PERGAMON

Pattern Recognition 32 (1999) 1783–1799

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

Learning affine transformations

George Bebis[a,*], Michael Georgiopoulos[b], Niels da Vitoria Lobo[c], Mubarak Shah[c]

[a]*Department of Computer Science, University of Nevada, Reno, NV 89557, USA*
[b]*Department of Electrical & Computer Engineering, University of Central Florida, Orlando, FL 32816, USA*
[c]*Department of Computer Science, University of Central Florida, Orlando, FL 32816, USA*

Received 19 February 1998; received in revised form 20 November 1998; accepted 20 November 1998

Abstract

Under the assumption of weak perspective, two views of the same planar object are related through an affine transformation. In this paper, we consider the problem of training a simple neural network to learn to predict the parameters of the affine transformation. Although the proposed scheme has similarities with other neural network schemes, its practical advantages are more profound. First of all, the views used to train the neural network are not obtained by taking pictures of the object from different viewpoints. Instead, the training views are obtained by sampling the space of affine transformed views of the object. This space is constructed using a single view of the object. Fundamental to this procedure is a methodology, based on singular-value decomposition (SVD) and interval arithmetic (IA), for estimating the ranges of values that the parameters of affine transformation can assume. Second, the accuracy of the proposed scheme is very close to that of a traditional least squares approach with slightly better space and time requirements. A front-end stage to the neural network, based on principal components analysis (PCA), shows to increase its noise tolerance dramatically and also to guides us in deciding how many training views are necessary in order for the network to learn a good, noise tolerant, mapping. The proposed approach has been tested using both artificial and real data.

Keywords: Object recognition; Artificial neural networks

1. Introduction

Affine transformations have been widely used in computer vision and particularly, in the area of model-based object recognition [1–5]. Specifically, they have been used to represent the mapping from a 2-D object to a 2-D image or to approximate the 2-D image of a planar object in 3-D space and it has been shown that a 2-D affine transformation is equivalent to a 3-D rigid motion of the object followed by orthographic projection and scaling (weak perspective). Here, we consider the case of real planar objects, assuming that the viewpoint is arbitrary. Given a known and an unknown view of the same planar object, it is well known that under the assumption of weak perspective projection [1,2], the two views are related through an affine transformation. Given the point correspondences between the two views, the affine transformation which relates the two views can be computed by solving a system of linear equations using a least-squares approach (see Section 3).

In this paper, we propose an alternative approach for computing the affine transformation based on neural

*Corresponding author.
E-mail address: bebis@cs.unr.edu (G. Bebis)

0031-3203/99/$20.00
PII: S0031-3203(98)00178-2

networks. The idea is to train a neural network to predict the parameters of the affine transformation using the image coordinates of the points in the unknown view. A shorter version of this work can be found in [6]. There were two main reasons which motivated us in using neural networks to solve this problem. First of all, it is interesting to think of this problem as a learning problem. Several other approaches have also been proposed [7,8] which treat similar problems as learning problems. Some of the issues that must be addressed within the context of this formulation are: (i) how to obtain the training views, (ii) how many training views are necessary, (iii) how long it takes for the network to learn the desired mapping, and (iv) how accurate are the predictions. Second, we are interested in comparing the neural network approach with traditional least squares used in the computation of the affine transformation. Given that neural networks are inherently parallelizable, the neural network approach might be a good alternative if it turns out that it is as accurate as traditional least-squares approaches. In fact, our experimental results demonstrate that the accuracy of the neural network scheme is as good as that of traditional least squares with the proposed approach having slightly less space and time requirements.

There are three main steps in the proposed approach. First, the ranges of values that the parameters of affine transformation can assume are estimated. We have developed a methodology based on singular-value decomposition (SVD) [9] and interval arithmetic (IA) [10] for this. Second, the space of parameters is sampled. For each set of sampled parameters, an affine transformation is defined which is applied on the known view to generate a new view. We will be referring to these views as *transformed views*. The transformed views are then used to train a single-layer neural network (SL-NN) [11]. Given the image coordinates of the points of the object in the transformed view, the SL-NN learns to predict the parameters of the affine transformation that align the known and unknown views. After training, the network is expected to *generalize*, that is, to be able to predict the correct parameters for transformed views that were never exposed to it during training.

The proposed approach has certain similarities with two other approaches [7,8]. In [7], the problem of approximating a function that maps any perspective 2-D view of a 3-D object to a standard 2-D view of the same object was considered. This function is approximated by training a generalized radial basis functions neural network (GRBF-NN) to learn the mapping between a number of perspective views (training views) and a standard view of the model. The training views are obtained by sampling the viewing sphere, assuming that the 3-D structure of the object is available. In [8], a linear operator is built which distinguishes between views of a specific object and views of other objects (orthographic projection is assumed). This is done by mapping every view of the object to a vector which uniquely identifies the object. Obviously, our approach is conceptually similar to the above two approaches, however, there are some important differences. First of all, our approach is different in that it does not map different views of the object to a standard view or vector but it computes the parameters of the transformation that align known and unknown views of the same object. Second, in our approach, the training views are not obtained by taking different pictures of the object from different viewpoints. Instead, they are affine transformed views of the known view. On the other hand, the other approaches can compute the training views *easily* only if the structure of the 3-D object is available. Since this is not always available, the training views can be obtained by taking different pictures of the object from various viewpoints. However, this requires more effort and time (edges must be extracted, interest point must be identified, and point correspondences across the images must be established). Finally, our approach does not consider both the x- and y-coordinates of the object points during training. Instead, we simplify the scheme by decoupling the coordinates and by training the network using only one of the two (the x-coordinates here). The only overhead from this simplification is that the parameters of the affine transformation must be computed in two steps.

There are two comments that should be made at this point. First of all, the reason that a SL-NN is used is because the mapping to be learned is linear. This should not be considered, however, as a trivial task since both the input (image) and output (parameter) spaces are continuous. In other words, special emphasis should be given on the training of the neural network to ensure that the accuracy in the predictions is acceptable. Second, it should be clear that the proposed approach assumes that the point correspondences between the unknown and known views of the object are given. That was also the case with [7,8]. Of course, establishing the point correspondences between the two views is the most difficult part in solving the recognition problem. Unless the problem to be solved is very simple, using the neural network approach without any a priori knowledge about possible point correspondences is not efficient in general (see [12,13] for some example). On the other hand, combining the neural network scheme with an approach which returns possible point correspondences will be ideal. For example, we have incorporated the proposed neural network scheme in an indexing-based-object recognition system [14]. In this system, groups of points are chosen from the unknown view and are used to retrieve hypotheses from a hash table. Each hypothesis contains information about a group of object points as well as information about the order of the points in the group. This information can be used to place the points from the unknown view into a correct order before they are fed to the network.

There are various issues to be considered in evaluating the proposed approach such as, how good is the mapping computed by the SL-NN, what is the discrimination power of the SL-NNs, and how accurate are the predictions of the SL-NN assuming noisy and occluded data. These issues have been considered in Section 5. The quality of the approximated mapping depends rather on the number of training views used to train the neural network. The term "discrimination power" means the capability of a network to predict wrong transformation parameters, assuming that it is exposed to views which belong to different objects than the one whose views were used to train the network (*model-specific networks*). Our experimental results show that the discrimination power of the networks is very good. Testing the noise tolerance of the networks, we found that it was rather poor. However, we were able to account for it by attaching a front-end stage to the inputs of the SL-NN. This stage is based on principal components analysis (PCA) [15] and its benefits are very important. Our experimental results show a dramatic increase in the noise tolerance of the SL-NN. We have also noticed some improvements in the case of occluded data, but the performance degrades rather rapidly even with 2–3 points missing. In addition, it seems that PCA can guide us in deciding how many training views are necessary in order for the SL-NN to learn a "good", noise tolerant, mapping.

The organization of the paper is as follows: Section 2 presents a brief review of the affine transformation. Section 3 presents the procedure for estimating the ranges of values that the parameters of the affine transformation can assume. In Section 3, we describe the procedure for the generation of the training views and the training the SL-NNs. Our experimental results are given in Section 4 while our conclusions are given in Section 5.

2. Affine transformations

Let us assume that each object is characterized by a list of "interest" points $(p'_1, p'_2, \ldots, p'_m)$, which may correspond, for example, to curvature extrema or curvature zero-crossings [16]. Let us now consider two images of the same planar object, each one taken from a different viewpoint, and two points $p = (x, y)$, $p' = (x', y')$, one from each image, which are in correspondence; then the coordinates of p can be expressed in terms of the coordinates of p', through an affine transformation, as follows:

$$p = Ap' + b, \tag{1}$$

where A is a non-singular 2×2 matrix and b is a two-dimensional vector. A planar affine transformation can be described by six parameters which account for translation, rotation, scale, and shear. Writing Eq. (1) in terms of the image coordinates of the points we have

$$x = a_{11}x' + a_{12}y' + b_1, \tag{2}$$

$$y = a_{21}x' + a_{22}y' + b_2. \tag{3}$$

The above equations imply that given two different views of an object, one known and one unknown, the coordinates of the points in the unknown view can be expressed as a linear combination of the coordinates of the corresponding points in the known view. Thus, given a known view of an object, we can generate new, affine transformed views of the same object by choosing various values for the parameters of the affine transformation. For example, Fig. 1b and d shows affine transformed views of the planar object shown in Fig. 1a. These views were generated by transforming the known view using the affine transformations shown in Table 1. Thus, for any affine transformed view of a planar object, there is a point in the 6-dimensional space of 2-D affine transformations which corresponds to the transformation that can bring the known and unknown views into alignment (in a least-squares sense).

3. Estimating the ranges of the parameters

Given a known view I' and an unknown affine transformed view I of the same planar object, as well as the point correspondences between the two views, there is an affine transformation that can bring I' into alignment with I. In terms of equations, this can be written as follows:

$$I' \begin{bmatrix} A \\ b \end{bmatrix} = I \tag{4}$$

or

$$\begin{bmatrix} x'_1 & y'_1 & 1 \\ x'_2 & y'_2 & 1 \\ \cdots & \cdots & \cdots \\ x'_m & y'_m & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \\ b_1 & b_2 \end{bmatrix} \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \cdots & \cdots \\ x_m & y_m \end{bmatrix}, \tag{5}$$

where $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$ are the coordinates of the points corresponding to I, while (x'_1, y'_1), $(x'_2, y'_2), \ldots, (x'_m, y'_m)$ are the coordinates of the points corresponding to I'. We assume that both views consist of the same number of points. To achieve this, we consider only the points that are common in both views. Eq. (5) can be split into two different systems of equations, one for the x-coordinates and one for the y-coordinates

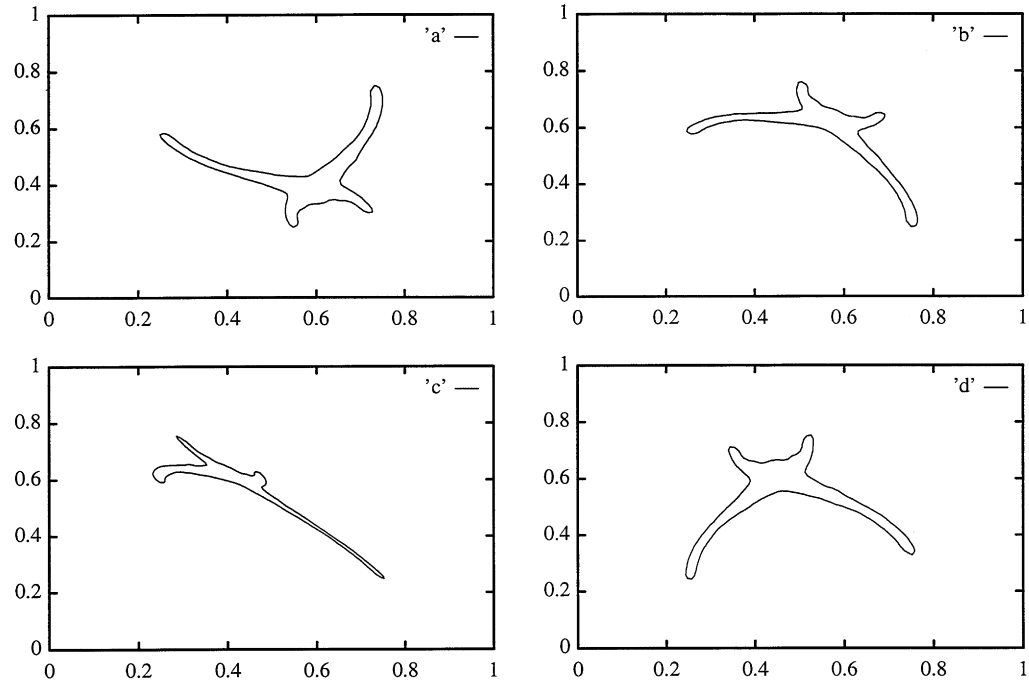

Fig. 1. (a) A known view of a planar object; (b)–(d) some new, affine transformed, views of the same object generated by considering the affine transformations shown in Table 1.

Table 1
The affine transformations used to generate Fig 1b and d

Parameters of the affine transformations			
Parameters	Fig. 1b	Fig. 1c	Fig. 1d
a_{11}, a_{12}, b_1	0.992 0.130 −0.073	−1.010 −0.079 1.048	0.860 0.501 −0.255
a_{21}, a_{22}, b_2	−0.379 −0.878 1.186	0.835 −0.367 0.253	0.502 −0.945 0.671

of I, as follows:

$$\begin{bmatrix} x'_1 & y'_1 & 1 \\ x'_2 & y'_2 & 1 \\ \cdots & \cdots & \cdots \\ x'_m & y'_m & 1 \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{12} \\ b_1 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_m \end{bmatrix}, \tag{6}$$

$$\begin{bmatrix} x'_1 & y'_1 & 1 \\ x'_2 & y'_2 & 1 \\ \cdots & \cdots & \cdots \\ x'_m & y'_m & 1 \end{bmatrix} \begin{bmatrix} a_{21} \\ a_{22} \\ b_2 \end{bmatrix} = \begin{bmatrix} y_2 \\ y_2 \\ \cdots \\ y_m \end{bmatrix}. \tag{7}$$

Using matrix notation, we can rewrite Eqs. (6) and (7) as $Pc_1 = p_x$ and $Pc_2 = p_y$ correspondingly, where P is the matrix formed by the x- and y-coordinates of the points of the known view I', c_1 and c_2 represent the parameters of the transformation, and p_x, p_y, are the x- and y-coordinates of the unknown view I. Both Eqs. (6) and (7) are overdetermined (the number of points is usually larger than the number of parameters, that is, $m \gg 3$), and can be solved using a least-squares approach such as SVD [9]. Using SVD, we can factor the matrix P as follows:

$$P = UWV^{\mathrm{T}}, \tag{8}$$

where both U and V are orthogonal matrices ($m \times 3$ and 3×3 size correspondingly), while W is a diagonal matrix (3×3 size) whose elements w_{ii} are always non-negative and are called the singular values of P. The solutions of

Eqs. (6) and (7) are then given by $c_1 = P^+ p_x$ and $c_2 = P^+ p_y$, where P^+ is the pseudo-inverse of P which is equal to $P^+ = VW^+U^T$, where W^+ is also a diagonal matrix with elements $1/w_{ii}$, if w_{ii} greater than zero (or a very small threshold in practice), and zero otherwise. Taking this into consideration, the solutions of Eqs. (6) and (7) are given by [14]

$$c_1 = \sum_{i=1}^{3} \left(\frac{u_i^T p_x}{w_{ii}} \right) v_i, \tag{9}$$

$$c_2 = \sum_{i=1}^{3} \left(\frac{u_i^T p_y}{w_{ii}} \right) v_i, \tag{10}$$

where u_i denotes the ith column of matrix U and v_i denotes the ith column of matrix V. Of course, the sum should be restricted for those values of i for which $w_{ii} \neq 0$.

To determine the range of values for the parameters of affine transformation, we first assume that the image of the unknown view has been scaled so that the x- and y-coordinates of the object points belong within a specific interval. This can be done, for example, by mapping the image of the unknown view to the unit square. In this way, its x- and y-coordinates are mapped to the interval [0,1]. To determine the range of values for the parameters of affine transformation, we need to consider all the possible solutions of Eqs. (6) and (7), assuming that the components of the vectors on the right-hand side of the equations are always restricted to belong to the interval [0,1]. Trying to calculate the range of values using mathematical inequalities did not yield "good" results in the sense that the novel views corresponding to the ranges computed were not spanning the whole unit square but only a much smaller sub-square within it. Therefore, we consider interval arithmetic [10]. In IA, each variable is actually represented as an interval of possible values. Given two interval variables $t = [t_1, t_2]$ and $r = [r_1, r_2]$, then the sum and the product of these two interval variables is defined as follows:

$$t + r = [t_1 + r_1,\ t_2 + r_2], \tag{11}$$

$$t^* r = [\min(t_1 r_1, t_1 r_2, t_2 r_1, t_2 r_2), \max(t_1 r_1, t_1 r_2, t_2 r_1, t_2 r_2)]. \tag{12}$$

Obviously, variables which assume only fixed values can still be represented as intervals, trivially though, by considering the same fixed value for both left and right limits. Applying interval arithmetic operators to Eqs. (9) and (10) instead of the standard arithmetic operators, we can compute interval solutions for c_1 and c_2 by setting p_x and p_y equal to [0,1]. In interval notation, we want to solve the systems $Pc_1^I = p_x^I$ and $Pc_2^I = p_y^I$, where the superscript I denotes an interval vector. The solutions c_1^I and c_2^I should be understood to mean $c_1^I = [c_1 \colon Pc_1 = p_x,\ p_x \in p_x^I]$ and $c_2^I = [c_2 \colon Pc_2 = p_y,\ p_y \in p_y^I]$. It should be noted that since both interval systems involve the same matrix P and p_x, p_y assume values in the same interval, the interval solutions c_1^I and c_2^I will be the same. For this reason, we consider only the first of the interval systems in our analysis.

A lot of research has been done in the area of interval linear systems [17]. In more complicated cases, the matrix of the system of equations is also an interval matrix, that is, a matrix whose components are interval variables. Our case here is simpler, since the elements of P_{xy} are the x- and y-coordinates of the known object view which are fixed. However, if we merely try to evaluate (9) using the interval arithmetic operators described above, most likely we will obtain a non-sharp interval solution. The concept of non-sharp interval solutions is very common in IA. When we solve interval systems of equations, not all of the solutions obtained satisfy the problem at hand [17,18]. We will be referring to these solutions as invalid solutions. An interval solution is considered to be sharp if it includes as few invalid solutions as possible. The reason that sharp interval solutions are very desirable in our approach is because the generation of the training views can be performed faster (see the next section). The sharpness of the solutions obtained using IA depends on various factors. One well-known factor that affects sharpness is when a given interval quantity enters into a computation more than once [18]. This is actually the case with Eq. (9). To make it clear, let us consider the solution for the ith component of c_1, $1 \leqslant i \leqslant 3$,

$$\begin{aligned} c_{i1} = {} & \frac{v_{i1}}{w_{11}} (u_{11}x_1 + u_{21}x_2 + \cdots + u_{m1}x_m) \\ & + \frac{v_{i2}}{w_{22}} (u_{12}x_1 + u_{22}x_2 + \cdots + u_{m2}x_m) \\ & + \frac{v_{i3}}{w_{33}} (u_{13}x_1 + u_{23}x_2 + \cdots + u_{m3}x_m). \end{aligned} \tag{13}$$

Clearly, each x_j $(1 \leqslant j \leqslant m)$ enters in the computations of c_{i1} more than once. To avoid this, we factor out the x_j's. Then, Eq. (13) takes the form

$$c_{i1} = \sum_{j=1}^{m} x_j \left(\sum_{k=1}^{3} \frac{v_{ik} u_{jk}}{w_{kk}} \right). \tag{14}$$

The interval solution of c_{i1} can now be obtained by applying interval arithmetic operators in Eq. (14) instead of Eq. (13). Similarly, we can obtain interval solutions for the remaining elements of c_1^I as well as for c_2^I. It should be mentioned that given the solutions c_1^I and c_2^I, then $p_x^I \subseteq Pc_1^I$ and $P_y^I \subseteq Pc_2^I$. In other words, not every solution in c_1^I and c_2^I corresponds to p_x and p_y that belong in p_x^I and p_y^I, respectively. This issue is further discussed in the next section.

4. Learning the mapping

In order to train the SL-NN, we first need to generate the training views. This is performed by sampling the space of affine transformed views of the object. This space can be constructed by transforming a known view of the object, assuming all the possible sets of values for the parameters of affine transformation. Since it is impossible to consider all the possible sets, we just sample the range of values of each parameter and we consider only a finite number of sets. However, it is important to keep in mind that not all of the invalid solutions contained in the interval solutions of Eq. (9) might have been eliminated completely. As a result, when we generate affine transformed views by choosing the parameters of affine transformation from the interval solutions obtained, then not all of the generated views will lie in the unit square completely (invalid views). These views correspond to invalid solutions and must be disregarded. Fig. 2a illustrates the procedure. It might be clear now why it is desirable to compute sharp interval solutions. Sharp interval solutions imply narrower ranges for the parameters of affine transformation and consequently, the sampling procedure of Fig. 2a can be implemented faster.

```
for (a11=min_a11; a11 ≤ max_a11; a11 += s_a11)
 for (a12=min_a12; a12 ≤ max_a12; a12 += s_a12)
  for (b1=min_b1; b1 ≤ max_b1; b1 += s_b1)
   for (a21=min_a21; a21 ≤ max_a21; a21 += s_a21)
    for (a22=min_a22; a22 ≤ max_a22; a22 += s_a22)
     for (b2=min_b2; b2 ≤ max_b2; b2 += s_b2) {
        x_i = a11 x_i + a12 y_i + b1
        y_i = a21 x_i + a22 y_i + b2
        if x_i or y_i ∉ [0,1], do not consider
        the current affine transformed view as a training view.
     }
```

(a)

```
for (a11=min_a11; a11 ≤ max_a11; a11 += s_a11)
 for (a12=min_a12; a12 ≤ max_a12; a12 += s_a12)
  for (b1=min_b1; b1 ≤ max_b1; b1 += s_b1) {
      x_i = a11 x_i + a12 y_i + b1
      if x_i ∉ [0,1], do not consider
      the current affine transformed view as a training view.
  }
```

(b)

Fig. 2. A pseudo-code description of the sampling procedure for the generation of the training views.

It is important to observe at this point that both equations for computing x_i and y_i (Eqs. (2) and (3) which appear in Fig. 2a) involve the same basis vector (x',y'). Also, given that the ranges of (a_{11}, a_{12}, b_1) will be the same with the ranges of (a_{21}, a_{22}, b_2), as we discussed in Section 3, the information to be generated for the x_i coordinates will be exactly the same as the information to be generated for the y_i coordinates. Hence, we decouple the x- and y-coordinates of the views and we generate information only for one of the two (the x-coordinates here). This is illustrated in Fig. 2b. This observation offers great simplifications since the sampling procedure shown in Fig. 2a can now take a much more simplified form as shown in Fig. 2b. Consequently, the time and space requirements of the procedure for generating and storing the training views are significantly reduced. Furthermore, the size of the SL-NN is reduced in half. Assuming $\bar{m}$ interest points per view on the average, the sampling scheme of Fig. 2a requires a network with $2\bar{m}$ input nodes and 6 output nodes (Fig. 3a) while the sampling scheme of Fig. 2b requires only $\bar{m}$ input nodes and 3 output nodes (Fig. 3b). It should be noted that although we consider only one of the two image point coordinates of the training views, we are still referring to them as training views and this should not cause any confusion.

The decoupling of the point coordinates of the views and the consideration of only one of the two, imposes an additional cost during the recovery of the transformation parameters: they must now be predicted in two steps: First, we need to feed to the network the x-coordinates of

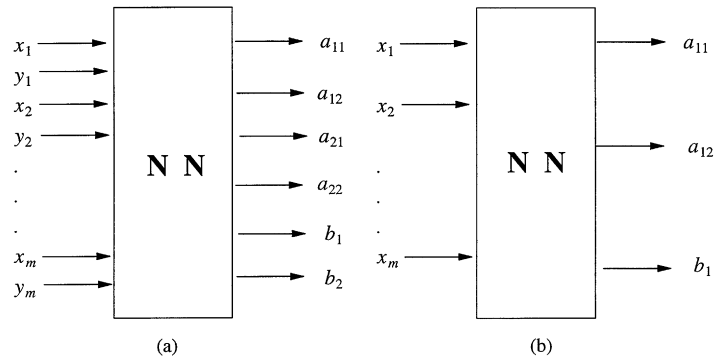

Fig. 3. (a) The original neural network scheme, (b) the simplified neural network scheme.

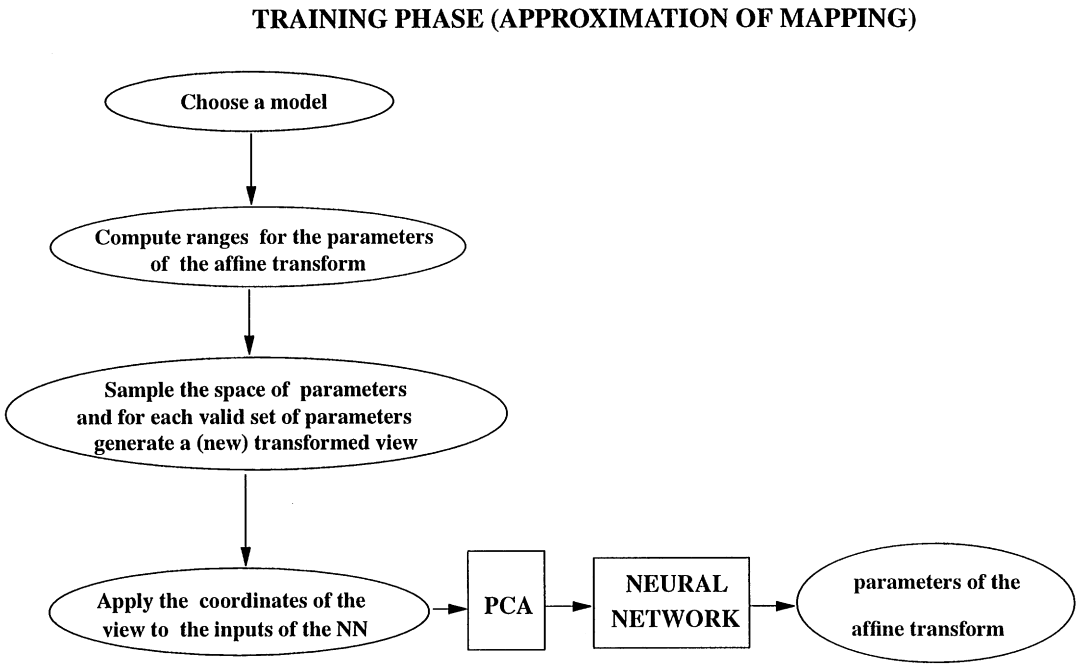

Fig. 4. The steps involved in training the SL-NN to approximate the mapping from the space of object's image coordinates to the space of affine transformations.

the unknown view in order to predict (a_{11}, a_{12}, b_1) and then we need to feed to the network the y-coordinates to predict (a_{21}, a_{22}, b_2). However, given the fast response time of the SL-NN after training has been completed, this additional cost is not very important. Fig. 4 presents an overview of the procedure for training a SL-NN to approximate the mapping between the image coordinates of an object's points and the space of parameters of the affine transformation. The meaning of the box labeled "PCA" will be discussed later.

5. Experiments

In this section, we report a number of experiments in order to demonstrate the strengths and weakness of the proposed approach. We have considered various issues such as accuracy in the predictions, discrimination power, and tolerance to noisy and occluded data.

5.1. Evaluation of the SL-NNs' performance

First, we evaluated how "good" the mapping computed by the SL-NN is. The following procedure was applied: first, we generated random affine transformed views of the object by choosing random values for the parameters of affine transformation. Then, we normalized the generated affine transformed views so that their x- and y-coordinates belong to the unit square. This was performed by choosing a random sub-square within the unit square and by mapping the square enclosing the view of the object (define by its minimum and maximum x- and y-coordinates) to the randomly chosen sub-square. After normalization, we applied the x-coordinates of the normalized unknown view first, and then its y-coordinates, to the SL-NN in order to predict the affine transformation that can align the known view with the normalized unknown view.

To judge how good the predictions yielded by the SL-NN were, we performed two tests: First, we compared the predicted values for the parameters of the affine transformation with the actual values which were computed using SVD. Second, we computed the mean square error between the normalized unknown view of the object and the back-projected known view, which was obtained by simply applying the predicted affine transformation on the known view. This is the most commonly used test in hypothesis generation-verification methods [1,2]. Fig. 5 summarizes the evaluation procedure. Fig. 6 shows the four different objects used in our experiments. For each object, we have identified a number of boundary "interest" points, which correspond to curvature extrema and zero-crossings [16]. These points are also shown in Fig. 6. The training of the SL-NN is based only on the coordinates of these "interest" points;

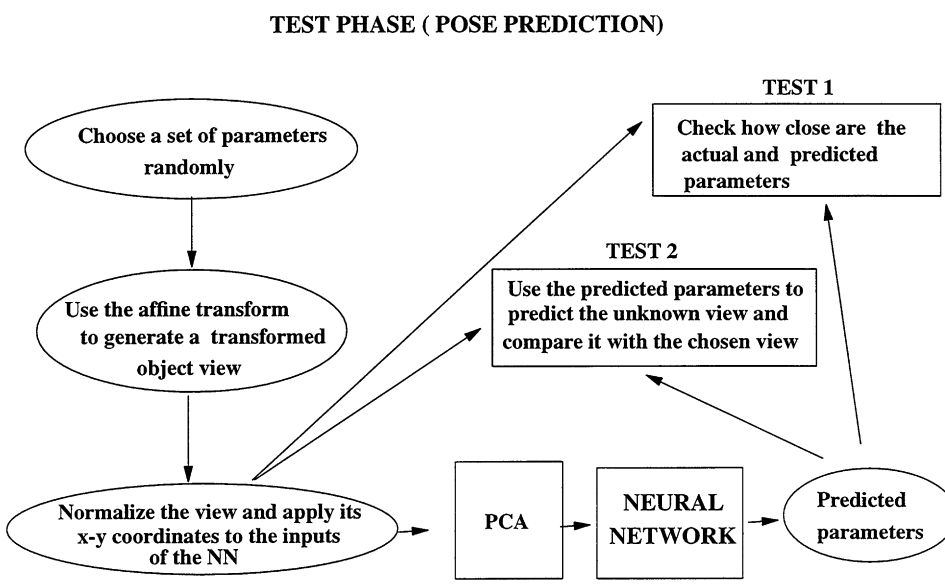

Fig. 5. The procedure used for testing SL-NN's ability to yield accurate predictions.

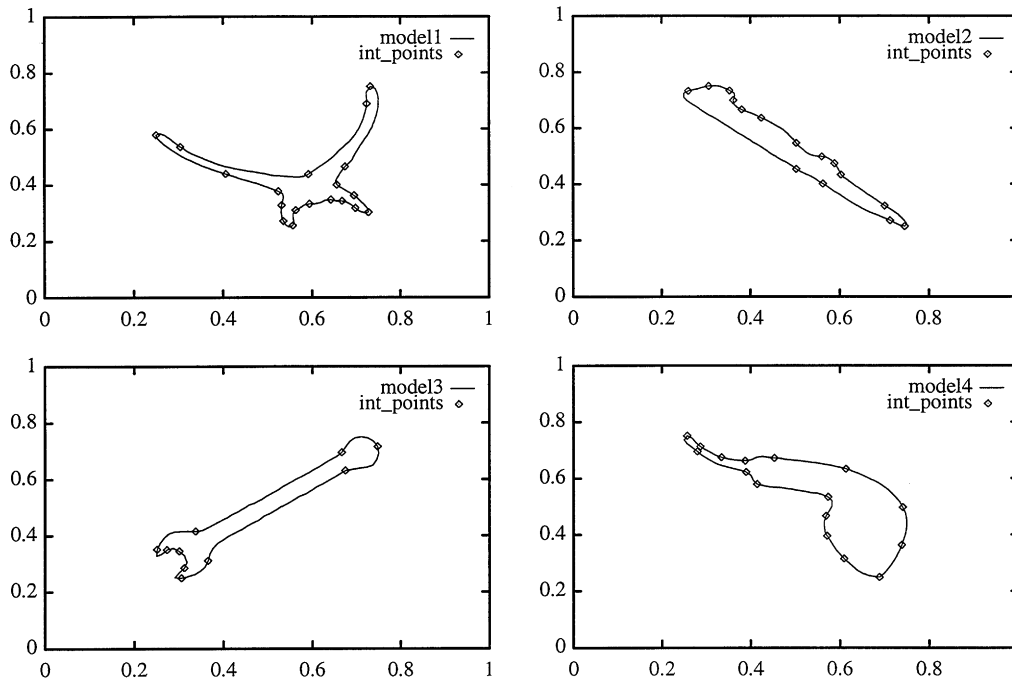

Fig. 6. The test objects used in our experiments along with the corresponding interest points.

however, the computation of the mean square error between the back-projected view and the unknown view utilizes all the boundary points for better accuracy. First, we estimated for each object the ranges of values that the parameters of affine transformation can assume. Only the interest points of each object were used for this estimation. Table 2 shows the ranges computed.

For each object, we generated a number of training views by sampling the space of affine transformed views of the object and we trained a SL-NN to learn the desired

Table 2
Ranges of values for the parameters of affine transformation

Model	Ranges of values		
	Range of a11	Range of a12	Range of b1
1	[− 2.953, 2.953]	[− 2.89, 2.89]	[− 1.662, 2.662]
2	[− 12.14, 12.14]	[− 11.45, 11.45]	[− 11.25, 12.25]
3	[− 8.22, 8.22]	[− 8.45, 8.45]	[− 0.8, 1.8]
4	[− 4.56, 4.45]	[− 4.23, 4.23]	[− 4.08, 5.08]

Table 3
Actual and predicted affine transformations

	Actual parameters		
a_{11}, a_{12}, b_1	0.6905 − 1.4162 0.8265	0.4939 − 0.8132 0.7868	− 0.3084 − 1.1053 1.3546
a_{21}, a_{22}, b_2	− 0.1771 − 0.8077 1.2053	0.8935 0.8684 − 0.4050	0.2782 − 1.2115 1.0551
	Predicted parameters (4 training views)		
a_{11}, a_{12}, b_1	0.6900 − 1.4156 0.8265	0.4935 − 0.8127 0.7867	− 0.3079 − 1.1058 1.3537
a_{21}, a_{22}, b_2	− 0.1768 − 0.8080 1.2045	0.8921 0.8698 − 0.4042	0.2781 − 1.2114 1.0547
	Predicted parameters (73 training views)		
a_{11}, a_{12}, b_1	0.6906 − 1.4167 0.8269	0.4942 − 0.8134 0.7871	− 0.3082 − 1.1053 1.3550
a_{21}, a_{22}, b_2	− 0.1768 − 0.8076 1.2055	0.8938 0.8682 − 0.4052	0.2783 − 1.2118 1.0554

mapping. One layer architectures were used because the mapping to be approximated is linear. The number of nodes in the input layer was determined by the number of interest points associated with each object while the number of nodes in the output layer was set to three (equal to the three parameters a_{11}, a_{12}, and b_1). Linear activation functions were used for the nodes in the output layer. Training was performed using the back-propagation algorithm [4]. Back propagation is an iterative algorithm which in each step adjusts the connection weights in the network, minimizing an error function. This is achieved using a gradient search which corresponds to a steepest descent on an error surface representing the weight space. The weight adjustment is determined by the current error and a parameter called *learning rate* which determines what amount of the error sensitivity to weight change will be used for the weight adjustment. In this study, a variation of the back-propagation algorithm (back-propagation with momentum) was used [11]. This is a simple variation for speeding up the back-propagation algorithm. The idea is to give each weight change some momentum so that it accelerates in the average down-hill direction. This may prevent oscillations in the system and help the system escape local error function minima. It is also a way of increasing the effective learning rate in almost-flat regions of the error surface. In all of our experiments, we used the same learning rate (0.2) and the same momentum term (0.4). We assumed that the network had converged when the sum of squared errors between the desired and actual outputs was less than 0.0001. Larger values ($\sim$0.01) can still lead to a well trained network, however, we found that the network becomes more sensitive to noise if we choose a more relaxed stopping criterion.

Table 3 shows some affine transformations predicted by a network trained with only four training views for the case of Model 1. These views were generated by sampling each parameter's range at six points. Views with image coordinates outside the interval [0,1] were not considered as training views, according to our discussion in Section 4. This is why although we sampled each parameter at six points, we finally ended up with only four training views. The actual parameters are also shown for comparison purposes. In addition, Table 3 shows the parameters predicted, for the same set of test affine transformed views, by a network trained with 73 views which were generated by sampling each parameter's range at 15 points. It can be observed that the predictions made by the network trained with the 73 views are not significantly better than the predictions made by the network trained with the four views.

Table 4 presents results for all of the planar objects, using various numbers of training views. For each case, we report the number of samples per parameter's range and the generated number of training views. Since it is not very easy to evaluate the quality of the predictions by

Table 4
Number of training views and average back-projection mse

Samples	Views	Avg-mse	SD	Epochs	CPU time (s)
Model 1					
6-6-6	4	0.122	0.003	7883	4.47
8-8-8	**14**	**0.01**	**0.003**	**20547**	**29.10**
15-15-15	73	0.003	0.001	18736	116.48
Model 2					
20-20-20	10	49.48	8.1	8876	9.33
26-26-26	**18**	**0.001**	**0.0**	**8798**	**13.83**
30-30-30	32	0.002	0.001	8566	24.97
Model 3					
6-6-6	6	35.065	6.825	19462	10.38
10-10-10	**14**	**0.006**	**0.002**	**26914**	**29.37**
15-15-15	49	0.005	0.001	23237	75.43
Model 4					
6-6-6	2	69.392	18.252	6024	1.88
10-10-10	8	0.005	0.001	5774	5.07
14-14-14	**20**	**0.002**	**0.001**	**20262**	**33.20**

simply examining the predicted parameter values, we also report the average mean square back-projection error and standard deviation. These were computed using 100 randomly transformed views for each object. Also, to get an idea of the training time, we report the number of training epochs required for convergence. These results indicate that the SL-NN is capable of approximating the desired mapping very accurately, it does not require many training views, and training time is fast. Increasing the number of training views did not yield a significant improvement in the case of noise-free data.

We also examined the computational requirements of the neural network approach. In our comparison, we assume that the training of the network is done off-line. If $\bar{m}$ is the average number of interest points per model, the neural network approach requires $3\bar{m}$ multiplications and $3\bar{m}$ additions to predict a_{11}, a_{12} and b_1. The same number of operations are required for predicting the other three parameters, so it requires $6\bar{m}$ multiplications and $6\bar{m}$ additions totally. For comparison, we also examined the computational requirements of a traditional least-squares approach. Specifically, we chose the SVD approach. Assuming that the factorization of P_{xy} is also done off-line, SVD requires $12\bar{m}$ multiplications, $6\bar{m}$ divisions, and $6(\bar{m}+6)$ additions. Given that these computations are repeated hundred of times during verification in object recognition, the neural network approach turns out to have less computational requirements. Also, the neural network approach has lower memory requirements than the traditional approach. Specifically, the neural network approach requires to store only $6\bar{m}$ values per network (i.e., weights) while the traditional approach requires to store $6\bar{m}+6+6^2$ values (for the elements of U, W, and V matrices). To avoid confusion, we need to emphasize again that the above comparison assumes that training and decomposition have been performed off-line. When this assumption is not true, then the SVD approach is faster than the neural network approach.

5.2. Discrimination power

Next, we investigated the discrimination power of each of the networks. For each object, we used the SL-NN trained with the numbers of training views shown highlighted in Table 4. These networks are noise tolerant and require a minimum number of training views to learn the mapping. Since each neural network has a different number of input nodes, depending on the number of interest points associated with the objects, it is practically impossible to present views of different objects, with different number of interest points, to the same network. To overcome this problem, we have attached a front-end stage to the SL-NN which actually reduces the dimensionality of the input vector, formed by the coordinates of the interest points of the views. In this way, we could use the same network architecture for each object. The front-end stage is based on principal components analysis (PCA) [15]. PCA is a multivariate technique which transforms a number of correlated variables, to a smaller set of uncorrelated variables. PCA might have important benefits for the performance of the neural network since less inputs, which are also uncorrelated, imply faster training

and probably better generalization. PCA works as follows: first, we compute the covariance matrix associated with our correlated variables and then we find the eigenvalues of this matrix. Then, we sort them and we form a new matrix whose columns consist of the eigenvectors to the largest eigenvalues. Deciding how many eigenvalues are significant depends on the problem at hand. The matrix formed by the eigenvectors corresponds to the transformation which is applied on the correlated variables to yield the new uncorrelated variables.

In our problem, the correlated variables are the training views associated with each SL-NN. For each training set, we applied the PCA and we kept the most significant principal components, three principal components were kept since only three eigenvalues were non-zero. The new training examples are now linear combinations of the old training views with dimensionality three. A separate network per object was used, having 3 nodes in the input layer and 3 nodes in the output layer. After training, we tested each network's discrimination ability. The results (average back-projection error and standard deviation over 100 randomly chosen affine transformed views for each model) are presented in Table 5. Clearly, each network predicts the correct affine transformation only for the affine transformed views of the object whose views were used to train the network. The discrimination power of the networks can be very useful during recognition. For example, suppose that we are given an unknown view. In order to recognize the object which has produced this view, it suffices to present the view to all of the networks. Each network will predict a set of transformation parameters, however, only one network (corresponding to the object which has produced the unknown view) will predict correct parameters.

5.3. Noise tolerance

In this subsection, we investigate how tolerant the networks' predictions are, assuming uncertainty in the locations of the object points. In particular, we assume that the location of each object point can be anywhere within a disc centered at the real location of the point and having a radius equal to ε (bounded uncertainty) [19]. Various ε values were chosen in order to evaluate the networks' ability to predict the correct transformation parameters. To test the networks, we used a set of 100 random affine transformed views and we computed the average mean square back-projection error. The results obtained, assuming that the front-end stage is inactive, show that the performance of the networks is rather poor. Fig. 7 (solid lined) shows a plot of the average mean square back-projection error versus ε. Also, we show the minimum and maximum errors observed. Trying to improve performance by using more training views did not help significantly. For instance, assuming $\varepsilon = 0.2$ and 4 training views for Model 1 (first row in Table 4) resulted in a mean square back-projection error equal to 1.622 with a standard deviation equal to 1.692. Assuming the same value for ε and 14 views, resulted in a mean square back-projection error equal to 1.62 with a standard deviation equal to 1.69. Using more views did not yield much better results.

Then, we tested the performance of the networks, assuming that the front-end stage is active. What we observed is quite interesting. For a small number of training views, the performance was not significantly better than the performance obtained using the SL-NNs trained with the original views (i.e., having the front-end stage inactive). However, a dramatic increase in the noise tolerance was observed by training the SL-NNs using more views. For instance, assuming Model 1, $\varepsilon = 0.2$ and 4 training views, resulted in a mean square back-projection error equal to 1.659 with a standard deviation equal to 1.39. These results are slightly better than those obtained using the original training views. However, assuming the same ε value and 14 views, resulted in a mean square back-projection error equal to 0.338 with a standard deviation equal to 0.244, a dramatic decrease, Fig. 7 (dashed line) shows a plot of the average mean square back-projection error vs. ε, as well as the minimum-maximum errors observed in this case. Some specific examples are shown in Fig. 8, where the figures in the left column show the matches achieved without using the PCA front-end stage, while the figures in

Table 5
Some results illustrating the discrimination power of the networks

	Model 1		Model 2		Model 3		Model 4	
	Avg-mse	SD	Avg-mse	SD	Avg-mse	SD	Avg-mse	SD
nn1	0.01	0.003	61.78	21.1	25.6	5.08	51.67	4.42
nn2	292.24	125.31	0.001	0.0	210.21	79.75	187.78	28.06
nn3	114.08	44.96	313.59	79.86	0.006	0.002	48.79	4.88
nn4	110.29	13.35	66.68	20.05	95.77	13.52	0.002	0.001

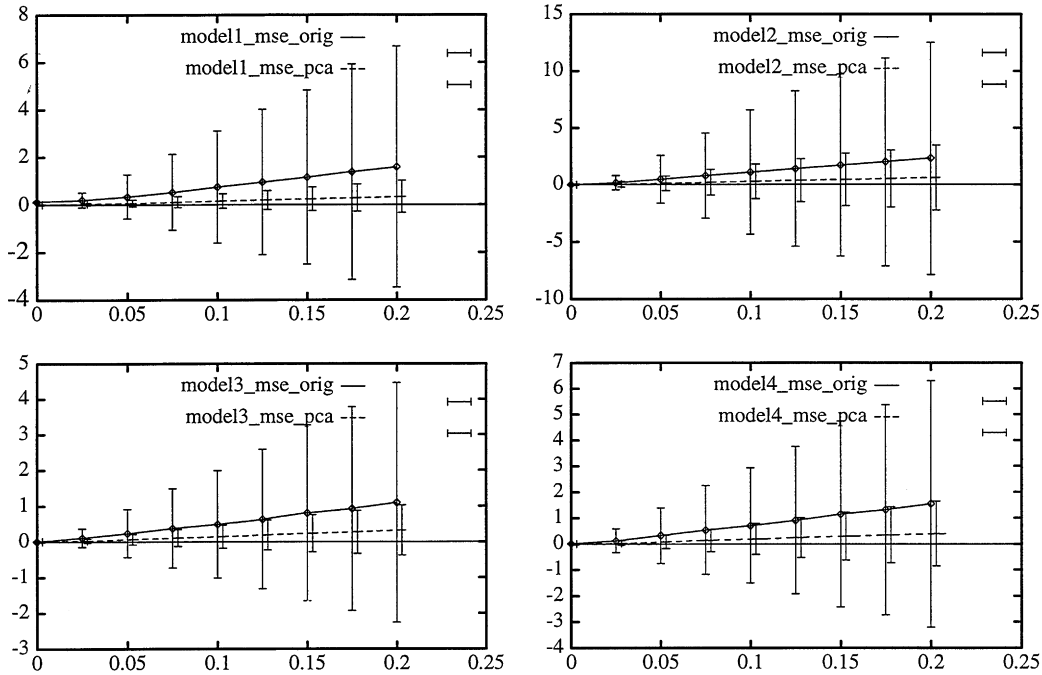

Fig. 7. The average mean square back-projection error vs. ε and the ranges of values observed. The solid line corresponds to the original data while the dashed line corresponds to the PCA transformed data.

the right column show the matches achieved using the PCA front-end stage. The solid line represents the unknown view and the dashed line represents the back-projected view which was computed using the predicted parameters. The actual and predicted parameters are shown in Table 6.

In particular, we observed that in the cases where the number of training views was not enough for the network to be noise tolerant, the number of non-zero eigenvalues associated with the covariance matrix of the training views was consistently less than three. Assuming more training views did not improve noise tolerance as long as the number of non-zero eigenvalues was less than three. However, utilizing enough training views so that the number of non-zero eigenvalues was three, resulted in a dramatic error decrease. Including more training views after this point did not improve noise tolerance significantly, and the number of non-zero eigenvalues remained three. The same observations were made for all of the four objects used in our experiments. The reason we finally end up with three non-zero eigenvalues is related to the fact that only three points are necessary to compute the parameters of the affine transformation. On the other hand, the training views obtained by sampling the space of transformed views might not span the space satisfactorily because of degenerate views. However, PCA can guide us in choosing a sufficient number of training views so that the networks can compute good, noise tolerant, mappings.

5.4. Occlusion tolerance

We have also performed a number of experiments assuming that some points are occluded. The performance of the SL-NNs trained with the original views was extremely bad, even with one point missing. Incorporating the PCA front-end stage improved the performance in cases where 2–3 points were missing. However, the performance was still poor when more points were removed. This suggests that in order for someone to apply the proposed method in cases where data occlusion is present, training of different networks for each object, using subsets of points rather than on all the object points, is more appropriate. The simplest way to select subsets of points is randomly. This, however, is not very efficient since the number of subsets increases exponentially with the number of points. A more efficient approach would be to apply a grouping approach [20,21] to detect groups of points which belongs to a particular object.

5.5. Performance using real scenes

In this section, we demonstrate the performance of the method using real scenes. Fig. 9a and b shows two of the

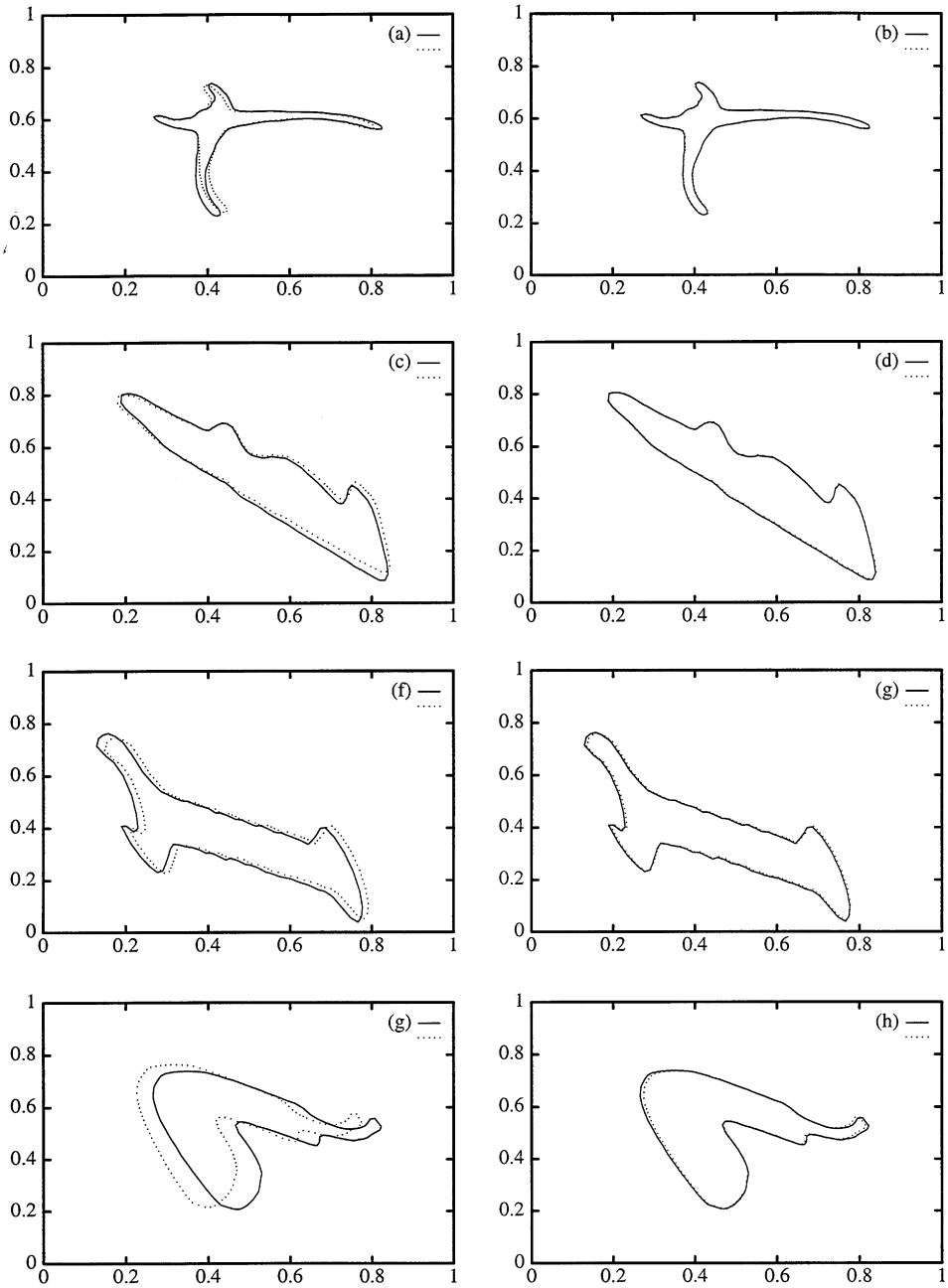

Fig. 8. The predictions obtained without using the PCA front-end stage (left) and using the PCA front-end stage (right).

scenes used in our experiments. The first scene contains Model 1, Model 2, and Model 3 while the second scene contains Model 1 and Model 4 as well as another object that we have not considered in our experiments. In Scene 1, we have intentionally left out the inner contour to make recognition more difficult. Point correspondences were established by hand. In cases that a model point did not have an exact corresponding scene point, we chose the closest possible scene point. Also, in cases that a model point did not have a corresponding scene point because of occlusion (for example, 1 point is occluded in Model 1 in Scene 1 and 2 points are occluded in Model

Table 6
Actual and predicted parameters (planar)

Actual parameters (Figs. 8a, b and c, d)		
a_1, a_2, a_3	−0.063063 −0.120558 0.438347	0.003732 −0.206111 0.530190
b_1, b_2, b_3	0.112543 −0.059775 0.574292	−0.080763 0.152122 0.528324
Predicted parameters (without using PCA)		
a_1, a_2, a_3	−0.066537 −0.112491 0.434431	0.001166 −0.211314 0.530985
b_1, b_2, b_3	0.107072 −0.058731 0.573535	−0.076414 0.145209 0.535657
Predicted parameters (using PCA)		
a_1, a_2, a_3	−0.063082 −0.120546 0.438362	0.003774 −0.206166 0.530237
b_1, b_2, b_3	0.112580 −0.059761 0.574307	−0.080823 0.152155 0.528266
Actual parameters (Fig. 8e, f and g, h)		
a_1, a_2, a_3	−0.227311 0.017821 0.363087	−0.073239 −0.143645 0.570144
b_1, b_2, b_3	0.132470 −0.133993 0.428320	0.116026 −0.063228 0.492276
Predicted parameters (without using PCA)		
a_1, a_2, a_3	−0.226994 0.017980 0.376362	−0.067705 −0.151540 0.520810
b_1, b_2, b_3	0.126998 −0.132556 0.424176	0.121118 −0.063729 0.507291
Predicted parameters (using PCA)		
a_1, a_2, a_3	−0.227329 0.017830 0.363073	−0.073270 −0.143638 0.570132
b_1, b_2, b_3	0.132497 −0.133942 0.428366	0.115976 −0.063170 0.492217

2 in Scene 1), we just picked the point (0.5, 0.5) (the center of the unit square) to be the corresponding scene point. The models were back-projected onto the scenes using the parameters predicted by the networks. The results are shown in Fig. 9e and f. As it can be seen, the models present in the scene have been recognized and aligned fairly well with the scene. It should be noted that in addition to the noise we have introduced by substituting missing "interest" points by neighboring points or even artificial points, there is also noise in the location of the rest scene points due to lack of robustness in the edge detection or/and "interest" point extraction. The best alignment was achieved in the case of Models 1 and 3 where most of their interest points were visible. The alignment of Model 2 has some problems at the non-sharp end of the object because there were missing "interest" points in this area as well as noise in the location of the rest points. Finally, Model 4 has been aligned with the scene quite satisfactorily. In the area of the boundary where the alignment is not very good, there was an "interest" point which was not detected and thus was replaced by the point (0.5, 0.5) in the prediction of the affine transformation.

6. Conclusions

In this paper, we considered the problem of learning to predict the parameters of the transformation that can align a known view of an object with unknown views of the same object. Initially, we compute the possible range of values that the parameters of the alignment (affine) transformation can assume. This is performed using singular-value decomposition (SVD) and interval analysis (IA). Then, we generate a number of novel views of the object by sampling the space of its affine transformed views. Finally, we train a single-layer neural network (SL-NN) to learn the mapping between the affine transformed views and the parameters of the alignment transformation. A number of issues related to the performance of the neural networks were considered such as accuracy in the predictions, discrimination power, noise tolerance, and occlusion tolerance.

Although our emphasis in this paper is to study the case of planar objects and affine transformations, it is important to mention that the same methodology can be extended and applied to the problem of learning to recognize 3-D objects from 2-D views, assuming orthographic or perspective projection. The linear model combinations scheme proposed by Basri and Ullman [8] and the algebraic functions of views proposed by Shashua [22] serve as a basis for this extension. In this case, novel orthographic or perspective views can be obtained by combining the image coordinates of a small number of reference views instead of a single reference view. The training views can be obtained by sampling the space of orthographically or perspectively transformed

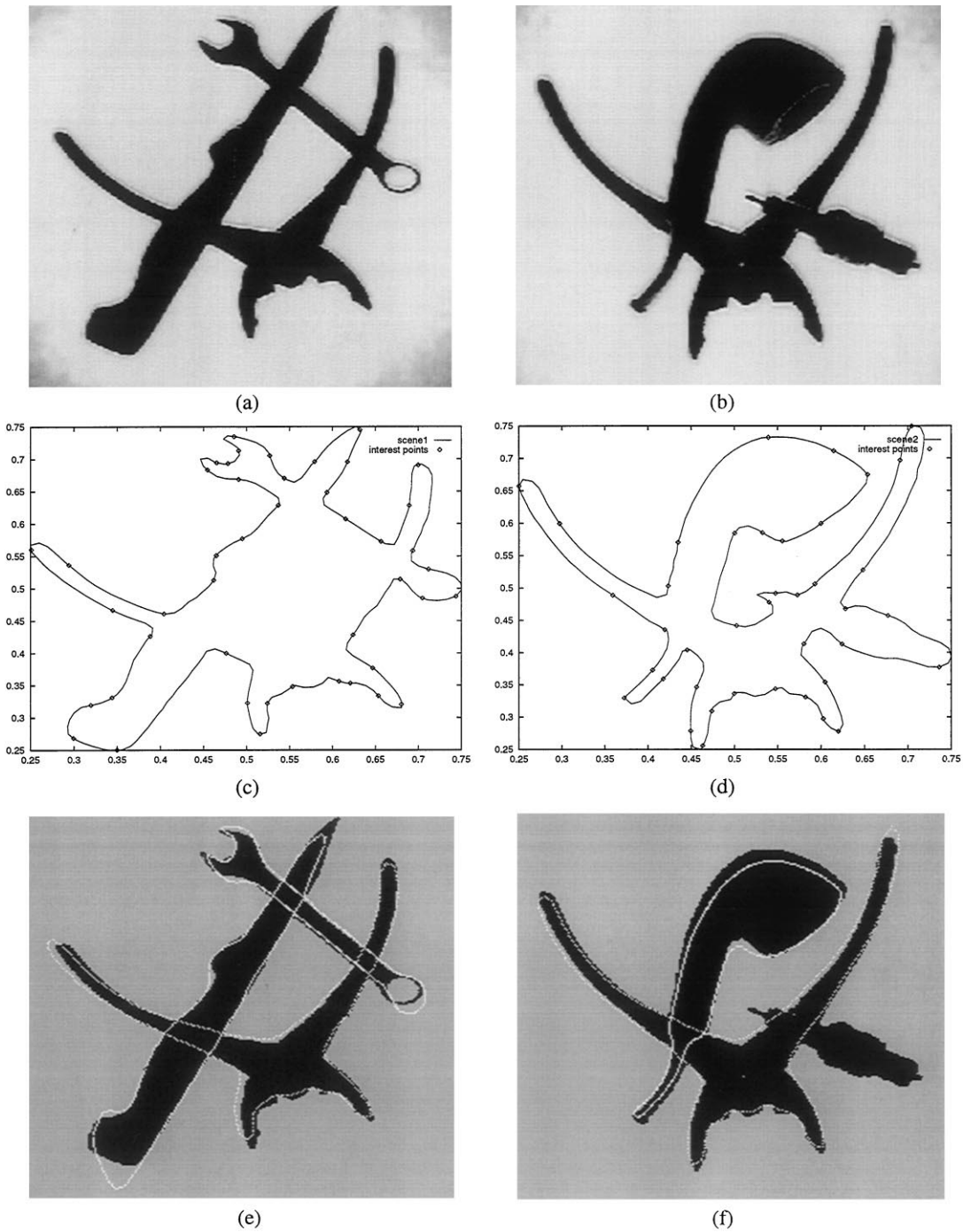

Fig. 9. The real scenes used to test the performance of the proposed method ((a), (b)) and the interest points extracted ((c), (d)). The back-projection results by using the affine transformation predicted by the SL-NN are also shown ((e), (f)).

views which can be constructed using a similar methodology. Interestingly, the decoupling of image point coordinates is still possible, even for the case of perspective projection (assuming that the known views are orthographic [23]). Some results for the orthographic case can be found in [23].

References

[1] Y. Lamdan, J. Schwartz, H. Wolfson, Affine invariant model-based object recognition, IEEE Trans. Robotics Automat. 6 (5) (1990) 578–589.

[2] D. Huttenlocher, S. Ullman, Recognizing solid objects by alignment with an image, Int. J. Comput. Vision 5 (2) (1990) 195–212.

[3] I. Rigoutsos, R. Hummel, Several results on affine invariant geometric hashing, in: Proc. 8th Israeli Conf. on Artificial Intelligence and Computer Vision, December 1991.

[4] D. Thompson, J. Mundy, Three dimensional model matching from an unconstrained viewpoint, in: Proc. IEEE Conf. on Robotics and Automation, 1987, pp. 208–220.

[5] E. Pauwels et al., Recognition of planar shapes under affine distortion, Int. J. Comput. Vision 14 (1995) 49–65.

[6] G. Bebis, M. Georgiopoulos, N. da Vitoria Lobo, M. Shah, Learning affine transformations of the plane for model based object recognition, in: Proc. 13th Int. Conf. on Pattern Recognition, Vienna, Austria, August 1996.

[7] T. Poggio, S. Edelman, A network that learns to recognize three-dimensional objects, Nature 343 (1990).

[8] S. Ullman, R. Basri, Recognition by linear combination of models, IEEE Pattern Anal. Machine Intell. 13 (10) (1991) 992–1006.

[9] G. Forsythe, M. Malcolm, C. Moler, Computer Methods for Mathematical Computations, Chapter 9, Prentice-Hall, Englewood Cliffs, NJ, 1977.

[10] R. Moore, Interval Analysis, Prentice-Hall, Englewood Cliffs, NJ, 1966.

[11] Hertz, A. Krogh, R. Palmer, Introduction to the Theory of Neural Computation, Addison Wesley, Reading, MA, 1991.

[12] H. Liao, J. Huang, S. Huang, Stock-based handwritten Chinese character recognition using neural networks, Pattern Recognition Lett.

[13] Alan M. Fu, Hong Yan, Object representation based on contour features and recognition by a Hop-field-Amari network, Neurocomputing 16 (1997) 127–138.

[14] G. Bebis, M. Georgiopoulos, M. Shah, N. La Vitoria Lobo, Indexing based on algebraic functions of views, Computer Vision Image Understanding (CVIU), 72 (3) (1998) 360–378.

[15] W. Press et. al, Numerical Recipes in C: the Art of Scientific Programming, Cambridge University Press, Cambridge, 1990.

[16] F. Mokhtarian, A. Mackworth, A theory of multiscale, curvature-based shape representation for planar curves, IEEE Trans. Pattern Anal. Machine Intell. 14 (8) (1992) 789–805.

[17] A. Neumaier, Interval Methods for Systems of Equations, Cambridge Univ. Press, Cambridge, 1990.

[18] E. Hansen, R. Smith, Interval arithmetic in matrix computations: Part II, SIAM J. Numer. Anal. 4 (1) (1967).

[19] W. Grimson, D. Huttenlocher, D. Jacobs, A study of affine matching with bounded sensor error, Int. J. Comput. Vision 13 (1) (1994) 7–32.

[20] J. Feldman, Efficient regularity-based grouping, in: Computer Vision and Pattern Recognition Conf. (CVPR'97) 1997, pp. 288–294.

[21] P. Havaldar, S. Medioni, F. Stein, Extraction of Groups for Recognition, Lecture Notes in Computer Science, vol. 800, Springer, Berlin, pp. 251–261.

[22] A. Shashua, Algebraic functions for recognition, IEEE Trans. Pattern Anal. Machine Intelli. 17 (8) (1995) 779–789.

[23] G. Bebis, M. Georgiopoulos, S. Bhatia, Learning orthographic transformations for object recognition, in: IEEE Int. Conf. Systems Man Cybernet. Vol. 4, Orlando, FL, October 1997, pp. 3576–3581.

About the Author—GEORGE BEBIS received the B.S. degree in Mathematics and the M.S. degree in Computer Science from the University of Crete, Greece, in 1987 and 1991, respectively, and the Ph.D. degree in Electrical and Computer Engineering from the University of Central Florida, Orlando, in 1996. Currently, he is an Assistant Professor in the Department of Computer Science at the University of Nevada, Reno (UNR) and director of the Computer Vision and Robotics Laboratory (CVRL) at UNR. From 1996 until 1997 he was a Visiting Assistant Professor in the Department of Mathematics and Computer Science at the University of Missouri, St. Louis while from June 1998 to August 1998 he was a summer faculty in the Center for Applied Scientific Research (CASC) at Lawrence Lovermore National Laboratory (LLNL). His research interests include computer vision, image processing, artificial neural networks, and genetic algorithms. Dr Bebis has served on the program committees of several national and international conferences and has organized and chaired several conference sessions. He is a member of the IEEE Society and the Program Chair of the IEEE local section.

About the Author—MICHAEL GEORGIOPOULOS received the Diploma in Electrical Engineering from the National Technical University of Athens, Greece, in 1981, and the M.Sc. and Ph.D. degrees in Electrical Engineering from the University of Connecticut, Storrs, in 1983 and 1986, respectively. In 1987, he joined the University of Central Florida, Orlando, where he is currently an Associate Professor at the Department of Electrical and Computer Engineering. His research interests are in the areas of neural networks, pattern recognition, and stochastic processes. Dr. Georgiopoulos is a member of the Technical Chamber of Greece and the International Neural Network Society.

About the Author—NIELS DA VITORIA LOBO completed the B.Sc (Honors) degree at Dalhousie University, Canada, in 1982, and the M.Sc. and Ph.D. degrees at the University of Toronto in 1985 and 1992 respectively. He is currently an Associate Professor at the University of Central Florida in the Department of Computer Science. His research interests are in vision and in physical modeling for

graphics. He receives funding from the National Science Foundation and the U.S. Department of Defense. He has several patents, numerous publications, and currently supervises a number of graduate students. He is a member of the Computer Society of the Institute of Electrical and Electronic Engineers.

About the Author—MUBARAK SHAH received his B. E. degree in 1979 in Electronics from Dawood College of Engineering and Technology, Karachi, Pakistan with the highest grades in the whole University, and was awarded a five year scholarship for his Ph.D. He spent 1980 at Philips International Institute of Technology, Eindhoven, The Netherlands, where he completed an E.D.E diploma. Dr. Shah received his M.S. and Ph.D. degrees both in Computer Engineering from Wayne State University, Detroit, Michigan, respectively in 1982 and 1986. Since 1986 he has been with the University of Central Florida, where he is currently a Professor of Computer Science, and the director of Computer Vision lab. Prof. Shah has published 50 research papers in refereed journals and conferences on topics including visual motion, gesture recognition, lipreading, edge and contour detection, multisensor fusion, shape from shading and stereo, and hardware algorithms for computer vision.