Non-Uniformity of a Pattern and "the Best" Single View 3D Pose Estimator

Georgii KHACHATUROV
Universidad Autónoma Metropolitana
Av. San Pablo 180, Azcapotzalco, D.F.,C.P. 02200, MEXICO
xgeorge@hp9000a1.uam.mx

ABSTRACT

Visual model of an object is presented as composition of a model-shape function and a model-pattern function. Non-uniformity of a pattern is defined as a quadratic form related to the gradient of intensity. For a planar model, it is proved that a more precise estimation responds to a greater non-uniformity. This is applied to develop a class of single-view 3D-pose estimators exploring extremely non-uniform patterns. The estimators are of high precision, fast, robust, and algorithmically simple. Possible application is spacecraft docking, or any robot problem permitting usage of a visual mark (target) on remote object.

Keywords: model-based single view 3D-pose estimation, spacecraft docking, robot control, visual servoing.

1. INTRODUCTION

"More non-uniform picture painted on a known object yields a higher precision for visual tracking of the object pose". This affirmation might intuitively seem true for the readers. First part of the paper represents a strict mathematical basement for this affirmation. It will be shown that this affirmation holds only for planar objects. It turns out that in general case the slope variation of an object may contribute more Fisher's information than a pattern painted on the object.

The case of planar objects is developed furthermore to a new 3D-pose estimator. As it follows from the main theoretic result of the paper, such an estimator has the highest possible precision for planar targets. An important property of method presented in the paper is that it does not require preliminary feature extraction from the target image. In addition, it is robust, fast, and algorithmically simple.

Past Works on Visual Estimators of 3D-Pose

The main application area for the visual 3D-pose tracking is robot control. The mutual pose of a pair of 3D coordinate systems is determined as a set of 6 scalar parameters. The problem is to estimate these 6 parameters by processing an image of the remote object.

Typically, the processing can be subdivided into two steps:

(***i***) Extracting features of the remote object from its image

(***ii***) Estimating 3D-pose by the features.

A reliable, fast, and precise method for the first step presents the main difficulty of visual processing of a 3D-pose estimator. Assuming that it is performed successfully, for a known geometrical model of the object, the reconstruction of 3D-pose itself may be considered as a respectively easier operation. Indeed, it is reduced to a strict mathematical problem of inversion of a known map (say, perspective projection) between two Euclidean spaces of small dimensions. [Surprisingly, many works dedicated to tracking of 3D-pose by visual information are still investigating new methods for the second part of the problem, leaving the fist one out of consideration.]

There exist various classifications of estimators of 3D-pose. The subdivision of visual-based estimators of 3D pose as the local and the global ones was proposed by Faugeras et al (1984).

The classic triangulation, which represents the main tool of the star navigation, is a simple example of local approach. Jarvis (1983) developed a computer vision triangulation method. However, dealing with a few local objects of reference, the precision is low since they represent a small part of visual information. Another property of local approaches is that they depend on the performance of extraction of local objects. A similar problem arises for the approach by Abidi and Chandra (1995) proposing intrinsic distances of the remote object to estimate its pose. On the other hand, working with many local features, the complicated logic of tracking of features makes this approach inconvenient for practice. Hel-Or and Werman (1995) although developed a kind of local approach. They propose a technique of uncertainty matrixes, which allows to process and fuse in a uniform way the data of range and intensity images.

Kriegman (1992) developed a model-based pose estimator, which finds 3D pose solving polynomial equations for surfaces represented as algebraic equations.

The methods of global approach map the image into an Euclidean space where matching with a model is performed. A mayor part of such methods must extract local features before mapping. As examples of the "shape from contour" technique applied for pose estimation may be mentioned works by Dhome et al (1989), and by Dunker et al (1996). Other class of global pose estimators makes use of Hough transform. Tanaka et al (1985) apply Hough transform to find known model in a 3D scene. Some estimators work with Extended Gaussian Image. The works by Brou (1984) and by Kang and Ikeuchi (1993) follow this way. Vinther

and Cipolla (1994) proposed affine invariants to find 3D-pose.

A planar target for 3D-pose estimate was proposed by Khachaturov et al (1987). It allows the measurement of parameters of its image without preprocessing of image and extraction of local features.

DeMenthon and Davis (1995) developed a numeric method for inverting perspective projection based on linear algebra. It works when a model of the object shape is known and the object features are already extracted from an input image. [So this is a method devoted to the second step of processing in the classification above.] The advantage of this method with respect to approaches of Lowe (1985, 1991) and Yuan (1989) is that it does not require an initial pose estimate and does not require matrix inversion in its iteration loop.

The work by Laurin and Rioux (1995) makes a contribution into the first part of the problem. It apply a sine-coding technique to range images. Then the Fourier transform (FT) is applied to coded images to estimate 3D pose of objects.

This technique is close to the one presented in the actual paper, since both apply FT and investigate the peaks.

The primary difference of these two approaches is that they use different hardware. It generates the semantic difference of the input information to process: the range images and, respectively, the intensity ones.

Then, there is a difference in the qualitative properties of the two approaches.

The approach by Laurin and Rioux uses assumption that there are a number of planar objects on the scene. This approach depends strongly on the performance of preliminary segmentation of range images. The aim of the segmentation is the separation of the planar objects from their background. A wrong segmentation leads to bad tracking of 3D-pose.

The method presented below performs processing of intensity images of a known target. An occlusion and other perturbations of a wide range do not destroy the method. However, not all applications allow using the target. So, the presented method is more special, but more reliable.

2. NON-UNIFORMITY OF A PATTERN AND PRECISION OF ESTIMATES

A Mathematical Model of Input Information

A complete *mathematical model of the target image* is defined as the quadruple $\{g, \mathbf{x}, I, G\}$, where each component is described below:

Let $\Omega \in R^L$ denote an unknown vector of parameters, and $\mathbf{y}$ denote a point of *image space* R^M. In particular, Ω may describe 3D-pose of remote object, so L=6 in this case, and $\mathbf{y}$ may be two-dimensional vector of a point on the visual image frame, M=2.

By definition, the *model of shape* of an object is a map $\mathbf{x}(\mathbf{y},\Omega)$ representing a point $\mathbf{x}$ of a manifold G of the same dimension M as the one of the image space. [In particular, for the case of usual gray-scale images, a model of shape recovers explicitly a point of G, which is a 2D manifold representing surface of a 3D object.] For a known vector Ω, as $\mathbf{y}$ spans entire image space, $\mathbf{x}(\mathbf{y},\Omega)$ spans a domain S of G. [For the usual case of 2D images, we can understand S as the domain of visibility on G. To construct it, we can place a view-point on the image of G, and while the view-point spans the entire image, all visible points of G are attributed to S.]

By definition, the *model of pattern* painted on the model object is a known real function $g(\mathbf{x})$, where $\mathbf{x}$ is a point of G. [So, if G is the surface of a 3D object, $g(\mathbf{x})$ represents the gray scale intensity of a pattern painted on G at $\mathbf{x}\in G$.]

The image $I(\mathbf{y})$ is defined as a random function with expectation $g(\mathbf{x}(\mathbf{y},\Omega))$. The random variables $I(\mathbf{y}_1)$, $I(\mathbf{y}_2)$ are assumed to be independent for $\mathbf{y}_1\neq\mathbf{y}_2$, and variance $\sigma^2=\sigma^2(I(\mathbf{y}))$ does not depend on $\mathbf{y}$. (Say, I is the signal plus Gaussian white noise).

Estimate of Pose and Its Precision

Under an estimate $\tilde{\Omega}$ of Ω we understand a point estimate [all used concepts and facts of statistics can be found in a book by Cox and Hinkley (1974)], i.e. any vector valued functional $\tilde{\Omega}=\tilde{\Omega}(I)\in R^L$ defined on input images.

An estimate is called *efficient* if it provides minimum to the functional $\int_{R^L} E(\|\Omega-\tilde{\Omega}(I)\|^2)d\Omega$. So, in a sense, an efficient estimate has the best possible precision.

In terms of the model $\{g, \mathbf{x}, I, G\}$, the least squares estimate is given by

$$\tilde{\Omega}=\arg\min_{\Omega}\Phi(I,\Omega),$$
$$\Phi(I,\Omega)=\int_D [I(\boldsymbol{y})-g(\boldsymbol{x}(\boldsymbol{y},\Omega))]^2 d\boldsymbol{y}, \quad (1)$$

where $I(\mathbf{y})$ is an image, and the domain D coincides with the image of G.

It is known that for the described model $\{g, \mathbf{x}, I, G\}$, the least squares estimate is asymptotically efficient. For the case under consideration, it means

$$\frac{\{\text{variance of efficient estimate}\}}{\{\text{variance of least squares estimate}\}}\xrightarrow[\sigma^2\to 0]{}1$$

for each component of the estimate. Interpreting this property for a practical application, we may accept that the least squares method reaches the best possible precision.

For any estimate, the lower bounds of its variances are established by the Cramér-Rao inequality. These bounds coincide with variances of an efficient estimate. Hence, for the variances of the least squares estimate, we

may, practically, accept values given by bounds of the Cramér-Rao inequality.

Due to the Cramér-Rao inequality, the covariance matrix of the efficient (i.e. having lowest variances) estimate $\tilde{\Omega}$ is given as $\mathrm{cov}(\tilde{\Omega}, \tilde{\Omega}^T) = F^{-1}$, where F is the matrix of Fisher's information. For model $\{g, \mathbf{x}, I, G\}$, the elements of Fisher's matrix $F = \{F^{k,l}\}_{k,l=1,\ldots L}$ may be expressed as

$$F^{k,l} = \int_D \sum_{m,n=1,\ldots,M} \frac{\partial g}{\partial x_m} \frac{\partial x_m}{\partial \omega_k} \frac{\partial g}{\partial x_n} \frac{\partial x_n}{\partial \omega_l} d\boldsymbol{y},$$

where $\Omega=\{\omega_k\}_{k=1,\ldots,L}$, and $\mathbf{x}=\{x_n\}_{n=1,\ldots,M}$.

Comparison of Precision for Different Non-Uniformities of the Target-Pattern

Using the above representation of the Fisher's matrix, we will study the precision of least squares method in dependence of a property of the pattern function g.

Let define *non-uniformity of a pattern g* as a quadratic form with the matrix $Q_g = \int_S P(\boldsymbol{x}) P^T(\boldsymbol{x}) d\boldsymbol{x}$, where $P(\mathbf{x})$ = grad g = $\{\frac{\partial g}{\partial x_l}\}_{l=1,\ldots,M}$, $\mathbf{x} \in S$, and S is defined above.

Let $\{g_i\}_{i=1,2}$ be a pair of pattern functions and Q_{g_i} be their non-uniformities. Let $\tilde{\Omega}_i$, i=1,2, denote the least squares estimates of Ω given, in accordance with (1), as

$$\tilde{\Omega}_i = \arg min \Phi_{g_i}(I, \Omega),$$

$$\Phi_{g_i}(I, \Omega) = \int_D [I(\boldsymbol{y}) - g_i(\boldsymbol{x}(\boldsymbol{y}, \Omega))]^2 d\boldsymbol{y}.$$

Let domain of visibility S does not change for a sufficiently small variation of Ω.

Under these conditions, the matrix of Fisher's information F_{g_i}, i=1,2 is represented as $F_{g_i}^{k,l} = \int_D \sum_{m,n=1,\ldots,M} \frac{\partial g_i}{\partial x_m} \frac{\partial x_m}{\partial \omega_k} \frac{\partial g_i}{\partial x_n} \frac{\partial x_n}{\partial \omega_l} d\mathbf{y}$.

For quadratic forms with the matrixes A and B of the same dimension, we use generally accepted notation $A \geq B$ denoting that for any vector $\mathbf{z}$, $\mathbf{z}^T A \mathbf{z} \geq \mathbf{z}^T B \mathbf{z}$ holds.

Using just presented notation and assumptions, the following theorem is true.

Theorem 1. Let the model of shape $\mathbf{x}(\mathbf{y},\Omega)$ span a planar object G in 3D space and the images of this object be produced as plane perspective projections. Let the size of G be much less than the distance between G and the projection plane. If non-uniformities of two patterns $\{g_i\}_{i=1,2}$ painted on G are related as $Q_{g_1} \geq a Q_{g_2}$ for a positive value a, then $F_{g_1}^{-1} \leq a^{-1} F_{g_2}^{-1}$.

Proof. We can rewrite the given above representation of F_{g_i} in the form $\int_D X^T P_i P_i^T X d\mathbf{y}$, where MxL-matrix X is defined as $X = \{\frac{\partial x_m}{\partial \omega_k}\}_{k=1,\ldots,L}^{m=1,\ldots,M}$, M=2, L=6. Since $\mathbf{x}(\mathbf{y},\Omega)$ represents a planar object and the target-size is much less than the distance to the target, then X almost does not depend on $\mathbf{y}$. Hence,

$$F_{g_i} = \int_D X^T P_i P_i^T X d\boldsymbol{y} \approx X^T [\int_D P_i P_i^T d\boldsymbol{y}] X = X^T [c \int_S P_i P_i^T d\boldsymbol{x}] X = c X^T Q_{g_i} X$$

, where c is a constant of Jacobian used in the change of variable. Hence, for any $\mathbf{z} \in R^L$,

$$\mathbf{z}^T [X^T Q_{g_1} X] \mathbf{z} \geq a \mathbf{z}^T [X^T Q_{g_2} X] \mathbf{z} \qquad (2)$$

is true, as far as it is the same as

$$[\mathbf{z}^T X^T] Q_{g_1} [X \mathbf{z}] \geq a [\mathbf{z}^T X^T] Q_{g_2} [X \mathbf{z}]$$

which, in turn, follows from $Q_{g_1} \geq a Q_{g_2}$.

However, (2) just means $F_{g_1} \geq a F_{g_2}$ from which $F_{g_1}^{-1} \leq a^{-1} F_{g_2}^{-1}$ follows. ■

For a=1, $Q_{g_1} \geq Q_{g_2}$ means that the pattern g_1 is more non-uniform than the one of g_2. For $\mathbf{z}=\{z_j\}$ with $z_k=1$ and $z_j=0$ for $j \neq k$, the inequality $F_{g_1}^{-1} \leq F_{g_2}^{-1}$ implies that the variance of efficient estimate of each component ω_k given by $\tilde{\Omega}_1$ is not greater than the one given by $\tilde{\Omega}_2$. In other words, one has obtained

Corollary 2. More non-uniform pattern on a planar object provides higher precision of efficient estimates of 3D-pose of the object.

Now let consider non-uniformity for periodic patterns. Let $\mathbf{x}=\{x_1,x_2\}$ and a pattern function $g_2(x_1,x_2)$ be a bi-periodic function restricted in a planar domain G. Let for both variables the number of periods in G be much greater than 1. Let the non-uniformity of pattern g_2 defined inside G be a positive definite form and Q_{g_2} be the corresponding matrix. Let define the pattern $g_1(x_1,x_2)$, $\{x_1,x_2\} \in G$, as $g_1(x_1,x_2) = g_2(\alpha x_1, \alpha x_2)$ with an $\alpha > 1$. Since G contains many periods of g_2 along as x_1 as x_2, then, asymptotically for a large α, $Q_{g_1} \approx \alpha^2 Q_{g_2}$ holds. Repeating literally the scheme of proof of the theorem, one obtains $F_{g_1} \approx \alpha^2 F_{g_2}$. Inverting it, $F_{g_1}^{-1} \approx \alpha^{-2} F_{g_2}^{-1}$, and using Corollary 2, one has $\{F_{g_1}^{-1}\}_{k,k} \approx \alpha^{-2} \{F_{g_2}^{-1}\}_{k,k}$ where $\{\bullet\}_{k,k}$ are diagonal elements of $F_{g_i}^{-1}$ for i=1,2. Since $\{F_{g_1}^{-1}\}_{k,k}$ and $\alpha^{-2}\{F_{g_2}^{-1}\}_{k,k}$ are the variances of components of $\tilde{\Omega}_i$, i=1,2, we come to

Corrolary 3. Let all assumptions preceding the theorem 1 be true. Let the model of shape correspond to a planar object in 3D space and the model of pattern be a periodic pattern function, then, asymptotically, proportional increasing of the frequencies of the pattern function yields inversely proportional decreasing of standard deviations of the efficient estimates of all components of 3D-pose.

Remark 4. At first glance, as the theorem as its corollaries could seem trivial for practical application, however it worth mentioning that in general case of a non-planar model of object, an affirmation similar to the theorem is not true: More exactly, using previous notation, the property

$$\int_S P_1(\boldsymbol{x}) P_1^T(\boldsymbol{x}) d\boldsymbol{x} \geq \int_S P_2(\boldsymbol{x}) P_2^T(\boldsymbol{x}) d\boldsymbol{x}$$

in general case, does not imply

$$\int_D X^T(\boldsymbol{y})P_1(\boldsymbol{x}(\boldsymbol{y}))P_1^T(\boldsymbol{x}(\boldsymbol{y}))X(\boldsymbol{y})d\boldsymbol{y} \geq$$

$$\int_D X^T(\boldsymbol{y})P_2(\boldsymbol{x}(\boldsymbol{y}))P_2^T(\boldsymbol{x}(\boldsymbol{y}))X(\boldsymbol{y})d\boldsymbol{y}.$$

A counter-example of this kind can be simply presented for P_i and X considered as scalar functions of one variable.

So, for non-planar G the affirmation of the theorem is not true. The explanation is that the slope variation for a model of shape may contribute more Fisher's information than non-uniformity of pattern.

3. "THE BEST" SINGLE-VIEW 3D-POSE ESTIMATOR

The Theorem and, specially, its Corollary 3 hint a straightforward method for increasing of the precision of a model-based 3D-pose visual estimator: the non-uniformity of a pattern should be increased.

Theoretically, the only limitation for this increasing is given by the pixel size and can be found according to the classic Kotel'nikov-Shannon-Whittaker sampling theorem, provided a size of digital photos a target to be fixed.

[In this respect, a natural question is how to join variable distance to the target with a constant picture size? Practically, it can be done by the variation of camera-zoom.

Namely, the zoom parameter must be included in a coordinated way into both mechanisms:

- Into the control loop of TV-camera: to provide a fixed picture size independently of distance.
- Into the algorithm estimating 3D pose: to update characteristics of the optical system performing perspective projection.

Of course, maintaining a constant picture-size can be provided only for a certain diapason of distances between the target and TV-camera. It specifies the working range of the corresponding system. The phrase at the beginning of this section is valid as long as such a virtual system stays in its working range of distances.

The rest of paper deals only with 3D-pose estimation. So, we are disregarding here as the control of camera-zoom as the interaction between both mechanisms, leaving these points for practical developments.]

However, it is not clear how to process an essentially non-uniform pattern: the least squares method given by (1) does not work in this case.

The spectral methods instead of the straightforward least squares method overwhelm this difficulty.

An Example Of "The Best" Estimator With A Planar Target

This example follows the work by Khachaturov (1998).

[To prevent misunderstanding, note that this example presents an estimator processing digital images of a planar target with an utterly non-uniform pattern. The only justification of the prefix “the best” used for such an estimator is given by Corollary 2.]

The Target and Parameters of Its Image

We use a square target. If a target size is much smaller than the distance between the camera and the target, then one may approximately consider the perspective projection of the square target as a parallelogram. So, the target image parameters (depicted on Fig.4[a]) may be defined as the triple of vectors $\{\boldsymbol{i}_x, \boldsymbol{i}_y, \boldsymbol{i}_c\}$ in the following way.

The pair $\{\boldsymbol{i}_x, \boldsymbol{i}_y\}$ consists, by definition, of two vectors oriented as two connected sides of the target-image parallelogram. The lengths of $\boldsymbol{i}_x$ and $\boldsymbol{i}_y$ coincide, respectively, with lengths of the related sides. $\boldsymbol{i}_c$ is defined as the vector connecting the image-frame origin with the target image center.

In other words, $\{\boldsymbol{i}_x, \boldsymbol{i}_y\}$ represents the form, size, and orientation of the corresponding target image; $\boldsymbol{i}_c$ is the target-image translation inside the frame.

The Pattern of Target

The model of pattern is given functionally as

$$g(x_1, x_2) = A[1 + \cos(2\pi f_1 x_1)\cos(2\pi f_2 x_2)], \tag{3}$$

where $\{x_1,x_2\} \in [0,1]\times[0,1]$. The term "1" is introduced to make the model of pattern non-negative to look like the intensity of a gray-scale image. Varying f_1 and f_2, one can reach non-uniformity as high as the pixel size permits.

Properties of the Pattern

An explicit calculation of Q_g for g of (3) shows that as f_1, $f_2 \to \infty$, Q_g tends to a diagonal matrix with its diagonal elements to be proportional to squares of the frequencies f_1, f_2. So, by increasing of f_1, f_2, due to the corollary 3, Q_g can be made as large as necessary for the precision of pose estimates; the pixel size is the only limitation for the increasing.

If g of (3) is defined on the entire plane, its FT is given by G:

$$G(u_1, u_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x_1, x_2) e^{-i2\pi(x_1u_1 + x_2u_2)} dx_1 dx_2 =$$

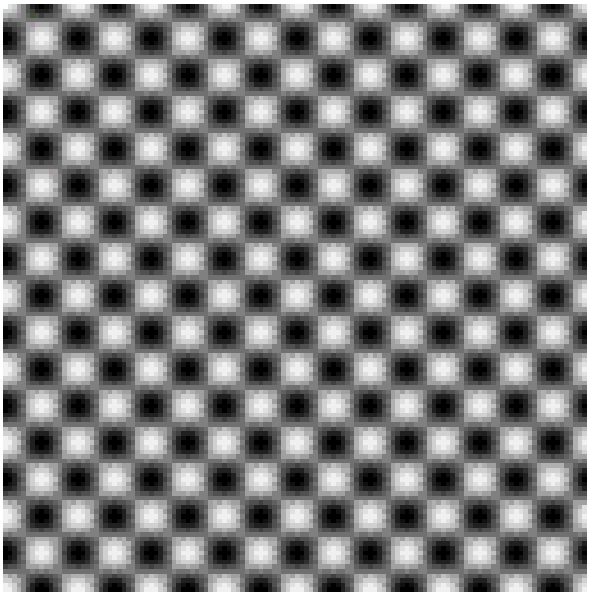
$$A\{\delta(u_1)\delta(u_2) + \frac{[\delta(u_1+f_1)+\delta(u_1-f_1)][\delta(u_2+f_2)+\delta(u_2-f_2)]}{4}\} \tag{4}$$

It is equal to zero everywhere except the five points

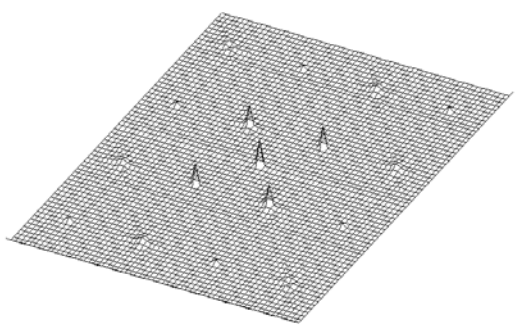
$$\mathbf{u}^{(0)}=\{0,0\},\ \mathbf{u}^{(1)}=\{f_1,f_2\},\ \mathbf{u}^{(2)}=\{-f_1,f_2\},\ \mathbf{u}^{(3)}=\{f_1,-f_2\},\ \mathbf{u}^{(4)}=\{-f_1,-f_2\}. \tag{5}$$

(For details of the reduction of (4), we refer to Brigham (1974)). The theoretic result of (4) is not getting much worse after restriction of the

entire plane of definition of g on the unit square $[0,1]\times[0,1]$: Fig. 1[a] gives visualization of g of (3) defined in the unit square for $f_1=f_2=8$, and Fig. 1[b] is the plot of $log(1+|G|)$, where G represents FT of g.



(a)



(b)

Figure 1. A simulated pattern (a) and (b) *log* of its Fourier transform

Due to Parseval's theorem (Pratt (1978)), $||g||^2=||G||^2$. So, all distributed in spatial domain energy of the functional pattern (3) passes in the spectral domain into the energy concentrated at the points (5). As it follows from the next well-known property, a similar relation between the model of pattern and its FT is valid if a linear transformation H is applied to the plane of definition of g:

Proposition 5. (Egorov and Shubin, 1992). *Let* $\mathbf{x}$ *and* $\mathbf{u}$ *be vectors of the same dimension,* $G(\mathbf{u})$ *be the FT of the function* $g(\mathbf{x})$*, then the FT of* $g(H(\mathbf{x}))$*, is* $\langle H^{-1}\rangle G(H^{-1T}\mathbf{u})$ *where* H^{-1} *is the inverse of* H*,* $\langle\bullet\rangle$ *is determinant of* $\bullet$*, and* $\bullet^T$ *is transposition of matrix* $\bullet$*.*

Estimation of Size, Form, and Orientation of the Target Image

The estimation of $\{\boldsymbol{i}_x, \boldsymbol{i}_y\}$ follows the block-operations of Fig.2.

The FT of the first block of Fig.2 is applied to a small fragment of the frame containing the target image.

The extraction of non-zero high-energy points in the second block can use (4) in the following way. Let $G(\mathbf{v})$ be the FT entering the second block. Due to (4), at any of the four unknown non-zero high-energy points $\mathbf{v}^{(k)}$, the values $|G(\mathbf{v}^{(k)})|_{k=1,...,4}$ are about four times less than $|G(\mathbf{v}^{(0)})|$, $\mathbf{v}^{(0)}=\{0,0\}$. So, for an ε being a small positive value, the algorithm can use $\frac{1-\varepsilon}{4}|G(\boldsymbol{v}^{(0)})|$ as a threshold to localize $\mathbf{v}^{(k)}$, $k=1,...,4$. In other words, if at a non-zero point $\mathbf{v}$, $|G(\mathbf{v})|>\frac{1-\varepsilon}{4}|G(\boldsymbol{v}^{(0)})|$holds, then $\mathbf{v}$ is a good candidate to be one of $\mathbf{v}^{(k)}$, $k=1,...,4$.

The estimation of matrix H in the third block is based on the relation (Proposition 5)

between the high-energy points $\{\boldsymbol{v}^{(k)}\}_{k=0,\ldots,4}$ of $g(H(\mathbf{x}))$ and those $\{\boldsymbol{u}^{(k)}\}_{k=0,\ldots 4}$ of a known calibrating image $g(\mathbf{x})$.

Let $\{\boldsymbol{i}_x^0, \boldsymbol{i}_y^0, \boldsymbol{i}_c^0\}$ be the target-image parameters for a calibrating 3D-pose, and $\boldsymbol{u}^{(k)} = \{u_1^{(k)}, u_2^{(k)}\}, k = 0,\ldots,4,$ be its high-energy points. Due to proposition 5, one may write the equations $[\mathbf{u}^{(k)}]^{\mathrm{T}}=[\mathbf{v}^{(k)}]^{\mathrm{T}}H$, k=0,…,4, where H is unknown and $\mathbf{u}^{(k)}$, $\mathbf{v}^{(k)}$ are known. The equation for k=0 gives no information for the search of H. The remaining four equations spawn eight equations in coordinates to find four elements of H. An application of least squares method terminates the estimation of H.

.

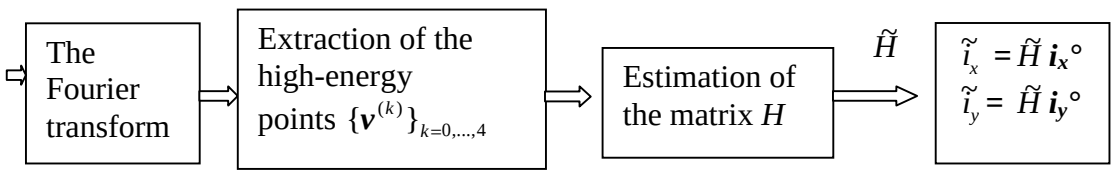


Figure 2. **The block-diagram of the estimation of $\boldsymbol{i}_x$ and $\boldsymbol{i}_y$**

Estimation of the Translation Vector of a Target Image

Unlike the search of $\{\boldsymbol{i}_x, \boldsymbol{i}_y\}$, which works only with *coordinates* of high-energy points, the estimation of the translation vector $\boldsymbol{i}_c$ explores the *values* of energy at these points. It is assumed that $\{\boldsymbol{i}_x, \boldsymbol{i}_y\}$ are known already.

Let $\mathbf{w}$ be the center of a window mask M of the same form, size, and orientation as the target image, $G_{\mathbf{w}}$ be the FT of the image inside M, and function F be defined as $F(\mathbf{w})=|G_{\mathbf{w}}(\mathbf{v}^{(1)})|^2 + |G_{\mathbf{w}}(\mathbf{v}^{(2)})|^2 + |G_{\mathbf{w}}(\mathbf{v}^{(3)})|^2 + |G_{\mathbf{w}}(\mathbf{v}^{(4)})|^2$, where the points $\mathbf{v}^{(\mathrm{i})}$, i=1,...,4 are four non-zero high-energy points of $G_{\mathbf{w}}(\mathbf{v})$. So, the value of $F(\mathbf{w})$ contributes a part of the whole energy of the image inside M.

The estimate $\tilde{\boldsymbol{i}}_c$ of $\boldsymbol{i}_c$ is defined as

$$\tilde{\boldsymbol{i}}_c = \mathbf{arg}\ max\ F(\mathbf{w}) \qquad (6)$$

The rest of item gives a justification of this rule and reduces it to a simple algorithm.

Let find how F depends on $\mathbf{w}$ and $\boldsymbol{i}_c$

Proposition 6. *Let D_1 denote the domain of intersection of the test window and the target image. Then for the target pattern given by (3) with f_1, f_2 >>1, the mathematical expectation E_F of F is asymptotically proportional to the area of D_1, provided the following assumption to be true:*

Assumption. *The image inside the difference of two sets {a window mask \the target image} makes no contribution into the magnitudes of* $\{G_{\mathbf{w}}(\mathbf{v}^{(\mathrm{i})})\}_{\mathrm{i}=1,\ldots,4}$.

Note that practically the assumption means that the image inside the complement of

the target image has no energetic noise with respect to main frequencies of the target pattern.

Proof. In spatial coordinates x_1, x_2, the energy of the image $I(x_1, x_2)$ inside a domain D is $E = \int_D I^2(x_1, x_2)dx_1dx_2$. Let D be the test window and the intensity function inside the background of the target image be zero, then $E = \int_{D_1} I^2(x_1, x_2)dx_1dx_2$. Substituting $g(H(\mathbf{x}))$ instead of I, we see that, for sufficiently large D_1, E is asymptotically proportional to the area of D_1 due to periodicity of g. Returning to the representation of energy in the spectral plane (Parseval's theorem), all energy of FT $G_{\mathbf{w}}(\mathbf{v})$ of the mask is concentrated at the points $\{\mathbf{v}^i\}_{i=0,...,4}$. Hence, for the case of zero-background, $F(\mathbf{w})$ is also proportional to the area of D_1. To finish the proof, note that the assumption implies the same effect as zero-valued background of the target image.■

The area of D_1 can easily be found as an explicit function of the $\mathbf{w}$ and i_c for known $\boldsymbol{i}_x$, $\boldsymbol{i}_y$.

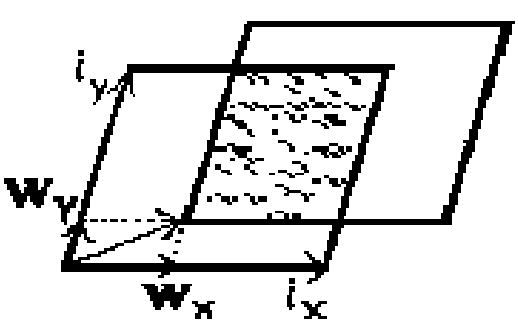


Figure 3. The auxiliary vectors to find the area of intersection of the mask and the target image. The target image corresponds to the parallelogram formed by vectors $\boldsymbol{i}_x$ and $\boldsymbol{i}_y$. Another congruent parallelogram represents a mask. Translations of target-image and mask are i_c and $\mathbf{w}$, respectively. So, the vector connecting origins of two parallelograms is $i_c - \mathbf{w}$. Its projections on $\boldsymbol{i}_x$ and $\boldsymbol{i}_y$ are denoted by $\boldsymbol{w}_x$ and $\boldsymbol{w}_y$.

Indeed, decomposing the vector i_c - $\mathbf{w}$ into w_x+w_y, where the vectors w_x and w_y are parallel with $\boldsymbol{i}_x$ and $\boldsymbol{i}_y$, respectively, Fig. 3, we see that the area in question is equal to the module of vector product $|(\boldsymbol{i}_x - w_x)\times(\boldsymbol{i}_y - w_y)|$.

So, if the intersection is not empty, E_F has the explicit form $E_F(\mathbf{w}, i_c) = B|(i_x - w_x)\times(i_y - w_y)|$, where the constant B is unknown. Otherwise, $E_F(\mathbf{w}, i_c) = 0$. The analytical geometry technique gives expressions for w_x and w_y:

$$w_x = \frac{|i_c - \boldsymbol{w} - i_x^*(i_c - \boldsymbol{w}, i_x^*)|^2}{(i_y^*, i_c - \boldsymbol{w} - i_x^*(i_c - \boldsymbol{w}, i_x^*))} i_x^*,\ w_y = \frac{|i_c - \boldsymbol{w} - i_y^*(i_c - \boldsymbol{w}, i_y^*)|^2}{(i_x^*, i_c - \boldsymbol{w} - i_y^*(i_c - \boldsymbol{w}, i_y^*))} i_y^*,$$

where unit vectors i_x^*, i_y^* are defined as

$$i_x^* = \frac{i_x}{|i_x|}, i_y^* = \frac{i_y}{|i_y|}.$$

At last, treating the values of $F(\mathbf{w})$ as random observations of the known function $E_F(i_c, \mathbf{w})$, the unknown vector i_c (and, by the way, the value B) can be estimated by usual least squares method. Thus, the estimation (6) of the vector i_c is reduced to

$$\tilde{i}_c = \arg min \sum_n |E_F(i_c, \mathbf{w}_n) - F(\mathbf{w}_n)|^2,$$

where E_F is a known function and the set $\{\mathbf{w}_n\}$ consists of known test positions of the mask center. We are omitting further details of this well-known problem.

3D-Pose Reconstruction

For given parameters {$\boldsymbol{i}_x$, $\boldsymbol{i}_y$, $\boldsymbol{i}_c$} of a target-image, the recovering of 3D-pose is well-known problem of inversion of the perspective projection: find space pose of the square by characteristics of its projection.

For example, 3D-pose may be represented as the triple {$\boldsymbol{e}_x$, $\boldsymbol{e}_y$, $\boldsymbol{r}_M$} defined in the caption of Fig. 4. The image parameters {$\boldsymbol{i}_x$, $\boldsymbol{i}_y$, $\boldsymbol{i}_c$} are perspective projections of {$\boldsymbol{e}_x$, $\boldsymbol{e}_y$, $\boldsymbol{r}_M$}. The reconstruction of 3D-pose means building {$\boldsymbol{e}_x$, $\boldsymbol{e}_y$, $\boldsymbol{r}_M$} by {$\boldsymbol{i}_x$, $\boldsymbol{i}_y$, $\boldsymbol{i}_c$}.

The methods by DeMenthon and Davis (1995) and by Abidi and Chandra (1995) deal with such a problem. Since they have good computational properties, the rest of 3D-pose estimation can follow either method. Nevertheless, we present here another scheme (perhaps, it is not new) just for methodical purposes: maybe, it not so good as mentioned methods, but it is very short and explicit.

Let e_A, e_B and e_C (e_C is omitted on Fig.4) be the unit vectors outgoing from *P* to the directions *PA*, *PB*, and *PC*, containing the end-points of i_x, i_y, and i_c, correspondingly. For known vectors i_x, i_y, i_c and the position of *P* [which is a known characteristic of the optical system], e_A, e_B and e_C may also be considered as known (we are omitting here their trivial explicit expressions). Representing r_M as $s\,e_C$, where *s* is unknown distance between *P* and the target center, the next equations hold obviously

$$e_x = t_s e_B - s e_C, e_y = u_s e_A - s e_C, \qquad (7)$$

where t_s and u_s are unknown factors depending on *s*.

The property of being unitary vector for e_x and e_y yields square equations for t_s and u_s: $1 = (e_x, e_x) = (e_y, e_y)$, where (•,•) is scalar product. Reducing it to $1 = t_s^2 - 2t_s s(e_B, e_C) + s^2 =$ $u_s^2 - 2u_s s(e_A, e_C) + s^2$, one has the roots

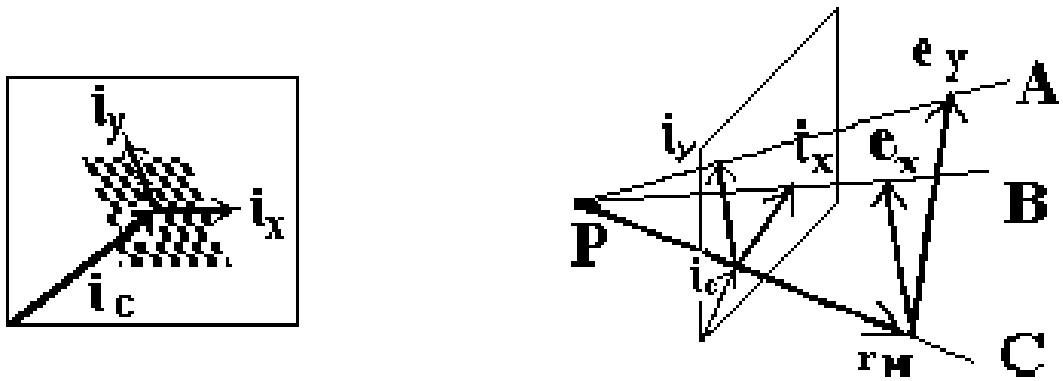
$$\begin{aligned} (t_s)_{1,2} &= s(e_B, e_C) \pm \sqrt{1 - s^2 + s^2 (e_B, e_C)^2}, \\ (u_s)_{1,2} &= s(e_A, e_C) \pm \sqrt{1 - s^2 + s^2 (e_A, e_C)^2} \end{aligned} \qquad (8)$$

where *s* is unknown parameter. Due to (7), orthogonality of e_x and e_y yields the equation for *s*:

$$\begin{aligned} 0 &= (t_s e_B - s e_C, u_s e_A - s e_C) = \\ & s^2 - s[t_s (e_B, e_C) + u_s (e_A, e_C)] + t_s u_s (e_B, e_A) \end{aligned} \qquad (9)$$

Finally, substitution of the roots (8) of t_s and u_s into (9) yields four polynomial equations for possible values of *s*. Real positive roots of these equations give, by means of (7-8), all options for { e_x, e_y, r_M }.

[Every reconstruction of 3D-pose by perspective image of a planar target is ambiguous due to the number of roots of these equations. To kill this ambiguity for a practical problem, for instance, one can use a few planar targets with different orientations.]



(a) **(b)**

Figure 4. (a) **Image frame with an image of the target**. The target image parameters $\{\boldsymbol{i}_x, \boldsymbol{i}_y, \boldsymbol{i}_c\}$ are depicted schematically inside the frame. (b) **The same** $\{\boldsymbol{i}_x, \boldsymbol{i}_y, \boldsymbol{i}_c\}$ **as projections**: Orthogonal vectors $\boldsymbol{e}_x$, $\boldsymbol{e}_y$ belong to the target plane, are attached to the target center, are parallel and of the same length with corresponding sides of the target square; $\boldsymbol{i}_x$ and $\boldsymbol{i}_y$ are planar perspective projections of $\boldsymbol{e}_x$ and $\boldsymbol{e}_y$. Vector $\boldsymbol{r}_M$ connects the center P of perspective projection with the target center

Remark: Expressing the Control Law in Sensor Space

Note that for an application in robot control, in fact, it is not obligatory to perform the inversion $\{\boldsymbol{i}_x, \boldsymbol{i}_y, \boldsymbol{i}_c\} \rightarrow \{\boldsymbol{e}_x, \boldsymbol{e}_y, \boldsymbol{r}_M\}$ developed above. Moreover, there is no need to find target image parameters $\{\boldsymbol{i}_x, \boldsymbol{i}_y, \boldsymbol{i}_c\}$. Instead, one can express the control goal directly in the spectral Sensor Space and introduce it in robot control loop. (See in this relation the work by Martrinet et al (1997), which, in particular, contains necessary references on Visual Servoing). For the problem under consideration, the sensor space can be formed as $R^{15} = R^3 \times R^3 \times R^3 \times R^3 \times R^3$, where each R^3 corresponds to possible values of a peak of the function $|G(\mathbf{v}_i)|$ and its $\mathbf{v}_i$. It means practically, that having some relation between 3D-pose and behavior of peaks in spectral domain, the aim of control is to force peaks to be in proper places and with proper magnitudes.

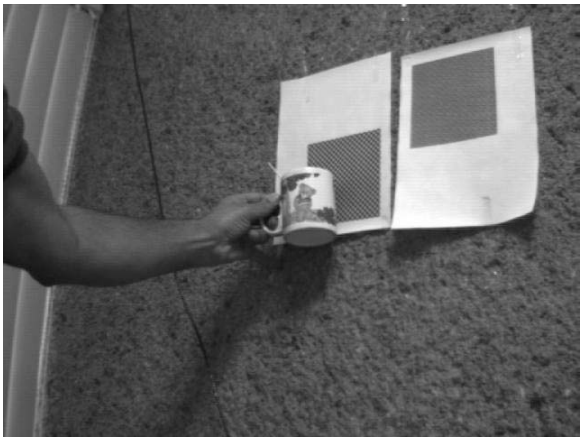
4. EXPERIMENTS AND DISCUSSION

Experiments

The necessary and sufficient condition, under which the processing of p.3 works, is a good extraction of non-zero high-energy points.

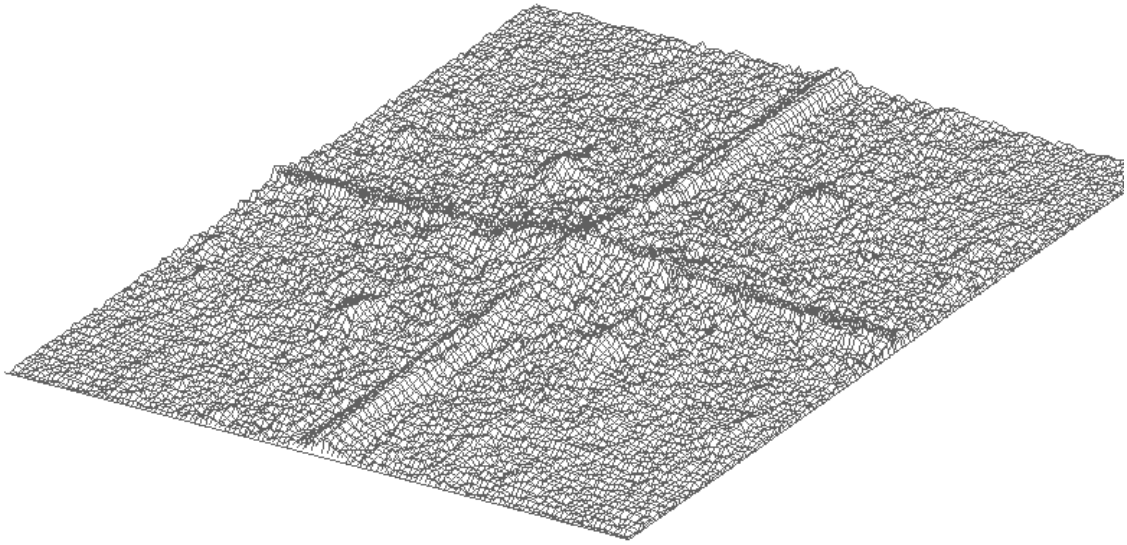
An experiment demonstrates this ability of the method. The Fig. 5[a] displays a photo that was performed by the standard camera of O_2™ workstation of Silicon Graphics. The left target on the photo has frequencies f_1=f_2=16 of the pattern function g corresponding to (3). The right one has f_1=f_2=32, but its pattern is hardly seen due to the resolution of camera. The Fig.5[b] cuts 128x128-fragment of this photo. The

log(1+|Fourier_Transform(*Image_fragment*)|) is plotted in Fig.5[c]. The four non-zero high-energy points are clearly seen on the plot. The method works well in spite of a considerable occlusion of the target and relatively bad light condition.



(a)

(b)



(c)

Figure 5. (a) - A real digital photo; (b) - its 128x128 fragment with the target image; (c) - the Fourier transform of the fragment.

Comparison with the Least Squares Estimates

Potential precision of the method presented in p.3 coincides with the least squares method (LSM) given by (1).

Indeed, FT is reversible and, hence, the FT of a target image contains the same information as the image itself. Then, after FT of an image of our planar target, all Fisher's information of the original pattern is hidden in the positions of FT-peaks and their values. Consequently, the method of p.3, based on the processing of the peaks, takes into consideration the whole information of the original image.

Hence, as far as LSM is asymptotically efficient, the presented method also has this property. Consequently, *in the class of single-view model-based estimators using planar model, the method reaches the best possible precision* increasing frequencies of the pattern.

Let compare the influence of non-uniformity on the presented method and LSM.

We already mentioned that high non-uniformity of a pattern is a strong obstacle for practical application of least squares. Indeed, for such a pattern, the least squares functional has many local extremums, and in addition, a slight distortion of image due to optical transformation of camera and secondary effects does not permit to reach a good accordance between an image $I(\mathbf{y})$ and its model $g(\mathbf{x}(\mathbf{y},\Omega))$ even for the case of exactly known vector Ω.

On the contrary, the presented method works in such a case. [Provided the control of zoom mentioned at the beginning of p.3.] So, in terms of the model $\{g, \mathbf{x}, I, G\}$ where g is variable and $\mathbf{x}, I, G$ are fixed, we may affirm to the following: *unlike the presented approach,* LSM *can not reach in practice the highest precision that the model permits.*

Let consider the influence of perturbations. We can interpret perturbations (integrating altogether all kinds of them: an image distortion, an occlusion of target, a partial shadow, etc.) as points in some functional space. The norm of an element of this space evaluates the magnitude of corresponding perturbation. For a pattern corresponding to (3), a perturbation transforms delta-wise FT-peaks into bell-shaped peaks. Irrespectively of the exact definition of this functional space and the norm, a perturbation affects FT-peaks in dependence of the value of its norm: A wider and lower bell responds to a perturbation with greater norm.

In these terms, the presented method is valid in such a range of values of a perturbation norm, which provides distinguishability of FT-peaks. We may say that the presented method is robust for perturbations belonging to this range of values of the norm.

Although we do not express explicitly this range, we can compare the presented method and LSM on the example of Fig. $5^{b\text{-}c}$. An occlusion like on Fig. 5^{b} surely destroys

estimation based on LSM. It is not the case for the presented method: a good extractability of the relevant FT-peaks on Fig. 5^c proves that this perturbation leaves the image inside the range of robustness.

Computational Aspect

We discuss here only the novel part of the presented algorithm, i.e., its first part producing $\{\boldsymbol{i}_x, \boldsymbol{i}_y, \boldsymbol{i}_c\}$.

Its massive part consists of few (4 or 5) operations of the Fourier transform. The complexity of fast FT of a $n \times n$-fragment is $O(n^2 \log n)$. For instance, for n=128 as in experiments of Fig. 1, 5, it yields about 10^5 operations per FR. This value is close to provide all necessary computations inside the picture acquisition cycle by a single processor. However, FT has excellent properties to be processed by a parallel computation, and there are FT-co-processors able to perform this routine part of the presented method practically instantly.

All the rest of algorithm requires a few thousand operations. So, for development of a real-time system based on the presented method, the computational complexity does not present any obstacle.

The presented method is straightforward: it does not require any improving of input image; it does not depend on extraction of local features (edgels, for example), or their posterior aggregation, etc. Such a property provides reliability of the method. On the other hand, if real technical factors will violate the mathematical model of p.2, it permits a relatively simple analysis of the influence of such factors.

An Extension of the Presented Method

The choice of the function g of (3) is based on properties of its FT. In fact, any pattern function with appropriate FT may be chosen. This property is that the FT must be zero almost everywhere except a few points of spectral plane, at which it must have delta-wise character. Choosing such a function in spectral domain, one can construct a new pattern function by means of the inverse FT of this function. [However, it should be taken in account, that the inverse may turn out to be a complex function!] Then, printing the inverse, one obtains a good target pattern itself. We can use it instead of g, with all the rest of processing of p.3 to be the same.

A Problem: Extension of the Approach to Non-Planar Targets

As it follows from the results of p.2, the slope-variation of a target contributes an additional Fisher's information. Theoretically, it can improve considerably the precision of 3D pose estimates based on a planar target.

For instance, one can imagine the model of shape of the target as a kind of a "fractal"

quasi-planar object similar to light reflectors of a car. An example of such a target is presented on Fig. 6. It can be designed to reach a very high slope-variation staying the target in a small volume.

However, the technique of p.3 is based on the supposition that the Fourier transform of a target image is just sum of a few δ-wise peaks. Otherwise [i.e., if as the intensity image of a target as its FT is rather complicated function], the technique proposed in p. 3 does not work.

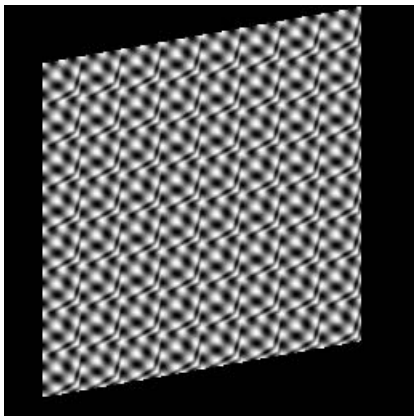


Figure 6. An example of non-planar target.

Our preliminary experiments with such non-planar targets gave a negative result: the FT of a target image has too many peaks. It does not permit a simple processing of target-images.

So, the problem is how to choose an appropriate target shape and its pattern for its FT to have a small number of peaks? Or, if it is impossible, how to process practically such kind of targets, which potentially can increase significantly the performance of 3D pose estimators?

[Of course, LSM corresponding to the formula (1) is a formal solution of such a problem. However, it works only for the case when the process of measurements corresponds exactly to the model $\{g, \mathbf{x}, I, G\}$ of p.2. Otherwise, by the same reasons as above, it does not work in practice for the case of high Fisher's information.]

5. CONCLUSION

The paper presents a new approach to a numerical comparison of precision of the 3D pose estimates for any patterns painted on a target. For the special case of planar targets, a method based on the Fourier transform if developed to reach the highest possible precision. It is algorithmically simple, fast, and robust. The approach can be applied to any mutual navigation problem that allows usage of a visual target on the tracked object, for example, to spacecraft docking, or to some problems of industrial robotics. The problem of tracking 3D-pose by processing of non-planar fractal targets with a high slope-variation is formulated.

6. ACKNOWLEDGMENTS

I am grateful to my colleague H. Moncayo for his help in experiments on real and simulated images and to A. Poznyak from

Cinvestav-IPN for a useful discussion on the topic of paper. The work was partially supported by CONACYT (Ref. 400200-5-1453PA).

REFERENCES

Abidi, M.A. and Chandra, T. "A New Efficient and Direct Solution for Pose Estimation Using Quadrangular Targets", *IEEE Trans. PAMI*, Vol.17, pp.534-538, May 1995.

Brigham, E.O. *The Fast Fourier Transform*, pp.20-21, Prentice-Hall,1974.

Brou, P., "Using the Gaussian Image to Find the Orientation of Objects", *IJRR*(3), No. 4, 1984, 89-125.

Cox, D.R. and Hinkley, D.V., *Theoretic Statistics*, pp. 254-270, Chapman and Hall, London, 1974.

DeMenthon D.F., Davis, L.S., "Model-Based Object Pose in 25 Lines of Code", *IJCV*, Vol. 15, n.1-2, pp.123-141, June 1995.

Dhome, M., Richetin, M., LaPreste, J.T., and Rives, G., "Determination of the Attitude of 3-D Objects from a Single Perspective View," *PAMI*(11), No. 12, December 1989, pp. 1265-1278.

Dunker, J., Hartmann, G., Stoehr, M., "Single View Recognition and Pose Estimation of 3D Objects Using Sets of Prototypical Views and Spatially Tolerant Contour Representations", *ICPR96*, 1996.

Egorov, Yu.V., Shubin, M.A. *Partial Differntial Equations I*. Foundations of the Classical Theory. (Encycl. of math. sci.; v.30). p.67. Springer, 1992.

Faugeras, O.D., Ayache, N., Faverjon, B., "A Geometric Matcher for Recognizing and Positioning 3-D Rigid Objects", *CAIA84*, pp.218-224, 1984.

Hel-Or, Y., Werman, M., "Pose Estimation by Fusing Noisy Data of Different Dimensions", *PAMI*(17), No. 2, February 1995, pp. 195-201. And: Correction: PAMI(17), No. 5, May 1995, pp. 544.

Jurvis, R.A. "A Perspective on Range Finding Technique for Computer Vision", *IEEE Trans. PAMI*, Vol. 5, No. 5, pp.122-139, 1983

Kang, S.B., and Ikeuchi, K., "The Complex EGI: A New Representation for 3-D Pose Determination", *PAMI*(15), No. 7, July 1993, pp. 707-721.

Khachaturov, G.A., Tsyganov, O.N., Ryzhkov, V.S., Stupin, K.N. "The Method and Device for Measurement of Position and Orientation of Visible Object by Processing of Visual Information", invention of the USSR No. 252388, *Bulletin of inventions of the USSR* (ISSN 0208-287X), 1987.

Khachaturov, G., “A Monocular 3D Pose Estimator Based on Fourier Technique”, Proceedings of III Iberoamerican Workshop on Pattern Recongnition, (Memorias del III Taller Iberoamericano de Reconocimiento de Patrones), March 23-27, 1998, pp.113-129, ICIMAF (Cuba), CIC IPN, IPN (Mexico), 1998.

Kriegman, D.J., "Computing Stable Poses of Piecewise Smooth Objects", *CVGIP*(55), No. 2, March 1992, pp.109-118.

Laurin D.G., Rioux M., "Three-Dimensional Object Tracking Using a Fourier Transform of Sine-Coded Range Images", in *Optical Engineering*, June 1995, Vol. 34, No. 6, pp. 1789-1798.

Lowe, D.G., *Perceptual Organization and Visual Recognition*. Kluwer Academic Publishers, 1985.

Lowe., D.G. “Fitting Parametrized Three-Dimensional Model to Images”, IEEE Trans. On Pattern Analysis and Machine Intelligence, Vol. 13, pp. 441-450, 1991.

Martinet P., Daucher N., Gallice J., Dhome M. "Robot Control Using 3D Monocular Pose Estimation", *Proc. of the Workshop on New Trends in Image Based Robot Servoing*, IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, IROS'97, pp 1-12, Grenoble, France, 7-11 September,1997.

Pratt, W.K *Digital image processing*, p.14, John Wiley \&Sons, Inc., 1978.

Tanaka, H.T., Ballard, D.H., Tsuji, S., and Curtiss, M., "Parallel Polyhedral Shape Recognition", *CVPR85*(491-496). (Osaka Univ. and Univ. of Rochester), 1985.

Vinther, S., Cipolla, R., "Active 3D Object Recognition Using 3D Affine Invariants," *ECCV94*(B:15-24),1994.

Yuan, J.S.C., “A General Photogrammetric Method for Determining Object Position and Orientation”, Vol. 5, pp. 129-142, 1989.