グラフィクスとＣＡＤ　66－20
(1993. 12. 17)

長時間画像時系列中における動物体の切出しと追跡

陳 暁静、　白井 良明

大阪大学 工学部 電子制御機械工学科
〒565 大阪府吹田市山田丘 2-1

あらまし

本報告では、連続画像から動物体を切り出して，それを追跡する研究を述べる。まず，２枚の連続画像からオプティカル・フローを抽出し，一様なフロー・ベクトルをパッチにまとめる。各々のパッチに含まれるフロー・ベクトルの平均を求めることによってパッチの動きベクトルを計算する。パッチの動きに基づいて，フロー場を複数の動きに分割する。分割によって動物体を切り出し，追跡を行なう。抽出したフロー・ベクトルに含まれるノイズの影響と、動き情報に基づく分割において生じた曖昧さを取り除くため、２枚の画像ではなく，長い時系列を利用する。

和文キーワード:

動画像処理、動きに基づく分割、動きの記述、追跡、長時間画像時系列

Detection and Tracking of Moving Objects in Long Image Sequences

Hsiao-Jing CHEN and Yoshiaki SHIRAI

Dept. of Mechanical Engineering for Computer-Controlled Machinery
Osaka University, Suita, Osaka 565, Japan
Email: chen@cv.ccm.eng.osaka-u.ac.jp

Abstract

A method is presented for detecting and tracking moving objects in long image sequences. An optical flow field is computed from each pair of consecutive images by using multiple filters. The former one of the two images is divided into small patches in such a way that optical flow in each patch is homogeneous. Flow vectors in each patch are spatially and temporally combined to compute correctly the local image motion of the patch. Incremental segmentation is performed to group patches into object segments based on the estimated local image motions. In order to remove inherent ambiguities in motion-based segmentation due to local or global similarity between apparent motions of different object, temporal coherence between the apparent motion of an object and the local image motion of each patch in the object segment is investigated over a long sequence.

Key words:

Motion analysis, motion-based segmentation, motion description, tracking, long image sequences

I. Introduction

Apparent motion is an important visual cue for various vision tasks [6]. It can be exploited to divide a retina image into regions of homogeneous motion properties. Such kind of (motion-based) segmentation is helpful for recognition of moving objects.

A. *Issues*

The apparent motion cue derived from only a pair of consecutive images, however, is unreliable because a flow field computed from the images is usually noisy. The noise in flow vectors will influence upon the segmentation of images and estimation of object motions.

Furthermore, there are several cases where the motion-based segmentation becomes inherently ambiguous. When apparent motions of at least two objects are locally similar at some locations in an image, evidences suggesting the one of those objects to which the locations belong are ambiguous. Fig.1 illustrates such an example. A circular region in the image center rotates in clockwise direction around the region center. The background moves to right direction between consecutive image frames. Apparent motions of the two parts are *locally* similar near the upper side of the circular region.

Another case occurs when apparent motions of different objects become ***globally*** similar during a time period. Fig.2 gives such an example. Upper and lower parts of man's left leg can be interpreted as two rigidly moving objects when the two parts rotate differently. However, the motions of those two parts between each pair of the last several frames become globally similar. Then, if the instantaneous apparent motion cue obtained from only a pair of consecutive frames is used, the two parts are interpreted as an individual object.

In human visual system, it is likely that a correct result of motion-based segmentation is not obtained at only one time instant. Instead the performance of the segmentation is accumulative (or say, incremental). By temporally integrating (smoothing) local velocity signals, effective signal-to-noise ratio is increased [6]. A correct segmentation is gradually obtained as the noise in local signals is reduced by the temporal integration.

A key role exploited to remove the ambiguities is the fact that a temporal coherence exists between the apparent motion of an object and those flow vectors at the locations which truly belong to the object. Flow vectors at the locations are coherent with the apparent motion over long time. But they are coincidental with apparent motions of other objects during only a short period. A decision of grouping each location into an object segment is made by accumulatively investigating whether the flow vector at the location and the apparent motion of the object are temporally coherent, rather than whether they are coincidental at only some instant. When apparent motions of two objects are globally similar in a short period, and flow vectors computed at locations belonging to one of the objects are temporally incoherent with the apparent motion of the other one, the two objects are interpreted as individual ones.

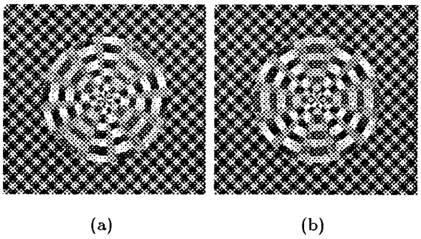


Fig. 1: The 1st (left) and 30th image frame (right) in a synthetic image sequence. Intensity at every pixel is corrupted by a random noise drown from a Gaussian distribution $N(0, 15)$.

B. *Outline of Our Method*

The goal of our research is to detect and track multiple moving objects by performing motion-based image segmentation. Each object rigidly moving in a scene is approximated by a plane. To deal with the problems mentioned above, our method accumulatively observes apparent motion in a long image sequence.

There are five stages in the method: (a) Computing an optical flow field; (b) Dividing the image at time t into small patches. In order to reduce noise in the computed flow vectors, a local image motion vector of each patch is estimated by averaging flow vectors at locations in the corresponding patches in several successive images; (c) Obtaining initial segments. In the first image, segments are initially obtained by grouping patches of similar local image motion. In each of succeeding images, segments are initially obtained based on the segments detected in the previous image; (d) Updating the obtained initial segments by iterating an attempt to merge two segments or split a segment. If the image motions of two segments are not temporally coherent each other up to time $t-1$, the corresponding segments at time t are not merged; (e) Investigating the temporal coherence between the local image motion of a patch and the apparent motion of every plane up to time t. Then each patch is reorganized into that segment of the plane whose apparent motion is temporally most coherent with the local image motion of the patch.

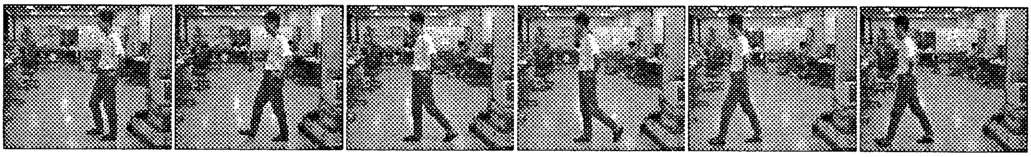


Fig. 2: Six images in a real image sequence: (from left to right) the 1st, 10th, 25th, 30th, 55th, and 60th image frame, respectively.

II. Estimating Local Image Motions

Several sophisticated methods have been proposed to estimate local image motion [7, 4]. Our method groups edge locations in each image into small patches, and estimate a local image motion vector of each patch by spatially and temporally combining flow vectors.

Let an edge location in the image frame at time t and a flow vector at the location be denoted by $\boldsymbol{x}$ $(= (x, y)^T)$ and $\boldsymbol{u}(\boldsymbol{x})$ $(= (u, v)^T$, where superscript T denotes vector or matrix transpose). By using more than two filters, the flow vector $\boldsymbol{u}(\boldsymbol{x})$ and a covariance matrix $\mathbf{C}(\boldsymbol{x})$ of the vector are computed [1].

At each location near a corner point, a flow vector is uniquely determined, and is named a "unique vector". At each location near a straight edge, only a velocity component normal to the edge orientation is determined, and is named a "normal vector".

Then edge locations in the image at time t are grouped into two kinds of patches. One kind is the patch which contains locations with similar unique flow vectors. Another kind contains only locations of normal vectors. Each patch contains only a certain number of spatially connected locations (Fig.3(a)). Let a patch in the image at time t be denoted by $U^{(t)}$.

In the first image, patches are obtained by grouping locations with similar flow vectors [1]. In each of succeeding images, patches are obtained based on corresponding patches in the previous image. Establishing a correspondence between a patch $U^{(t)}$ and a patch $U^{(t-1)}$ will be described in Section IV. Coordinates of the center of a patch $U^{(t)}$ are calculated as $\bar{\boldsymbol{x}} = n_f^{-1} \sum_{\boldsymbol{x} \in U^{(t)}} \boldsymbol{x}$, where n_f denotes the number of locations in $U^{(t)}$. The local image motion at $U^{(t)}$ is represented by a vector $\bar{\boldsymbol{u}}(\bar{\boldsymbol{x}})$ (Fig.3(b)).

Let $\{U^{(t-c)}, \ldots, U^{(t-1)}, U^{(t)}\}$ denote corresponding patches in several successive images at time $\{t - c, \ldots, t-1, t\}$, respectively, where c is a positive constant. The local image motion vector $\bar{\boldsymbol{u}}$ of $U^{(t)}$ is calculated by averaging flow vectors in the patches $\{U^{(t-c)}, \ldots, U^{(t-1)}, U^{(t)}\}$. The covariance matrix $\bar{\mathbf{C}}$ of residual errors between $\bar{\boldsymbol{u}}$ and flow vectors $\boldsymbol{u}$ in $U^{(t)}$ is also estimated (see [2] for details).

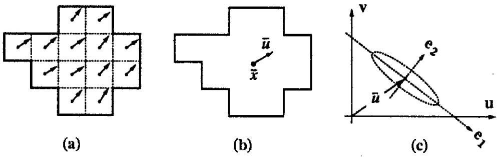


Fig. 3: Illustration of a patch. (a) The locations of similar flow vectors which are grouped into a patch. (b) Boundary, center, and local image motion vector of the patch. (c) Local image motion vector $\bar{\boldsymbol{u}}$ of a patch $U^{(t)}$ and an ellipse defined by the covariance matrix of the vector.

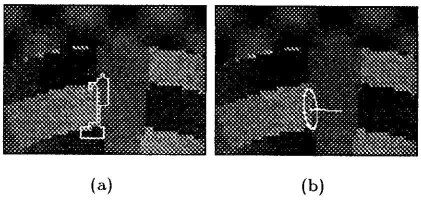


Fig. 4: A patch in the 5th image. (a) Its boundary; (b) Local image motion estimated at the patch.

Let the larger one of two eigenvalues of the matrix $\bar{\mathbf{C}}$ be denoted by σ_1^2, and the other one by σ_2^2. Two unit eigenvectors corresponding to σ_1^2 and σ_2^2 are denoted by $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ (Fig.3(c)). Usually the vector $\bar{\boldsymbol{u}}$ of a patch near a straight edge has its eigenvector $\boldsymbol{e}_1$ parallel to the edge direction, and has $\boldsymbol{e}_2$ normal to the direction. Uncertainties of $\bar{\boldsymbol{u}}$ along the directions of $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ are measured by the two eigenvalues σ_1^2 and σ_2^2, respectively.

Fig.4(a) shows a patch located at the upper side of the circular region in the image. Fig.4(b) shows the local image motion vector of the patch and an ellipse determined by the covariance matrix $\bar{\mathbf{C}}$.

III. Motion-Based Segmentation

Segmentation of each image is performed by grouping patches into segments. A segment is defined as a group of patches belong to the same plane.

A. Estimating a Description of Apparent Motion

A kind of apparent motion representation is used to describe the apparent image motion of a moving plane.

At the center $\bar{\boldsymbol{x}}$ $(= (\bar{x}, \bar{y})^T)$ of a patch belonging to the plane, the image flow $\tilde{\boldsymbol{u}}(\bar{\boldsymbol{x}})$ is represented by $\tilde{\boldsymbol{u}}(\bar{\boldsymbol{x}}) = \mathbf{J}(\bar{\boldsymbol{x}})\,\tilde{\boldsymbol{a}}$, where $\tilde{\boldsymbol{a}} = (\tilde{a}_7 \ldots \tilde{a}_0)^T$, and

$$\mathbf{J}(\bar{\boldsymbol{x}}) = \begin{pmatrix} \bar{x}^2 & \bar{x}\bar{y} & \bar{x} & \bar{y} & 1 & 0 & 0 & 0 \\ \bar{x}\bar{y} & \bar{y}^2 & 0 & 0 & 0 & \bar{x} & \bar{y} & 1 \end{pmatrix} \qquad (1)$$

(see literatures [5, 8] for details). The parameter vector $\tilde{\boldsymbol{a}}$ is used as a model to describe the apparent motion of a plane.

Suppose that there exist a certain number of planes detected at time t. In the image at time t, let the segment and the motion model of the i-th detected plane be denoted by $R_i^{(t)}$ and $\tilde{\boldsymbol{a}}_i^{(t)}$, respectively. In order to make the estimation of a motion model as insensitive to the noise as possible, $\tilde{\boldsymbol{a}}_i^{(t)}$ is estimated by spatially and temporally combining local image motion vectors of patches in corresponding segments $R_i^{(\tau)}$ for $\tau = t, t-1, \ldots$, where $R_i^{(\tau)}$ denotes the segment of the i-th plane at time τ. Establishing a correspondence between two segments $R_i^{(\tau-1)}$ and $R_i^{(\tau)}$ will be described in Section IV.

The estimation of $\tilde{\boldsymbol{a}}_i^{(t)}$ is represented as

$$\hat{\boldsymbol{a}}_i^{(t)} = \arg \min_{\tilde{\boldsymbol{a}}_i^{(t)}} \sum_{\tau} w_g^{(\tau)} \sum_{U^{(\tau)} \in R_i^{(\tau)}} e_g^2(U^{(\tau)}), \qquad (2)$$

where $e_g^2(U^{(\tau)}) = (\bar{\boldsymbol{u}} - \mathbf{J}(\bar{\boldsymbol{x}})\tilde{\boldsymbol{a}}_i^{(t)})^T \bar{\mathbf{C}}^{-1} (\bar{\boldsymbol{u}} - \mathbf{J}(\bar{\boldsymbol{x}})\tilde{\boldsymbol{a}}_i^{(t)})$, and $\bar{\boldsymbol{x}}$, $\bar{\boldsymbol{u}}$, and $\bar{\mathbf{C}}$ denote the center, local image motion vector, and the vector's covariance matrix of $U^{(\tau)}$, respectively. $w_g^{(\tau)}$ denotes a relative weight assigned to the local image motion vectors of all patches in $R_i^{(\tau)}$. Although the model varies between different images, it can be approximately regarded as being constant if the image motion is not too large.

B. Obtaining Initial Segments in an Image

After the computation of the flow field at time t, initial segments in the image at time t are obtained.

In the first image, patches are grouped into initial segments in such a way that local image motions of patches in each segment are uniform. The next steps are repeated. (i) Find a patch which has not been grouped. If such a patch can not be found then stop; else increase i by one, where i is used to indicate the number of a group. Let the found patch be the first member of the i-th group. A motion vector $\bar{\boldsymbol{u}}_i$ is initialized as $\bar{\boldsymbol{u}}_i = \bar{\boldsymbol{u}}$, where $\bar{\boldsymbol{u}}$ is the local image motion vector of the found patch. (ii)Find a patch which has not been grouped and is spatially connected to a patch in the i-th group. In order to identify whether the found patch can be grouped into the i-th group, a condition is defined as

$$(\bar{\boldsymbol{u}} - \bar{\boldsymbol{u}}_i)^T \bar{\mathbf{C}}^{-1} (\bar{\boldsymbol{u}} - \bar{\boldsymbol{u}}_i) < \mu\,, \qquad (3)$$

where $\bar{\boldsymbol{u}}$ denotes the local image motion vector of the found patch, $\bar{\mathbf{C}}$ denotes the covariance matrix, and the μ is a threshold (which is predefined to be 0.2 in our experiments). If Eq.(3) is satisfied, the found patch is grouped into the i-th group, and $\bar{\boldsymbol{u}}_i$ is modified by averaging local motion vectors of all patches in the group. This step is repeated until no patch can be grouped. Go to (i).

When the grouping stopped, the first image is divided into groups of patches. Let the i-th group of patches be the i-th initial segment in the first image.

Segments in the image at time t for $t > 1$ are firstly determined from their corresponding segments obtained in the image at time $t-1$ (Section IV).

C. Updating Segments in an Image

1) Merging Segments

While as the apparent motion of a plane is not spatially constant, over-segmentation will be derived from the first image by the way described above. Furthermore, patches belonging to the same plane may not be grouped up to the previous image because of the noise in local image motion vectors. When local image motions of patches are correctly estimated by temporally combining flow vectors at locations in corresponding patches, it becomes possible to group the patches. Therefore, it is necessary to try to merge segments in each image.

Segmentation (an attempt to merge two segments or split a segment) is performed iteratively. In the k-th iteration, an attempt to merge a segment $R_i^{(t)}$ into another segment $R_j^{(t)}$ is performed. A segmentation criterion is needed to identify whether the two segments can be merged (i.e., whether they belong to the same plane). In our method, the segmentation criterion is achieved by a hypothesis testing scheme. The null hypothesis H_0 is defined as: the two segments belong to the same plane. The alternative one H_1 is: they belong to different ones. The null hypothesis is tested by the log-likelihood ratio defined as

$$l_{ij} = -2\log(L(H_0)/L(H_1))\,, \qquad (4)$$

where $L(H_0)$ and $L(H_1)$ denote the likelihood functions of H_0 and H_1, respectively.

Let the segment obtained by merging $R_i^{(t)}$ and $R_j^{(t)}$ be denoted by $R_m^{(t)}$. If all patches in $R_m^{(t)}$ truly belong to the same plane, a residual error between $\bar{\boldsymbol{u}}(\bar{\boldsymbol{x}})$ of each patch $U^{(t)} \in R_m^{(t)}$ and the apparent motion of the plane,

$\hat{\boldsymbol{u}}_m$ $(= \mathbf{J}(\bar{\boldsymbol{x}})\hat{\boldsymbol{a}}_m^{(t)})$, should be close to zero. As described above, $\bar{\boldsymbol{u}}$ has uncertainties σ_1^2 and σ_2^2 along directions of the two unit eigenvectors $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ (see Fig.3(c)). Therefore, the residual error $\bar{\boldsymbol{u}} - \hat{\boldsymbol{u}}_m$ is decomposed to two components along $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$, i.e.,

$$\Delta u_{m1} = \boldsymbol{e}_1^T(\hat{\boldsymbol{u}}_m - \bar{\boldsymbol{u}}), \quad \Delta u_{m2} = \boldsymbol{e}_2^T(\hat{\boldsymbol{u}}_m - \bar{\boldsymbol{u}}),$$

and the obtained components are normalized based on the uncertainties

$$\boldsymbol{\delta}_m = \left(\frac{1}{\sigma_1}\Delta u_{m1}\right)\boldsymbol{e}_1 + \left(\frac{1}{\sigma_2}\Delta u_{m2}\right)\boldsymbol{e}_2 . \qquad (5)$$

It is assumed that the distribution of $\boldsymbol{\delta}_m$ is Gaussian of mean $\hat{\boldsymbol{\delta}}_m$ and covariance matrix $\hat{\boldsymbol{\Sigma}}_m$, where

$$\hat{\boldsymbol{\delta}}_m = \frac{1}{n_m}\sum_{U^{(t)}\in R_m^{(t)}} \boldsymbol{\delta}_m ,$$

$$\hat{\boldsymbol{\Sigma}}_m = \frac{1}{n_m}\sum_{U^{(t)}\in R_m^{(t)}} (\boldsymbol{\delta}_m - \hat{\boldsymbol{\delta}}_m)(\boldsymbol{\delta}_m - \hat{\boldsymbol{\delta}}_m)^T ,$$

and n_m is the number of patches in $R_m^{(t)}$.

Let $\boldsymbol{\theta}_m$ denote a distribution parameter which contains $\hat{\boldsymbol{\delta}}_m$ and $\hat{\boldsymbol{\Sigma}}_m$. The likelihood function of $\boldsymbol{\theta}_m$ is

$$L(\boldsymbol{\theta}_m) = \prod_{U^{(t)}\in R_m^{(t)}} p(\boldsymbol{\delta}_m \mid \boldsymbol{\theta}_m) \qquad (6)$$

where $p(\boldsymbol{\delta}_m \mid \boldsymbol{\theta}_m)$ denotes the distribution density function of $\boldsymbol{\delta}_m$. The likelihood functions of $\boldsymbol{\theta}_i$ of $R_i^{(t)}$ and $\boldsymbol{\theta}_j$ of $R_j^{(t)}$ are obtained in the same way.

Then the functions $L(H_0)$ and $L(H_1)$ defined in Eq.(4) are substituted by $L(\boldsymbol{\theta}_m)$ and $L(\boldsymbol{\theta}_i)L(\boldsymbol{\theta}_j)$, respectively. If l_{ij} is less than a threshold α which is related to a significance level of the test, the null hypothesis H_0 is accepted (i.e., $R_i^{(t)}$ is merged into $R_j^{(t)}$).

2) Splitting a Segment into Different Motion Parts

Patches belonging to different planes may be grouped into a segment due to the global similarity between apparent motions of the planes up to the previous image. When apparent motions of the planes at time t become different, the segment is split to different motion parts.

In order to identify whether a segment $R_m^{(t)}$ contains patches belonging to different planes, an attempt is made to split $R_m^{(t)}$ to two new segments $R_i^{(t)}$ and $R_j^{(t)}$. As depicted in Fig.5(a), the normalized residual $\boldsymbol{\delta}_m$ of each patch $U^{(t)} \in R_m^{(t)}$ defined in Eq.(5) is transformed to the u'-v' space whose axes are identical with the eigenvectors of $\hat{\boldsymbol{\Sigma}}_m$. Only the u' component of $\boldsymbol{\delta}_m$ is investigated because it has larger variance than the v' component.

In the segment $R_m^{(t)}$, patches are firstly classified into a class C^+ if their $\boldsymbol{\delta}_m$ have non-negative u' component, and C^- otherwise (Fig.5(b)).

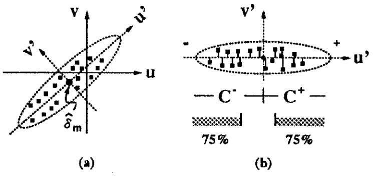


Fig. 5: Transforming residuals to a space u'-v'. (a) Distribution of the residuals and two eigenvectors u' and v' of the distribution's covariance matrix. (b) Selecting the patches whose residuals' u' components have relatively larger absolute values.

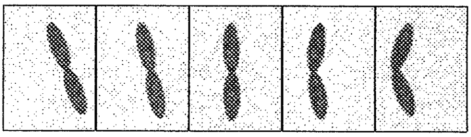


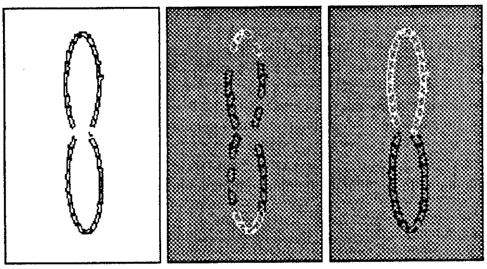
Fig. 6: Five images in a synthetic image sequence: (from left to right) the 1st, 5th, 10th, 15th, and 20th image frame, respectively. Intensity at each pixel is corrupted by a random noise $N(0, 15)$.

As described previously, a motion model can be estimated more reliably when many local image motion vectors are used. Therefore, it is desirable to divide all patches in $R_m^{(t)}$ into several groups. Each group has to contain patches belonging to the same plane, and the number of patches in the group is enough for the model estimation. We found that this can be achieved by (1) selecting a certain percent of patches in each of classes C^+ and C^-, respectively, (our method selects 75% of the patches in each class) such that $|u'|$ of the selected patches are greater than other patches in the class (other patches are neglected); (2) Collecting spatially connected patches with the same sign of u' into a group.

A synthetic image sequence shown in Fig.6 is used to illustrate the splitting of a segment. In the sequence, two elliptical regions simulate two planes perpendicular to the sensor's optical axis. They have the same rotation during the 1st to 10th frame, and then rotate differently. Two planes can not be separated up to the 10th image. Fig.7(a) shows obtained patches in the 10th image. The patches are predicted as belonging to the same segment, even though their image motions become different.

Fig.7(b) shows the selected patches in the 10th image. It can be seen that the selected patches which are spatially connected and have the same sign of u' constitute a group of patches belonging to the same plane.

Then, each of the neglected patches is regarded as



(a) (b) (c)

Fig. 7: An example of splitting of a segment. (a) Boundaries of patches in a segment. (b) White and black lines indicate boundaries of the selected patches with positive and negative u' component, respectively. (c) Boundaries of patches in two segments which are obtained by splitting the predicted segment shown in (a).

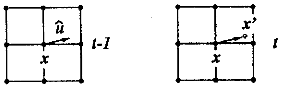


Fig. 8: Establishing the correspondence between a location x in the image at time $t-1$ (left) and a location x' in the image at time t (right).

an individual group, and the next three steps are repeated: (i) Obtained groups are sorted into ascending order of the number of patches in each group; (ii) The group with smallest number of patches, denoted by G_0, is merged into a larger group G such that

$$\min_{G} \frac{1}{n_p} \sum_{U^{(t)}\in G_0 \cup G} (\bar{\boldsymbol{u}}-\hat{\boldsymbol{u}})^T \bar{\mathbf{C}}^{-1}(\bar{\boldsymbol{u}}-\hat{\boldsymbol{u}})$$

where n_p denotes the number of patches in $G_0 \cup G$, $\bar{\boldsymbol{u}}$ and $\bar{\mathbf{C}}$ denote the local image motion vector and covariance matrix of $U^{(t)}$, and $\hat{\boldsymbol{u}}$ is calculated at $U^{(t)}$ by a motion model estimated from the local image motion vectors of all patches in $G_0 \cup G$; (iii) Go to (i) if the number of groups is larger than two, otherwise stop.

Final two groups of patches are regarded as $R_i^{(t)}$ and $R_j^{(t)}$, respectively. Fig.7(c) shows two segments obtained by splitting the segment shown in Fig.7(b). The null hypothesis that $R_i^{(t)}$ and $R_j^{(t)}$ belong to the same plane is tested. The attempt to split $R_m^{(t)}$ is accepted if the null hypothesis is rejected.

A segment may not be perfectly split after only one iteration. Instead a correct splitting can be obtained after several iterations of attempts to split and merge segments.

IV. Tracking

At first, each patch in the image at time t is obtained according to a patch in the image at time $t-1$.

Let a grid location in a patch $U^{(t-1)} \in R_i^{(t-1)}$ be denoted by $\boldsymbol{x}$. As illustrated in Fig.8(a), a location $\boldsymbol{x}'$ in the image at time t is determined as $\boldsymbol{x}' = \boldsymbol{x} + \hat{\boldsymbol{u}}_i$, where $\hat{\boldsymbol{u}}_i = \mathbf{J}(\boldsymbol{x})\hat{\boldsymbol{a}}_i^{(t-1)}$. Because $\boldsymbol{x}'$ usually does not lie on a grid location, the grid location which is nearest to $\boldsymbol{x}'$ is grouped into $U^{(t)}$. All grid locations in $U^{(t-1)}$ are used to group locations in the image at time t. The grouped locations constitute $U^{(t)}$ which corresponds to $U^{(t-1)}$.

Each initial segment $R_i^{(t)}$ in the image at time t is determined by grouping the patches $\{U^{(t)}\}$ which correspond to the patches $\{U^{(t-1)}\}$ in $R_i^{(t-1)}$. In this way, the correspondence between $R_i^{(t-1)}$ and $R_i^{(t)}$ is established. By establishing the correspondence between a patch in the image at time $t-1$ and a patch in the image at time t, patches and segments are tracked over a long sequence.

In order to reduce the computational cost, our method performs a few iterations at each image. Then initial segments in the succeeding image are obtained, and the attempt to merge or split segments is also iterated at the image, and so on.

By tracking each patch, flow vectors in the corresponding patches at different times are combined to reduce the noise in flow vectors. When the local image motions are correctly estimated, each image is incrementally segmented.

V. Removing Ambiguities in Segmentation

A. Removing the Ambiguity due to Local Similarity

The ambiguity due to local similarity between apparent motions of different objects is removed by investigating temporal coherence between the local image motion of a patch and the apparent motion of each of the detected planes.

When the number of detected planes becomes stable during several successive images up to time $t_0 - 1$, the investigation is carried out from time t_0. Let N denote the number of the detected planes. At each patch $U^{(t)}$ $(t \geq t_0)$, a set of state vectors $\{\boldsymbol{\omega}_i, i = 1, \ldots, N\}$ is defined, where $\boldsymbol{\omega}_i = (\omega_i^{(t_0)}, \ldots, \omega_i^{(t)})^T$, $\omega_i^{(\tau)}$ denotes a state: $U^{(\tau)}$ very likely belongs to the i-th plane at time τ. A matrix $\mathbf{V}$ is defined as $\mathbf{V} = (\bar{\boldsymbol{u}}^{(t_0)}, \ldots, \bar{\boldsymbol{u}}^{(t)})$ where $\bar{\boldsymbol{u}}^{(\tau)}$ denotes the local image motion of $U^{(\tau)}$ in the image at time τ.

Here we define a set of discriminant functions

$\{\gamma_i(U^{(t)}),\ i = 1, \ldots, N\}$ at every patch $U^{(t)}$ in the image at time t. A function $\gamma_i(U^{(t)})$ represents a likelihood of grouping $U^{(t)}$ into $R_i^{(t)}$. The functions are determined as $\{\gamma_i(U^{(t)}) = P(\boldsymbol{\omega}_i|\mathbf{V})\,, i = 1, \ldots, N\}$, where $P(\boldsymbol{\omega}_i|\mathbf{V})$ represents the *a posteriori* probability of $\boldsymbol{\omega}_i$ given $\mathbf{V}$ (see [3]).

By applying the compound Bayes decision rule [3], $P(\boldsymbol{\omega}_i|\mathbf{V})$ is estimated as

$$P(\boldsymbol{\omega}_i|\mathbf{V}) = p(\mathbf{V}|\boldsymbol{\omega}_i)P(\boldsymbol{\omega}_i)/P(\mathbf{V})\,, \tag{7}$$

where $p(\mathbf{V}|\boldsymbol{\omega}_i)$ is the conditional density function of $\mathbf{V}$ given the total state $\boldsymbol{\omega}_i$, $P(\boldsymbol{\omega}_i)$ is the *a priori* probability of $\boldsymbol{\omega}_i$, and $P(\mathbf{V}) = \sum_{j=1}^{N} p(\mathbf{V}|\boldsymbol{\omega}_j)P(\boldsymbol{\omega}_j)$. Without any prior knowledge about the probability of each state $\boldsymbol{\omega}_i$, we have $\{P(\boldsymbol{\omega}_i) = 1/N, i = 1, \ldots, N\}$.

The function $p(\mathbf{V}|\boldsymbol{\omega}_i)$ in Eq.(7) is estimated as

$$p(\mathbf{V}|\boldsymbol{\omega}_i) = \prod_{\tau=t_0}^{t} p(\bar{\boldsymbol{u}}^{(\tau)}|\omega_i^{(\tau)})\,, \tag{8}$$

where $p(\bar{\boldsymbol{u}}^{(\tau)}|\omega_i^{(\tau)})$ denotes the conditional probability of $\bar{\boldsymbol{u}}^{(\tau)}$ given the state $\omega_i^{(\tau)}$. The probability $p(\bar{\boldsymbol{u}}^{(\tau)}|\omega_i^{(\tau)})$ can be determined by comparing $\bar{\boldsymbol{u}}^{(\tau)}$ with the apparent motion of the i-th detected plane, i.e., with $\hat{\boldsymbol{u}}_i$ $(= \mathbf{J}(\bar{\boldsymbol{x}})\hat{\boldsymbol{a}}_i^{(\tau)})$. Because the estimated local image motion $\bar{\boldsymbol{u}}^{(\tau)}$ of $U^{(\tau)}$ may be uncertain in some direction, the uncertainty should be considered. Therefore, we define $p(\bar{\boldsymbol{u}}^{(\tau)}|\omega_i^{(\tau)})$ as

$$p(\bar{\boldsymbol{u}}^{(\tau)}|\omega_i^{(\tau)}) = \frac{1}{2\pi|\bar{\mathbf{C}}|^{1/2}} \exp\{-\varepsilon_i^2(U^{(\tau)})\}, \tag{9}$$

$\varepsilon_i^2(U^{(\tau)}) = (\hat{\boldsymbol{u}}_i - \bar{\boldsymbol{u}}^{(\tau)})^T \bar{\mathbf{C}}^{-1}(\hat{\boldsymbol{u}}_i - \bar{\boldsymbol{u}}^{(\tau)})/2$,, and $\bar{\mathbf{C}}$ denotes the covariance matrix of $\bar{\boldsymbol{u}}^{(\tau)}$. Note that the covariance of $\hat{\boldsymbol{u}}_i$ is not considered in Eq.(9). We expect that the model $\hat{\boldsymbol{a}}_i^{(\tau)}$ can be estimated reliably by spatially and temporally combining local image motion vectors of many patches.

The discriminant functions of each patch can be easily evaluated by tracking the patch up to the image at time t. By grouping $U^{(t)}$ with $\gamma_i(U^{(t)}) > \gamma_j(U^{(t)})$ for all $j \neq i$ into $R_i^{(t)}$, a stable performance of segmentation is obtained when the ambiguity absences from several images. The performance can still be retained even though the ambiguity appears again.

The ambiguity induced by the local similarity between apparent motions of different planes is removed as follows. When the update of segments and evaluations of the discriminant functions of each patch have been accomplished, reorganization of patches is performed in the image at time t. If a patch $U^{(t)}$ has been grouped into a segment $R_j^{(t)}$ and $\gamma_i(U^{(t)}) > \gamma_j(U^{(t)})$, it is removed from $R_j^{(t)}$ and is grouped into $R_i^{(t)}$.

B. Removing the Ambiguity due to Global Similarity

It is also necessary to address another case of ambiguity in which apparent motions of different moving objects become globally similar during a period.

Suppose that two segments $R_i^{(t)}$ and $R_j^{(t)}$ belong to different planes which have identical apparent motions at time t. The apparent motions of the two planes are not coherent over the entire sequence. In order to avoid merging the two segments when the apparent motions become identical, the temporal coherence between the apparent motions is also investigated.

As described above, the temporal coherence between the local image motion of a patch $U^{(t-1)} \in R_i^{(t-1)}$ and apparent motion of the j-th detected plane up to the image at time $t-1$ is represented by the discriminant function $\gamma_j(U^{(t-1)})$. It is obvious that, if apparent motions of the i-th and j-th planes are not temporally coherent in the period $[t_0, t-1]$, $\gamma_i(U^{(t-1)})$ and $\gamma_j(U^{(t-1)})$ of almost every patch $U^{(t-1)} \in R_i^{(t-1)}$ are different. We have $\gamma_i(U^{(t-1)}) \gg \gamma_j(U^{(t-1)})$, or say, $\gamma_i(U^{(t-1)}) - \gamma_j(U^{(t-1)})$ is close to 1.0.

The values of $\gamma_i(U^{(t-1)})$ of all patches $U^{(t-1)}$ in $R_i^{(t-1)}$ are averaged

$$\bar{\gamma}_i^{(t-1)} = \frac{1}{n_p} \sum_{U^{(t-1)} \in R_i^{(t-1)}} \gamma_i(U^{(t-1)})\,,$$

where n_p is the number of patches in $R_i^{(t-1)}$. Also

$$\bar{\gamma}_j^{(t-1)} = \frac{1}{n_p} \sum_{U^{(t-1)} \in R_i^{(t-1)}} \gamma_j(U^{(t-1)})\,.$$

Clearly $\bar{\gamma}_i^{(t-1)}$ and $\bar{\gamma}_j^{(t-1)}$ are also significantly different, i.e., $\gamma_i(U^{(t-1)}) \gg \gamma_j(U^{(t-1)})$. Thus, merging $R_i^{(t)}$ into $R_j^{(t)}$ is avoided if $\bar{\gamma}_i^{(t-1)} - \bar{\gamma}_j^{(t-1)}$ is larger than a threshold T_γ. Since the apparent motions of the two planes may be locally similer up to the image at time $t-1$, $\gamma_i(U^{(t-1)})$ may be approximately equal to $\gamma_j(U^{(t-1)})$ at several patches in $R_i^{(t-1)}$. Hence we can expect that $\bar{\gamma}_i^{(t-1)} > \bar{\gamma}_j^{(t-1)}$, but $\bar{\gamma}_i^{(t-1)} - \bar{\gamma}_j^{(t-1)}$ may not be close to 1.0. From results of various experiments, we found that predefining the threshold T_γ to be 0.5 is significant.

VI. Experimental Results

The first example is the result obtained from the sequence shown in Fig.1. Fig.9(a) shows segments of the foreground and background obtained at the 5th image in the sequence. The investigation of temporal coherence is still not performed. It can be seen that a part of the foreground is grouped into the segment of background because of the local similarity between image motions of the two segments. Fig.9(b) gives the segment of the foreground detected in several succeeding images. Influences of the ambiguity are gradually removed, and the foreground is detected almost correctly in the 25th image.

Fig.10 shows the segments of the upper and lower parts of the man's left leg which are obtained in several images in the sequence shown in Fig.2. Each of the

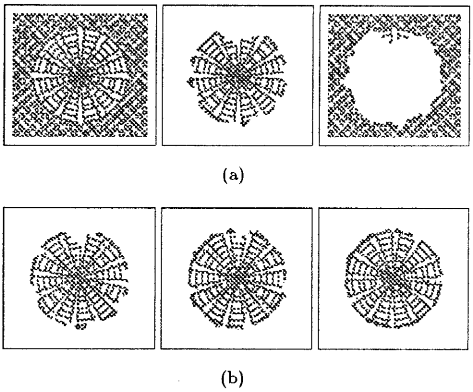


Fig. 9: Results of segmentation at two images. (a) Points (left, depicted in gray) indicating locations where flow vectors are computed in the 5th image, and the two segments of the foreground and background obtained in the image. (b) The segment of the foreground in the 10th, 15th, and 25th image, respectively, obtained by removing the ambiguity.

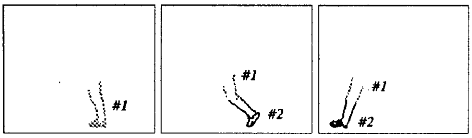


Fig. 10: Segments of the upper and lower parts of the man's left leg obtained in (from left to right) 5th, 30th, and 59th image, respectively.

two parts is approximated by a rigid plane. It can be seen from that the two parts are detected as a single segment in the 5th image (#1 in Fig.10(a)) due to the global similarity between their image motions. They have been separated to two segments in the 30th image (#1 and #2 in Fig.10(b)). In the 59th image, the segments of the two parts are still interpreted as individual planes (#1 and #2 in Fig.10(c)), even though their image motions become globally similar again.

VII. Conclusion

A distinctive feature of the proposed method is to accumulatively observe apparent motion in a long image sequence. By tracking patches between several images, flow vectors at locations in corresponding patches are combined to obtain a correct estimate of local image motion. Furthermore, the estimated local image motion vectors are spatially and temporally combined to estimate the model for describing the apparent motion of each object. When local image motion vectors of patches are correctly estimated, incremental segmentation of each image becomes possible.

In order to remove the inherent ambiguities, temporal coherence between the local image motion of a patch and the apparent motion of every object is investigated over long time. Each patch is grouped into a segment whose image motion is most temporally coherent with the local image motion observed at the patch. When apparent motions of two objects are not temporally coherent each other up to the previous image, segments of the two objects are not merged. In this way, segmentation become stable when the ambiguities absence during a certain number of images.

Acknowledgements

A part of this work was supported by the Grant-In-Aid for Scientific Research under Grant No.04452194 from the Ministry of Education, Science, and Culture, Japanese Government, and also by a grant for the research of technology development from the Broadcasting Culture Foundation of Japan.

References

[1] Chen, H. -J., Shirai, Y. and Asada, M., "Detecting multiple rigid image motions from an optical flow field obtained with multi-orientation, multi-scale filters," *IEICE Trans. Infor. & Syst.* **E76-D**, 1253-1262 (1993).

[2] Chen, H. -J. and Shirai, Y., "Incremental partitioning and tracking of moving objects in a long sequence," *Proc. ACCV'93*, Osaka, Nov. 1993; see also ——, "Segmentation based on accumulative observation of apparent motion in long sequences," *IEICE Trans. Infor. & Syst.* (submitted).

[3] Duda, R. O. and Hart, P. E., *Pattern Classification and Scene Analysis*, Chapter 2, NY: Wiley, 1973.

[4] 栄藤、白井, "色, 位置, 輝度勾配に基づく領域分割による2次元動き推定," 信学論 D-II (1993).

[5] Kanatani, K., "Structure and motion from optical flow under perspective projection," *CVGIP* **38**, 122-146 (1987).

[6] Nakayama, K., "Biological image motion processing: A review," *Vision Res.* **25**, 625-660 (1985).

[7] Singh, A., "Incremental estimation of image-flow using a Kalman filter," *Proc. IEEE Workshop on Visual Motion*, 36-43 (1991).

[8] Subbarao, M. and Waxman, A. M., "Closed form solutions to image flow equations for planar surfaces in motion," *CVGIP* **36**, 208-228 (1986).