Pose Estimation Using Linear or Nonlinear Composite Correlation Filters and a Neural Network

María-Albertina Castro[a*], Yann Frauel[b**], Eduardo Tepichín[a***], Bahram Javidi[c****]
[a] Instituto Nacional de Astrofisica, Optica y Electrónica, Puebla, México.
[b] IIMAS, Universidad Nacional Autónoma de México, México.
[c] Electrical and Computer Eng. Dept., University of Connecticut, USA.

ABSTRACT

Cameras provide only bi-dimensional views of three-dimensional objects. These views are projections that change depending on the spatial orientation or pose of the object. In this paper we propose a technique to estimate the pose of a 3D object knowing only a 2D picture of it. The proposed technique explores both the linear and the nonlinear composite correlation filters in a combination with a neural network. We present results in estimating two orientations: in-plane and out-of-plane rotations within an 8 degree square range.

Keywords: Pattern recognition, pose estimation, composite correlation filters, neural networks, nonlinear correlation.

1. INTRODUCTION

The evaluation of the three-dimensional (3-D) orientation or pose of an object is an important issue in the area of object recognition. This problem has applications in various areas such as face recognition[1,2], robotic vision[3], and biomedical imaging[4,5]. Several approaches to the problem of pose estimation have been proposed[6-9]. In this paper, we present a technique to estimate the orientation of a 3-D object from a two-dimensional (2-D) view obtained with a camera. The reference object is known through a set of 2-D projections obtained from various points of view. The method we present was originally motivated by the work of Monroe and Juday[9]. The basic idea is to compare an unknown view of the object to all the known views. For greater efficiency, the reference views are combined into a composite correlation filter. Different weights are assigned to the views in such a way that the value of the correlation peak between the filter and the tested image depends on the orientation of the object under study. By using several composites filters, it is possible to simultaneously estimate several pose parameters, for instance several rotation angles for various axes. The problem is to retrieve these pose parameters from the correlation values obtained with the set of composite filters. In ref. 9, the relation is assumed to be linear and a set of training images is used to estimate the best fitting conversion. Although this technique gives acceptable results when a linear correlation is used to compare the unknown view to the known ones, it is largely inefficient when a nonlinear correlation is used. In the present paper, we propose to retrieve the pose parameters from the correlation values using a neural network. We show with experimental data that a two-layer backpropagation network improves the performance of the pose estimation for linear and, even more, for nonlinear correlation. The organization of this work is as follow, we first explain the construction of linearly weighted composite filters for two cases: linear correlation and *k*th-law nonlinear correlations. Section 3 presents the estimation of the pose for both the linear and nonlinear correlations using a linear method similar to the one proposed in Ref. 9. In section 4 we present the results of pose estimation using our proposal, a two-layer neural network. Section 5 is dedicated to the conclusions of this work.

[*] betina@inaoep.mx; phone: +52-222-266 3100; fax: +52-222-247 2940; Instituto Nacional de Astrofísica, Óptica y Electrónica, Apdo. Postal 51, Puebla 72000, (México).

[**] yann@leibniz.iimas.unam.mx; phone: +52-55-5622 3573; fax: +52-55-5622 3620; IIMAS, Universidad Nacional Autónoma de México, Apdo. postal 20-726 Admon. No. 20. Del. de Alvaro Obregón 01000 México, D.F.(México).

[***] tepichin@inaoep.mx; phone: +52-222-266 3100 Ext 1223; fax: +52 222 247 2940; Instituto Nacional de Astrofísica, Óptica y Electrónica, Apdo. Postal 51, Puebla 72000, (México).

[****] bahram@engr.uconn.edu; phone: +1-860-486 2867; fax: +1-860-486 2447; Electrical and Computing Engineering Dept., University of Connecticut, 260 Glenbrook Road, U-157, Storrs, CT 06269-2157 (USA).

2. CONSTRUCTION OF THE COMPOSITE FILTERS

We combine several images with known orientations into synthetic-discriminant-function (SDF) filters[10-12]. These reference images are called the construction set of the filter. The main advantage of a composite filter compared to a bank of single filters is the reduction of time in the processing step. A single correlation is sufficient to compare a given image with the whole set of rotated versions of the target to be recognized. In order to estimate the pose, the elements of the construction set must be weighted according to the amount of distortion.

2.1 Linear correlation

The expression for a basic SDF filter is

$$H = S\left(S^{+}S\right)^{-1} c \,, \tag{1}$$

where + stands for the transpose conjugate, S represents a matrix whose columns contain the pixel values of the reference images, namely the construction set; *c* denotes the design constraint. It denotes a vector containing the desired correlation peak values for each element of the construction set. Figure 1 shows some elements of the set we use here for composing a filter. We include images with two orientations: out-of-plane rotation θ and in-plane rotation ϕ. Specifically, the construction set consists of 9 images rotated from 0 to 8 degrees in steps of 4 degrees in both θ and ϕ. Figure 1a) shows the reference object, a 256 x 256 image of an F15 airplane. Figures 1b), and 1c) are respectively out-of-plane and in-plane rotated versions of the target.

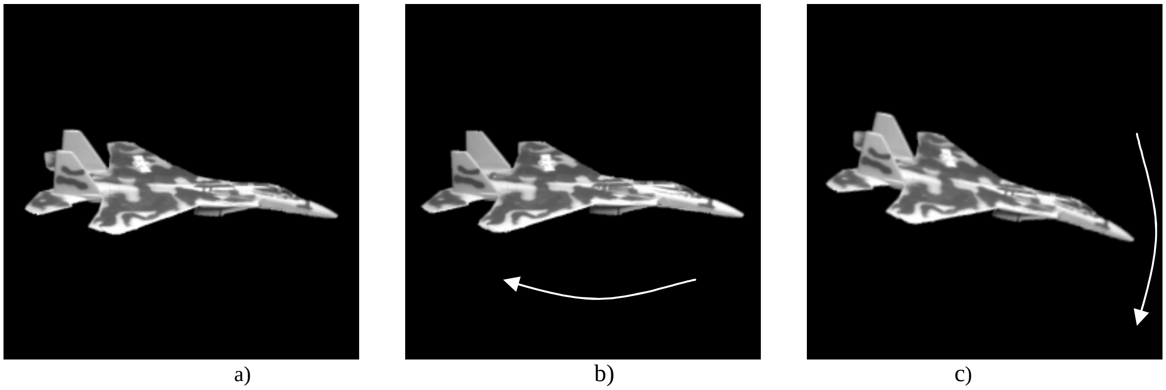
a) b) c)

Figure 1. Some elements of the construction set. a) Reference object, an F15 airplane, b) out-of-plane rotated, and c) in-plane rotated versions of the plane.

In order to create one SDF filter useful in the pose estimation problem, it is necessary to assign different weights to each image of the construction set. We choose a linear relation between the poses and the correlation peak values *c*. That is, if the filter is built with just one distortion parameter, then the *c* values fall into a line. As we want to estimate two distortion parameters, out-of-plane rotation θ, and in-plane rotation ϕ, the correlation peak values *c* for this set must fall into a plane. Considering that we need to retrieve three parameters: two orientation angles plus a recognition flag, then we use 3 SDF filters containing the same construction set but with different weights in order to create 3 different planes. That is, for a given image of the construction set the correlation peak value in one SDF filter will be different than the ones in the other two filters. So that when this image is presented as input and correlated with the 3 SDF filters, we obtain three different correlation-peak values. Then, knowing the definition of the 3 planes we can retrieve the orientation information as well as the recognition flag just by solving a 3 x 3 equation system. The term pose gathers, besides the orientation, the recognition flag. That is, for an image I1, element of the construction set, with orientation (0, 8), its pose is represented by

$$p1 = [\theta 1 \quad \phi 1 \quad 1]^{\mathrm{T}} = [0 \quad 8 \quad 1]^{\mathrm{T}}, \tag{2}$$

where the third element equal to 1 stands for the true targets class. This is done in order to achieve the recognition of the object at the same time as the pose estimation. In this way, the retrieved orientation will only make sense if the studied object gets a quantity close to 1 in the third element of the pose vector.

To construct the 3 SDF filters with correlation peaks falling into different planes we denote with *T* the 3x3 matrix that contains, by row, the coefficients that define each plane – see appendix for details – and with *Pconst* the 3x9 matrix containing in a vector form the poses of the 9 images of the construction set. Then, all the correlation peak values needed for generating these filters are computed by:

$$C_{const} = T \;\; P_{const} \;\; . \tag{3}$$

The resulting matrix C_{const} contains by row the correlation peak values for the 9 images of the construction set, for each plane. SDF1 is generated by taking the first row of C_{const} as the constraint *c* required in Eq. (1). SDF2 and SDF3 are generated in the same way, but now using rows 2 and 3 as *c*, respectively.

2.2 Nonlinear correlation

The problem of pose estimation, as we are stating here, deals first with the issue of the proper recognition of the object by itself. Our interest in considering nonlinear correlations is that they have proved to have better discrimination capabilities than the linear one when similar object are presented [13,14]. Nonlinear correlations apply non-linearities in the Fourier domain. Specifically the nonlinear *k*th-law correlators raise to the *k*-th power the magnitudes of the Fourier spectra of both the analyzed image and the filter, while keeping the phase information with no change. That is, if we denote by $\tilde{S}$ the matrix which contains in a vector form all the Fourier transforms of the reference images then

$$\tilde{S}^{k} = \left|\tilde{S}\right|^{k} \exp\left[i\theta_{\tilde{S}}\right], \qquad\qquad 0 \le k \le 1. \tag{4}$$

The value of *k* controls the strength of the nonlinearity. For *k*=1 a linear or conventional correlation is performed, whereas *k*=0 sets the magnitudes to unity for all the frequencies, which leads to a correlation based only on phase information, resulting in a highly discriminant system. Intermediate values of *k* permit to vary the features of the correlator, such as its discrimination capabilities. Now, the modified expression for constructing one *k*th-law SDF filter[15] is,

$$\tilde{H}^{k} = \tilde{S}^{k}\left(\tilde{S}^{k+}\tilde{S}^{k}\right)^{-1} c\,. \tag{5}$$

Again *c* denotes the vector that contains the desired correlation peak value for each reference image and we choose it to have the same (real) values than for the linear SDF filters, which means that we are defining again a linear relation between correlation peak values and orientation.

3. ESTIMATION OF THE POSE PARAMETERS: LINEAR FIT

3.1 Linear correlation

The retrieval stage is based on the known information about the relationship between the correlation peak values for the training objects and their orientation. This information is kept in the definition of the three planes represented by matrix *T*. Subsequently, when an image is presented to the system, it is correlated with SDF1, SDF2 and SDF3 filters and we obtain three correlation peak values, c1, c2 and c3 respectively, one for each filter. Now it is straightforward to deduce the pose parameters for a tested image as:

$$p_{estim} = T^{-1}\begin{bmatrix} C1 \\ C2 \\ C3 \end{bmatrix} = \begin{bmatrix} \theta_{estim} \\ \phi_{estim} \\ RF \end{bmatrix}, \tag{6}$$

where T^{-1} denotes the inverse of matrix *T*. As already mentioned the third element of the pose estimation vector contains the recognition flag. When it is 1or roughly 1, then the tested image is of the true class and the two first elements of the pose estimation vector indicate the amount of out-of-plane rotation θ and in-plane rotation ϕ of the identified object. Figure 2 shows the pose estimation results for 81 rotated images from 0 to 8 degrees in steps of one degree in both θ, and ϕ. If the dashed lines were perfectly straight passing all the way through the dots, then the estimation would be exact. The error analysis for this procedure shows in θ a standard deviation of 0.280 degrees and a maximum error in degrees of 0.854, while in ϕ the values are 0.162 degrees and 0.343 degrees, respectively.

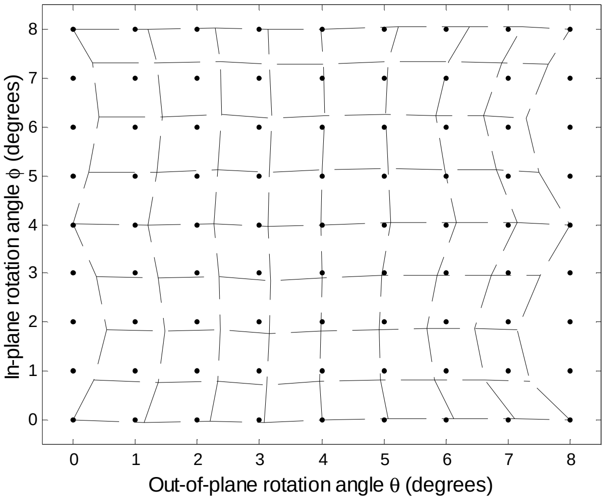

Figure 2. Pose estimation results by using linear SDF filters and the inverse matrix procedure (Eq. 6).

These results show that the retrieval of the pose for rotated images that are included in the filters is very accurate, but is not for the rest. This implies that the real relationship between the correlation peak values and the orientation is not perfectly linear. Ref. 9 proposed alternatively to redefine the matrix *T* by performing a least square fit in an evaluation process where a set of images - set2 - with known orientations were tested. The testing set Set2 contains the 9 images of the construction set plus another 16 rotated images that are not included in the design of the filters. Explicitly Set2 contains 25 images from 0 to 8 in step of 2 degrees in both θ and ϕ. The error analysis with this procedure presented in θ a standard deviation of 0.202 degrees and a maximum error of 0.641 degrees. Parameter ϕ gave a standard deviation of 0.112 degrees and a maximum error of 0.201 degrees. This technique requires the inversion of matrix *T* after the least-square fit. We propose to approximate directly the function that transforms the correlation vector into the pose vector. We begin with using a one layer neural network that we train with the same testing set, Set2. The results are shown on figure 3. The error analysis gives a standard deviation for out-of-plane rotation estimation θ of 0.203 degrees and a maximum error in degrees of 0.664 and a standard deviation for in-plane rotation estimation ϕ of 0.112 and a maximum error of 0.195 degrees.

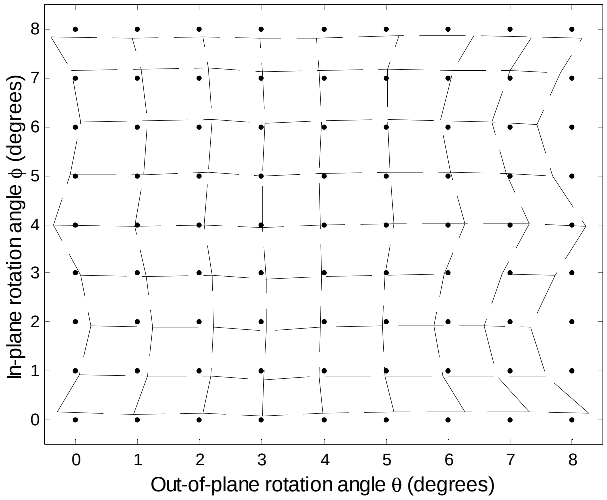

Figure 3. Pose estimation results by using linear SDF filters and a neural network with one layer.

Comparing the error analysis, it is evident that the results are similar to the ones obtained with the method in Ref. 9. This is because if we do not take into account the final threshold operation, a one layer neural network is linear by definition[16] and can therefore be modeled as a mere linear transformation. The only advantage of this method with respect to the one proposed in Ref. 9 is that we obtain directly the pose estimation, that is, we do not need to reverse the process. However, we can still improve the results by modifying the neural network (see section 4).

3.2 Nonlinear correlation

With the aim of a superior performance in term of discrimination capabilities, we now turn to nonlinear correlations. In this context when an image is presented to the system, we perform a *k*th-law nonlinear correlation of this image with each *k*th-law nonlinear SDF filter. The chosen value of *k* is 0.5 in order to have an intermediate discrimination. Again we use a one layer neural network that we train with the same testing set, Set2. Figure 4 presents the results in the pose estimation for the same square portion of 8 degrees in both θ and ϕ.

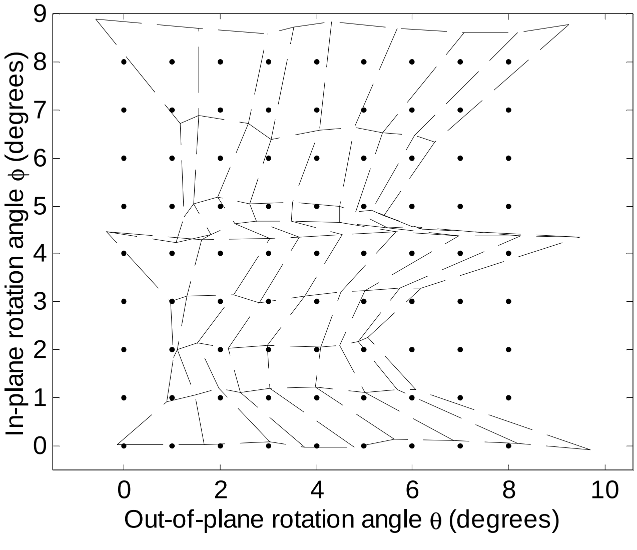

Figure 4. Pose estimation results by using nonlinear SDF filters and a neural network with one layer.

It is evident from figure 4 that this procedure is not good at all, so we did not even perform the error analysis for this case.

4. ESTIMATION OF THE POSE PARAMETERS: TWO-LAYER NEURAL NETWORK

The previous section has shown that the relationship between correlation peak values and orientation is not really linear for none of the cases. With the intention of getting rid of the linear limitation of the previously proposed neural network, we add a second layer. We train this two-layer feedforward network using the backpropagation rule[17] and we explore the two cases: feeding the neural network with correlation peak values obtained first from linear correlation and second from nonlinear *k*th-law correlations.

4.1 Linear correlation

Figure 5 shows a schematic diagram of the net we use here. The hidden layer contains 20 neurons and the output layer 3, each of which corresponds to one parameter to estimate. In order to find out the proper weights for each neuron we train the network by taking as inputs the correlation peak values obtained for each of the 3 linear correlation SDF filters for each image of the testing set, Set2. Their known pose parameters were the desired outputs.

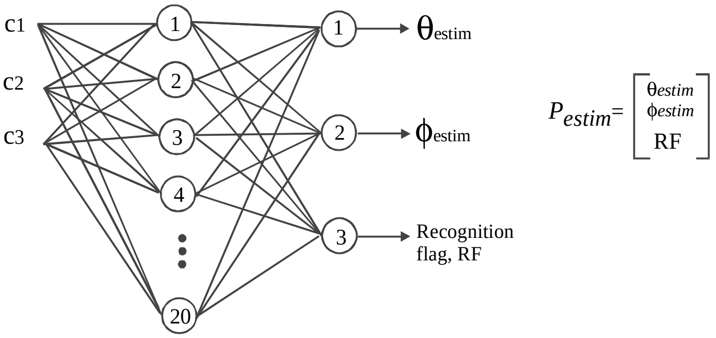
c1
c2
c3
1
2
3
4
20
1
2
3
θ_{estim}
ϕ_{estim}
Recognition flag, RF
$P_{estim} = \begin{bmatrix} \theta_{estim} \\ \phi_{estim} \\ RF \end{bmatrix}$

Figure 5. The proposed neural network to estimate the pose. A two-layer feedforward backpropagation neural network.

The learning stage uses 100 epochs. The number of epochs, as well as the number of hidden neurons, has been found heuristically. Figure 6 shows the pose estimation results. It is obvious in this figure that the estimation is improved compared to a single layer network.

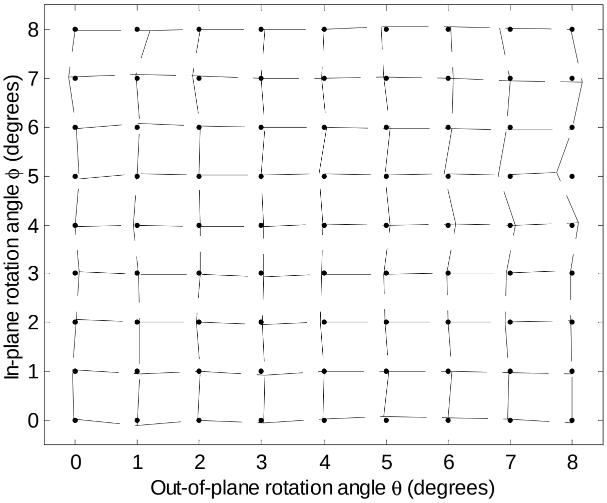

Figure 6. Pose estimation results using linear SDF filters and the proposed two-layer neural network.

For this last procedure the error analysis presents a typical standard deviation for the out-of-plane rotation estimation of 0.065 degrees and a maximum error of 0.235 degrees. As for in-plane rotation estimation we find a typical standard deviation of 0.034 degrees and a maximum error of 0.106 degrees.

4.2 Nonlinear correlation

For the case of nonlinear SDF filters, we use an ANN similar to the one used in section 4.1 except that this one has 40 hidden neurons and that the training stage lasts for only 30 epochs. Now we obtain the pose estimation vector by feeding this neural network with the correlation peak values obtained from correlating the input image with the 3 nonlinear *k*th-law filters. Here also the parameters have been found heuristically. Figure 7 presents a typical example of pose estimation for the evaluation set. The error for the out-of-plane rotation θ typically has a standard deviation of 0.25 degrees and a maximum of 0.97 degree. The typical standard for the in-plane rotation ϕ is 0.07 degrees and a maximum of 0.3 degrees of error.

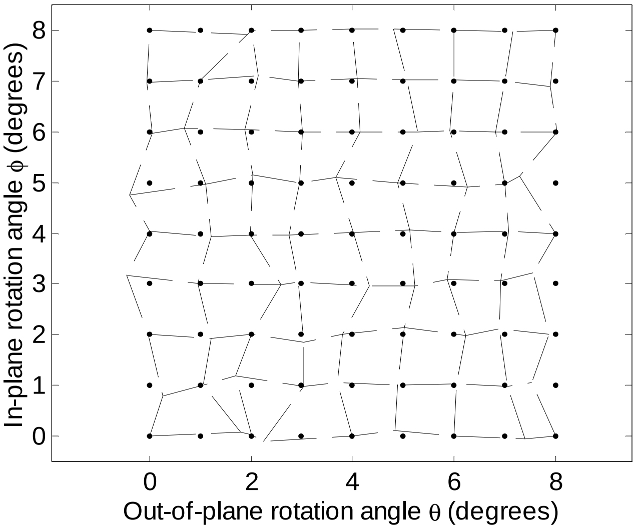

Figure 7. Pose estimation results using *k*th-law nonlinear SDF filters and the proposed two-layer neural network.

If we just observe figures 6 and 7 we can infer that these last results with nonlinear correlation are less accurate than the ones obtained with the two-layer neural network and linear correlation. They are nevertheless acceptable considering that the recognition capability has been improved. Now if we compare the estimation results presented for the two nonlinear correlation cases – figures 4 and 7 – we can say that the ones obtained with the two-layer neural network are a lot better than the first ones provided by least-square fitting.

5. CONCLUSIONS

In this paper we described a technique to estimate the pose of a 3-D object knowing only a 2-D picture of it. This technique combines synthetic-discriminant-function correlation filters and a neural network. We considered both linear and *k*th-law nonlinear SDF filters. We showed that a good tradeoff between the discrimination in the recognition process and the pose estimation is a combination of 3 *k*th-law nonlinear SDF filters and a two-layer backpropagation neural network. The pose estimation was performed for out-of-plane rotations and in-plane rotations for 81 rotated images within an 8 degree square range. It also possible to estimate more than two distortion parameters. The required number of SDF filters is equal to the number of distortion parameters plus one. This last one is related to the recognition flag.

ACKNOWLEDGEMENTS

This work was partially supported by Consejo Nacional de Ciencia y Tecnología, of México (CONACyT), under the project I39180-E.

REFERENCES

1. T. Horprasert, Y. Yacoob, L.S. Davis, "Computing 3-D head orientation from a monocular image sequence," *Proc. SPIE* **2962**, pp. 244-252, 1997.
2. I. Shimizu, Z. Zhang, S. Akamatsu, K. Deguchi, "Head pose determination from one image using a generic model," *Proc. 3rd IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp. 100-105, 1998.
3. W. Wilson, "Visual servo control of robots using Kalman filter estimates of robot pose relative to work-pieces," in *Visual Servoing*, K. Hashimoto, ed., pp. 71-104, World Scientific, 1994.
4. M. Bajura, H. Fuchs, R. Ohbuchi, "Merging virtual objects with the real world: seeing ultrasound imagery within the patient," *Proc. SIGGRAPH*, pp. 203-210, 1992.
5. N. Ezquerra, R. Mullick, "An approach to 3D pose determination," *ACM Trans. on Graphics* **15**, pp. 99-120, 1996.
6. R.M. Haralick, C.-N. Lee, K. Ottenberg, M. Nolle, "Review and analysis of solutions of the three point perspective pose estimation problem," *Int. Journal of Computer Vision* **13**, pp. 331-56, 1994.
7. C.-P. Lu, G.D. Hager, E. Mjolsness, "Fast and globally convergent pose estimation from video images," *IEEE Trans. on Pattern Anal. and Machine Intell.* **22**, pp. 610-622, 2000.
8. D.P. Huttenlocher, S. Ullman, "Recognizing solid objects by alignment with an image," *Int. Journal of Computer Vision* **5**, pp. 195-212, 1990.
9. S.E. Monroe Jr., R.D. Juday, "Multidimensional synthetic estimation filter,*" Optical Information-Processing Systems and Architectures II, SPIE* **1347**, pp. 179-185, 1990.
10. H. J. Caulfield, W. T. Maloney, "Improved discrimination in optical character recognition," *Appl. Opt.*, **8**, pp. 2354-2356, 1969.
11. H. J. Caulfield, "Linear combinations of filters for character recognition: a unified treatment," *Appl. Opt.*, **19**, pp. 3877-3879, 1980.
12. D. Casasent, "Unified synthetic discriminant function computational formulation," *Appl. Opt.*, **23**, pp. 1620-1627, 1984.
13. B. Javidi, "Nonlinear joint power spectrum based optical correlation," *Appl. Opt.*, **28**, pp.2358-2367, 1989.
14. B. Javidi, J. L. Horner, A. Fazlollahi, J. Li, "Illumination-invariant pattern recognition with a binary nonlinear joint transform correlator using spatial frequency dependent threshold function," *Proc. SPIE*, **2026**, pp. 100-106, 1993.
15. B. Javidi, D. Painchaud, "Distortion-invariant pattern recognition with Fourier-plane nonlinear filters," *Appl. Opt.*, **35**, pp. 318-331, 1996.
16. R. Beale and T. Jackson, *Neural Computing: An introduction* ,IOP Publishing Ltd, 1992.
17. D.E. Rumelhart, G.E. Hinton, R.J. Williams, *Learning internal representations by error propagation, in Parallel distributed processing*, D.E. Rumelhart, J.L. McClelland, eds., MIT Press, Cambridge, MA, 1986.

APPENDIX

We define the relation between the correlation peak value *c* and the pose *p* to be linear, so that for a particular image with pose $p = (\theta, \phi, 1)^T$ the correlation peak values for each of the three filters will be given by

$$c = T\ p\ , \tag{A1}$$

where matrix T contains by row the coefficients that defines the planes. That is,

$$\begin{bmatrix} c1 \\ c2 \\ c3 \end{bmatrix} = \begin{bmatrix} a1 & b1 & d1 \\ a2 & b2 & d2 \\ a3 & b3 & d3 \end{bmatrix} \begin{bmatrix} \theta \\ \phi \\ 1 \end{bmatrix}. \qquad \text{(A2)}$$

Now, in order to find these coefficients, we use three points or poses (θ, ϕ,1): (0,0,1) , (0, 8, 1) y (8, 8, 1). Each filter will give a correlation peak value of 1 at one of these three points and it is designed to have 0.8 at the other two points. Explicitly matrix *T* is found as

$$T = \begin{bmatrix} 1 & 0.8 & 0.8 \\ 0.8 & 1 & 0.8 \\ 0.8 & 0.8 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 8 \\ 0 & 8 & 8 \\ 1 & 1 & 1 \end{bmatrix}^{-1}. \qquad \text{(A3)}$$