Experimental study on performance of view-based pose estimation

Toru Tamaki†, Hiroyuki Okugawa†, Toshiyuki Amano‡, Kazufumi Kaneda†
† Hiroshima University, Japan, tamaki@ieee.org
‡ NAIST, Japan, amano@is.naist.jp

Abstract In this paper, we report experimental results on the performance of a view-based pose estimation method called EbC (Estimation-by-Completion) for objects rotating about single axis in 3D. Estimating rotating angle θ (pose) of an object from an image $\boldsymbol{x}$ (view or appearance) of the object is formulated to find a matrix F such that $\theta = F\boldsymbol{x}$ for a given set of learning samples $\{\theta_j, \boldsymbol{x}_j\}$. EbC learns F and estimates pose based on eigenspace of learning samples. Experiments shown in this paper use images in COIL-20 dataset. A learning set is defined by not only the number of images in the set but also which images are included. As a result, some objects keeps good performance even when only three images is used, and performance of several objects is remarkably worse when a learing set with images of 40 degrees separated from each other is used.

1 Introduction

In this paper, we report experimental results on the performance of view-based pose estimation of a 3D object rotating about single axis. Estimating rotating angle θ (pose) of an object from an image $\boldsymbol{x}$ (view or appearance) of the object is widely studied, and the problem is formulated to find a function f such that $\theta = f(\boldsymbol{x})$ for a given set of learning samples $\{\theta_j, \boldsymbol{x}_j\}$. Kernel methods or manifold learning are often used for a nonlinear function, however, a linear function (or a matrix) F is also used to relate pose and image as $\theta = F\boldsymbol{x}$. This linear method is very attractive because 1) estimation is simple, 2) computing F is intuitive (linear subspace methods such as PCA, CCA, and LDA can be used).

The main concern of this paper is to investigate the performance of a view-based linear pose estimation method. We have proposed a liner method [1, 2] called *EbC* (Estimation-by-Completion) and shown that EbC estimates three pose parameters (3DOF) including 1DOF rotation in 3D, and 2DOF translation in 2D image plane. Nevertheless 1DOF rotation of an object about any axis in 3D (this rotation is called *off-the-plane rotation*) is still a hot topic [3] for view-based pose estimation, the performance has not been studied well. Especially, how much does the performance change as the number of images for learning decreases? This is important to know how many images the method requires to achieve desired performance. And also, even when the number of images is fixed, can the method still achieve the same performance? If not, we have to think what learning set of fixed number of images is good or not.

To investigate the questinos above, we describe in this paper experiments on estimation performance of 1DOF off-the-plane rotation. In section 2, we describe EbC, a view-based linear pose estimation method to be investigated. A set of sample images for learning is defined in section 3, and experimental results is shown in section 4.

2 EbC: a linear pose estimation

Let $\boldsymbol{x}_j \in \mathbb{R}^N$ be an image corresponding to pose θ_j, and X be a matrix that has the images and poses column-wise as

$$X = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_n \\ \boldsymbol{p}_1 & \boldsymbol{p}_2 & \cdots & \boldsymbol{p}_n \end{bmatrix} \in \mathbb{R}^{(N+w)\times n}, \quad (1)$$

where n is the number of learning sample images, and

$$\boldsymbol{p}_j = \begin{bmatrix} \cos(0 - \theta_j), \\ \cos\left(\frac{2\pi}{w} - \theta_j\right), \\ \cos\left(2\frac{2\pi}{w} - \theta_j\right), \\ \vdots \\ \cos\left(\frac{2(w-1)\pi}{w} - \theta_j\right) \end{bmatrix} \in \mathbb{R}^w, \quad (2)$$

where $w \geq 2$ [2]. That is, each column of the matrix X is an sample image $\boldsymbol{x}_j$ augmented by corre-

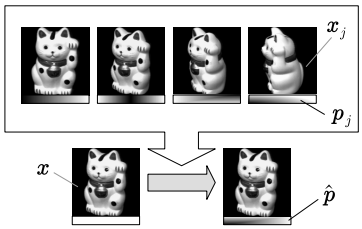

Fig. 1: Overview of the estimation by EbC

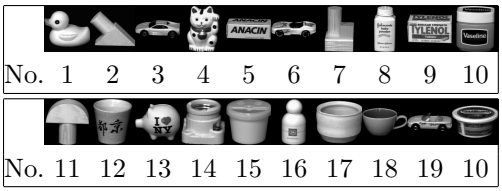

Fig. 3: Images of 20 objects in COIL-20 dataset. Each image is grayscale, 128×120 in size.

sponding parameter vector $\boldsymbol{p}_j$. Then X is decomposed by SVD (Singular Value Decomposition) as $X = EDV^T$, then the eigenvectors $\{\boldsymbol{e}_k\} \in \mathbb{R}^{N+w}$ are obtained in E as its columns. Next E is decomposed into upper and lower parts:

$$E = [\boldsymbol{e}_1\ \boldsymbol{e}_2\ \ldots\ \boldsymbol{e}_{n'}] = \begin{bmatrix} E_{\text{in}} \\ E_{\text{out}} \end{bmatrix} \in \mathbb{R}^{(N+w)\times n'}, \quad (3)$$

where $n' \leq n$. $E_{\text{in}} \in \mathbb{R}^{N\times n'}$ and $E_{\text{out}} \in \mathbb{R}^{w\times n'}$ correspond to images (input) and poses (output), respectively. Then the pose estimation of a test image $\boldsymbol{x}$ is done as follows:

$$\hat{\theta} = \tan^{-1}\left(\frac{\boldsymbol{\Omega}_s^T \boldsymbol{x}}{\boldsymbol{\Omega}_c^T \boldsymbol{x}}\right), \quad (4)$$

$$\boldsymbol{\Omega}_s^T = \boldsymbol{\omega}_s^T E_{\text{out}} (E_{\text{in}}^T E_{\text{in}})^T E_{\text{in}}, \quad (5)$$

$$\boldsymbol{\Omega}_c^T = \boldsymbol{\omega}_c^T E_{\text{out}} (E_{\text{in}}^T E_{\text{in}})^T E_{\text{in}}, \quad (6)$$

$$\boldsymbol{\omega}_s = \left[\sin 0, \sin\left(\tfrac{2\pi}{w}\right), \ldots, \sin\left(\tfrac{2(w-1)\pi}{w}\right)\right]^T, \quad (7)$$

$$\boldsymbol{\omega}_c = \left[\cos 0, \cos\left(\tfrac{2\pi}{w}\right), \ldots, \cos\left(\tfrac{2(w-1)\pi}{w}\right)\right]^T. \quad (8)$$

Here, $\boldsymbol{\Omega}_s$ and $\boldsymbol{\Omega}_c$ are learined matrices (actually vectors) that make relationship between images and poses. Any test image $\boldsymbol{x}$ does not its augmented part $\boldsymbol{p}$ that shows a parameter of pose, of course. However, the part $\boldsymbol{p}$ is first completed with $\boldsymbol{x}$ as $\hat{\boldsymbol{p}} = E_{\text{out}}(E_{\text{in}}^T E_{\text{in}})^T E_{\text{in}}$, then pose is estimated with $\boldsymbol{\omega}_s$ and $\boldsymbol{\omega}_c$ by the equations above (see Fig.1). Therefore, the method has been named as *Estimation-by-Completion, EbC*.

Note that EbC is identical to a linear regression[4] of an image $\boldsymbol{x}_j$ to a pose $(\sin\theta_j, \cos\theta_j)^T$ when the dimension of the eigenspace E is not reduced, i.e., $n = n'$ (see [1, 2]). We fixed $n = n'$ for all experiments described in section 4.

3 Learning image set

The linear estimation method described in the previous section requires a set of learning sample images before pose of a test image is estimated. If we have many samples for an object, the estimated pose becomes accurately. For example, 360 images of an object rotating about single axis allows us to estimate pose so that the error in the estimate is less than one degree, and if 3600 images the error may be smaller than 0.1 degrees.

However, the small number of learning samples is better because taking many pictures of an object is a hard task, and the learning stage becomes computationally expensive for calculating eigenvectors. Therefore, we have to investigate how many images are enough as a learning sample set to achieve desired accuracy of estimates.

Another question is what is the appropriate images when we fix the number of images in a learning set. Four different image sets are shown in Fig.2. The number of images is same for each sets, but the performance of estimation may differ because different appearances of an object is learned. In the following sections, we will discuss about the two topics.

Now we define a learning set. In the following experiments, we use images in COIL-20[5] datasets (Fig.3). There are 72 images for an object rotated by 5 degrees each: $\boldsymbol{x}_0$ for $\theta = 0$ [deg], $\boldsymbol{x}_5$ for $\theta = 5$ [deg], ..., $\boldsymbol{x}_{355}$ for $\theta = 355$ [deg]. Some images of an object are shown at the top row in Fig.2. For each object, we take some images from 72 images to make *a learning set* $S_{i,s} = \{\boldsymbol{x}_{ik+s}\}$ for $k = 0, 1, \ldots, n_i - 1$. Here, i is a sample span (in degree), s is a start angle (also in degree), and $n_i = \frac{360}{i}$ is the number of images in the set.

For example, if we take images every 20 degrees $(\boldsymbol{x}_{20}, \boldsymbol{x}_{40}, \ldots)$ started from 0 degrees $(\boldsymbol{x}_0)$, $i = 20$ and $s = 0$, then the learning set is $S_{20,0} = \{\boldsymbol{x}_0, \boldsymbol{x}_{20}, \boldsymbol{x}_{40}, \ldots, \boldsymbol{x}_{340}\}$ and 18 $(= \frac{360}{20} = n_{20})$ images are stored in $S_{20,0}$.

But when we take images every 20 degrees started from 5 degrees $(i = 20, s = 5)$, the learning set $S_{20,5} = \{\boldsymbol{x}_5, \boldsymbol{x}_{25}, \boldsymbol{x}_{45}, \ldots, \boldsymbol{x}_{345}\}$ is different with $S_{20,0}$, whereas the number of images is still 18 $(= \frac{360}{20} = n_{20})$.

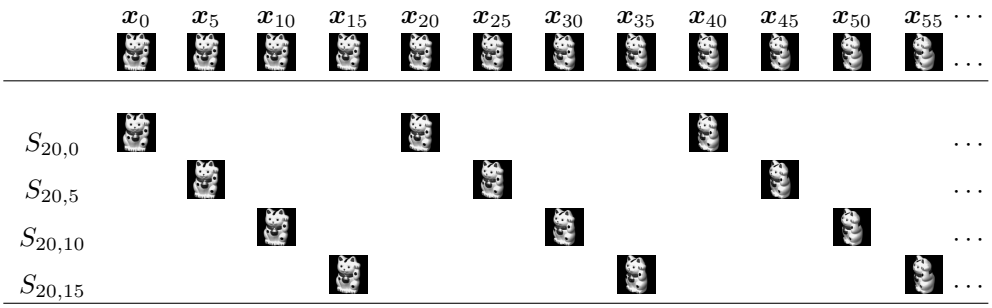

Fig. 2: Examples of images for learning sets $S_{20,0}, S_{20,5}, S_{20,10}$, and $S_{20,0}$. These sets are different, but they have the same number of images in which the object is rotated by 20 degrees each. Note that $\boldsymbol{x}_j$ is an image when the corresponding pose $\theta_j = j$.

4 Experiments

In this section, we describe results of some experiments to see how the performance of the estimation depends on learning sets. For each learning set $S_{i,s}$, we defined the following measurement of performance, RMSE (root mean square error):

$$\mathrm{RMSE}_{i,s} = \sqrt{\frac{1}{72 - n_i} \sum_{\boldsymbol{x}_j \notin S_{i,s}} (\hat{\theta}_j - \theta_j)^2}, \qquad (9)$$

where θ_j is a true angle for an image $\boldsymbol{x}_j$, and $\hat{\theta}$ is an estimated angle. Note that errors for only images that were not learned are considered, therefore, $(72-n_i)$ images are tested among all 72 images when n_i images are used for learning.

Fig.4, 5, 6, and 7 show RMSE for each object. Horizontal axis is the sample span i [deg] for a learning set $S_{i,s}$, and vertical axis shows RMSE. We use divisors of 360 as sample spans: $i = 5, 10, 15, 20, 30, 40, 45, 60, 90, 120$. $\mathrm{RMSE}_{i,s}$ with different s are plotted for same i, in other words, performances of learning sets with the fixed number of images are marked as vertically spread dots.

In the figures, the performance is degraded as the sample span increases, that is, as the number of images used for learning decreases. Of course, the rate how much RMSE increases depends on an object.

We can see two interesting characteristics. First, some object have very small error even when $i = 120$ [deg], that is, using only three images for learning. This occurred when the shape of an object is round and the appearance does not change drastically: object 12, 15, and 20 keep good performance. We may think that estimating pose is difficult when appearance is similar across images, but in fact the estimation is difficult for object of which appearance greatly changes (for example, object 2).

Second point is that errors for only $S_{40,s}$ of several objects were worse than other sets. Especially, RMSEs for object 5, 6, 9, 11, 14, 19 had a "performance dip" at 40 degrees. The learning set $S_{40,s}$ used more images than other sets such as $S_{45,s}$, $S_{60,s}$ and even $S_{90,s}$, but the performance was worse. This does not support an intuitive thinking: more the images, the better the performance.

5 Conclusions

We have reported experimental results on the performance of EbC, a view-based linear pose estimation method. Experiments using images in COIL-20 dataset showed that the estimation error increased as the number of images used for learning decreases. However, some objects kept good performance even when only three images (120 degrees separated from each other) was enough to achieve good performance. Moreover, performance of several objects was remarkably worse when 9 images (40 degrees separated from each other) were used as a learning set.

"Keeping good performance" and "performance dip at 40 degrees" are very interesting topics, and future work is to further investigate these topics as well as to find a way how we effectively choose images for a learing set.

References

[1] Toshiyuki Amano, Toru Tamaki: "A fast linear pose estimation method of 3D object using EbC image pair," IEICE Trans. on Info. Sys., Vol. J90-D, No. 8, pp. 2060-2069, 2007. (in Japanese) `http://search.ieice.org/bin/summary.php?id=j90-d_8_2060&category=D&year=2007&lang=J&abst=`

[2] Toru Tamaki, Toshiyuki Amano: "Multi-port Eigenspace Method," Proc. of Subspace2006, pp.

7-15, 2006. (in Japanese)
http://ir.lib.hiroshima-u.ac.jp/00017333

[3] Toru Tamaki, Toshiyuki Amano, Kazufumi Kaneda: "The secret of rotating object images – Using cyclic permutation for view-based pose estimation –," Proc. of Subspace2007, pp. 24-31, 2007.
http://ir.lib.hiroshima-u.ac.jp/00020419

[4] T. Okatani, K. Deguchi: "Yet another appearance base method for pose estimation based on a linear model," Proc. of MVA2000, pp. 258-261, 2000.

[5] S. A. Nene, S. K. Nayar, H. Murase: "Columbia Object Image Library (COIL-20)," Technical Report CUCS-005-96, 1996.
http://www1.cs.columbia.edu/CAVE/software/softlib/coil-20.php

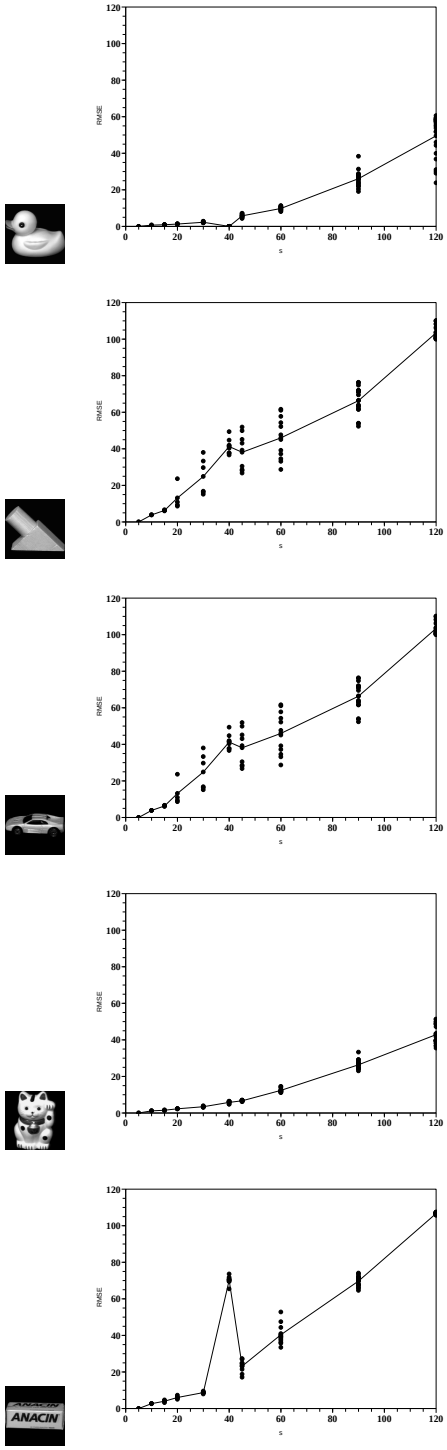

Fig. 4: RMSE for object 1, 2, 3, 4, 5 (from top to bottom) in COIL-20.

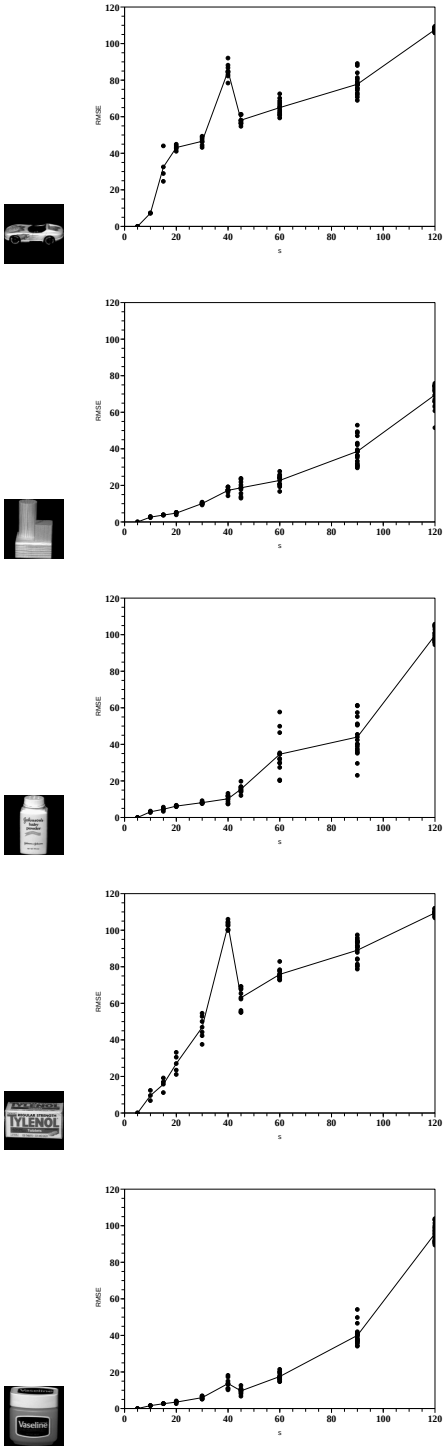

Fig. 5: RMSE for object 5, 6, 7, 8, 9, 10 (from top to bottom) in COIL-20.

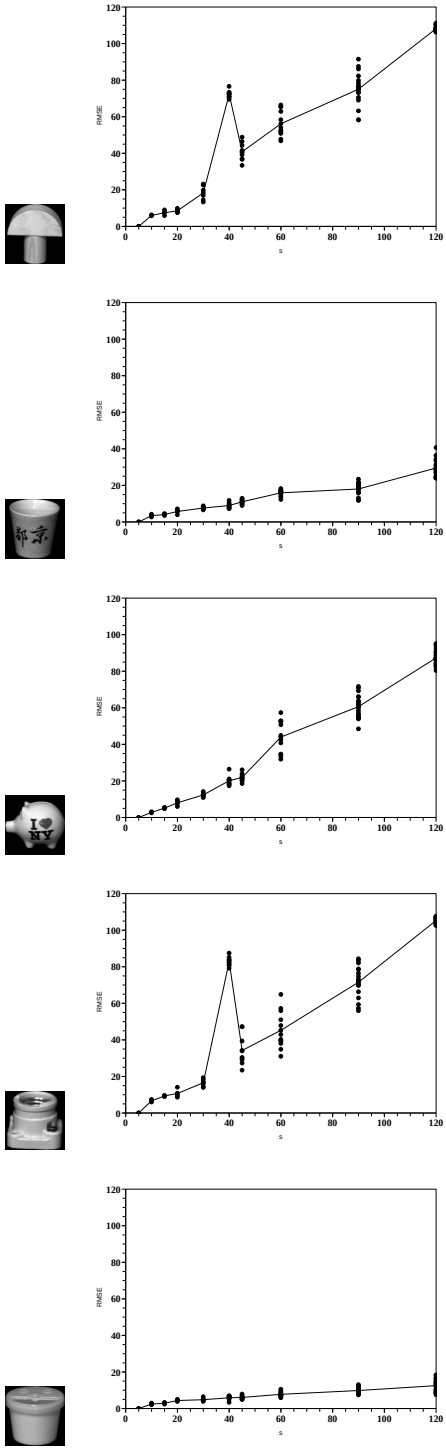

Fig. 6: RMSE for object 10, 11, 12, 13, 14, 15 (from top to bottom) in COIL-20.

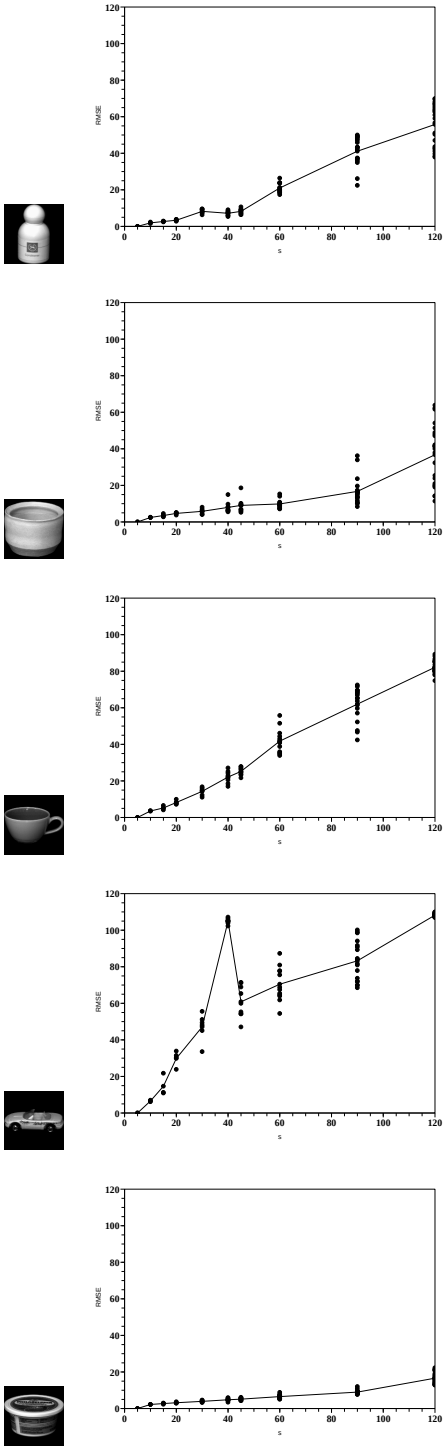

Fig. 7: RMSE for objects 15, 16, 17, 18, 19, 20 (from top to bottom) in COIL-20.