Distance Metric Between 3D Models and 2D Images for Recognition and Classification

Ronen Basri and Daphna Weinshall

Abstract—Similarity measurements between 3*D* objects and 2*D* images are useful for the tasks of object recognition and classification. We distinguish between two types of similarity metrics: metrics computed in image-space (*image metrics*) and metrics computed in transformation-space (*transformation metrics*). Existing methods typically use image metrics; namely, metrics that measure the difference in the image between the observed image and the nearest view of the object. Example for such a measure is the Euclidean distance between feature points in the image and their corresponding points in the nearest view. (This measure can be computed by solving the *exterior orientation calibration problem*.) In this paper we introduce a different type of metrics: *transformation metrics*. These metrics penalize for the deformations applied to the object to produce the observed image.

In particular, we define a *transformation metric* that optimally penalizes for "affine deformations" under weak-perspective. A closed-form solution, together with the nearest view according to this metric, are derived. The metric is shown to be equivalent to the Euclidean image metric, in the sense that they bound each other from both above and below. It therefore provides an easy-to-use closed-form approximation for the commonly-used least-squares distance between models and images. We demonstrate an image understanding application, where the true dimensions of a photographed battery charger are estimated by minimizing the transformation metric.

Index Terms—Affine deformations, 3D-to-2D metric, object recognition, exterior orientation calibration.

1 Introduction

Object recognition is a process of selecting the object model that best matches the observed image. A common approach to recognition uses features (such as points or edges) to represent objects. An object is recognized in this approach if there exists a viewpoint from which the model features coincide with the corresponding image features, e.g., [4], [7], [9]. Since images often are noisy and models occasionally are imperfect, it is rarely the case that a model aligns perfectly with the image. Moreover, in problems such as classification and recognition of non-rigid objects, the agreement between model and image is even less predictable. Systems therefore look for a model that "reasonably" aligns with the image. Consequently the general problem of recognition requires measures that provide a robust assessment of the similarity between objects and images. In this paper we describe two such measures, and develop a rigorous solution to the minimization problem that each measure entails.

A common measure for comparing 3D objects to 2D images is the Euclidean distance between feature points in the actual image and their corresponding points in the nearest view of the object. The assumption underlying this measure is that images are significantly less reliable than models, and so perturbations should be measured in the image plane. This assumption often suits recognition tasks. Other measures may better suit different assumptions. For example, when classifying objects, there is an inherent uncertainty in the structure of the classified object. One may therefore attempt to minimize the amount of deformations applied to the object to account for this uncertainty. Such a distance is measured in transformation space rather than in image space. A definition of these two types of measures is given in Section 3.

Measures to compare 3D models and 2D images generally are desired to have metrical properties; that is, they should monotonically increase with the difference between the measured entities. The Euclidean distance between the image and the nearest view defines a metric. (We refer to this measure as the *image metric*.) The difficulty with employing this measure is that a closed-form solution to the problem has not yet been found, and therefore currently numerical methods must be employed to compute the measure. A common method to achieve a closed-form metric is to extend the set of transformations that objects are allowed to undergo from the rigid to the affine one. The problem with this measure is that it bounds the rigid measure from below, but not from above. Other methods either achieve only suboptimal distances, or they do not define a metric. The existing approaches are reviewed in Section 2.

This paper presents a closed-form distance metric to compare 3D models and 2D images. The metric penalizes for the non-rigidity induced by the optimal affine transformation that aligns the model to the image under weak-perspective projection. Specifically, if A is the affine transformation that best aligns the model with the image, and $\mathcal{R}ig$ represents the set of all rigid transformations, then the metric is defined as

$$N_{tr} = \min_{R \in \mathcal{R}ig} \|A - R\|^2 \tag{1}$$

where the norm taken is the sum of squared elements. This metric is shown to bound the least-square distance between the model and the image both from above and below. We foresee three ways to use the metric developed in this paper:

1) Obtain a direct assessment of the similarity between 3D models and 2D images.
2) Obtain lower and upper bounds on the *image metric*. In many cases such bounds may suffice to unequivocally determine the identity of the observed object.
3) Provide an initial guess to be then used by a numerical procedure to solve the image distance.

The rest of this paper is organized as follows: In Section 2, we review related work. In Section 3, we define the concepts used in this paper. In Section 4, we provide the main results of this paper. Finally, in Section 5, we illustrate the outcome of using the new transformation metric on real images. In addition, we demonstrate an application, in which the true dimensions of a photographed battery charger are estimated by minimizing the transformation metric.

2 Previous Approaches

Previous approaches to the problem of model and image comparison using point features are divided into two major categories: least-square minimization in image space, and suboptimal methods using correspondence subsets.

The traditional photometric approach to the problem of model and image comparison involves retrieving the view of the object that minimizes the least-square distance to the image. This problem is referred to as the *exterior orientation calibration problem* and is defined as follows. Given a set of n 3D points (model points) and a corresponding set of n 2D points (image points), find the rigid transformation that minimizes the distance in the image plane

- *R. Basri is with the Department of Applied Mathematics, the Weizmann Institute of Science, Rehovot 76100, Israel. E-mail: ronen@wisdom.weizmann.ac.il.*
- *D. Weinshall is with the Institute of Computer Science, the Hebrew University of Jerusalem, 91904 Jerusalem, Israel. E-mail: daphna@cs.huji.ac.il.*

Manuscript received Jan. 25, 1993; revised Aug. 29, 1995. Recommended for acceptance by S. Levine.
For information on obtaining reprints of this article, please send e-mail to: transactions@computer.org, and reference IEEECS Log Number P95180.

between the transformed model points and the image points. An analytic solution to this problem has not yet been found. Consequently, numerical methods are employed [11], [8]. Such solutions often suffer from stability problems, they are computationally intensive, and they require a good initial guess.

To avoid using numerical methods, frequently the object is allowed to undergo affine transformations instead of just rigid ones. Affine transformations are composed of general linear transformations (rather than rotations) and translations, and they include in addition to the rigid transformations also reflection, stretch, and shear. The solution in the affine case is simpler than that of the rigid case because the quadratic constraints imposed in the rigid case are not taken into account, enabling the construction of a closed-form solution. At least six points are required to find an affine solution under perspective projection [4], and four are required under orthographic projection [9]. The affine measure bounds the rigid measure from below. The rigid measure, however, is not bounded from above, and so the actual rigid measure may sometimes be significantly larger than the computed affine measure.

A second approach to comparing models to images, often called *alignment*, involves the selection of a small subset of correspondences (*alignment key*), solving for the transformation using this subset, and then transforming the other points and measuring their distance from the corresponding image points. The obtained distances are clearly suboptimal. However, by relying on small subsets of correspondences alignment can overcome occlusion and clutter. Three [4] or four [6] corresponding points are required under perspective projection, and three points under weak perspective [7].

3 Definitions and Notation

In the following discussion, we assume weak-perspective projection. Namely, the object undergoes a 3D transformation that includes rotation, translation, and scaling, and is then orthographically projected onto the image. Perspective distortions are not accounted for and treated as noise. The weak-perspective projection model is particularly useful when objects are observed from a relatively long distance.

In order to define a similarity measure for comparing 3D objects to 2D images, as discussed in Section 1, we first define the **best-view** of a 3D object given a 2D image:

DEFINITION 1 [best-view]. *Let ∂ denote a difference measure between two 2D images of n features. Given a 2D image of an object composed of n features, the* **best-view** *of a 3D object (model) composed of n corresponding features, is the view for which the smallest value of ∂ is obtained. The minimization is performed over all the possible views of the model; the views are obtained by applying a transformation T, taken from the set of permitted transformations Λ, and followed by a projection, Π.*

We compute ∂, the difference between two 2D images of n features, in two ways:

Image metric: We measure position differences in the image, namely, it is the Euclidean distance between corresponding points in the two images, summed over all points.

Transformation metric: The images are considered to be instances of a single 3D object. The metric measures the difference between the two transformations that align the object with the two images. This difference can be measured, for instance, by computing the Euclidean distance between the matrices that represent the two transformations (when the two transformations are linear).

As is mentioned above, the measure ∂ is applied to the given image and to the views of the given model. These views are generated by applying a transformation from a set Λ of permitted transformations. The view that minimizes the distance ∂ to the image is considered as the **best view**, and the distance between the best view and the actual image is considered as the distance between the object and the image.

We consider in this paper two families of transformations: rigid transformations[1] and affine transformations, and we discuss the following metrics:

N_{im}: a metric that measures the image distance between the given image and the best *rigid* view of the object.

N_{af}: a metric that measures the image distance between the given image and the best *affine* view of the object.

N_{tr}: a *transformation metric*. We assume that the image is an affine view of the object. (When it is not, we substitute the image by the best affine view.) We look for the rigid view of the object so as to minimize the difference between the two transformations: the affine transformation (between the object and the image) and the rigid transformation (between the object and its possible rigid view). In other words, we look for a view so as to minimize the amount of "affine deformations" applied to the object.

To illustrate the difference between *image metrics* and *transformation metrics*, Fig. 1 shows an example of three 2D images, whose similarity relations reverse, depending on which kind of metric is used. Consider the planar object in Fig. 1b as a reference object, and assume Λ contains the set of rigid transformations in 2D. The images in Fig. 1a and Fig. 1c are obtained by stretching the object horizontally (by 9/7) and vertically (by 3/2) respectively. (The image in Fig. 1b is obtained by applying a unit matrix to the object.)

- The *image metric* between the images in (b) and (a) is 4, two pixel at each of the left corners of the rectangle. The *image metric* between the images in (b) and (c) is 2, one pixel at each of the upper corners of the rectangle.
- Therefore, according to the *image metric*, Fig. 1c is closer to Fig. 1b than Fig. 1a is.
- To compute the *transformation metric* consider the planar object illustrated in (b). We compute the difference between the matrices that represent the affine transformation from (b) to both (a) and (c) and the matrix that represent the best rigid transformation (in this case it is the unit matrix): (a) is obtained from (b) by a horizontal stretch of 9/7. The *transformation metric* between (a) and (b) is therefore $2/7 = 9/7 - 1$. (c) is obtained from (b) by a vertical stretch of 3/2. The *transformation metric* in this case is $1/2 = 3/2 - 1$.
- Therefore, according to the *transformation metric*, Fig. 1a is closer to Fig. 1b than Fig. 1c is.

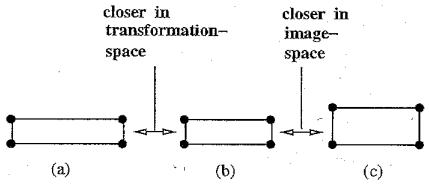

Fig. 1. The 2D image shown in (b) is closer to the image in (a) when the difference is computed in transformation space, and closer to the image in (c) when the difference is the Euclidean difference between the two images.

1. Note that a rigid transformation under weak perspective is equivalent to a similarity transformation followed by an orthographic projection.

3.1 Derivation of N_{im} and N_{af}

We now define the rigid and the affine *image metrics* precisely. Under weak-perspective projection, the position in the image, $\vec{q}_i = (x_i, y_i)$, of a model point $\vec{p}_i = (X_i, Y_i, Z_i)$ following a rigid transformation is given by

$$q_i = \Pi(R\vec{p}_i + \vec{t}) \tag{2}$$

where R is a scaled, 3×3 rotation matrix, $\vec{t}$ is a translation vector, and Π represents the orthographic projection operator. More explicitly, denote by $\vec{r}_1^T$ and $\vec{r}_2^T$ the top two row vectors of R, and denote $\vec{t} = (t_x, t_y, t_z)$; we have that

$$\begin{aligned} x_i &= \vec{r}_1^T \cdot \vec{p}_i + t_x \\ y_i &= \vec{r}_2^T \cdot \vec{p}_i + t_y \end{aligned} \tag{3}$$

where

$$\begin{aligned} \vec{r}_1^T \cdot \vec{r}_2 &= 0 \\ \vec{r}_1^T \cdot \vec{r}_1 &= \vec{r}_2^T \cdot \vec{r}_2 \end{aligned} \tag{4}$$

The rigid metric, N_{im}, minimizes (over all R and $\vec{t}$) the difference between the two sides of (3) subject to the constraints (4).

When the object is allowed to undergo affine transformations, the rotation matrix R is replaced by a general 3×3 linear matrix (denoted by A) and the constraints (4) are ignored. That is

$$q_i = \Pi(A\vec{p}_i + \vec{t}) \tag{5}$$

Denote by $\vec{a}_1^T$ and $\vec{a}_2^T$ the top two row vectors of A, we obtain

$$\begin{aligned} x_i &= \vec{a}_1^T \cdot \vec{p}_i + t_x \\ y_i &= \vec{a}_2^T \cdot \vec{p}_i + t_y \end{aligned} \tag{6}$$

The affine metric, N_{af}, minimizes (over all A and $\vec{t}$) the difference between the two sides of (6).

To define the rigid and the affine metrics, we first note that the translation component of both the best rigid and affine transformations can be ignored if the centroid of both model and image points are moved to the origin. In other words, we begin by translating the model and image points so that

$$\sum_{i=1}^{n} \vec{p}_i = \sum_{i=1}^{n} \vec{q}_i = 0 \tag{7}$$

We claim that now $\vec{t} = 0$ obtains the minimum (a proof is given in [2]).

Denote

$$P = \begin{pmatrix} X_1 & Y_1 & Z_1 \\ & \vdots & \\ X_n & Y_n & Z_n \end{pmatrix} \tag{8}$$

a matrix of model point coordinates, and denote

$$\vec{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \qquad \vec{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \tag{9}$$

the location vectors of the corresponding image points. A rigid metric that reflects the desired minimization is given by

$$N_{im} = \min_{\vec{r}_1, \vec{r}_2 \in \mathcal{R}^3} \left\| \vec{x} - P\vec{r}_1 \right\|^2 + \left\| \vec{y} - P\vec{r}_2 \right\|^2 \quad \text{s.t. } \vec{r}_1^T \cdot \vec{r}_2 = 0,\ \vec{r}_1^T \cdot \vec{r}_1 = \vec{r}_2^T \cdot \vec{r}_2 \tag{10}$$

The corresponding affine metric is given by

$$N_{af} = \min_{\vec{a}_1, \vec{a}_2 \in \mathcal{R}^3} \left\| \vec{x} - P\vec{a}_1 \right\|^2 + \left\| \vec{y} - P\vec{a}_2 \right\|^2 \tag{11}$$

In the affine case the solution is simple. We assume that the rank of P is 3 (the case for general, not coplanar, 3D objects). Denote $P^+ = (P^T P)^{-1} P^T$, the pseudo-inverse of P; we obtain that

$$\begin{aligned} \vec{a}_1 &= P^+ \vec{x} \\ \vec{a}_2 &= P^+ \vec{y} \end{aligned} \tag{12}$$

And the affine distance is given by

$$N_{af} = \left\| (I - PP^+)\vec{x} \right\|^2 + \left\| (I - PP^+)\vec{y} \right\|^2 \tag{13}$$

Since the solution in the rigid case is significantly more difficult than the solution in the affine case, often the affine solution is considered, and the rigidity constraints are used only for verification [9].

The constraints (4) (substituting $\vec{a}_i$ for $\vec{r}_i$, and using (12)) can be rewritten as

$$\begin{aligned} \vec{x}^T (P^+)^T P^+ \vec{y} &= 0 \\ \vec{x}^T (P^+)^T P^+ \vec{x} &= \vec{y}^T (P^+)^T P^+ \vec{y} \end{aligned} \tag{14}$$

Denote

$$B = (P^+)^T P^+ \tag{15}$$

we obtain that

$$\begin{aligned} \vec{x}^T B \vec{y} &= 0 \\ \vec{x}^T B \vec{x} &= \vec{y}^T B \vec{y} \end{aligned} \tag{16}$$

where B is an $n \times n$ symmetric, positive-semidefinite matrix of rank 3. (The rank would be smaller if the object points are coplanar.)

We call B the **characteristic matrix** of the object. B is a natural extension to the 3×3 model-based invariant Gramian matrix defined in [10].

3.2 Derivation of N_{tr}

We now define the *transformation metric*. Consider the affine solution. The nearest "affine view" of the object is obtained by applying the model matrix, P, to a pair of vectors, $\vec{a}_1$ and $\vec{a}_2$, defined in (12). In general, this solution is not rigid, and so the rigid constraints (4) do not hold for these vectors. The metric described here is based on the following rule. We are looking for another pair of vectors, $\vec{r}_1$ and $\vec{r}_2$, which satisfy the rigid constraints, and minimize the Euclidean distance to the affine vectors $\vec{a}_1$ and $\vec{a}_2$. $P\vec{r}_1$ and $P\vec{r}_2$ define the best rigid view of the object under the defined metric. The metric, N_{tr}, is defined by

$$N_{tr} = \min_{\vec{r}_1, \vec{r}_2 \in \mathcal{R}^3} \left\| \vec{a}_1 - \vec{r}_1 \right\|^2 + \left\| \vec{a}_2 - \vec{r}_2 \right\|^2 \quad \text{s.t. } \vec{r}_1^T \cdot \vec{r}_2 = 0,\ \vec{r}_1^T \cdot \vec{r}_1 = \vec{r}_2^T \cdot \vec{r}_2 \tag{17}$$

where $\vec{a}_1$ and $\vec{a}_2$ constitute the optimal affine solution, therefore

$$N_{tr} = \min_{\vec{r}_1, \vec{r}_2 \in \mathcal{R}^3} \left\| P^+\vec{x} - \vec{r}_1 \right\|^2 + \left\| P^+\vec{y} - \vec{r}_2 \right\|^2 \quad \text{s.t. } \vec{r}_1^T \cdot \vec{r}_2 = 0,\ \vec{r}_1^T \cdot \vec{r}_1 = \vec{r}_2^T \cdot \vec{r}_2 \tag{18}$$

4 Results

We derive the following results (the proof is given in [2]):

4.1 Transformation Space:

The *transformation metric* defined in (18) has the following solution

$$N_{tr} = \frac{1}{2}\left(\vec{x}^T B \vec{x} + \vec{y}^T B \vec{y} - 2\sqrt{\vec{x}^T B \vec{x} \cdot \vec{y}^T B \vec{y} - \left(\vec{x}^T B \vec{y} \right)^2} \right) \tag{19}$$

where B is defined in (15), and $\vec{x}, \vec{y}$ in (9). The **best view** according to this metric is given by

$$\begin{aligned} \vec{x}^* &= PP^+(\beta_1 \vec{x} + \beta_2 \vec{y}) \\ \vec{y}^* &= PP^+(\gamma_1 \vec{x} + \gamma_2 \vec{y}) \end{aligned} \tag{20}$$

where

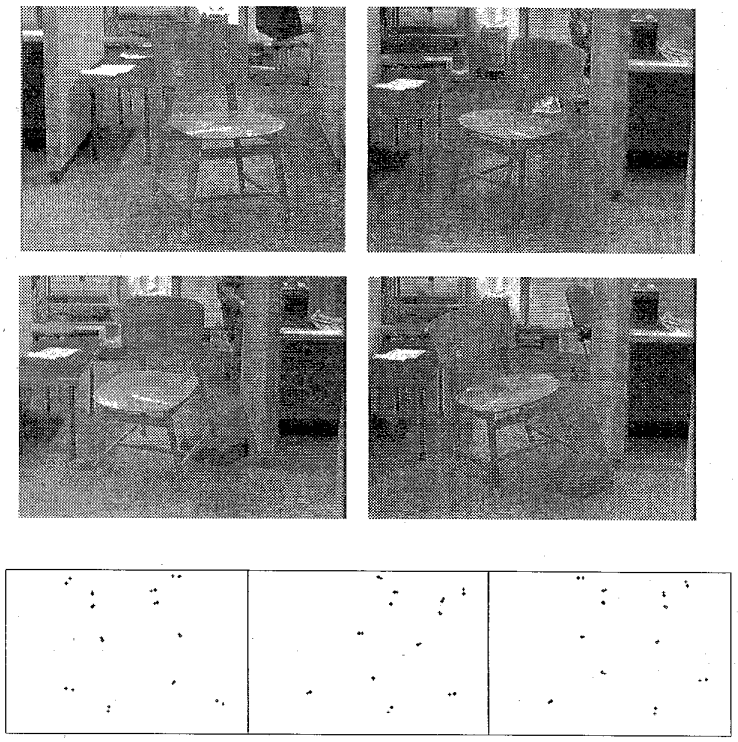

Fig. 2. Top and middle rows: Four images of a chair, with feature points marked on one of them for illustration, for which the 3D coordinates of the points are known (i.e., the model is given). Bottom row: For three of the images, the original feature points of the image are marked by +; for comparison, the feature points of the model, in the closest image according to N_{tr}, are marked by diamonds.

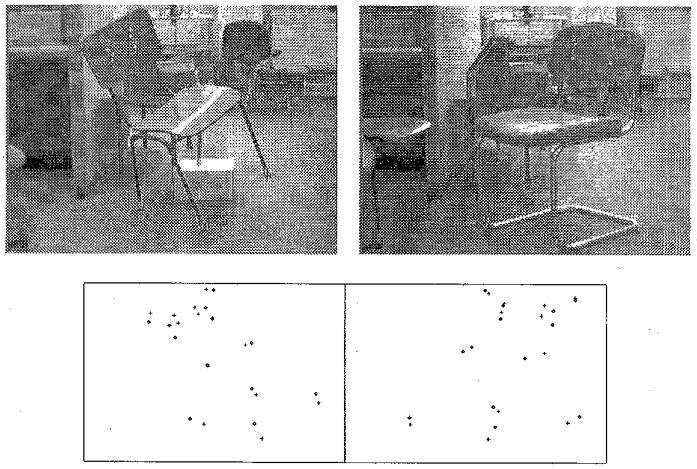

Fig. 3. Top row: Two images of different chairs, with feature points marked on one of them for illustration. Bottom row: The original feature points of the image are marked by +; for comparison, the feature points of the model chair (shown in Fig. 2), the closest image according to N_{tr}, are marked.

TABLE 1

THE TRANSFORMATION METRIC VALUES N_{tr}, THE AFFINE METRIC, AND THE VARIOUS BOUNDS COMPUTED FOR THE FOUR CHAIRS IN FIG. 2 AND IN FIG. 3.

	Same chair (Fig. 2)				Other chairs (Fig. 3)	
	Top left	Top right	Bottom left	Bottom right	Left	Right
N_{tr}	0.060	0.048	0.053	0.040	1.080	0.410
N_{aff}	1.745	1.352	1.240	1.240	4.008	5.364
Lower bound (22)	1.971	1.582	1.510	1.450	5.587	5.876
Upper bound (22)	2.730	2.319	2.332	2.122	9.777	7.680
Tighter (26)	2.336	1.941	1.916	1.778	7.719	6.731
Tightest (25)	2.313	1.884	1.878	1.755	7.350	6.514

Except for the transformation metric, the values are normalized so they reflect the average distortion in pixels of a single feature point.

$$\beta_1 = \frac{1}{2}\left(1 + \frac{\vec{y}^T B \vec{y}}{\sqrt{\vec{x}^T B \vec{x} \cdot \vec{y}^T B \vec{y} - \left(\vec{x}^T B \vec{y}\right)^2}}\right)$$

$$\beta_2 = \gamma_1 = -\frac{\vec{x}^T B \vec{y}}{2\sqrt{\vec{x}^T B \vec{x} \cdot \vec{y}^T B \vec{y} - \left(\vec{x}^T B \vec{y}\right)^2}}$$

$$\gamma_2 = \frac{1}{2}\left(1 + \frac{\vec{x}^T B \vec{x}}{\sqrt{\vec{x}^T B \vec{x} \cdot \vec{y}^T B \vec{y} - \left(\vec{x}^T B \vec{y}\right)^2}}\right)$$

4.2 Image Space:

Using N_{tr} we can bound the *image metric* from both above and below. Denote

$$N_{af} = \left\|(I - PP^+)\vec{x}\right\|^2 + \left\|(I - PP^+)\vec{y}\right\|^2 \tag{21}$$

we show that

$$N_{af} + \lambda_1 N_{tr} \le N_{im} \le N_{af} + \lambda_3 N_{tr} \tag{22}$$

where $\lambda_1 \le \lambda_2 \le \lambda_3$ are the eigenvalues of $P^T P$. A suboptimal solution to N_{im} is given by

$$N_{af} + \frac{2\mu_1\mu_2}{\mu_1 + \mu_2} N_{tr} \tag{23}$$

where $\sqrt{\mu_1} \le \sqrt{\mu_2}$ are the principal axes of the ellipse, defined by the intersection of the ellipsoid B with the plane $span\{\vec{x}, \vec{y}\}$. A tighter upper bound is deduced from this suboptimal solution

$$N_{im} \le N_{af} + h(\lambda_2, \lambda_3) N_{tr} \le N_{af} + 2\lambda_2 N_{tr} \tag{24}$$

where $h(\lambda_2, \lambda_3) = \frac{2}{\frac{1}{\lambda_2} + \frac{1}{\lambda_3}}$ is the Harmonic mean of λ_2, λ_3.

To summarize, we have

$$N_{af} + \lambda_1 N_{tr} \le N_{im} \le N_{af} + \frac{2\mu_1\mu_2}{\mu_1 + \mu_2} N_{tr} \tag{25}$$

$$N_{af} + \lambda_1 N_{tr} \le N_{im} \le N_{af} + h(\lambda_2, \lambda_3) N_{tr} \tag{26}$$

The suboptimal solution (23) (the upper bound in (25)) is proposed as an initial guess for an iterative algorithm to compute N_{im}.

5 Applications

In this section we show examples of applying the transformation metric to the problems of recognition and shape reconstruction. In the first example (Section 5.1) we compare a model of a chair to several images. The results are then compared to the results obtained when a different chair is used as a model. In the second example (Section 5.2) we use the transformation metric to determine the dimensions of a battery charger from a single image.

5.1 Recognition

In this experiment a 3D model of a chair, including twelve of its feature points, was given. Four images of the chair at different orientations, as well as two more images of two different chairs, were photographed (see Figs. 2 and 3). Twelve feature points corresponding to the model points were manually extracted from these images. The model was compared to the six pictures using the *transformation metric*. Figs. 2 and 3 also show the model points in the *best view* according to the *transformation metric*, overlaid on their corresponding points in the reference image. We can see that albeit the model chair is compared in Fig. 3 to different chairs, the matching obtained is relatively good. Note that in Fig. 2 the matching between the model and the images of the same chair is not perfect due to errors in the 3D measurements and the weak perspective approximation.

The distances between the model of the reference chair (condition number 5.25) and the six images of Figs. 2 and 3 are given in Table 1. It can be seen that the transformation metric values obtained for the images of the same chair (range between 0.04 and 0.06) are significantly smaller than those of the other chairs (range between 0.41 and 1.08). Similar results are obtained for the affine metric and the various bounds. As is expected, the affine metric always underestimates the image metric. The tightest upper bound is 10%-30% larger than the lower bound, and the worst upper bound for the same chair (2.73 for the top left image in Fig. 2) is still much lower than the lowest upper bound for the other chairs (5.587 for the left image in Fig. 3). Thus, the bounds suffice to discriminate between the images of the same chair from the images of the other chairs.

5.2 Reconstruction

In this experiment we attempt to infer the dimensions of an object from a single view. We will use an image of a battery charger as an input (Fig. 4). Suppose that we can identify the object either by recognizing it as a box of some arbitrary dimensions or by identifying certain surface markings on the object. Our task now is to estimate the dimensions of the box from the image coordinates of the seven visible corners of the charger.

To find the actual dimensions of the battery charger, we search the parameter space $u \times v \times w$, where u is the depth of the charger (the width of the left face), v is its height, and w is the length of the front face. Since under the weak-perspective projection model we can infer the dimensions of objects up to a scale factor only, we may set one of these parameters to be constant and search the space of the other two measurements. In our experiment we set w to its true value, 28 cm, and searched the space of the other two parameters, u and v.

In Fig. 5 the upper limit on N_{im}, given in (26), is plotted for each pair of parameters. The first search was done on a coarse scale (Fig. 5a). The minimum of the error bound is obtained for $u = 22.6$, $v = 19.1$, which is the (correct) answer with certainty of ±2 cm for u, and ±1 cm for v. The second search was done on a finer scale (Fig. 5b). The minimum of the error bound is obtained for $u = 22.06$, $v = 19.1$, which is the answer with certainty of ±0.28 cm in each dimension. This final result provides a reasonably good estimate of the dimensions of the battery charger, with an error of ≈ 0.5 cm in one dimension (u).

6 Conclusion

We have proposed a *transformation metric* to measure the similarity between 3D models and 2D images. The *transformation metric* measures the amount of affine deformation applied to the object to produce the given image. A simple, closed-form solution for this metric has been presented. This solution is optimal in transformation space, and it is used to bound the *image metric* from both above and below. The *transformation metric* presented in this paper can be used to obtain a direct assessment of the similarity between models and images or as a mean to evaluate the image metric. The proposed metric can be used in several different ways in the recognition and classification tasks. We conclude the paper with a brief discussion of possible applications of the metric.

The *transformation metric* provides a suboptimal closed-form estimate for the *image metric*. A scheme which uses this measure will prefer "symmetric" objects, objects whose convex-hull is close to a sphere, over other objects which are significantly stretched or contracted along one spatial dimension. This solution can also be used as an initial guess in an iterative process that computes the optimal value of the image metric numerically. The suboptimal solution derived using the image metric provides a better estimate

for the image metric than the affine solution, which has been used for example in [3] as the initial guess for computing the perspective image metric numerically.

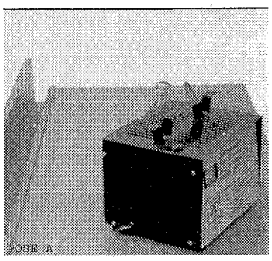

Fig. 4. A picture of a battery charger, whose dimensions are: depth: 22.5 cm, length: 28 cm, and height: 19 cm.

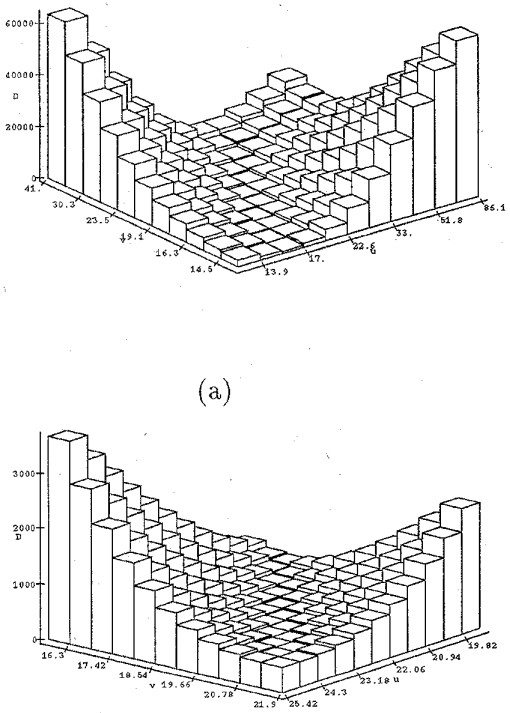

Fig. 5. Plots of the upper bound on N_{im} (26), when comparing the picture in Fig. 4 to a box model whose dimensions are $u \times v \times 28$ cm. (a) Plot of the bound, for coarse sampling in log scale of $u \in$ [13.9 – 86.1 cm] and $v \in$ [14.5 – 41 cm]. (b) Plot of the bound, for fine sampling in linear scale of $u \in$ [19.82 – 25.42 cm] and $v \in$ [16.3 – 21.9 cm].

Another potential application of the metric is in evaluating hypothesized correspondences in an alignment algorithm. Alignment is a method for evaluating the similarity between models and images based on a small number of correspondences. While the use of few correspondences is advantageous for recognizing objects in polynomial time complexity while overcoming partial occlusion, it may often yield errors in estimating the distance between models and images (see, e.g., [5]). Therefore, typical algorithms often try, after obtaining an initial alignment, to extend the match with additional correspondences (e.g., [4]). The bounds derived on the image metric may be used at this stage to evaluate potential correspondences. Simulations [2] show that these bounds often provide better estimates than those provided by using alignment.

Finally, our transformation metric can be used in schemes that attempt to classify objects. A scheme for classification was recently proposed [1], in which classes contain objects that share the same basic features in distorted positions. Our metric can be used under such a scheme to evaluate the amount of affine distortion applied to the object relative to a prototype object in order to determine its class identity.

REFERENCES

[1] R. Basri, "Recognition by Prototypes," *Computer Vision and Pattern Recognition (CVPR-93)*, New York, 1993.
[2] R. Basri and D. Weinshall, "Distance Metric Between 3D Models and 2D Images for Recognition and Classification," MIT AI Memo #1373, 1992.
[3] D.F. DeMenthon and L.S. Davis, "Model-Based Object Pose in 25 Lines of Code," *Proc. Second European Conf. Computer Vision*, Santa Margherita Ligure, Italy. Springer-Verlag, 1992.
[4] M.A. Fischler and R.C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Comm. ACM*, vol. 24, pp. 381-395, 1981.
[5] W.E.L. Grimson, D.P. Huttenlocher, and T.D. Alter, "Recognizing 3D Objects from 2D Images: An Error Analysis," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Urbana, Ill., 1992.
[6] R. Horaud, B. Conio, O. Leboulleux, and B. Lacolle, "An Analytic Solution for the Perspective 4-Point Problem," *Computer Vision, Graphics, and Image Processing*, vol. 47, pp. 33-44, 1989.
[7] D.P. Huttenlocher and S. Ullman, "Recognizing Solid Objects by Alignment with an Image," *Int'l J. Computer Vision*, vol. 5, no. 2, pp. 195-212, 1990.
[8] T.Q. Phong, R. Horaud, A. Yassine, and D.T. Pham, "Optimal Estimation of Object Pose from a Single Perspective View," *Proc. Fourth Int'l Conf. Computer Vision*, pp. 534-539, Berlin, 1993.
[9] S. Ullman and R. Basri, "Recognition by Linear Combinations of Models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, pp. 992-1,006, 1991.
[10] D. Weinshall, "Model-Based Invariants for 3D Vision," *Int'l J. Computer Vision*, vol. 10, no. 1, pp. 27-42, 1993.
[11] J.S.C. Yuan, "A General Photogrammetric Method for Determining Object Position and Orientation," *IEEE Trans. Robotics and Animation*, vol. 5, no. 2, pp. 129-142, 1989.