densities [4]. Assuming that each moving object has its own motion parameter \( q \) and that Eq. (4) is the distribution of optical flow within the object region, the following algorithm uses the EM algorithm [7] to perform estimation and segmentation:

1. Compute optical flow \( \dot{p}_j = (u_j, v_j)^T \) at each point \( p_j = (x_j, y_j)^T \) (j = 1, ..., N).
   Perform initial segmentation of the optical flow based on the direction of velocity to obtain initial clusters \( R_i \) (i = 1, ..., M) (see section 5).
   Then, set initial values of weights \( w_{ij} \) as probabilities that point \( p_j \) belongs to cluster \( R_i \),

\[
w_{ij} = \begin{cases} 1 & (p_j \in R_i) \\ 0 & (p_j \notin R_i) \end{cases}.
\]

2. Normalize weights \( w_{ij} \);

\[
w_{ij}' = \frac{w_{ij}}{\xi_i},
\]

where \( \xi_i = \frac{1}{N} \sum_j w_{ij} \).

3. Find parameters \( q_i = (\alpha_i, \beta_i, \gamma_i, \delta_i) \) of each cluster \( R_i \) by solving the following overdetermined system of equations:

\[
\begin{pmatrix} \sqrt{w_{i1}'} \dot{p}_1 \\ \sqrt{w_{i2}'} \dot{p}_2 \\ \vdots \end{pmatrix} = \begin{pmatrix} \sqrt{w_{i1}'} A_1 \\ \sqrt{w_{i2}'} A_2 \\ \vdots \end{pmatrix} q_i,
\]

with QR decomposition [9].

The solution \( q_i \) is the weighted mean of the motion parameters for cluster \( R_i \).

4. Compute weighted variances \( \sigma_x^2 \) and \( \sigma_y^2 \) for each
and results of a scene with two moving arms are shown in only one limb in a scene. However, we can determine the

3. Direct method using nonlinear optimization

Tamaki, T.

For simplicity, we assume in the last step that there is only one limb in a scene. However, we can determine the number of movements in a scene [4] and the background with no motion. The method described in the following sections can be applied to each moving region separately, and results of a scene with two moving arms are shown in section 5.

For simplicity, we assume in the last step that there is only one limb in a scene. However, we can determine the number of movements in a scene [4] and the background with no motion. The method described in the following sections can be applied to each moving region separately, and results of a scene with two moving arms are shown in section 5.

3. Direct method using nonlinear optimization

This section describes the direct method that is applicable to the motion edge shown in Fig. 2 where the indirect method fails. Although optical flows are incorrect at the motion edge, we have extracted a rough region of rotating limb $R_\Omega$ and estimated motion parameters $q_\Omega$ of the limb with a motion model approximated by the indirect method. We use these estimates to help the direct method improve computational cost and the initialization problem.

Instead of the simplified motion model in the previous section with the four parameters, the direct method models motion using eight parameters [5, 6] to model a motion in an image. The model has been widely used to handle a motion of a plane in three-dimensional space. A human arm is not a plane, but it can be approximated enough as a planar object because usually the range of depth of an arm is rather smaller than the distance between the camera and the arm.

Now we introduce the motion model with eight parameters. Let $I_t$ and $I_{t+1}$ be images at times $t$ and $t+1$. Point $p_j$ in $I_t$ corresponds to $p_j + u(p_j; \theta)$ in $I_{t+1}$, where $u$ is a motion vector and $\theta$ represents motion parameters $\theta = (\theta_1, \ldots, \theta_8)^T$. Note that $\theta$ and $u$ are unknown and estimated below, so the correspondence of points between successive frames is also unknown.

The estimation of the motion parameters $\theta$ is done by minimizing the square of residuals of intensities $r_j$ of the corresponding two points. The cost function is a summation of residuals over a region $R$ in $I_t$,

$$\min_{\theta} \sum_{p_j \in R} r_j^2,$$

where

$$r_j \equiv r(p_j) = I_t(p_j) - I_{t+1}(p_j + u(p_j; \theta)),$$

$$u(p; \theta) = M\theta \equiv \begin{pmatrix} x y 0 0 1 0 x^2 xy & 
0 x y 0 1 0 y^2 \end{pmatrix} \theta.$$

To seek $\theta$ minimizing the cost function in Eq. (14), we use the Gauss-Newton method [8], which is an iterative gradient-based nonlinear optimization. Starting with some initial value, estimates are modified as $\theta \leftarrow \theta + \delta \theta$ at each step of the iteration, and modification term $\delta \theta = (\delta \theta_1, \ldots, \delta \theta_8)^T$ is obtained by solving the following system of equations [6]:

$$\sum_{p_j \in R} \begin{pmatrix}
\frac{\partial r_j}{\partial \theta_1} & \ldots & \frac{\partial r_j}{\partial \theta_8} \\
\frac{\partial r_j}{\partial \theta_1} & \ldots & \frac{\partial r_j}{\partial \theta_8} \\
\vdots & \ddots & \vdots \\
\frac{\partial r_j}{\partial \theta_1} & \ldots & \frac{\partial r_j}{\partial \theta_8}
\end{pmatrix}
\begin{pmatrix}
\delta \theta_1 \\
\delta \theta_2 \\
\vdots \\
\delta \theta_8
\end{pmatrix}
= \sum_{p_j \in R} \begin{pmatrix}
\frac{\partial r_j}{\partial \theta_1} \\
\frac{\partial r_j}{\partial \theta_2} \\
\vdots \\
\frac{\partial r_j}{\partial \theta_8}
\end{pmatrix},$$

where

$$\left(\begin{array}{c}
\frac{\partial r}{\partial \theta_1} \\
\vdots \\
\frac{\partial r}{\partial \theta_8}
\end{array}\right)^T = \frac{\partial r}{\partial \theta} = \frac{\partial u}{\partial \theta} \frac{\partial r}{\partial u} \frac{\partial u}{\partial \theta}$$

$= -M^T \nabla I_{t+1}(p + u(p; \theta))(18)$

The estimation procedure is repeated to solve the system of equations and update estimates until the estimation converges, and final estimates for $\theta$ (we write as $\hat{\theta}$) are obtained.

Optimization requires appropriate initial values for $\theta$ to start updating estimates, and region $R$ to make the summation for the cost function. Usually an entire image is
used for \( R \) and all parameters are set to zero at the begin-
ing. However, this initialization makes the number of iterations increase and estimates tends to fall into local minima.

Instead, we use extracted arm region \( R_\Omega \) as region \( R \) to make the optimization focus on the region of interest and reduce the computational cost. To use \( \mathbf{q}_\Omega \) obtained by the indirect method for initializing \( \theta \), we substitute \( \alpha = \theta_2 \), \( \beta = \theta_3 \), \( \delta = \theta_4 \), and \( \gamma = \theta_5 \) according to Eqs. (2) and (16). The rest of \( \theta \) is initialized to 0.

4. Extraction by Bayesian classifier

Since the estimation in the previous section is performed only in region \( R_\Omega \), the extracted limb region with a rough boundary, we need to determine limb region \( \hat{R} \) with a more precise boundary according to the final estimates \( \hat{\theta} \).

To extract the limb region, we use a Bayesian classifier that maximizes a posterior probability. Assuming that the density of residual \( r_j \) for each pixel is Gaussian, conditional probabilities are defined as follows:

\[
p(r_j|\theta_i) = \frac{1}{\sqrt{2\pi}\sigma^2_i} \exp\left( -\frac{r_j^2}{2\sigma^2_i} \right),
\]

where \( i = 1 \) means human limb region \( \hat{R} \) and \( \theta_1 = \hat{\theta} \), and \( i = 0 \) is the rest of the image and \( \theta_0 = \theta \) (no movement). \( \sigma^2_i \) is calculated within \( R_\Omega \). Then we define prior probabilities of \( \hat{R} \) and \( \theta \) by the ratio of areas,

\[
P(\hat{\theta}) = \frac{|R_\Omega|}{N},
\]

\[
P(\theta) = 1 - P(\hat{\theta}),
\]

where \( |R_\Omega| \) is the number of pixels in region \( R_\Omega \) and \( N \) be the number of all pixels in the image.

Finally, posterior probabilities are

\[
P(\hat{\theta}|r_j) \propto p(r_j|\hat{\theta})P(\hat{\theta}),
\]

\[
P(\theta|r_j) \propto p(r_j|\theta)P(\theta),
\]

and the classification for each pixel \( p_j \) is performed as

\[
p_j \in \begin{cases} \hat{R}, & P(\hat{\theta}|r_j) > P(\theta|r_j) \\ \text{background}, & \text{otherwise} \end{cases}
\]

5. Experimental results

The proposed method has been implemented on PC-linux (2.4 GHz CPU) using C++. Computation of optical flow (algorithm step 1) was performed by code released by \[10, 11\], and initial clusters (step 1) were made by simple histogram clustering, which divides directions of flow vector into 24 sections and finds peaks in the direction histogram as the center of the clusters.

Figure 3 shows the experimental result for a real image sequence of arm bending toward the shoulder while fixing the elbow position. Figure 3(a) is the first frame and Fig. 3(b) shows the superimposed optical flow between the first and second frames. \( R_\Omega \) is the result of the indirect method, is shown in Fig. 3(c), and the area of the top of the arm where the motion is large is not extracted because of the inaccurate optical flow at the motion edge. The boundary of the extracted region is not identical to that of the actual arm.

Figure 3(d) is \( \hat{R} \), the result of extraction by the direct method, which uses the result (estimated parameters and extracted region) of Fig. 3(e) as the initial value. Compared to Fig. 3(c), we can see that that top of the lower arm region is extracted and the whole boundary is close to the actual contour of the arm in Fig. 3(d). However, the region around the elbow is not extracted because the posterior probability around the area where intensity is flat is small in both classes. Which class the point should belong to is therefore ambiguous. Figure 3(e) and (f) illustrate posterior probabilities, and high probability is in white and low in black. In Fig. 3(e), the arm area is white except around the elbow, and the background with uniform intensity is the same gray level in both (e) and (f).

Another experiment is shown in Fig. 4. The arm moves downward, and Fig. 4(b) shows that the region extracted...
by the indirect method becomes narrow at the top of the arm because of the motion edge problem. In contrast, extraction of the direct method in Fig. 4(c) is better in terms of accuracy of the boundary of the extracted region. Ambiguity at flat intensity (near the shoulder) also occurs in this case.

As mentioned at the end of section 2, the proposed method can extract multiple arms with different motions. Figure 5 shows that the two arms move simultaneously: the left arm moves downward and the right arm moves upward. The two regions of the arms have the largest and second largest angular velocities among several regions. Note that the improvement of extraction for the left arm in Figs. 5(b) and (c), and the boundary around the hand of the right arm becomes clear (Figs. 5(d) and (e)) while the under part of the arm is not extracted well because there are many areas with flat intensities.

Note that small regions in extracted regions were removed and holes buried in Figs. 3(c)(d), Figs. 4(b)(c), and Figs. 5(b)–(e).

6. Discussions

Here we discuss how the proposed method reduces computational cost. Since the direct method uses the nonlinear minimization of the cost function, estimation takes much time to converge unless appropriate initial values of the parameter, initial segmentation, and the number of classes are given.

Figure 6 shows the computation time consumed by just using the direct method for Fig. 4 without the indirect method. Compared to the proposed indirect-direct method, the eight motion parameters for each class are randomly initialized, and weights for each pixel are decided based on the initialized parameters. Fixing the number of clusters, this trial was performed 20 times. Computation time increases in proportion to the number of clusters, and it took at least 145±29 [s] for only two clusters.

In contrast, the proposed method requires less time because of the initialization of the indirect method. For Fig. 4, it took 1.465±0.0241 [s] for the computation of optical flow, and 88.3±0.520 [s] for the indirect and direct method in an average of 20 trials. Although initial flow segmentation produced six clusters, the EM algorithm in the indirect method requires about 35 [s]. In general, it is difficult to theoretically compare the cost of computing optical flow and that of nonlinear optimization by the direct method. However, the proposed method was faster to process than the direct method with random initialization for all sequences used in experiments.

The extraction of the proposed method is slightly better than the result of the direct method without the indirect method as shown in Fig. 7. The problem of flat intensities around the shoulder mentioned before severely affects the extraction of the upper arm. Because the indirect method extracts the shoulder as in Fig. 4(b), the successive direct method mostly avoids the problem except for the armpit (Fig. 4(c)).

7. Conclusions

We have proposed a method to extract human limb regions by a direct method that uses two frames directly with initial values that is the result of the indirect method based on optical flow. At first, rotating limb regions that
Human Limb Extraction based on Motion Estimation using Optical Flow and Image Registration

Fig. 6. Computation time of the direct method without the indirect method.

Fig. 7. Extraction result of the direct method without the indirect method.

have rough boundaries are extracted and motion parameters of the approximated model are estimated. Then the extracted region and estimated parameters are used as initial values of nonlinear optimization that minimizes residuals of two successive frames, and estimates motion parameters. According to estimated parameters, the limb region is extracted by a Bayesian classifier to obtain an accurate boundary of the region. Experimental results using real images show that the result of the direct method is better than that of the indirect method from the viewpoint of dealing with the motion edge. However, there is still difficulty in determining whether an area is in motion or not when the area has flat intensity. We plan to handle this problem by applying shape model of an arm or using results of several frames.

Acknowledgements
The author thanks Noboru Ohnishi and Tsuyoshi Yamamura for invaluable comments.

References:
Name:
Toru Tamaki

Affiliation:
Research Associate, Graduate School of Science and Technology, Niigata University

Address:
8050 Ikarashi2nocho, Niigata City, Niigata Pref. 950-2181, Japan

Brief Biographical History:
2001: Department of Information Engineering, Faculty of Engineering, Niigata University
2003: Research Associate, Graduate School of Science and Technology, Niigata University

Main Works:

Membership in Learned Societies:
- IEEE Computer Society
- The Institute of Electronics, Information and Communication Engineers (IEICE)
- The Institute of Image Information and Television Engineers (ITEJ)