Motion Segmentation Based on Factorization Method and Discriminant Criterion

Naoyuki ICHIMURA

Electrotechnical Laboratory
1-1-4, Umezono, Tsukuba
Ibaraki, 305-8568 Japan
ichimura@etl.go.jp

Abstract

A motion segmentation algorithm based on factorization method and discriminant criterion is proposed. This method uses a feature with the most useful similarities for grouping, selected using motion information calculated by factorization method and discriminant criterion. A group is extracted based on discriminant analysis for the selected feature's similarities. The same procedure is applied recursively to the remaining features to extract other groups. This grouping is robust against noise and outliers because features with no useful information are automatically rejected. Numerical computation is simple and stable. No prior knowledge is needed on the number of objects. Experimental results are shown for synthetic data and real image sequences.

1 Introduction

Segmentation is fundamental processing in computer vision. Motion segmentation has attracted great attention, because it can be used for applications such as modeling by shape from motion, video coding, and the analysis of movement.

Many algorithms have been proposed for motion segmentation based on Hough transformation [1], mixture model [2], random field model [3] [4], and epipolar constraints between views [5] [6] etc. Basically, these involve two concepts. One is optimization based on maximum likelihood method, e.g., the EM algorithm. The other is sampling and verification: some data are randomly sampled and then it is verified whether constraints for data are satisfied. The former requires appropriate initial guesses of motion information. The latter incorporates spatial proximity into random sampling to increase the possibility for extracting data that belong to the same group. These facts mean that information about "each group" is necessary in advance to obtain information used in segmentation – the chicken and egg problem.

To avoid this problem, the algorithm based on multiple epipolar constraints has been proposed [7]. The epipolar equation for multiple objects is constructed by the tensor product of the epipolar constraint of each object. The essential matrix of each object is calculated directly using the epipolar equation for multiple objects, and the motion and shape of each object are obtained from the essential matrix. Thus segmentation can be carried out without initial segmentation. This algorithm is difficult to apply, however, if the number of objects exceeds two.

Factorization method [8]-[11] has also been used to avoid the chicken and egg problem. A measurement matrix with the coordinates of feature correspondences as entities is factorized into two matrices only once. Initial segmentation is not needed, since these matrices contain motion parameters and 3D coordinates of features of multiple objects. This is easy to apply if the number of objects varies.

In factorization-based procedures, optimization based on the energy represented by the sum of entities of the shape interaction matrix [9] and bipartite graph [10] have been used for grouping noisy data. This optimization is needed to solve the combinatorial problem in grouping, but may involve local minima and high computational cost. The lack of robustness for noisy data is also a problem [11].

A motion segmentation algorithm using the similarity matrix on motion obtained from factorization method and discriminant criterion is proposed for grouping noisy data. A feature with the most useful similarities for grouping is selected based on a discriminant criterion. A group is extracted based on the result of discriminant analysis for the selected feature's similarities. The same procedure is applied recursively to remaining features to extract other groups. Since features with no useful information are rejected automatically, the method is robust against noise and outliers. No combinatorial problem occurs in grouping, since only one feature is used. The proposed method differs from the conventional in this systematic feature selection based on a discriminant criterion which is numerically stable and able to eliminate combinatorial problem. Additionally, no prior knowledge is needed on the number of objects because groups are extracted recursively.

Section 2 shows how to calculate the similarity matrix on motion, the so-called shape interaction matrix[9], using factorization. Section 3 describes the motion segmentation algorithm using feature selection based on the discriminant criterion. Section 4 shows experimental results of segmentation and weak calibration for synthetic data and real image sequences. Section 5 presents conclusions.

2 Similarity Matrix on Motion Obtained from Factorization Method

The features, e.g., points, in image sequence are tracked to obtain feature correspondence. Feature correspondences obtained from P features and F frames are collected in measurement matrix $\boldsymbol{W}$ $(2F \times P)$. Under the affine projection model, the measurement matrix is decomposed by SVD[11][12].

$$\underset{2F\times P}{\boldsymbol{W}} = \underset{2F\times r}{\boldsymbol{U}_r}\ \underset{r\times r}{\boldsymbol{\Sigma}_r}\ \underset{r\times P}{\boldsymbol{V}_r^t} \tag{1}$$

where $r = 1, \ldots, P$. If r is identical to the number of nonzero singular values, it is the rank of the measurement matrix.

The shape interaction matrix is defined as follows[9]:

$$\underset{P\times P}{\boldsymbol{X}} = \boldsymbol{V}_r \boldsymbol{V}_r^t = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_P)^t \tag{2}$$

If data are not contaminated by noise, entities x_{ij} of $\boldsymbol{X}$ have the following property:

$$x_{ij} \begin{cases} \neq 0, & \text{If features corresponding to the } i\text{-th row and } j\text{-th column belong to the same object.} \\ = 0, & \text{If features corresponding to the } i\text{-th row and } j\text{-th column belong to a different object.} \end{cases} \tag{3}$$

Since the shape interaction matrix has the above property, it is regarded as the similarity matrix of motions among objects.

The proof on the property of shape interaction matrix shown above has been given by Kanatani[13]. The summary of the proof is shown here. The key of the proof is the exchange of the basis of Ker$\boldsymbol{W}$. Ker$\boldsymbol{W}$ is spaned by orthonormal basis $\{\boldsymbol{v}_{r+1}, \ldots, \boldsymbol{v}_P\}$, where $\boldsymbol{v}_i$ $(i = 1, \ldots, P)$ are the columns of matrix $\boldsymbol{V}_P = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_P\}$ and r is the rank of $\boldsymbol{W}$.

Let us assume that segmentation of tracked features are given: for N objects, $P^1, \ldots, P^N \left(\sum_{i=1}^N P^i = P\right)$ features belong to each object and the dimensions of linear spaces V^i corresponding to objects spaned by the columns of V_r are $r^1, \ldots, r^N \left(\sum_{i=1}^N r^i = r\right)$. The dimension of null space of V^i is $\mu^i = P^i - r^i$ and one can construct basis of null space of V^i using following vector:

$$\tilde{\boldsymbol{n}}^i = \left(\mathbf{o}_1^t, (\boldsymbol{n}^i)^t, \mathbf{o}_2^t\right)^t \tag{4}$$

where $\mathbf{o}_1$ and $\mathbf{o}_2$ are the vectors with $\sum_{j=1}^{i-1} P^j$ and $\sum_{j=i+1}^{N} P^j$ zeros, and $\boldsymbol{n}^i$ is a P^i dimensional vector in the orthonormal basis of null space of V^i; $\tilde{\boldsymbol{n}}^i$ is constructed by adding $\left(P - P^i\right)$ zeros to $\boldsymbol{n}^i$. The set of the vectors $\left\{\tilde{\boldsymbol{n}}_1^1, \ldots, \tilde{\boldsymbol{n}}_{\mu^1}^1, \ldots, \tilde{\boldsymbol{n}}_1^N, \ldots, \tilde{\boldsymbol{n}}_{\mu^N}^N\right\}$ is the orthonormal basis of Ker$\boldsymbol{W}$ with $P - r\left(= \sum_{i=1}^N \mu^i\right)$ dimension.

The relationship between two orthonormal bases of Ker$\boldsymbol{W}$, $\{\boldsymbol{v}_{r+1}, \ldots, \boldsymbol{v}_P\}$ and $\left\{\tilde{\boldsymbol{n}}_1^1, \ldots, \tilde{\boldsymbol{n}}_{\mu^1}^1, \ldots, \tilde{\boldsymbol{n}}_1^N, \ldots, \tilde{\boldsymbol{n}}_{\mu^N}^N\right\}$, can be represented as follows:

$$\{\boldsymbol{v}_{r+1}, \ldots, \boldsymbol{v}_P\} = \left\{\tilde{\boldsymbol{n}}_1^1, \ldots, \tilde{\boldsymbol{n}}_{\mu^1}^1, \ldots, \tilde{\boldsymbol{n}}_1^N, \ldots, \tilde{\boldsymbol{n}}_{\mu^N}^N\right\} \boldsymbol{C} \tag{5}$$

where $\boldsymbol{C}$ is the exchange matrix of two bases. Since both bases are orthonormal, $\boldsymbol{C}$ is orthogonal matrix.

The correlation matrix of $\{\boldsymbol{v}_{r+1}, \ldots, \boldsymbol{v}_P\}$ is the block diagonal matrix.

$$\boldsymbol{X}' = \{\boldsymbol{v}_{r+1}, \ldots, \boldsymbol{v}_P\}\{\boldsymbol{v}_{r+1}, \ldots, \boldsymbol{v}_P\}^t = \mathrm{diag}(\boldsymbol{D}^1, \ldots, \boldsymbol{D}^N) \tag{6}$$

Above result is derived from Eq.(5) and the orthogonality of exchange matrix $\boldsymbol{C}$. Each block matrix $\boldsymbol{D}^i$ has the size $P^i \times P^i$.

The correlation matrix of $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_P\}$ is as follows:

$$\boldsymbol{V}_P \boldsymbol{V}_P^t = \boldsymbol{X} + \boldsymbol{X}' = \boldsymbol{I}_P \tag{7}$$

where $\boldsymbol{I}_P$ is the $P \times P$ identity matrix. This shows that shape interaction matrix $\boldsymbol{X}$ is block diagonal. This result is valid under arbitrary permutation of the basis of Ker$\boldsymbol{W}$. Therefore shape interaction matrix has the property shown in Eq.(3).

The property, however, is not exactly satisfied for data with noise and outliers. The lack of robustness for such data is a problem for the method using shape interaction matrix[11].

3 Motion Segmentation Based on Discriminant Criterion

3.1 Algorithm

All entities of a matrix that show similarities among motions are used in conventional methods [9][10]. On the other hand, only a feature with the most useful information for grouping is selected in the present method. This feature selection reduces the effect of noise and outliers automatically, and eliminates the combinatorial problem in grouping.

The usefulness of feature selection can be explained using the example shown in Fig.1. A group containing features with similar motion is extracted if changes in similarities in the row of matrix Eq.(2) are as shown in Fig.1(b) after sorting; features with large similarity are detected clearly. Only part of the features, however, may have information useful for grouping. For example, Fig.1 (b) and (c) show the change in distance from data A and B (Fig.1(a)) to others, and Fig.1(c) has no useful information for grouping, while Fig.1(b) does. Data such as B, i.e. outlier, is normally observed in real image sequences. Thus feature selection is introduced to extract a feature with the most useful information for grouping, and to reject features with no useful information.

Given r in Eq.(1), similarity matrix $\boldsymbol{X}$ is computed. Entities of each row $\boldsymbol{x}_k$ of $\boldsymbol{X}$ $(k = 1, \ldots, P)$ are then

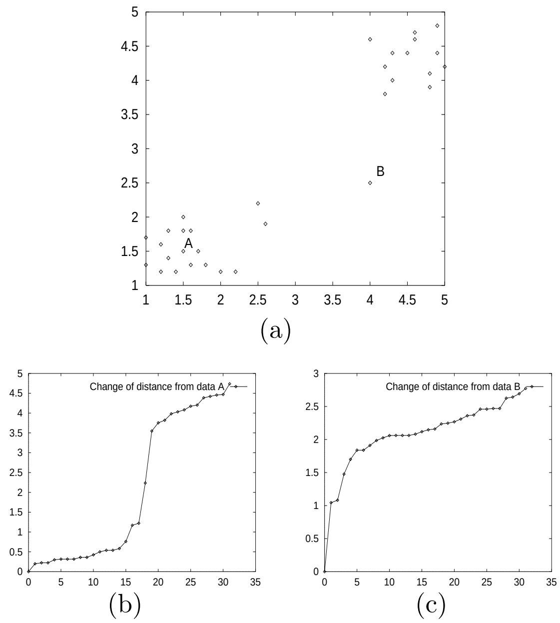

Figure 1: Typical example of data with information useful for grouping (data A) and without (data B).

sorted. The following discriminant criterion[14] is used to separate entities of row vector $\boldsymbol{x}_k$ into two groups:

$$\begin{aligned} \lambda &= \frac{\sigma_B^2}{\sigma_W^2} && (8) \\ \sigma_B^2 &= N^1 N^2 \left(\bar{\varepsilon_1} - \bar{\varepsilon_2}\right)^2 && (9) \\ \sigma_W^2 &= N^1 \sigma_1^2 + N^2 \sigma_2^2 && (10) \end{aligned}$$

where $\bar{\varepsilon_1}$, $\bar{\varepsilon_2}$, σ_1^2, σ_2^2, N^1, and N^2 are the mean, variance, and the number of entities of each group, σ_B^2 and σ_W^2 are the variance between groups and one within groups. To find the grouping that maximizes λ, $x_{kl}\,(l = 1, \ldots, P)$ are used as a threshold for entities of $\boldsymbol{x}_k = \{x_{k1}, x_{k2}, \ldots, x_{kP}\}$.

Let λ_k be the maximum of the discriminant criterion for the k-th row vector. The feature with the most useful information for grouping is selected as follows:

$$k_{select} = \arg\max_k \lambda_k \qquad (11)$$

A group is extracted using the threshold of similarities of row vector $\boldsymbol{x}_{k_{select}}$ maximizing the discriminant criterion.

The same procedure is applied recursively to remaining features to extract other groups. If the mean and discriminant criterion of similarities of the selected row is smaller than given thresholds, the procedure stops. Remaining features are regarded as a group because no useful information is found.

The meaning of the proposed algorithm is interpreted using the concept of the orthogonal projection matrix. Matrix $\boldsymbol{X}$ of Eq.(2) is computed from orthogonal vectors in matrix $\boldsymbol{V}_r$. Thus matrix $\boldsymbol{X}$ is the orthogonal projection matrix for the subspace spaned by r orthogonal vectors. This subspace is constructed by N subspaces corresponding to N objects in the scene, and thus conventional algorithms attempt to decompose this subspace using all similarities [8]- [11]. The present method extracts only one axis in $\boldsymbol{R}^P$ space of P tracked features, maximizing separation among the projections to each subspace corresponding to each object. This procedure coincides with that maximizing separation among subspaces, because the orthogonal projection matrix has a one-to-one correspondence to subspaces.

3.2 Determination of Parameter r

In Section 3.1, parameter r in Eq.(1) is the rank of the measurement matrix if it can be estimated correctly. The rank, however, is difficult to estimate for a real situation. In this research, parameter r is determined from the view on separation. The algorithm shown in Section 3.1 is applied using parameter r in some range, e.g., [2:10]. The segmentation result maximizing the following equation, sum of $\lambda_{k_{select}}$ of extracted g groups, is used:

$$\lambda\left(r\right) = \sum_{i=1}^{g} \lambda_{k_{select}}\left(i, r\right) \qquad (12)$$

4 Experimental Results

The experiments using synthetic data and real image sequences are shown to confirm the usefulness of the proposed method. For real image sequences, the calculation of epipolar constraint for each motion in a scene is also presented to show the validity of segmentation results.

4.1 Synthetic Data

Data included two objects − object 1 constructed of curves and object 2 constructed of lines (Fig.2 (a)). Rotations of objects were $(\psi_1, \theta_1, \phi_1) = (0.5, 0.5, 0.5)$ [deg] and $(\psi_2, \theta_2, \phi_2) = (1.0, 0.0, -1.0)$ [deg], where ψ, θ, ϕ are roll, pitch, and yaw. Translation vectors of objects were $(1.0, 1.0, 1.0)$ [mm] and $(0.0, -1.5, -1.5)$ [mm].

To simulate the effect of noise in images and false matches, Gaussian noise with standard deviation 5.0 was added only to data after motion (Fig.2 (a)). Data were projected on to a 2D plane using the perspective projection matrix of a lens with 8[mm] focal length. The distance between the camera and object was 2[m].

The thresholds to stop recursive procedure of segmentation were 1.5×10^{-3} and 6.0 for mean of similarities and discriminant criterion. The result for $r = 3$ was used in accordance with the changes in the value of Eq.(12) (Fig.2 (b)). The feature with a maximum discriminant criterion had useful information for grouping under fairly large noise (Fig.2 (c)), while one with a minimum discriminant criterion had no useful information (Fig.2 (d)). The two objects were segmented correctly using the information of selected features.

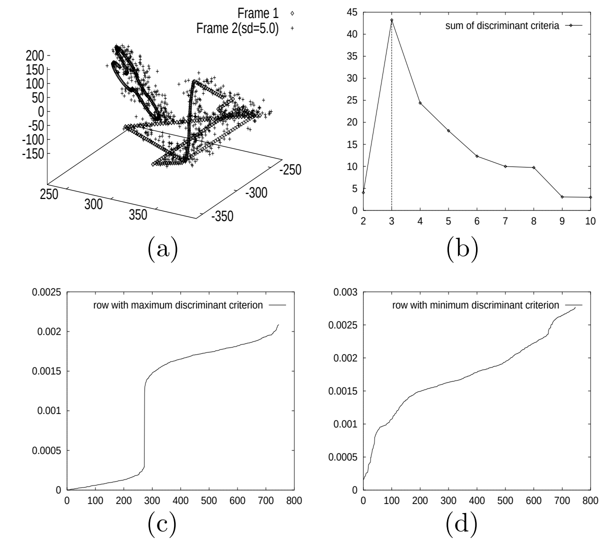

Figure 2: Segmentation result of synthetic data. (a) Data with Gaussian noise with 5.0 standard deviation. (b) Change in the sum of discriminant criteria as a function of r. (c) Change in entities of a row with maximum discriminant criterion. (d) Change in entities of a row with minimum discriminant criterion.

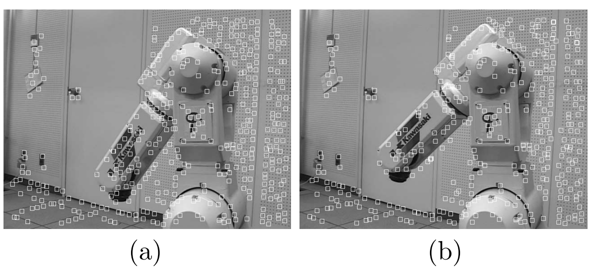

Figure 3: Robot arm image sequence. (a) Frame 1. (b) Frame 60.

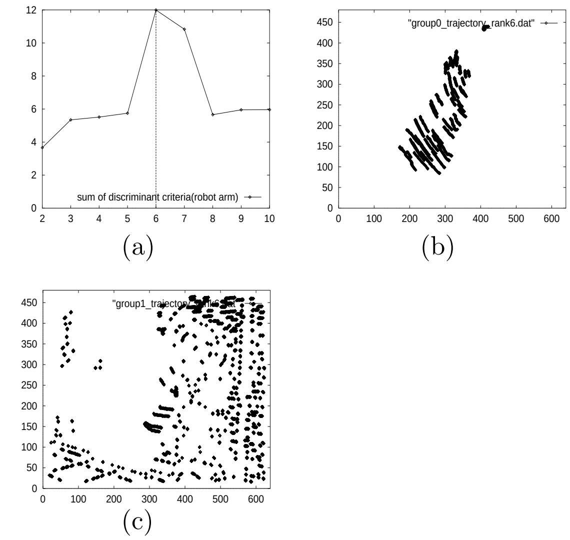

Figure 4: Segmentation result for a robot arm image sequence. (a) Change in sum of discriminant criteria as a function of r. (b),(c) Trajectories of motion for each group.

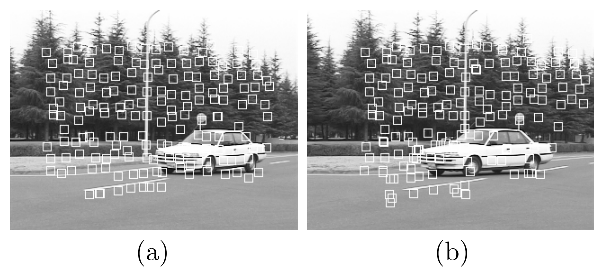

Figure 5: Car image sequence. (a) Frame 1. (b) Frame 10.

4.2 Real Image Sequences

The results for three image sequences containing a robot arm, car, and human being are shown. The feature points in images were extracted using the corner detector proposed in [15]. Detected features are tracked by block matching. Features with large matching error due to occlusion were removed in tracking, but no other processing was used to remove noise and outliers. Thresholds to stop recursive procedures of segmentation were 1.5×10^{-3} and 3.0.

4.2.1 Segmentation Results

For the robot arm sequence (Fig.3), 60 frames were used. The arm was rotated and the background was still. The result for $r = 6$ was used (Fig.4 (a)). The two groups extracted mostly corresponded to the arm and background (Fig.4 (b), (c)).

For the car sequence (Fig.5), 10 frames were used. Both the camera and car were moved; the background was not still for this sequence. The result for $r = 6$ was used (Fig.6 (a)). The two groups extracted mostly corresponded to the car and background (Fig.6 (b), (c)).

For the human sequence (Fig.7), 60 frames were used. The result for $r = 4$ was used (Fig.8 (a)). The three groups extracted mostly corresponded to the left hand, face and shoulders, and background (Fig.8 (b)-(d)).

Although the data used in these experiments had many errors due to simple block matching and homogeneous regions such as the white wall in the room, the proposed method gave adequate results.

4.2.2 Calculation of Epipolar Constraint of Each Group

Using segmentation results, we calculated the epipolar constraint for each group. Factorization method can be used to calculate epipolar constraints.

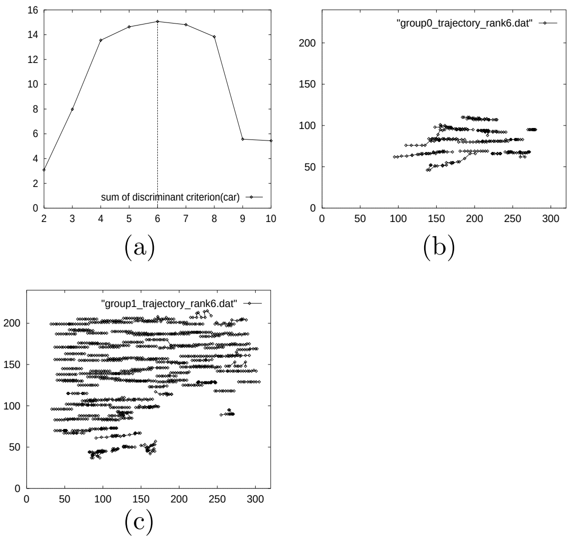

Figure 6: Segmentation results for a car image sequence. (a) Change in sum of discriminant criteria as a function of r. (b),(c) Trajectories of motion for each group.

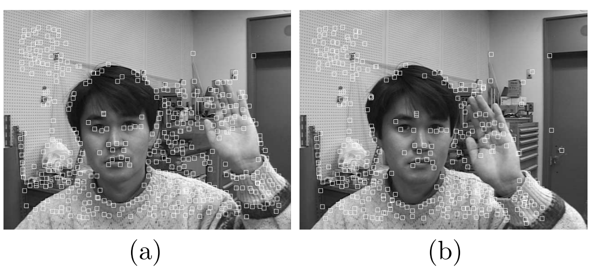

Figure 7: Human image sequence. (a) Frame 1. (b) Frame 60.

The use of factorization method under perspective projection is still difficult, however, particularly for noisy data, although some extensions have been considered [16]-[18]. A two-views algorithm was therefore used: the fundamental matrix for each group was calculated using the first and the last frames.

The algorithm proposed in [19] is used to calculate the fundamental matrix. Algorithms for the initial guess of epipole and outlier rejection were needed to obtain an adequate solution. An affine epipolar geometry algorithm [20] and outlier rejection based on the eigen value perturbation theory [21] were used for these purposes.

The calculated epipolar constraints clearly reflected the motion of each group (Figs.9-11), e.g., rotation of a robot arm. The epipolar constraints for the second group of robot arm image sequence and the third group of human image sequence (Fig.4 (c), Fig.8 (d)) could not be calculated, since these groups corresponded to a still background. This could be detected from entities of the fundamental matrix.

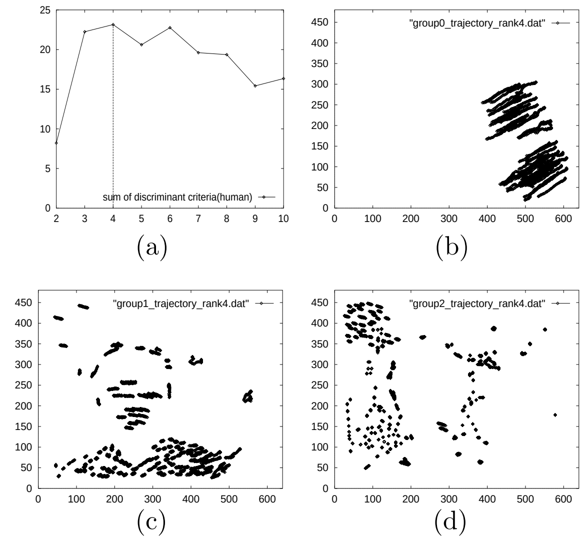

Figure 8: Segmentation result for a human image sequence. (a) Change in sum of discriminant criteria as a function of r. (b)-(d) Trajectories of motion for each group.

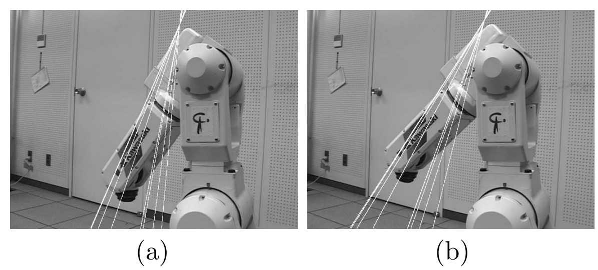

Figure 9: Epipolar constraints for a robot arm image sequence. (a),(b) Epipolar constraints for the first group. Epipolar constraints for the second group could not be calculated because no motion appeared.

Images containing multiple motions could thus be weakly calibrated using the segmentation results of the proposed method. The calibration results can be utilized to refine segmentation results and calculate 3D map of the scene using dence matching.

5 Conclusions

A motion segmentation algorithm based on factorization method and discriminant criterion features:
• Simultaneous calculation of motions based on factorization method.
• Robustness against noise and outliers due to feature selection based on discriminant criterion.
• Simple, stable numerical computation.
• No need for prior knowledge on the number of objects.
Experimental results using real image sequences demonstrated the usefulness of the proposed algorithm in practical situations.

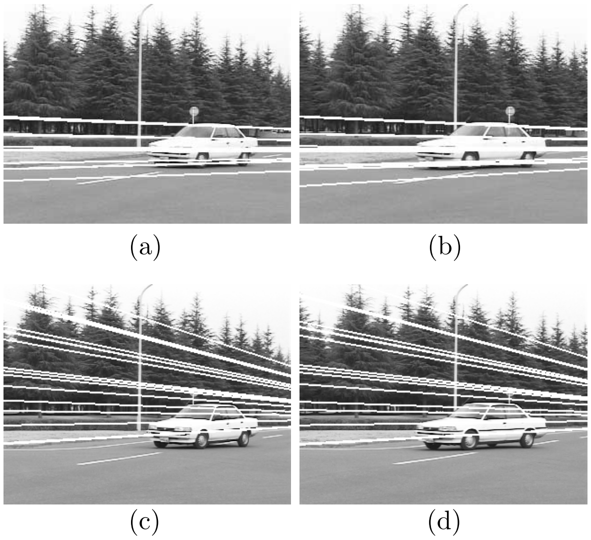

Figure 10: Epipolar constraints for a car image sequence. (a),(b) Epipolar constraints for the first group. (c),(d) Epipolar constraints for the second group.

The proposed method requires approximation of a camera model because it is based on factorization method under affine projection. Although factorization method under perspective projection has been proposed [16]-[18], all algorithms use a single motion assumption, preventing them from being used directly for multiple motions – problem for important future work.

Acknowledgements

The author would like to thank Dr. Nobuyuki Otsu, the director of the Electrotechnical Laboratory, for his encouragement.

References

[1] G. Adiv: "Determining three-dimensional motion and structure from optical flow generated by several moving objects," IEEE Trans. Pattern Anal. & Mach. Intell., Vol.7, No.4, pp.384-401, 1985

[2] Y. Weiss and E. H. Adelson: "A unified mixture framework for motion segmentation: Incorporating spatial coherence and estimating the number of models," Proc. CVPR, pp.321-326l, 1996

[3] D. W. Murray and B. F. Buxton: "Scene segmentation from visual motion using global optimization," IEEE Trans. Pattern Anal. & Mach. Intell., Vol.9, No.2, pp.220-228, 1987

[4] J. Konrad and E. Dubois: "Bayesian estimation of motion vector fields," IEEE Trans. Pattern Anal. & Mach. Intell., Vol.14, No.9, pp.910-927, 1992

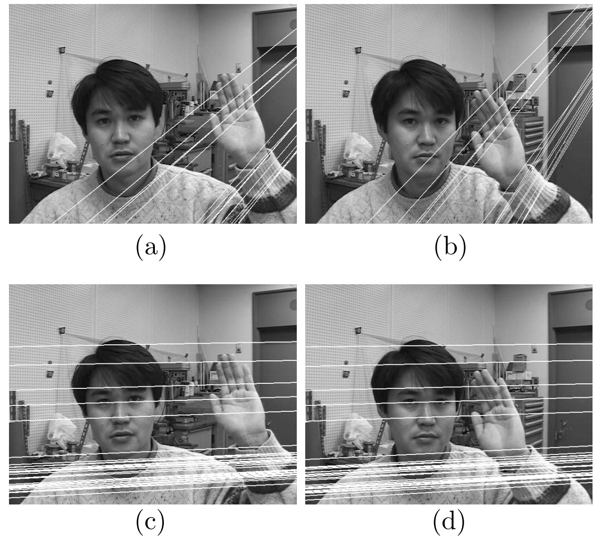

Figure 11: Epipolar constraints for a human image sequence. (a),(b) Epipolar constraints for the first group. (c),(d) Epipolar constraints for the second group. Epipolar constraints for the third group could not be calculated because no motion appeared.

[5] E. Nishimura, G. Xu and S. Tsuji: "Motion segmentation and correspondence using epipolar constraint," Proc. ACCV, pp.199-204, 1993

[6] P. H. S. Torr: "Geometric motion segmentation and model selection," Phil. Trans. R. Soc. Lond. A, Vol.356, pp.1321-1340, 1998

[7] M. Shizawa: "Transparent 3D motions and structures from point correspondences in two frames: A quasi-optimal, closed-form, linear algorithm and degeneracy analysis," Proc. ACCV, pp.329-334, 1993

[8] T. E. Boult and L. G. Brown: "Factorization-based segmentation of motions," Proc. IEEE Workshop on Vis. Mot., pp.179-186, 1991

[9] J. P. Costeira and T. Kanade: "A multi-body factorization method for independently moving objects," Internat. J. Comp. Vis., 29, 3, pp.159-179, 1998

[10] C. W. Gear: "Multibody grouping from motion images," Internat. J. of Comp. Vis., 29, 2, pp.133-150, 1998

[11] T. Kanade and D. D. Morris: "Factorization methods for structure from motion," Phil. Trans. R. Soc. Lond. A, 356, pp.1153-1173, 1998

[12] C. Tomasi and T.Kanade: "Shape and motion from image streams under orthography: a factorization method," Internat. J. Comp. Vis., 9, 2, pp.137-154, 1992

[13] K. Kanatani: "Factorization without factorization: multibody segmentation," Technical Report of IEICE, PRMU98-117, 1998 (in Japanese)

[14] N. Otsu: "A threshold selection method from gray-level histograms," IEEE Trans. Sys., Man, and Cybern., Vol.SMC-9, No.1, pp.62-66, 1979

[15] M. Trajkovic and M. Hedley: "Fast corner detection," Image and Vision Computing, 16, pp.75-87, 1998

[16] P. Sturn and B. Triggs: "A factorization based algorithm for multi-image projective structure and motion," Proc. ECCV, Vol.II, pp.709-720, 1996

[17] K. Deguchi: "Factorization method for structure from perspective multi-view images," IEICE Trans. Inf. & Syst., Vol.E81-D, No.11, pp.1281-1289, 1998

[18] T. Ueshiba and F. Tomita: "A factorization method for projective and Euclidean reconstruction from multiple perspective views via iterative depth estimation," Proc. ECCV, Vol.I, pp.296-310, 1998

[19] R. I. Hartley: "Minimizing algebraic error," Phil. Trans. R. Soc. Lond. A, Vol.356, pp.1175-1192, 1998

[20] L. S. Shapiro, A. Zisserman and M. Brady: "3D motion recovery via affine epipolar geometry," Internat. J. Comp. Vis., 16, pp.147-182, 1995

[21] L. S. Shapiro: "Affine analysis of image sequences," Cambridge University Press, 1995