SEGMENTATION OF PEOPLE IN MOTION

Akio Shio
NTT Human Interface Laboratories
Nippon Telegraph and Telephone Corp.
1-2356 Take, Yokosuka, 238-03 Japan

Jack Sklansky
Department of Electrical Engineering
University of California, Irvine
Irvine, CA, 92717 U.S.A.

Abstract
A method for segmenting monocular images of people in motion from a cinematic sequence of frames is described. This method is based on image intensities, motion, and an object model—i.e., a model of the image of a person in motion. Though each part of a person may move in different directions at any instant, the time averaged motion of all parts must converge to a global average value over a few seconds. People in an image may be occluded by other people, and usually it is not easy to detect their boundaries. These boundaries can be detected with motion information if they move in different directions, even if there are almost no apparent differences among object intensities or colors. Each image of a person in a scene usually can be divided into several parts, each with distinct intensities or colors. The parts of a person can be merged into a single group by an iterative merging algorithm based on the object model and the motion information because the parts move coherently. This merging is analogous to the property of perceptual grouping in human visual perception of motion. Experiments based on a sequence of complex real scenes produced results that are supportive of our approach to the segmentation of people in motion.

1 INTRODUCTION
Object segmentation from a sequence of images is very useful for many practical applications such as surveillance and measurement of automotive and pedestrian traffic, and driver assistance. Automating the counting and measurement of the directions of human and vehicular traffic would be extremely useful for traffic control and for the planning of road extensions and construction. The main concern of this paper is the segmentation and tracking of moving objects from a sequence of images which may include many objects. The objects can be nonrigid, such as a human body, which consists of several rigid parts connected to each other.

The conventional approach for object segmentation from scenes is based on recursive region splitting[6][15]. Ohlander et al. used histograms of feature values computed from an image to guide region splitting[15]. Another approach is based on clustering in a color space[12]. However, these methods do not produce satisfactory results when the object background includes complex structures, or when the color contrast between the object and the neighboring background is low. To overcome this difficulty, a time sequence of images may be needed to increase segmentation accuracy. Even if it is impossible to distinguish the boundary between two objects in a particular frame, most likely they can be cleanly distinguished in another frame. Motion information from a sequence of images also provides useful cues for detecting boundaries, merging regions which belong to the same object, and solving the problem of occlusion.

Many papers have been published on the restoration of the three dimensional structures of objects from observations of their motion[2]. However, almost all of these techniques deal with a single rigid object. Only a few of them deal with multiple objects[1][18][13]. Though the recovery of the three dimensional structure of an object is interesting from a theoretical standpoint, it is not necessary for many practical applications such as the measurement of pedestrian and automotive traffic. In these applications, each image may include many objects, e.g. fifty. The problem in recovering the 3-D structures of all these objects from a sequence of limited-resolution images seems to be too difficult to solve. In addition, we believe algorithm robustness against noise and changes of imaging conditions is more important than fidelity when recovering object structure. Therefore, we concentrate on recovering 2-D translational motion from visual information, and ignore the reconstruction of the object structure.

Even though research on the automation of motion analysis in machine vision has become very active in the last few years, most papers have been concerned with only rigid motion[10]. Interesting analyses of human body motion can be seen in early papers or in the papers of the cognitive science field. Johannson demonstrates a sophisticated example of the interpretation of nonrigid motion using moving light dis-
plays (MLD) formed by small lights attached to a human body at several points. Similar techniques are used by Braunstein et al. in demonstrating how people discriminate rigid from non-rigid motion with minimum cues and views. Another interesting approach was used by O'Rourke and Badler. They analyzed human motion using constraint propagation and a detailed model of the human body. However, these detailed analyses of human body motion are not practical if the person is surrounded by many objects as is common in a real scene. Moreover, these analyses may be much more difficult than just segmenting objects. This paper, therefore, does not try to analyze human body motion in detail; rather, it analyzes or models the average motion of the human body.

Though each part of a person may move in different directions at any instant, the time average of the motion of the various parts converge to a global average value over a few seconds. Suppose we obtain the motion information of each distinct part of a person. Because these parts move coherently, they can be merged into a single group. If a region is found to include different velocities, the region should be divided into subregions, one for each distinct object. These merging and splitting operations are analogous to the perceptual grouping in the human visual perception of motion. The form of perceptual grouping pertinent here is that in which people perceive several points as belonging to a single group if they move together. It must be noted that, when two or more objects move in step, there is no significant difference in their motions. Therefore, additional information such as an object model is necessary to distinguish the individual objects.

This paper describes a new algorithm for segmenting monocular images of people in motion from a cinematic sequence of frames. An overview of the algorithm is described in the following section. The algorithm consists of two phases: motion estimation and people segmentation. The first phase, motion estimation, is described in Section 3. The second phase, people segmentation, is described in Section 4. The experimental results from a sequence of complex real images are shown and discussed in Section 5. Section 6 presents comments and conclusions.

2 Algorithm Overview

This section presents an overview of our algorithm. The algorithm must have the capability of detecting multiple nonrigid objects such as people walking in different directions, and must be robust against the noise typically encountered in real scene images. To eliminate some of the problems discussed in the previous section, the camera is assumed to be stationary, and the camera parameters and an object model are assumed to be given. Only translational 2-D motion is modeled in this paper.

The algorithm for people segmentation is based on image intensities, motion, and an object model — i.e. a model of the image of a person in motion. The algorithm consists of two successive phases; (1) motion estimation and (2) people segmentation. In the first phase, a motion field is estimated from pairs of consecutive frames of a monocular image sequence. The motion estimation algorithm initially uses a conventional correlation technique. This initial process detects only motion around the edges, while no motion is detected inside of the detected regions. To improve this initial motion estimation, a three-valued region image is extracted. Each pixel in the region image has one of three values $-1, 0, 1$, where $-1$ corresponds to darker regions relative to the background, $1$ to brighter regions, and $0$ to the background regions. Since the background is usually not simple but includes complex structures, a background image showing only stationary objects is extracted from the sequence to allow the complex background structure to be removed from each frame. The initial motion is improved by spatial propagation of the motion field, using the three-valued region image as a guide. Since people in motion are not rigid, the estimated motion is not constant within the region of a single person. The motion is temporally smoothed over a few seconds to reduce the non-rigidity of people in motion. After this smoothing process, the motions of all parts of a single person tend to converge to a global average value. Details of this phase are discussed in Section 3.

In the second phase, the people segmentation phase, the temporally smoothed motion is used to find the boundary between objects and the background. The regions in which motion is detected are first divided into subregions according to a quantized direction of motion, where the magnitude of the motion is ignored. These subregions are iteratively merged based on a hypothesis for merging two arbitrary regions. The hypothesis is evaluated by an acceptance measure obtained from a probabilistic object model of the image of a person in motion. This model includes the information about object sizes and motion. The most likely hypothesis is accepted in each iteration. The details of this phase are discussed in Section 4.
3 Motion Estimation

Two distinct techniques have been developed for the computation of motion from image sequences[9][2]. The first technique is based on the spatial and temporal gradients of image sequences[8][7][14]. The second is based on features such as points, edges, corners, or regions. The gradient based technique needs additional constraints such as smoothness[8]. Though this technique produces good results for smooth object surfaces, the constraints increase the estimation error around the object boundary. On the other hand, the feature based technique needs stable feature extraction prior to establishing the correspondence of objects in consecutive frames. This correspondence problem becomes more difficult when the image includes many objects and feature extraction is ambiguous. Both techniques are sensitive to noise and do not produce satisfactory results for the real images typically encountered in practical applications.

We chose the feature based approach for the initial computation of motion flow because of its relative insensitivity to noise. Image intensities within a small aperture are selected as a feature because we don't require specific feature extraction. A similar approach was applied to real outdoor scenes for the computation of optical flow by Burt[4]. The method proposed here consists of the following four steps;

1. region growing based on image intensities,
2. initial motion estimation,
3. motion propagation,
4. temporal smoothing of motion.

The first step extracts coherent regions having similar gray level values. This region image is utilized in the third step to improve the initial motion extracted in the second step. The second step calculates the initial motion field based on a quasi-cross-correlation technique from two consecutive image frames, but the estimated field is obtained only around edges. The field is blurred around the edges because of the use of a finite aperture size. Therefore, the initial flow is spatially smoothed by motion field propagation only within the object boundary (Step 3). The motion inside an object boundary is recovered in this step. The spatially smoothed flow will be almost constant if the object is rigid. If it is non-rigid, the flow changes inside the boundaries because each part of the object moves in different directions at any instant. In Step 4, the flow is recursively smoothed so that the flow for the parts of a single person is made to converge to a global average value.

3.1 Region Growing

Image intensities are useful for object segmentation, but the complex structure of most backgrounds makes segmentation difficult. If the background image (which includes stationary objects) can be extracted from a sequence of images, the background structure can be removed from the region image. The background image is extremely useful for extracting objects from a cinematic sequence of complex images. The background image $R_{i,j}$ can be obtained by mode filtering in the time domain — i.e., by extracting the most frequent value of image intensity at $(i,j)$. The background image must be revised to track changes in the lighting conditions over extended periods of time. A difference image $D_{i,j}$ between an input $F_{i,j}$ and the background image $B_{i,j}$ shows only moving object regions, while the stationary background is suppressed.

$$D_{i,j} = F_{i,j} - B_{i,j}$$  \hspace{1cm} (1)

An edge image produces a more accurate object boundary between the regions with similar intensities which belong to different objects. The edge image $E_{i,j}$ is obtained by an appropriate edge operator such as Canny’s operator[5], where $E_{i,j} = 1$ for edge points and $E_{i,j} = 0$ otherwise. A three-valued region image $R_{i,j}$ is obtained by the following equation using $D_{i,j}$, $E_{i,j}$, and two thresholds, $-T_d$ and $T_d$ which can be obtained experimentally for the difference image.

$$R_{i,j} = \begin{cases} 1 & \text{if } T_d \leq D_{i,j} \text{ and } E_{i,j} = 0 \\ 0 & \text{if } -T_d < D_{i,j} < T_d \text{ or } E_{i,j} = 1 \\ -1 & \text{if } D_{i,j} \leq -T_d \text{ and } E_{i,j} = 0 \end{cases}$$ \hspace{1cm} (2)

In this region image, the pixels expressing object boundaries or background retain 0 values, and pixels within moving object regions take -1 or 1 values.

3.2 Initial Motion Estimation

An initial motion vector $\delta = (\delta_x, \delta_y)$ is estimated from each successive pair of frames $F^t$ and $F^{t+1}$ based on a quasi-cross-correlation within local regions of the frames, where $\delta_x, \delta_y$ are the discrete function of spatial coordinates $i,j$, and time, $t$, and $F^t, F^{t+1}$ are the function of $i$ and $j$.

$$\delta = (\delta_x, \delta_y) = \{(m,n) \mid \min_{m,n \in (-l,..,l)} C_{m,n}\},$$ \hspace{1cm} (3)

where

$$C_{m,n} = \sum_{i=-1}^{1} \sum_{j=-1}^{1} |F_{i,j}^{t+1} - F_{i+m,j+n}^{t}|.$$ \hspace{1cm} (4)

In equations (3) and (4), the estimated motion is restricted to an integer value between -1 and 1. A 3-by-3 aperture size is selected to reduce the computation
time required. This process can be easily implemented on a dedicated parallel computer.

This initial motion estimation has three major handicaps: (1) motion is detected only in the area near edges, while no motion is detected inside of the object boundary, (2) motion field is blurred around the edges, and (3) the limited aperture size causes the aperture problem. The following subsection discusses how to improve the initial motion estimation.

3.3 MOTION PROPAGATION

Spatial smoothing may improve the initial motion estimation by offsetting the first and third problems mentioned above. This smoothing, however, should be restricted to within object boundaries because there is originally no motion outside of the object boundary if the background is stationary. The region image discussed in Section 3.1 is used for spatial propagation of motion to preserve the object boundaries and remove the blur of the motion field around edges. The motion propagation step is implemented by iterative weighted averaging. In this step, all estimated flow at points external to the region boundaries, where

\[ R_{i,j}^t = 0, \]

is eliminated, and all estimated flow except for the boundary points where

\[ E_{i,j}^t = 0, \]

is propagated by iterative averaging using 3-by-3 mask operations as follows.

\[
v_x^t(i,j) = \frac{1}{k} \sum_{i-1}^{i+1} \sum_{j-1}^{j+1} \delta_x(i,j),
\]

\[
v_y^t(i,j) = \frac{1}{k} \sum_{i-1}^{i+1} \sum_{j-1}^{j+1} \delta_y(i,j),
\]

where

\[
k = \sum_{i-1}^{i+1} \sum_{j-1}^{j+1} a_{i,j},
\]

and

\[
a_{i,j} = \begin{cases} 1 & \text{if } \delta_x \neq 0 \text{ or } \delta_y \neq 0 \\ 0 & \text{otherwise.} \end{cases}
\]

3.4 TEMPORAL SMOOTHING OF MOTION

The instantaneous motion field \( v^t = (v_x^t, v_y^t) \) at time \( t \) is not constant within an object boundary if the object is non-rigid. If the estimated field is averaged over a reasonable period of time, the motion vector, \( \dot{v}^t = (\dot{v}_x^t, \dot{v}_y^t) \), for a single object may be converged to a global average value. Recursive temporal smoothing is applied to the motion field \( (v_x^t, v_y^t) \) in the following manner:

\[
\begin{align*}
v_x^t(i,j) &= \frac{1}{1 + w^{t-1}} \left( v_x^t(i,j) + w^{t-1}V_x^{t-1}(i^*, j^*) \right), \\
v_y^t(i,j) &= \frac{1}{1 + w^{t-1}} \left( v_y^t(i,j) + w^{t-1}V_y^{t-1}(i^*, j^*) \right),
\end{align*}
\]

where

\[
i^* = i - \lfloor w^{t-1}(i,j) \rfloor, \\
j^* = j - \lfloor w^{t-1}(i,j) \rfloor,
\]

\( w^{t-1} \) is a weight for the previous motion estimates at time \( t - 1 \), and the parentheses \([ \ ]\) in equations (10) and (11) round the content off to the nearest integer. In these equations, the previous instantaneous motion estimation, \((v_x^{t-1}, v_y^{t-1})\) is also used to shift the previous motion estimation before the averaging operations.

4 PEOPLE SEGMENTATION

This section describes our method for segmenting images of people in motion from an image sequence. There are three major problems in extracting objects from the motion images obtained in the previous section:

- merged regions: A single connected region where motion is detected sometimes includes two or more objects,
- split regions: Each object tends to be divided into several parts, each of which has distinct intensity,
- occlusion: Objects are sometimes occluded by other objects.

We have developed a new region split-merge algorithm which consists of two steps;

1. motion-based region splitting,
2. model-based region grouping.

In the first step, each region is divided into a set of sub-regions by digitizing the direction of motion (step 1). These subregions are iteratively merged based on their directions of motion and the sizes of the rectangle circumscribing them to form a single object region (step 2). This step is analogous to the property of perceptual grouping in human visual perception. In this step, a model of the image of a person in motion is employed to merge subregions obtained in the first step. The merging process is based on the most likely hypothesis of merging two arbitrary subregions.

4.1 MOTION-BASED REGION SPLITTING

A connected region containing motion detected in the first step sometimes includes more than one object because of occlusion. In this step, each region is split by digitizing the direction of motion. Though clustering techniques also can be used to detect the boundaries between persons, our method is good enough and much simpler than clustering.
The direction of motion, \( d_{i,j} \) for each point \((i,j)\) in an image can be calculated from the motion vector,
\[
\mathbf{v}^t = \left( V_x^t(i,j), V_y^t(i,j) \right),
\]
where \( 0 \leq d_{i,j} < 2\pi \). Each single region is divided into subregions by digitizing the direction \( d_{i,j} \). The occlusion boundaries among people are detected in this step.

4.2 Model-Based Region Grouping

Objects are often divided into several discrete parts because of inadequate intensity or color contrasts between the object and background. If these parts move together, however, people perceive them as parts of a single object. This property of human visual perception is called perceptual grouping of moving objects.

(1) An Object Model

A probabilistic object model is employed to merge neighboring subregions. The model is simple and described by the probability density function \( p(w, h, \Delta \theta) \), where \( w, h, \) and \( \Delta \theta \) are the object’s width, height, and its direction difference of motion within a single object region, respectively. If they are mutually independent,
\[
p(w, h, \Delta \theta) = p(w)p(h)p(\Delta \theta). \tag{12}
\]
Suppose they are respectively normally distributed around averages \((\mu_w, \mu_h, 0)\) with variances \((\sigma_w^2, \sigma_h^2, \sigma_{\Delta \theta}^2)\). The density functions \( p(w), p(h), \) and \( p(\Delta \theta) \) are described as follows;
\[
p(w) = \frac{1}{\sigma_w \sqrt{2\pi}} e^{-(w-\mu_w)^2/2\sigma_w^2}, \tag{13}
\]
\[
p(h) = \frac{1}{\sigma_h \sqrt{2\pi}} e^{-(h-\mu_h)^2/2\sigma_h^2}, \tag{14}
\]
\[
p(\Delta \theta) = \frac{1}{\sigma_{\Delta \theta} \sqrt{2\pi}} e^{-\Delta \theta^2/2\sigma_{\Delta \theta}^2}. \tag{15}
\]
The parameters, \( \mu_w, \mu_h, \sigma_w, \sigma_h, \) and \( \sigma_{\Delta \theta} \), in these equations can be estimated experimentally.

(2) Hypotheses of Merging

Let \( x_i^j \) and \( x_i^* \) \((x_i^j \leq x_i^*)\) be the minimum and maximum \( x \)-coordinates of subregion \( i \), and \( y_i^j \) and \( y_i^* \) \((y_i^j \leq y_i^*)\) be the minimum and maximum \( y \)-coordinates of subregion \( i \) as shown in Figure 1. The region width \( w_{ij} \) and height \( h_{ij} \) after merging subregion \( j \) into \( i \) are respectively expressed by
\[
w_{ij} = w_{ij} = \max(x_i^j, x_i^*) - \min(x_i^j, x_i^*), \tag{16}
\]
\[
h_{ij} = h_{ij} = \max(y_i^j, y_i^*) - \min(y_i^j, y_i^*). \tag{17}
\]
The direction difference of the motion after merging is
\[
\Delta \theta_{ij} = \Delta \theta_{ij} = \theta_j - \theta_i. \tag{18}
\]
If \( w_{ij} \) and \( h_{ij} \) are small enough compared to their average values respectively, \( \mu_w \) and \( \mu_h \), and if \( |\Delta \theta_{ij}| \) is almost zero, the hypothesis of merging region \( j \) to \( i \) can be reasonably accepted. But if \( w_{ij} \) or \( h_{ij} \) is much greater than \( \mu_w \) or \( \mu_h \) respectively, or if \( |\Delta \theta_{ij}| \) is large, the hypothesis must be rejected. The following probability, \( P \), is selected as an acceptance measure for the hypotheses:
\[
P = P(w_{ij} < w)P(h_{ij} < h)P(|\Delta \theta_{ij}| < |\Delta \theta|), \tag{19}
\]
where \( P(w_{ij} < w) \), \( P(h_{ij} < h) \), and \( P(|\Delta \theta_{ij}| < |\Delta \theta|) \) are supposed to be mutually independent, and expressed by the following equations:
\[
P(w_{ij} < w) = 1 - \frac{1}{\sigma_w \sqrt{2\pi}} \int_{-\infty}^{w_{ij}} p(w)dw \tag{20}
\]
\[
P(h_{ij} < h) = 1 - \frac{1}{\sigma_h \sqrt{2\pi}} \int_{-\infty}^{h_{ij}} p(h)dh \tag{21}
\]
\[
P(|\Delta \theta_{ij}| < |\Delta \theta|) = 1 - \frac{2}{\sigma_{\Delta \theta} \sqrt{2\pi}} \int_{0}^{\Delta \theta_{ij}} p(\Delta \theta)d(\Delta \theta) \tag{22}
\]

(3) Iterative Grouping

Since the merging process is affected by the previous merging results, the process should be done slowly and sequentially as listed below;
1. Hypotheses are created for all combination of subregions \( i \) and \( j \).
2. The most likely hypothesis regarding subregion \( i \) which maximize equation (19) for an arbitrary subregion \( j \) is accepted in each iteration.
3. If \( P \) for \( i \) and \( j \) is greater than a threshold \( P_{th} \), the subregions \( i \) and \( j \) are merged, and a new subregion \( k \) is created while \( i \) and \( j \) are deleted.
4. Repeat from Step 1 until no more mergers occur.

![Figure 1: Region Merging](image-url)
5 EXPERIMENTS

The proposed algorithm was tested using a cinematic sequence of monocular intensity images. They were acquired by a stationary CCD camera which recorded an outdoor urban scene of a pedestrian crossing. Each image usually included thirty to forty persons against a stationary background. Therefore, objects often occluded each other. Examples of original images $I_t$ are shown in Figure 2; (a)frame #064, and (b)frame #128. The CCD camera used interlaced scanning which yields odd and even fields acquired at different times. Since this time difference causes motion blur between the odd and even scanning lines, only the odd fields were used in the following experiment. The image size was reduced to $256 \times 240$ to equalize the vertical and horizontal sampling rates. The frame rate was also reduced by 50% for computational simplicity. The resulting frame rate was fifteen frames per second.

Figure 3 shows the extracted background image using 20 equally spaced image frames from 240 frames which corresponds to an image sequence eight seconds in length. The image was extracted based on mode filtering in time and space. The result shows that all moving objects were eliminated in Figure 3. The image is not perfect and includes some noise around the top-right and bottom-left corners because of an excessively high density of moving people. A longer sequence would yield better results. Figure 4 shows the difference images for frames #064 and #128, where the background structure has been eliminated. Figures 5(a) and (b) show the region images obtained by thresholding the images shown in Figures 4(a) and (b) respectively.

Figure 6 shows an example of the initial motion field (small white arrows) estimated from two consecutive frames #128 and #130. The motion information is propagated from the object boundaries into the object regions via the spatial smoothing of motion. The result of this smoothing is shown in Figures 7. This
Figure 6: Initial Motion Field

Figure 7: Spatially Smoothed Motion Field

Figure 8: Temporally Smoothed Motion Field

6 Conclusions

A method for segmenting monocular images of people in motion is described. The method consists of two phases: motion estimation and people segmentation. In the motion estimation phase, initial motion is improved by using the region image obtained by splitting the histogram of the differences between an intensity image and background image at each pixel point. The background image is extracted by mode filtering of a sequence of intensity images in the time domain. Experimental results show that only background and stationary objects are retained in the background image, while people in motion are eliminated. To remove the non-rigidity of people in motion, the motion field is temporally smoothed by recursive averaging of motion fields. A nearly uniform motion field is obtained for each person after ten iterations in the experiments.

In the first step of the people segmentation phase, each connected region containing detected motion is divided into subregions by digitizing the direction of motion. In the next step, the subregions are merged to form the regions of a person based on a model of a person in motion. The proposed method is quite accurate and results from real outdoor scenes confirm that our approach to the segmentation of people in motion is effective.

We believe the performance of our algorithm can be improved by extending the segmentation procedure into the time domain.

Acknowledgments

We are grateful to Drs. Takahiko Kamae, Yukio Kobayashi, Yasuhito Suenaga, Ken-ichiro Ishii, and Shigeru Akamatsu of NTT Human Interface Laboratories for giving one of the authors, Akio Shio, the
opportunity to work at the University of California at Irvine. We also thank Shouji Kurakake at the University of Southern California for commenting on the first draft of this paper. This research was carried out at the University of California, Irvine with the support of Nippon Telegraph and Telephone Corporation.

References


