Apply Autoassociative Learning to Recover the Motion and the Shape from Sequences of Scaled Orthographic Images

JUN FUJIKI[†], TAKASHI TAKAHASI[††] and TAKIO KURITA[†]
[†] Neuroscience Institute,
National Institute of Advanced Industrial Science and Technology,
Tsukuba-Central 2, 1-1-1 Umezono, Tsukuba-shi, Ibaraki 305-0035,
[††] Faculty of Science and Technology,
Ryukoku University,
Otsu-shi, Siga 520-2194,
JAPAN

Abstract: - It is well known that the feature of hidden layer of three-layered perceptron under autoassociative learning of some data is equivalent to the principal component analysis of the data. When we consider the autoassociative learning, the representation of the hidden layer is not so important and do not paid attention. However, when we consider the structure from motion problem under scaled orthographic projection, the hidden layer of three-layered perceptron is significant because the hidden layers represent the affine transformation of the feature points of the object. To fix the affine ambiguity of the hidden layer, we present an autoassociative learning of three-layered perceptron of its connecting coefficients constrained. After the present learning, the hidden layers represent the Euclidean coordinates of the feature points of the object. The present algorithm works well even if half of data are missing. We evaluate the ability of the algorithm through experiment with synthesize data.

Key-Words: - autoassociative learning, hidden layer, constraint, structure from motion, scaled orthographic projection, Euclidean reconstruction, missing data, occlusion

1 Introduction

It is well known that the feature of hidden layer of three-layered perceptron (3-MLP) under autoassociative learning of some data is equivalent to the principal component analysis (PCA) of the data[1]. When we consider the autoassociative learning, the representation of the hidden layer is not so important and do not paid attention. However, when we consider the structure from motion problem under scaled orthographic projection (which is a kind of affine approximation of the perspective projection), the hidden layer of three-layered perceptron is significant because the hidden layers are closely related to the three-dimensional coordinates of the feature points of the object, that is, the hidden layers represent the affine transformation of the feature points of the object. To fix the affine ambiguity of the hidden layer, we give the constraints on the three-layered perceptron. After autoassociative learning with constraints is finished, there is no ambiguity on the hidden layers and they represent the Euclidean coordinates of the feature points of the object.

From the view point of the computer vision, recovering the camera motion and the object shape from multiple images with point correspondences is the fundamental and important problem and many researches are investigated. Perspective camera model is suitable for this recovering problem because perspective camera represents pin-hole camera theoretically. However, pin-hole camera model derives non-linear inverse problem which has noise-sensibility and unstablity for numerical computation. Hence, affine approximation models such as scaled orthographic model are pre-

sented for camera model. Although affine approximation models has the limitation of accuracy for reconstruction, these models derives stable reconstruction and speedy calculation on the contrary to the pin-hole camera model, because these models are consist of linear inverse problem. The reconstruction under affine approximation model is also used for the initial value for the algorithm under pin-hole camera model. Then the study of affine approximation camera is important to improve the recovering problem. Therefore, many algorithm under affine approximation model are presented. However, these algorithms, which include the famous and excellent method named the factorization method[6], are not sufficient to overcome the lack of data caused by occlusion and/or failure of tracking points.

In this paper, we present the new method to recover the motion and the shape of a object from sequences of scaled orthographic images, which has robustness against missing data. The main structure of present method is the autoassociative learning by a 3-MLP, and we introduce the constraints on connecting coefficients of the 3-MLP. The autoassociative learning can estimates the missing data and it makes the present method robust against missing data. The key idea of the method is the singular value decomposition which appears in the factorization method is closely related to the PCA which appears in the autoassociative learning by 3-MLP, then the present method is the implementation of the factorization method for scaled orthographic model to a 3-MLP.

2 Scaled orthographic projection

Scaled orthographic projection is the orthographic projection considering the distance between camera center and the object, that is, composition of the orthographic projection and the extension of which rate is determined by the distance between camera center and the object.

In the context of the factorization method, we can set the object is stable and only the camera is moving without loss of generality because the images are determined only the relative position between camera and object.

Let $\{\boldsymbol{i}_f, \boldsymbol{j}_f\}$ be the orthonormal basis on the f-th image plane, $\boldsymbol{k}_f$ be the unit vector along optical axis, $C_f = (\boldsymbol{i}_f, \boldsymbol{j}_f, \boldsymbol{k}_f)^{\mathrm{T}}$ be the camera basis matrix and $\boldsymbol{s}_p$ be the world coordinate of the p-th feature point. We also define $\boldsymbol{X}_{fp} = (X_{fp}, Y_{fp}, Z_{fp})^{\mathrm{T}}$ and $\boldsymbol{x}_{fp} = (x_{fp}, y_{fp})^{\mathrm{T}}$ as the camera coordinate and the image coordinate of the p-th feature point on the f-th image plane, respectively. When considering the scaled orthographic projection, using the relative coordinate from the some feature point named $*$-th feature point (or center-of-mass of the object) is convinent. By using the relative coordinate $\boldsymbol{s}^*_p = \boldsymbol{s}_p - \boldsymbol{s}_*$, $\boldsymbol{X}^*_{fp} = \boldsymbol{X}_{fp} - \boldsymbol{X}_{f*}$ and $\boldsymbol{x}^*_{fp} = \boldsymbol{x}_{fp} - \boldsymbol{x}_{f*}$, there holds $\boldsymbol{X}^*_{fp} = C_f \boldsymbol{s}^*_p$ (see figure 1), and the representaion of the scaled orthographic projection is

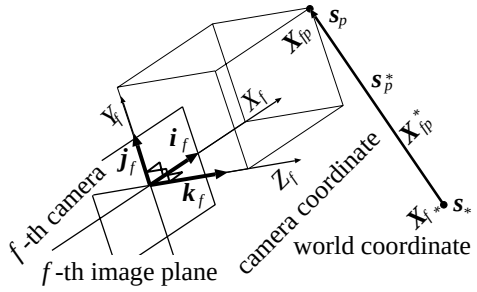

Figure 1. Camera coordinate and world coordinate.

$$\begin{aligned} \boldsymbol{x}^*_{fp} &= \frac{l}{Z_{f*}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \boldsymbol{X}^*_{fp} \\ &= \frac{l}{Z_{f*}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} C_f \boldsymbol{s}^*_p \end{aligned}$$

where Z_{f*} is the distance between camera center and the object (the $*$-th feature point) as shown in figure 2. In the projection, we cannot determine

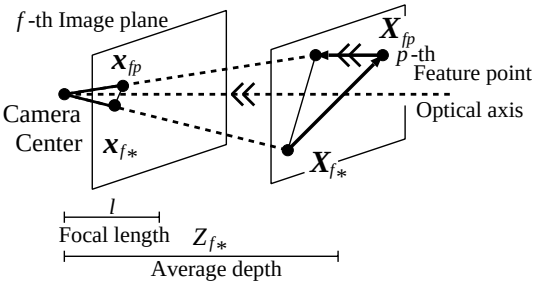

Figure 2. Scaled orthographic projection.

the value of $\{Z_{f*}\}_{f=1}^{F}$ but the ratio of $\{Z_{f*}\}_{f=1}^{F}$ because twice distance of twice sized object derives same image, for example. The ambiguity is parametrized as the ratio called global scale parameter.

To fix the ambiguity, the depth parameters which include global scale parameter are defined as $\lambda_{f*} = Z_{f*}/Z_{1*}$ (note that $\lambda_{1*} = 1$). By using the new depth parameter λ_{f*}, the new representation of the scaled orthographic projection is as follows:

$$\begin{aligned} \boldsymbol{x}_{fp}^{*} &= \frac{1}{\lambda_{f*}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} C_f \boldsymbol{s}_p^{*} \\ &= \lambda_{f*}^{-1} \begin{pmatrix} \boldsymbol{i}_f^{\mathrm{T}} \\ \boldsymbol{j}_f^{\mathrm{T}} \end{pmatrix} . \boldsymbol{s}_p^{*} \end{aligned} \tag{1}$$

Note that $\{\lambda_{f*}\}_{f=1}^{F}$ is proportional to the distance between camera center and the object, and fix the ambiguity comes from global scale parameter as $\lambda_{1*} = 1$.

Let measurement matrix W^*, motion matrix M and shape matrix S^* be defined as

$$\begin{aligned} W^{*} &= \begin{pmatrix} \boldsymbol{x}_{11}^{*} & \cdots & \boldsymbol{x}_{1P}^{*} \\ \vdots & \ddots & \vdots \\ \boldsymbol{x}_{F1}^{*} & \cdots & \boldsymbol{x}_{FP}^{*} \end{pmatrix}, \quad M = \begin{pmatrix} M_1 \\ \vdots \\ M_F \end{pmatrix} \\ M_f &= \begin{pmatrix} \boldsymbol{m}_f^{\mathrm{T}} \\ \boldsymbol{n}_f^{\mathrm{T}} \end{pmatrix} = \lambda_{f*}^{-1} \begin{pmatrix} \boldsymbol{i}_f^{\mathrm{T}} \\ \boldsymbol{j}_f^{\mathrm{T}} \end{pmatrix}, \\ S^{*} &= (\boldsymbol{s}_1^{*}, \ldots, \boldsymbol{s}_P^{*}), \end{aligned}$$

there holds $W^* = M_{(2F\times 3)}\, {S^*}_{(3\times P)}$. (Note that $\operatorname{rank} W^* \leq 3$).

We can easily compute

$$\begin{aligned} C_f &= (\boldsymbol{i}_f, \boldsymbol{j}_f, \boldsymbol{i}_f \times \boldsymbol{j}_f)^{\mathrm{T}}, \\ \lambda_{f*} &= ||\boldsymbol{m}_f||^{-1} = ||\boldsymbol{n}_f||^{-1} \end{aligned}$$

from $M_f = \lambda_{f*}^{-1}(\boldsymbol{i}_f, \boldsymbol{j}_f)^{\mathrm{T}}$, then the decomposition of W^* into MS^* attains the recover of the camera motion and the object shape. However, the decomposition of W^* into MS^* is not unique because the decomposition of $W^* = \widehat{M}_{(2F\times 3)}\, \widehat{S}^*_{(3\times P)}$ derives another decomposition $(\widehat{M}A)(A^{-1}\widehat{S}^*)$ where A is arbitraly 3×3 invertible matrix. Hence, $\widehat{M}$, $\widehat{S}^*$ are only the affine reconstruction. To upgrade the affine reconstruction to Euclidean reconstruction, the matrix A should be computed to satisfy

$$M_f M_f^{\mathrm{T}} = \lambda_{f*}^{-2}\, \mathrm{I}_2 \Longleftrightarrow \begin{cases} \boldsymbol{m}_f^{\mathrm{T}} \boldsymbol{m}_f - \boldsymbol{n}_f^{\mathrm{T}} \boldsymbol{n}_f = 0, \\ \boldsymbol{m}_f^{\mathrm{T}} \boldsymbol{n}_f = 0, \\ \boldsymbol{m}_1^{\mathrm{T}} \boldsymbol{m}_1 = 1 \end{cases} \tag{2}$$

($f = 1, \ldots, F$) which are named metric constraints. Note that λ_{f*} is computed by $\lambda_{f*} = ||\boldsymbol{m}_f||^{-1} = ||\boldsymbol{n}_f||^{-1}$.

After computed the matrix A, the Euclidean reconstruction is derived. This is the procedure of the factorization method.

Note that a pair of reconstructions are derived from affine approximation images under point corresponcences, and the pair is mutually reflection called Necker reversal. It is well known that we cannot chose the pair of which is true reconstruction only from point correspondences. Hence, we identify the pair reconstructions.

3 Autoassociative Learning

It is known that the feature of hidden layer of 3-MLP under autoassociative learning of some data is equivalent to the principal component analysis of the data[1]. In this section, we explain the autoassociative learning by a 3-MLP and that with constraints on its connecting coefficients.

Let us consider a 3-MLP which has N units in both the input and the output layers, respectively, and $H(< N)$ units in the hidden layer.

Let $\{\boldsymbol{x}_p = (x_{p1}, \ldots, x_{pN})^{\mathrm{T}} \in \boldsymbol{R}^N\}_{p=1}^{P}$ be a given input vector, $\boldsymbol{y}_p$ be a hidden layer vector associated with $\boldsymbol{x}_p$, and $\widehat{\boldsymbol{x}}_p = (\widehat{x}_{p1}, \ldots, \widehat{x}_{pN})^{\mathrm{T}} \in \boldsymbol{R}^N$ be a output vector associated with $\boldsymbol{x}_p$. When all units have a linear activation function, there exist the matrices U and W to represent the connection between input layer and hidden layer, and between hidden layer and output layer, respectively. Then there hold

$$\begin{aligned} \boldsymbol{y}_p &= \underset{(H\times N)}{U}\, \boldsymbol{x}_p, \\ \widehat{\boldsymbol{x}}_p &= \underset{(N\times H)}{W}\, \boldsymbol{y}_p = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_N)^{\mathrm{T}} \boldsymbol{y}_p. \end{aligned}$$

The autoassociative learning is to approximate each input data by using its output. Then the learn-

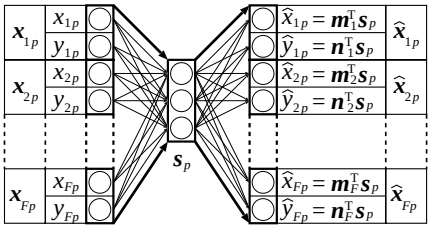

Figure 3. Factorization and autoassociative learning.

ing is realized by minimizing

$$E = \frac{1}{2}\varepsilon^2 = \frac{1}{2}\sum_{p=1}^{P} \|\widehat{\boldsymbol{x}}_p - \boldsymbol{x}_p\|^2.$$

Therefore, the learning rules are given as

$$\Delta \boldsymbol{w}_n = -\alpha \sum_p \Delta x_{pn} \boldsymbol{y}_p,$$
$$\Delta U = -\alpha \sum_{p,n} \Delta x_{pn} \boldsymbol{w}_n \boldsymbol{x}_p^{\mathrm{T}}$$

where $\Delta x_{pn} = \widehat{x}_{pn} - x_{pn}$ and α is the learning rate.

Note that a 3-MLP of connection $(A^{-1}U, WA, A^{-1}\boldsymbol{y}_p)$ is the same performance as that of connection $(U, W, \boldsymbol{y}_p)$ for arbitrary invertible $H \times H$ matrix A. Hence, the hidden layer vector under an autoassociative learning of $\boldsymbol{\chi}_p^* = (\boldsymbol{x}_{1p}^{\mathrm{T}}, \ldots, \boldsymbol{x}_{Fp}^{\mathrm{T}})^{\mathrm{T}} = (\chi_{p1}, \ldots, \chi_{p,2F})^{\mathrm{T}}$ is not $\boldsymbol{s}_p^*$ itself but the affine transformation of $\boldsymbol{s}_p^*$ (see figure 3).

To Determine the matrix A to represent the hiddel layer vector as $\boldsymbol{s}_p^*$, that is, to upgrade the affine reconstruction into the Euclidean reconstruction, we introduce the constraints

$$E_1 = \frac{1}{2}\sum_{f=1}^{F} \|\boldsymbol{m}_f^{\mathrm{T}}\boldsymbol{m}_f - \boldsymbol{n}_f^{\mathrm{T}}\boldsymbol{n}_f\|^2$$
$$E_2 = \frac{1}{2}\sum_{f=1}^{F} \|\boldsymbol{m}_f^{\mathrm{T}}\boldsymbol{n}_f\|^2$$

to a 3-MLP, which is equivalent to the equation (2). This autoassociative learning with constraints is what we presents in this paper. The presented learning is realized by minimizing

$$E = \frac{1}{2}\varepsilon^2 + \beta\,(E_1 + E_2)$$

where β is a weight. In this case, the learning rules are given as

$$\begin{aligned}\Delta \boldsymbol{m}_f &= -\alpha \sum_p \Delta\chi_{p,2f-1}\boldsymbol{s}_p \\ &\quad - \alpha\beta\{(\boldsymbol{m}_f^{\mathrm{T}}\boldsymbol{m}_f - \boldsymbol{n}_f^{\mathrm{T}}\boldsymbol{n}_f)\boldsymbol{m}_f \\ &\quad\quad + (\boldsymbol{m}_f^{\mathrm{T}}\boldsymbol{n}_f)\boldsymbol{n}_f\} \\ \Delta \boldsymbol{n}_f &= -\alpha \sum_p \Delta\chi_{p,2f}\boldsymbol{s}_p \\ &\quad - \alpha\beta\{(\boldsymbol{n}_f^{\mathrm{T}}\boldsymbol{n}_f - \boldsymbol{m}_f^{\mathrm{T}}\boldsymbol{m}_f)\boldsymbol{n}_f \\ &\quad\quad + (\boldsymbol{m}_f^{\mathrm{T}}\boldsymbol{n}_f)\boldsymbol{m}_f\}.\end{aligned}$$

where $\Delta\chi_{pn} = \widehat{\chi}_{pn} - \chi_{pn}$.

Note that $\beta = 0$ stands for the learning without connecting coefficients constrained.

To cope with missing data such as occlusions and tracking errors in real image, we replace the conventional squared error ε^2 with the following weighted error.

$$\tilde{\varepsilon}^2 = \sum_{p=1}^{P}\sum_{f=1}^{F} \mu_{fp}\|\widehat{\boldsymbol{x}}_{fp} - \boldsymbol{x}_{fp}\|^2$$

where μ_{fp} is a constant in $\{0, 1\}$. If $\boldsymbol{x}_{fp}$ is a missing data, μ_{fp} is set to 0, otherwise it is set to 1. Then the learning rules are modified so as to minimize

$$E = \frac{1}{2}\tilde{\varepsilon}^2 + \beta\,(E_1 + E_2)\,.$$

After learning converged, we can estimate the missing data as the output vector of 3-MLP.

The procedure of the estimation of missing data is as follows:

(0) Initialize missing data as zero.

(1) Initialize connecting coefficients of 3-MLP.

(2) Autoassociative learning to estimate of missing data.

(3) Initialize missing data as estimate **(2)** and go to **(1)**.

4 Experiment

In this section, we evaluate the presented algorithm by synthesized data. the data are genetated

by scaled orthographic projection, not perspective projection.

An object consists of 20 feature points $\{\boldsymbol{s}_p\}_{p=1}^{20}$ generated from uniform distribution on 200-pixel cube.

We use 10 image of the object. The first camera base matrix is set to identity matrix, $C_1 = \mathrm{I}_3$, and the rest of nine camera base matrix for each object $C_2 \sim C_{10}$ are generated independently by Euler angle representation of rotation matrix of three Euler angles are chosen from uniform distribution. The first and second depth parameters of both objects are set to $\lambda_{1*} = 1$, $\lambda_{2*} = 2$, and the rest of eight depth parameters of each object $\lambda_{3*} \sim \lambda_{10,*}$ are generated independently from uniform distribution on closed interval $[1, 2]$.

We set parameters as follows: learning rate $\alpha = 5 \times 10^{-7}$, weight for constrains term $\beta = 1000$.

The number of iteration for estimation of missing data as **(2)** as previous section is set to 1×10^6.

The shape errors and the depth errors are measured by relative error as

$$\left\| S^{*\mathrm{true}} - S^{*\mathrm{estimated}} \right\| \cdot \left\| S^{*\mathrm{true}} \right\|^{-1},$$

$$\frac{\lambda_{f*}^{\mathrm{estimated}} - \lambda_{f*}^{\mathrm{true}}}{\lambda_{f*}^{\mathrm{true}}}.$$

The motion errors are measured by the angle between true $\boldsymbol{k}_f$ and estimated $\boldsymbol{k}_f$, that is, $\cos^{-1} ||(\boldsymbol{k}_f^{\mathrm{estimated}})^{\mathrm{T}} \boldsymbol{k}_f^{\mathrm{true}}||$.

First, we investigate the reconstruction error against noise. In the computation of errors, we use true value and estimated value of missing data.

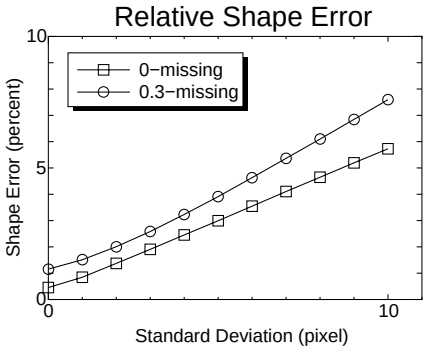

Figure 4. Shape error against Gauss noise

Figure 4-6 shows the reconstruction errors against

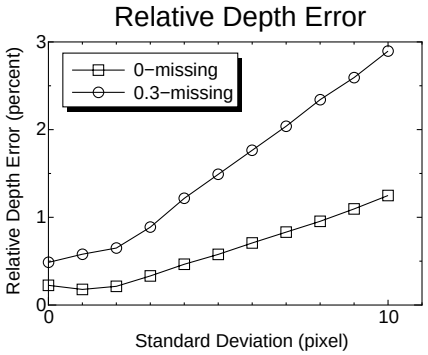

Figure 5. Depth error against Gauss noise

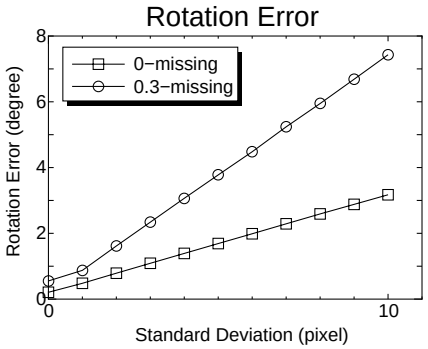

Figure 6. Motion error against Gauss noise

noise for missing rate is 0% and 30%. The reconstruction errors are inclease linearly. We can see high-performance of the reconstruction is achived when there is no missing data, and acceptable reconstruction is achived although 30% of the data were missing.

We investigated the robustness against data missing for the data added the Gauss noise of 0 or 5-pixel standard deviation. Figure 7-9 shows the reconstruction errors against missing rate. Noise has little effects on the breakdown point and the breakdown point of the algorithm is around $50\% \sim 60\%$. Then the presented method has the robustness against missing data.

5 Conlcusion

We presented an autoassociative learning of three-layered perceptron of its connecting coefficients constrained, and we applied the learning to recover the motion and the shape under scaled orthographic projection. The presented algorithm

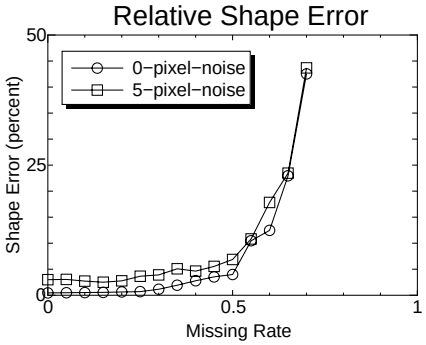

Figure 7. Shape error against missing data

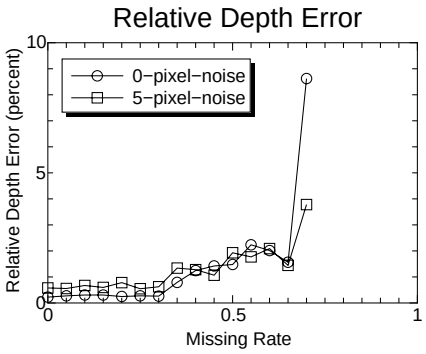

Figure 8. Depth error against missing data

works well even if half of data are missing. In the presented algorithm, we change only missing data and do not change the observed data. We expect that good changing of the observed data derive better performance of the reconstruction. Then the improvement of presented algorithm is needed.

References

[1] P. Baldi. and K. Hornik. Neural networks and principal component analysis. *Neural Networks*, 2:53–58, 1989.

[2] J. P. Costeira. and T. Kanade. A multibody factorization method for independently moving objects. *IJCV*, 29(3):159–179, 1998.

[3] C. W. Gear. Multibody grouping from motion images. *IJCV*, 29(2):133–150, 1998.

[4] N. Ichimura. Motion segmentation based on factorization method and discriminant criterion *ICCV*, 600–605, 1999.

[5] R. J. Rousseeuw. and A. M. Leroy. *Robust Regression and Outlier Detection*. John Wiley & Sons, 1996.

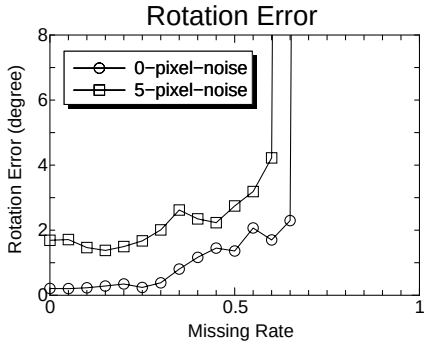

Figure 9. Motion error against missing data

[6] C. Tomasi. and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *IJCV*, 9(2):137–154, 1992.