On View Likelihood and Stability

Daphna Weinshall and Michael Merman

***Abstract*—**

We define two measures on views: view likelihood and view stability. View likelihood measures the probability that a certain view of a given $3D$ object is observed; it may be used to identify typical, or "characteristic", views. View stability measures how little the image changes as the viewpoint is slightly perturbed; it may be used to identify "generic" views. Both definitions are shown to be identical up to the prior probability of camera orientations, and determined by the $2D$ metric used to compare images. We analytically derive the stability and likelihood measures for two feature-based $2D$ metrics, where the most stable and most likely view is shown to be the flattest view of the $3D$ shape.

Incorporating view likelihood or stability in $3D$ object recognition and $3D$ reconstruction increases the chance of robust performance. In particular, we propose to use these measures to enhance $3D$ object recognition and $3D$ reconstruction algorithms, by adding a second step where the most likely solution is selected among all feasible solutions. These applications are demonstrated using simulated and real images.

***Keywords*—** **generic views, characteristic views, canonical views, view likelihood, view stability, object recognition, $3D$ reconstruction, Bayesian vision**

1 Introduction

In this paper we address in a systematic way the loose notions of "characteristic" views and "generic" views, by precisely defining and computing view likelihood and view stability. Incorporating these measures in object recognition and $3D$ reconstruction, we argue, increases the chance of robust and predictable performance. To illustrate this point, we start with an intuitive example:

Consider the three images shown in the top row of Fig. 1. Given three objects in the database (illustrated in the bottom row of Fig. 1): a cube, a flat box and an elongated box, a recognition system is asked to match an object to each image. The images were produced in such a way that the left image is actually a picture of the cube, the middle image is a picture of the flat box, and the right image is the elongated box. A typical (good) computer vision recognition system would correctly produce this output, shown in Fig. 1 with white arrows. However, a human looking at those images would prefer the following interpretation: left image ⇒ flat box, middle image ⇒ elongated box, and right image ⇒ cube, as shown in Fig. 1 with thin black arrows. Why would humans make this "mistake"? The answer seems to be: for a good computational reason!

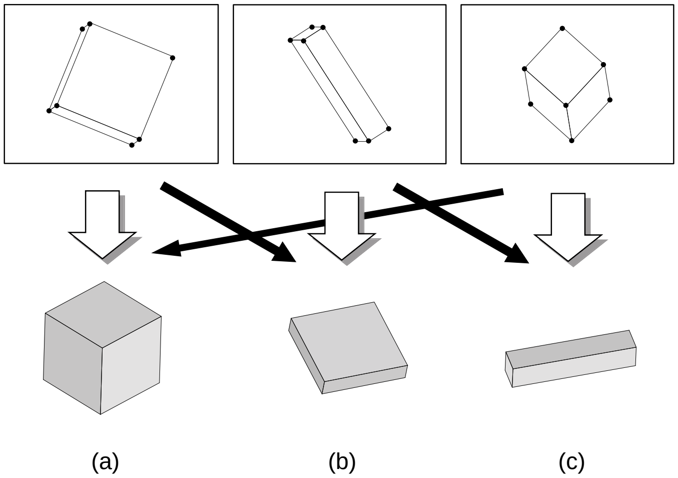

Figure 1: Three polyhedral objects: (a) a cube of dimensions $1 \times 1 \times 1$ cm., (b) a flat box of dimensions $5 \times 5 \times 1$ cm., (c) an elongated box of dimensions $5 \times 1 \times 1$ cm. Top: three images of these objects, obtained from special viewpoints; the extracted features are shown with dark circles. Bottom: the same polyhedral objects viewed from a more typical viewpoint. A geometry-based recognition system would "correctly" recognize each image in the top row as an instance of the object illustrated in the bottom row (this matching is shown by thick white arrows). A maximum likelihood recognition system would recognize the images in the top row as: left = flat box, middle = elongated box, right = cube (this matching is shown by thin black arrows.)

Thus we motivate our work by a paradoxical example: although wrong, the black arrows in Fig. 1 give the statistically *optimal* answer in this example. In other words, a system which gives the wrong answer in this example behaves better *overall*, and will not be mislead by solutions which imply very special viewing positions.

The problem illustrated in Fig. 1 is not limited to simple artificial objects, but generally applies to many complex natural objects whose images may be ambiguous. Take Fig. 2 for example. Based on matching alone, a recognition system that has in its database the models of water bottles, Frisbees, and glass saucers cannot determine which object is depicted in the image. An enhanced recognition system will examine all three solutions and choose the most likely one, either Frisbee or a glass saucer. It will reject the correct but less generic water-bottle solution (unless given a good reason to favor the existence of water bottles in the scene).

Thus the problem with which we are dealing here is gen-

D. Weinshall and M. Werman are with the Institute of Computer Science, Hebrew University of Jerusalem, 91904 Jerusalem, Israel; email: daphna@cs.huji.ac.il.

Manuscript received Apr. 13, 1995. Recommended for acceptance by J. Malik. For information on obtaining reprints of this article, please send e-mail to: transpami@computer.org, and reference IEEECS Log Number P96101.

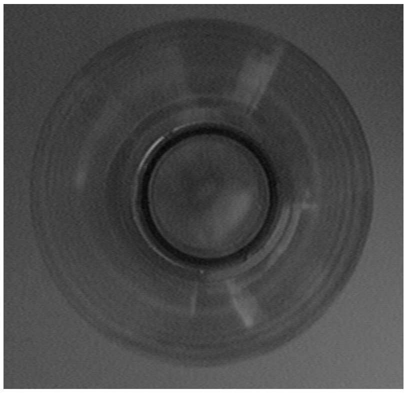

Figure 2: A non generic (unlikely) image of a water-bottle viewed from above, most often interpreted as the image of a glass saucer or Frisbee.

eral, and would plague any recognition system that deals with large databases. The problem arises because most $3D$ object recognition and $3D$ reconstruction algorithms address only the geometrical matching issue: which $3D$ interpretation, or which object, is described in a given $2D$ image. In many cases a good matching algorithm should correctly provide a list of feasible solutions (objects or reconstructed scenes), all consistent with the image up to some uncertainty value arising from measurement errors. To select among these solutions it is necessary to enhance the algorithm with disambiguation criteria, similar to those used by humans in interpreting the images in Fig. 1.

Below we propose to use for disambiguation measures of view likelihood and stability, and choose the more *typical* or *generic* solution. **View likelihood** measures the probability that a certain view of a given object is observed - it characterizes how typical a view is. **View stability** characterizes how generic a view is - it measures how little the image changes as the viewpoint is slightly perturbed. We show that both measures are identical up to the prior probability of camera orientations, and we show how to obtain view stability and view likelihood from the $2D$ metric used to compare images.

The rest of this paper is organized as follows: after reviewing related work in Section 2, we define the concepts of stability and likelihood for images of general objects in Section 3. For the general case we provide fairly simple expressions describing the likelihood and stability of views. One only needs to plug in the particular $2D$ metric which tells images apart. In Section 4 we develop explicit analytical expressions for the stability and likelihood of views of objects with feature points, which depend only on the three principal second moments of the object. The most stable and most likely view is shown to be the flattest view of the object. In Section 5 we demonstrate the usefulness of this theory to $3D$ reconstruction and object recognition. Using examples of simulated and real objects, we show how the measures of view likelihood and stability are: (1) easy to use, and (2) enhance performance when incorporated into existing $3D$ reconstruction and object recognition techniques.

2 Previous work: review and comparison

A few recent studies attempted to approach in a formal way issues related to the ones discussed in this paper:

Bayesian image understanding: Freeman [10, 11] suggested to use a measure of view stability in the interpretation of ambiguous scenes (see also [21]). In his Bayesian scheme, an interpretation which involves the more stable, or more generic, viewpoint is preferred. Freeman's approach is the closest to ours, with the following differences:
On the one hand Freeman uses a more complete probabilistic framework, where image errors (due to the fact that an image is not an instance of the object) are transformed into probabilities and taken into account; prior on models are also taken into account. On the other hand, Freeman computes image likelihood using the Jacobian of the transformation between the viewing parameters and the image measurements. This is the true view likelihood only when the image measurements (normalized by their uncertainty) form a Euclidean space, which is rarely the case (see section 3.5). Our definition of view likelihood is therefore more general, and allows the computation of view likelihood given only the $2D$ metric that is used to compare images. Using this observation, our likelihood measure can be incorporated into Freeman's Bayesian probabilistic scheme by taking the role of conditional image probability.

Measuring likelihood: View likelihood of angles was computed using numerical simulations by Ben-Arie [3] and Burns et al. [6]. Dickinson et al. [8] empirically found the more likely views of particular objects decomposed into geons. In this earlier work, the analysis of likelihood was carried out for simple image measurements: either discrete (qualitative) or 1-dimensional (angles). Note, however, that the general problem requires the numerical estimation of likelihood when the image measurements change continuously with the viewing parameters; this computation is harder, as it requires the numerical estimation of limits. Thus the simulation work described in [3, 6, 8] cannot be readily generalized to compute view likelihood of general objects. Below we provide a simple expression which can be used to numerically estimate the stability and likelihood profiles of general objects, and identify the most likely and stable views of any object.

Measuring stability: Binford & Levitt [4] defined the concept of quasi-invariants, or the local minima of the change in the image when changing the viewing parameters. Other studies proposed to measure stability via the Lie derivatives of the group of transformations describing the motion of the camera [15]. In most of this earlier work, the analysis of stability rarely went beyond the basic definitions (which were different from our definition).

Another line of work, that may superficially appear similar to ours, addresses a very different question. Practically all the probabilistic approaches to image understanding do not take into account image likelihood as defined here (with the exception of [10]). Rather, their goal is to transform an optimization problem to a maximum likelihood problem. This is achieved by defining a probability function which is large when the error is small, and vice versa (see, e.g., [17]); thus the most likely solution, which is selected at the end, is the solution which minimizes the measurement error.

Using the terminology introduced in the introduction, such probabilistic schemes still try to accomplish the first step of recognition: they seek the solution which minimizes the error between the predicted image and the observed image. They do not address the problem of how to choose among all equally plausible solutions, namely, how to choose among those solutions which achieve roughly the same error. Thus these approaches to object recognition can be readily enhanced by our view likelihood measure, using it to fine-tune their probability space to take into account both error minimization and high plausibility (or *genericity*).

The qualitative analysis presented here identifies the most stable and most likely views of objects, which are the most suitable images to be used as the object's *characteristic* views. The concept of characteristic views appears in viewer-centered approaches to $3D$ shape representation, where three dimensional information is not represented explicitly. Rather, the shape of object is represented implicitly by a list of $2D$ characteristic views (e.g., [7]). Our study is the first to give a computational analysis of what makes images *characteristic*.

3 Stability and likelihood of views: general

We define measures of view likelihood and stability, assigned to a general $3D$ object denoted $\mathcal{O}$ and its projection along a specific viewing direction. These measures depend on the variability of the images of object $\mathcal{O}$; thus they depend on the $2D$ metric used to compare those images (e.g., feature-based or intensity-based). In this section the general problem, where the $2D$ metric is not yet fixed, is addressed. We assume general three dimensional objects, including opaque objects with self occlusions.

3.1 Images and the viewing sphere

We first describe how to parameterize all the possible different views, or $2D$ images, of $3D$ objects. This is obtained from the parameterization of all the different viewpoints, or camera orientations, from which a $3D$ object can be observed.

The viewing sphere is an imaginary sphere around the centroid of the object. We assume weak perspective projection, and therefore the viewing sphere describes all the possible different orientations of the camera with respect to the object. Since each camera orientation corresponds to a point on the viewing sphere, all the object's images are completely parameterized by two angles. In the following, φ denotes the azimuth (longitude) and ϑ denotes the elevation (colatitude). The range $\vartheta \in [0, \frac{\pi}{2}]$, $\varphi \in [0, 2\pi]$ parameterizes half the viewing sphere. Let $V_{\vartheta,\varphi}$ denote a viewpoint on the viewing sphere with elevation ϑ and azimuth φ. Let $I_{\vartheta,\varphi;\mathcal{O}}$ denote the $2D$ image (or view) of object $\mathcal{O}$ obtained from viewpoint $V_{\vartheta,\varphi}$.

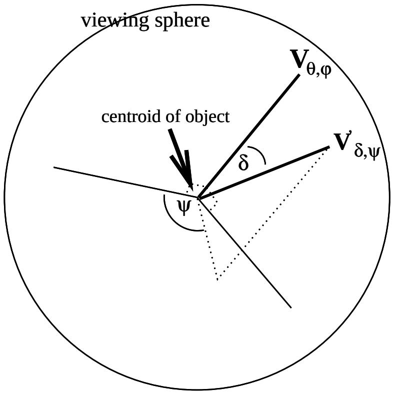

Figure 3: Two views on the viewing sphere, $V_{\vartheta,\varphi}$ and $V_{\delta,\psi}$.

We now define a second parameterization of the viewing sphere. This parameterization is **relative** with respect to a given viewpoint $V_{\vartheta,\varphi}$. In this relative parameterization, δ denotes the azimuth and ψ denotes the elevation, but this time they are both measured with respect to viewpoint $V_{\vartheta,\varphi}$ which serves as the pole (see Fig. 3). At the danger of notation abuse, we denote viewpoints such parameterized by $V_{\delta,\psi}$. Let $I_{\delta,\psi;\mathcal{O}}$ denote the $2D$ image (or view) of object $\mathcal{O}$ obtained from viewpoint $V_{\delta,\psi}$.

3.2 How images differ

The stability and likelihood of a view fundamentally depend on how similar it is to other views of the same object. Intuitively, a typical (or generic) view is one that is similar to many other views of the object, and vice versa. We therefore need to be able to measure the similarity between views. Henceforth we assume that the lack of similarity is measured by some distance measure $d()$, which takes two views as arguments and returns the distance between them. This distance measure can be any. In the present discussion (Section 3) we develop the dependence of the view likelihood and stability on the distance function, whichever it may be. Later on (Section 4) we compute view likelihood and stability for specific cases, substituting specific distance measures.

How does the view likelihood and stability of image $I_{\vartheta,\varphi;\mathcal{O}}$ can be measured? To answer this question, assume that the optical axis of the camera is initially oriented along viewpoint $V_{\vartheta,\varphi}$; then the camera is rotated by δ to observe the object from viewpoint $V_{\delta,\psi}$. Depending on the initial camera orientation $V_{\vartheta,\varphi}$, the difference between the initial and final images $I_{\vartheta,\varphi;\mathcal{O}}$ and $I_{\delta,\psi;\mathcal{O}}$ may be large or small. In other words, having rotated the camera by a fixed amount δ, the image of the object may change only slightly (Fig. 4a,b), or the change may be large (Fig. 4d,e). Intuitively, view $I_{\vartheta,\varphi}$ is stable if the difference between $I_{\vartheta,\varphi;\mathcal{O}}$ and $I_{\delta,\psi;\mathcal{O}}$ is small (Fig. 4a,b); $I_{\vartheta,\varphi}$ is un-stable if the difference is large (Fig. 4d,e).

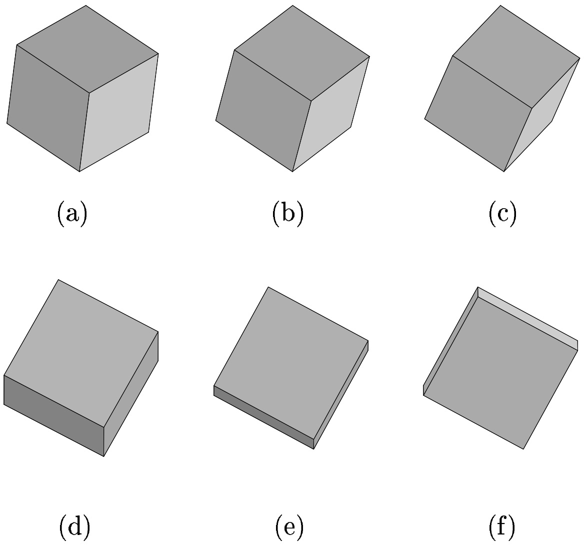

Figure 4: Top: (a) a stable view of a cube, (b) the view obtained when the camera is rotated by 10^o from (a), (c) the view obtained when the camera is rotated by 20^o from (a). Bottom: (d) a less stable view of a cube, (e) the view obtained when the camera is rotated by 10^o from (d), (f) the view obtained when the camera is rotated by 20^o from (d).

Rotating the camera away from the original viewpoint $V_{\vartheta,\varphi}$ by, say, $\delta = 20^o$ is likely to cause a much larger change than rotating it by $\delta = 10^o$ (Fig. 4a-c). Clearly this should not affect the magnitude of view likelihood or stability. We will therefore use the normalized distance $\frac{d()}{\delta}$, at the limit where δ vanishes.

Let $d(\vartheta,\varphi,\delta,\psi;\mathcal{O})$ denote the $2D$ distance between the 2 views $I_{\vartheta,\varphi;\mathcal{O}}$ and $I_{\delta,\psi;\mathcal{O}}$. Intuitively, a good indicator of view likelihood and stability can be obtained from the normalized distance $\frac{d(\vartheta,\varphi,\delta,\psi;\mathcal{O})}{\delta}$, at the limit $\delta = 0$; we denote this function by $D(\vartheta,\varphi,\psi;\mathcal{O})$. Note that, since $d(\vartheta,\varphi,\delta,\psi;\mathcal{O}) = 0$ at $\delta = 0$ (since when δ vanishes $V_{\vartheta,\varphi}$ and $V_{\delta,\psi}$ are the same viewpoint), the following holds:

$$d(\vartheta,\varphi,\delta,\psi;\mathcal{O}) = D(\vartheta,\varphi,\psi;\mathcal{O})\delta + O(\delta^2) \qquad (1)$$

This discussion of view stability and likelihood will now be made precise.

3.3 Stability of views:

The view stability of image $I_{\vartheta,\varphi;\mathcal{O}}$ measures how much the image changes as the viewpoint $V_{\vartheta,\varphi}$ is slightly perturbed. This quantity takes on larger values when the image does not change much as the camera's position is changed.

More specifically, $\frac{\delta}{d(\vartheta,\varphi,\delta,\psi;\mathcal{O})}$ measures the normalized similarity between 2 images obtained from 2 camera orientations separated by δ. The view stability $s(\vartheta,\varphi;\mathcal{O})$ is defined as the average normalized similarity over a small neighborhood of viewpoints, in the limit where the area of the neighborhood vanishes. More precisely:

$$s(\vartheta,\varphi;\mathcal{O}) = \lim_{\varepsilon\to 0} \frac{\int_0^{2\pi}\int_0^{\varepsilon}[\frac{\delta}{d(\vartheta,\varphi,\delta,\psi;\mathcal{O})}]^2 sin(\delta) d\delta d\psi}{\int_0^{2\pi}\int_0^{\varepsilon} sin(\delta) d\delta d\psi}$$

Since $\frac{\delta}{d(\vartheta,\varphi,\delta,\psi;\mathcal{O})} = \frac{1}{D(\vartheta,\varphi,\psi;\mathcal{O})} + O(\delta)$ near $\delta = 0$, and assuming $D(\vartheta,\varphi,\psi;\mathcal{O}) \neq 0 \ \ \forall\psi$:

$$\begin{aligned} s(\vartheta,\varphi;\mathcal{O}) &= \lim_{\varepsilon\to 0} \frac{\int_0^{2\pi} \frac{1}{D^2(\vartheta,\varphi,\psi;\mathcal{O})} \int_0^{\varepsilon} (sin(\delta) + O(\delta^2)) d\delta d\psi}{\int_0^{2\pi}\int_0^{\varepsilon} sin(\delta) d\delta d\psi} \\ &= \frac{1}{2\pi} \int_0^{2\pi} \frac{1}{D^2(\vartheta,\varphi,\psi;\mathcal{O})} d\psi \end{aligned}$$

3.4 Likelihood of views:

View likelihood is the probability induced on the images of object $\mathcal{O}$ by the projection process: for a given prior distribution of camera orientations (or viewpoints), a *different* probability is induced on the views of $\mathcal{O}$ by the projection process. This implies, for example, that even when all the different camera orientations are equally likely, the images are *not* equally likely: the induced probability depends on the $3D$ structure of the object $\mathcal{O}$ and is almost never uniform.

More specifically, we denote the prior distribution of the camera orientations $f(V_{\vartheta,\varphi}) = f(\vartheta,\varphi)$. This distribution induces - via the projection process - a different distribution on the images $I_{\vartheta,\varphi;\mathcal{O}}$ of $\mathcal{O}$, which we denote $l(\vartheta,\varphi;\mathcal{O})$. $l(\vartheta,\varphi;\mathcal{O})$ is the conditional probability to see the particular image $I_{\vartheta,\varphi;\mathcal{O}}$, given that the image is known to be of object $\mathcal{O}$; by our definition, therefore, $I_{\vartheta,\varphi;\mathcal{O}}$ is the view likelihood of $\mathcal{O}$.

First, note that $l(\vartheta,\varphi;\mathcal{O})$ is a density function. In order to compute the value of a density function $l()$ at some point $\mathbf{x}$, one can proceed by computing the cumulative distribution of the appropriate random variable, and then differentiate it to obtain the density of the random variable. The cumulative distribution is typically easier to compute, because it measures the probability of a real (rather than infinitesimal) event: the probability that a random variable obtains a value in some interval $[\mathbf{x}-\varepsilon,\mathbf{x}+\varepsilon]$. In our case the event is $\mid d(\vartheta,\varphi,\delta,\psi;\mathcal{O}) \mid \leq \varepsilon$, namely, we compute the probability that the camera obtains a viewpoint from which the object $\mathcal{O}$ appears different from $I_{\vartheta,\varphi;\mathcal{O}}$ by less than ε. Denoting by χ_b the function which returns 1 when b is true and 0 otherwise, we have:

$$Prob(\mid d(\vartheta,\varphi,\delta,\psi;\mathcal{O}) \mid \leq \varepsilon) = \qquad (2)$$

$$\int_0^{2\pi}\int_0^{\frac{\pi}{2}} f(\delta,\psi)\chi_{\{|d(\vartheta,\varphi,\delta,\psi;\mathcal{O})|\leq\varepsilon\}} sin(\delta)d\delta d\psi$$

From (1) it follows that near $\delta = 0$

$$| \, d(\vartheta,\varphi,\delta,\psi;\mathcal{O}) \, | \leq \varepsilon \quad \Longleftrightarrow \quad \delta \leq \frac{\varepsilon + O(\varepsilon^2)}{D(\vartheta,\varphi,\psi;\mathcal{O})}$$

Substituting this into (2), and assuming $D(\vartheta,\varphi,\psi;\mathcal{O}) \neq 0 \ \ \forall\psi$, we get:

$$Prob(| \, d(\vartheta,\varphi,\delta,\psi;\mathcal{O}) \, | \leq \varepsilon) = \tag{3}$$

$$\int_0^{2\pi} \int_0^{\frac{\varepsilon+O(\varepsilon^2)}{D(\vartheta,\varphi,\psi;\mathcal{O})}} (f(\vartheta,\varphi) + O(\delta))(\delta + O(\delta^2))d\delta d\psi$$

(3) measures the cumulative distribution, corresponding to an area on the viewing sphere. The corresponding density function is proportional to the rate of increase in this area. More precisely[1]:

$$l(\vartheta,\varphi;\mathcal{O}) = \tag{4}$$

$$\lim_{\varepsilon\to 0} \frac{Prob(| \, d(\vartheta,\varphi,\delta,\psi;\mathcal{O}) \, | \leq \varepsilon)}{\pi\varepsilon^2}$$

Substituting (3) into (4) and computing the value at the limit $\varepsilon \to 0$, we get the following expression for the view likelihood of object $\mathcal{O}$:

$$l(\vartheta,\varphi;\mathcal{O}) = \frac{1}{2\pi} f(\vartheta,\varphi) \int_0^{2\pi} \frac{1}{D^2(\vartheta,\varphi,\psi;\mathcal{O})} d\psi \tag{5}$$

(Recall that $f(\vartheta,\varphi)$ denotes the prior distribution of camera orientations.)

3.5 View likelihood: numerical evaluation

In the rest of this paper we primarily use (5) to obtain analytical expressions of view likelihood and stability. However, when the distance between images is complex and only numerical estimation of view likelihood is possible, the integral in (5) may be hard to evaluate. Thus we derive below another expression for view likelihood, which depends only on the derivatives of the $2D$ image distance with respect to the viewpoint parameters ϑ, φ. To keep the focus of this paper, the derivations are only briefly described.

Thinking about the projection from $3D$ space to a $2D$ image as transformation of coordinates, $l(\vartheta,\varphi;\mathcal{O})$ can be obtained from $f(\vartheta,\varphi)$ using the Jacobian of the transformation: in Euclidean spaces, the Jacobian measures how an area element in one coordinate system (on the viewing sphere) changes in another coordinate system (in the image space).

More specifically, if an image is represented by some vector $\mathbf{x} \in \mathcal{R}^n$, the transformation is:

$$T \; : \; \theta \longrightarrow \mathbf{x}, \qquad \theta = \begin{pmatrix} \vartheta \\ \varphi \end{pmatrix}, \quad \mathbf{x} = \{x_i\}_{i=1}^n = I_{\vartheta,\varphi;\mathcal{O}}$$

If the image space $\mathcal{R}^n$ is Euclidean, that is - when the distance between 2 images $\mathbf{x}, \mathbf{y}$ is the L_2 norm $\|\mathbf{x} - \mathbf{y}\|$, an area element on the viewing sphere $\sin\vartheta d\vartheta d\varphi$ locally deforms by

$$\sqrt{\det{(MM^T)}}, \quad M_{ij} = \frac{\partial x_j}{\partial \theta_i}$$

In the general metric case, when the distance between the 2 images $\mathbf{x}, \mathbf{y}$ is some arbitrary "good" metric[2] $d(\mathbf{x}, \mathbf{y})$, the local area deformation can be shown to be

$$\sqrt{\det{(MG(\mathbf{x})M^T)}}, \quad G(\mathbf{x})_{ij} = \frac{1}{2}\frac{\partial^2 d^2(\mathbf{x},\mathbf{y})}{\partial x_i \partial x_j}_{|\mathbf{y}=\mathbf{x}} \tag{6}$$

where elements of the matrix $MG(\mathbf{x})M^T$ are:

$$(MGM^T)_{ij} = \sum_{k,l} M_{ik}G_{kl}M_{jl} =$$

$$\sum_{k,l} \frac{1}{2}\frac{\partial^2 d^2(\mathbf{x}(\theta),\mathbf{y})}{\partial x_k \partial x_l}_{|\mathbf{y}=\mathbf{x}} \frac{\partial x_k}{\partial \theta_i}\frac{\partial x_l}{\partial \theta_j} = \frac{1}{2}\frac{\partial^2 d^2(\mathbf{x}(\theta),\mathbf{y})}{\partial \theta_i \partial \theta_j}_{|\mathbf{y}=\mathbf{x}}$$

Substituting $\theta = [\vartheta, \varphi]$ into (6), the area deformation becomes

$$\sqrt{\det{(MGM^T)}} = \frac{1}{2}\sqrt{\begin{vmatrix} \frac{\partial^2 d^2(\mathbf{x}(\theta),\mathbf{y})}{\partial\vartheta^2} & \frac{\partial^2 d^2(\mathbf{x}(\theta),\mathbf{y})}{\partial\vartheta\partial\varphi} \\ \frac{\partial^2 d^2(\mathbf{x}(\theta),\mathbf{y})}{\partial\vartheta\partial\varphi} & \frac{\partial^2 d^2(\mathbf{x}(\theta),\mathbf{y})}{\partial\varphi^2} \end{vmatrix}_{|\mathbf{y}=\mathbf{x}}}$$

The image likelihood corresponds to the area deformation caused by the *inverse* mapping from the image space to the viewing sphere; therefore

$$l(\vartheta,\varphi;\mathcal{O}) = \frac{f(\vartheta,\varphi)\sin\vartheta}{\sqrt{\det{(MGM^T)}}} = \tag{7}$$

$$\frac{2f(\vartheta,\varphi)\sin\vartheta}{\sqrt{\frac{\partial^2 d^2(\mathbf{x}(\vartheta,\varphi),\mathbf{y})}{\partial\vartheta^2}\frac{\partial^2 d^2(\mathbf{x}(\vartheta,\varphi),\mathbf{y})}{\partial\varphi^2} - (\frac{\partial^2 d^2(\mathbf{x}(\vartheta,\varphi),\mathbf{y})}{\partial\vartheta\partial\varphi})^2}_{|\mathbf{y}=\mathbf{x}}}$$

3.6 View likelihood and stability: summary

Let

$$d(\vartheta,\varphi,\delta,\psi;\mathcal{O}) = D(\vartheta,\varphi,\psi)\delta + O(\delta^2), \quad D(\vartheta,\varphi,\psi) \neq 0$$

denote the $2D$ distance between 2 images $I_{\vartheta,\varphi;\mathcal{O}}$ and $I_{\delta,\psi;\mathcal{O}}$. The view likelihood $l(\vartheta,\varphi;\mathcal{O})$ and view stability $s(\vartheta,\varphi;\mathcal{O})$ of image $I_{\vartheta,\varphi;\mathcal{O}}$ are the following:

$$s(\vartheta,\varphi;\mathcal{O}) = \frac{1}{2\pi}\int_0^{2\pi}\frac{1}{D^2(\vartheta,\varphi,\psi)}d\psi \tag{8}$$

$$l(\vartheta,\varphi;\mathcal{O}) = prior(\vartheta,\varphi)s(\vartheta,\varphi;\mathcal{O}) \tag{9}$$

View stability and likelihood can also be computed from the following expression:

$$s(\vartheta,\varphi;\mathcal{O}) = \tag{10}$$

$$\frac{2\sin\vartheta}{\sqrt{\frac{\partial^2 d^2(\vartheta,\varphi,\delta,\psi;\mathcal{O})}{\partial\vartheta^2}\frac{\partial^2 d^2(\vartheta,\varphi,\delta,\psi;\mathcal{O})}{\partial\varphi^2} - (\frac{\partial^2 d^2(\vartheta,\varphi,\delta,\psi;\mathcal{O})}{\partial\vartheta\partial\varphi})^2}_{|\delta=0}}$$

[1] Note that in (4) the cumulative probability is divided by $\pi\varepsilon^2$, not by ε. This is because the parameter space is 2-dimensional, and thus differentiation is obtained by dividing by an infinitesimal area element, not length.

[2] A "good" metric can be locally approximated to first order by the Euclidean distance.

We are also interested in 2 qualitative characterizations of views:

Most likely view: $I_{\bar{\vartheta},\bar{\varphi};\mathcal{O}}$ which satisfies

$$l(\bar{\vartheta},\bar{\varphi};\mathcal{O}) = \max_{\vartheta,\varphi} l(\vartheta,\varphi;\mathcal{O}) \tag{11}$$

Most stable view: $I_{\bar{\vartheta},\bar{\varphi};\mathcal{O}}$ which satisfies

$$s(\bar{\vartheta},\bar{\varphi};\mathcal{O}) = \max_{\vartheta,\varphi} s(\vartheta,\varphi;\mathcal{O}) \tag{12}$$

It follows from (9):

Result 1: *If the prior distribution of camera orientations of a given object is uniform, namely, all viewpoints are equally likely, then:*

1. *The view stability and likelihood functions are the same.*
2. *The* **most stable** *and* **most likely** *views are the same image.*

With this result we are able to restrict the discussion henceforth to view stability. In order to obtain the view likelihood, each expression should be multiplied by the prior distribution of camera orientations.

4 Stability and likelihood of views: feature-based metrics

In the previous section we derived the dependence of the view stability and likelihood on the $2D$ metric used to compare images. To show the usefulness of this general approach, we now derive explicit forms for the view likelihood and stability given two specific feature-based $2D$ matching metrics.

Given objects composed of feature points, there exist natural $2D$ metrics to compare the images of such objects [20]:

affine metric d_{aff}**:** the two images are first aligned with each other with the best $2D$ affine transformation, and then the sum of the squared distances between each pair of matching feature points is taken.

similarity metric d_{sim}**:** the two images are first aligned with each other with the best $2D$ similarity transformation, and then the sum of the squared distances between each pair of matching feature points is taken.

Simplified expressions for these metrics are derived next, using a coordinate system rotated such that $V_{\vartheta=0,\varphi=0}$ is the flattest view of the object. The flattest view is the view where the image achieves maximal spread: it is obtained from the viewing direction along which the three dimensional object has its minimal spread.

4.1 Image representation

Let $\{\hat{\mathbf{p}}_i = (\hat{x}_i, \hat{y}_i, \hat{z}_i)\}_{i=1}^n$ denote the coordinates of the object features in the initial camera coordinate system in $\mathcal{R}^3$. A three dimensional representation of the object is the $3\times n$ matrix $\hat{\mathbf{P}}$, whose i-th column is $\hat{\mathbf{p}}_i$ - the vector describing the world coordinates of the i-th feature of the object.

An image of the object is approximated by a rigid transformation (of the object or the camera), followed by weak perspective (or scaled orthographic) projection from three dimensional space to the two dimensional image. An image of the object is therefore the set of n image points $\{\mathbf{p}_i = (x_i, y_i)\}_{i=1}^n$. An equivalent representation of the image is the $2\times n$ matrix $\mathbf{P}$, whose i-th column is $\mathbf{p}_i$ - the vector of image coordinates of the i-th feature of the object.

4.2 $2D$ image comparison

Given two images, or the two matrices $\mathbf{P}$ and $\mathbf{Q}$, the question of comparing them is equivalent to matrix comparison. We are using the usual metric, which is the Frobenius norm of the difference matrix, and which is the same as the Euclidean distance between points in the images:

$$\|\mathbf{P}-\mathbf{Q}\|_F^2 = \sum_{i,j}(\mathbf{P}[i,j]-\mathbf{Q}[i,j])^2 = tr[(\mathbf{P}-\mathbf{Q})\cdot(\mathbf{P}-\mathbf{Q})^T]$$

(tr denotes the trace of a matrix). Henceforth we shall omit the subscript F, and a matrix norm will be the Frobenius norm by default.

Before taking the norm of the difference between the images, we want to remove differences which are due to irrelevant effects, such as the size of the image (which is arbitrary under scaled orthography) or the exact location of the object (e.g., due to an arbitrary translation and rotation of the object in the image). In particular, we consider as equivalent all images obtained from each other by one of the following two groups of two dimensional transformations: the **similarity** group, which includes $2D$ rotations, translations, and scale, or the **affine** group, which includes $2D$ linear transformations and translations.

It can be readily shown that the optimal translation when measuring the distance by the sum of square differences, under both the similarity and affine equivalence, puts the centroid of the object at the origin of the image. We therefore assume w.l.o.g. that the images are centered at the centroid of the object, so that the first moment of each image is 0. In [20] we defined image metrics, which compare the images $\mathbf{P}$ and $\mathbf{Q}$ while taking into account the desired image equivalences discussed above. We get the following expressions for the similarity-equivalence metric $d_{sim}(\mathbf{P},\mathbf{Q})$ and the affine-equivalence metric $d_{aff}(\mathbf{P},\mathbf{Q})$:

$$\begin{aligned} d^2_{sim}(\mathbf{P},\mathbf{Q}) &= 1 - \frac{\|\mathbf{QP}^T\|^2 + 2det(\mathbf{QP}^T)}{\|\mathbf{P}\|^2\|\mathbf{Q}\|^2} \\ d^2_{aff}(\mathbf{P},\mathbf{Q}) &= 2 - tr(\mathbf{P}^+\mathbf{P}\cdot\mathbf{Q}^+\mathbf{Q}) \end{aligned} \tag{13}$$

where $A^+ = (A^TA)^{-1}A^T$ denotes the pseudo-inverse of a matrix A.

4.3 The flattest view

Any image $\mathbf{P}$ is obtained from some viewpoint $V_{\vartheta,\varphi}$ of the object by the weak perspective projection of $\mathbf{V}_{\vartheta,\varphi}\hat{\mathbf{P}}$, where

$\mathbf{V}_{\vartheta,\varphi}$ is a $3D$ orthogonal matrix. Let $\mathbf{S}_{\vartheta,\varphi}$ denote the 3×3 symmetric autocorrelation scatter matrix of the object at viewpoint $V_{\vartheta,\varphi}$:

$$\mathbf{S}_{\vartheta,\varphi} = \mathbf{V}_{\vartheta,\varphi}\hat{\mathbf{P}} \cdot (\mathbf{V}_{\vartheta,\varphi}\hat{\mathbf{P}})^T$$

Definition 1: *[The flattest view $I_{\vartheta,\varphi;\mathcal{O}}$:] is the view obtained by parallel projection of the object $\mathcal{O}$ from viewpoint $V_{\vartheta,\varphi}$ whose scatter matrix $\mathbf{S}_{\vartheta,\varphi}$ is diagonal, and where the eigenvalues (the diagonal elements) are ordered in decreasing order. Let $\mathbf{S}_f$ denote the scatter matrix at the flattest view, then:*

$$\mathbf{S}_f = \begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix}$$

where $a \geq b \geq c > 0$ (a, b, c are the 3 principal second moments of the object).

Recall that a symmetric matrix can always be diagonalized by a similarity transformation with an orthogonal matrix. Such a diagonalization of the initial scatter matrix $\mathbf{S}_{\vartheta=0,\varphi=0}$ is equivalent to a rotation of the coordinate system defining the object shape matrix $\hat{\mathbf{P}}$. Thus it is easy to compute the rotation of the object from its initial representation, so that the flattest view will correspond to $V_{\vartheta=0,\varphi=0}$. This rotation is the orthogonal matrix which diagonalizes the original scatter matrix of the object $\mathbf{S}_{\vartheta=0,\varphi=0}$. It is unique (up to a rotation around the optical axis of the camera) if $b > c$. Henceforth we will assume w.l.o.g. that the viewing sphere is initially parameterized so that $V_{\vartheta=0,\varphi=0}$ is the flattest view.

As an example, consider a three dimensional straight corner, an object composed of the points: $\{(0,0,0),\ (1,0,0),\ (0,1,0),\ (0,0,1)\}$. After centering this object, its three principal second moments are $a = 1,\ b = 1,\ c = 0.25$ (note that they are not all 1!). The flattest view of this object is shown in Fig. 5.

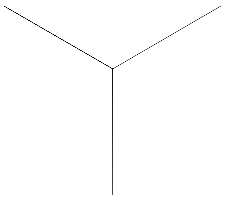

Figure 5: The flattest view of a straight corner.

4.4 The distance between two views

As defined in Section 3.1, let $V_{\vartheta,\varphi}$ and $V_{\delta,\psi}$ denote 2 viewpoints on the viewing sphere of a given object. Let $d(\vartheta,\varphi,\delta,\psi;\mathcal{O})$ denote the image distance between the corresponding images $I_{\vartheta,\varphi;\mathcal{O}}$ and $I_{\delta,\psi;\mathcal{O}}$, where $d(\vartheta,\varphi,\delta,\psi;\mathcal{O})$ is one of the two distance metrics defined in (13).

Simplified expressions for the affine distance: $d_{aff}(\vartheta,\varphi,\delta,\psi;\mathcal{O}) = D_{aff}(\vartheta,\varphi,\psi;\mathcal{O})\delta + O(\delta^2)$, and for the similarity distance: $d_{sim}(\vartheta,\varphi,\delta,\psi;\mathcal{O}) = D_{sim}(\vartheta,\varphi,\psi;\mathcal{O})\delta + O(\delta^2)$, are given in Appendix A. From the general expressions given in (16) and (17), derived for an object whose 3 principal moments (the 3 eigenvalues of the autocorrelation matrix of the feature points) are $a \geq b \geq c > 0$, it follows that:

$$D_{aff}(\vartheta,\varphi,\psi;\mathcal{O}) = \frac{\sqrt{abc}}{ab\cos^2\vartheta + ac\cos^2\varphi\sin^2\vartheta + bc\sin^2\varphi\sin^2\vartheta} \cdot \sqrt{as_1 + bs_2 + cs_3}$$

where

$$\begin{aligned} s_1 &= (\cos\psi\cos\varphi - \sin\psi\cos\vartheta\sin\varphi)^2 \\ s_2 &= (\cos\psi\sin\varphi + \sin\psi\cos\vartheta\cos\varphi)^2 \\ s_3 &= \sin^2\psi\sin^2\vartheta \end{aligned}$$

and

$$D_{sim}(\vartheta,\varphi,\psi;\mathcal{O}) = \frac{\sqrt{bc(1-\sin^2\varphi\sin^2\vartheta) + ac(1-\cos^2\varphi\sin^2\vartheta) + ab\sin^2\vartheta}}{a(1-\sin^2\varphi\sin^2\vartheta) + b(1-\cos^2\varphi\sin^2\vartheta) + c\sin^2\vartheta}$$

Clearly $D_{sim}(\vartheta,\varphi,\psi;\mathcal{O}) > 0$ and $D_{aff}(\vartheta,\varphi,\psi;\mathcal{O}) > 0$, $\forall\vartheta,\varphi,\psi$.

It immediately follows that:

Result 2: *For both the similarity and affine metrics, $d(\vartheta,\varphi,\delta,\psi;\mathcal{O})$ depends only on the 3 principal second moments (a,b,c) of the object $\mathcal{O}$, regardless of the number of features in the object or their distribution in space.*

This result shows that the 3 principal second moments of an object completely characterize the stability and likelihood of each of its views, regardless of the particular shape of the object. Note that this is a result, and not an assumption, of our analysis.

4.5 View likelihood and stability

Substituting $D_{aff}(\vartheta,\varphi,\psi;\mathcal{O})$ and $D_{sim}(\vartheta,\varphi,\psi;\mathcal{O})$ into the definition (8) (similar results are obtained by substituting (13) into (10)), we obtain the view stability of an object whose 3 principal moments are $a \geq b \geq c$. Each image metric defines a different measure:

$$s_{aff}(\vartheta,\varphi;\mathcal{O}) = \int_0^{2\pi} \frac{1}{as_1 + bs_2 + cs_3} d\psi \qquad (14)$$

$$\cdot \frac{1}{2\pi} \frac{(ab\cos^2\vartheta + ac\cos^2\varphi\sin^2\vartheta + bc\sin^2\varphi\sin^2\vartheta)^2}{abc} = \frac{(ab\cos^2\vartheta + ac\cos^2\varphi\sin^2\vartheta + bc\sin^2\varphi\sin^2\vartheta)^{\frac{3}{2}}}{abc}$$

$$s_{sim}(\vartheta,\varphi;\mathcal{O}) = \qquad (15)$$

$$\frac{(a(1-\sin^2\varphi\sin^2\vartheta) + b(1-\cos^2\varphi\sin^2\vartheta) + c\sin^2\vartheta)^2}{bc(1-\sin^2\varphi\sin^2\vartheta) + ac(1-\cos^2\varphi\sin^2\vartheta) + ab\sin^2\vartheta}$$

It follows from these expressions (e.g., by differentiation) that:

Result 3: *[Characteristic views:]*

- *In each aspect, under both the affine and similarity $2D$ metrics and if $a \geq b > c$, the most stable view is unique, and it is the flattest view $V_{\vartheta=0,\varphi=0}$.*
- *The stability of the flattest view is $\frac{a+b}{c}$ under the similarity metric, and $\frac{\sqrt{ab}}{c}$ under the affine metric.*
- *The stability at viewpoint $V_{\vartheta,\varphi}$ decreases monotonically with its geodesic distance from the flattest view ϑ for both the affine and similarity $2D$ metrics.*

Proof: The result follows immediately from (14)-(15), as $[\vartheta = 0, \varphi = 0]$ is clearly the only maximum point of both $s_{aff}(\vartheta, \varphi; \mathcal{O})$ and $s_{sim}(\vartheta, \varphi; \mathcal{O})$ for any object $\mathcal{O} = [a, b, c]$. By differentiating with respect to ϑ, it can be shown that both functions are everywhere monotonically decreasing with ϑ.
□

Corollary 4: *Given a $2D$ image, and without any additional information about its $3D$ structure, the most likely interpretation is the flattest view, implying that the image represents a fronto-parallel flat object.*

This corollary justifies the heuristic, which assigns depth 0 to points whose depth is unknown, as the optimal decision based on the available information in one image. In fact, this heuristic is typically employed by iterative reconstruction algorithms which assign depth 0 as the default value in the first iteration (e.g., [13]).

5 Applications: enhanced object recognition and reconstruction

The following examples illustrate various applications of view likelihood in object recognition and $3D$ reconstruction. We start by computing some characteristic views in Section 5.1. In Section 5.2 we demonstrate maximum likelihood $3D$ object recognition using simulated images. In Section 5.3 we demonstrate maximum likelihood $3D$ reconstruction.

These applications take advantage of the dependence of the view likelihood $l(\vartheta, \varphi; \mathcal{O})$ on the object $\mathcal{O}$. Treating $l(\vartheta, \varphi; \mathcal{O})$ as a function of $\mathcal{O}$, we can estimate $\mathcal{O}$ using maximum likelihood estimation: the object $\mathcal{O}$ which maximizes the view likelihood is chosen among all possible objects.

5.1 Characteristic views

It seems plausible to choose the characteristic view in each aspect of the object to be the most stable and likely view in the aspect. From Section 4 we identify the characteristic view of n features to be the flattest view of the features, or the fronto-parallel view of 3 features. To illustrate this result we shall compute the flattest view in each aspect of specific objects, where an aspect includes all the views of the object from which the same features are visible.

First, to obtain the representation in the canonical coordinate system assumed in Section 4.3, we: (1) translate the coordinate system so that the centroid of the visible feature points (in the aspect) is $(0,0,0)$, (2) rotate the coordinate system so that the scatter matrix of the visible feature points is a diagonal matrix, with the diagonal elements decreasingly ordered.

Given a square (non-transparent) pyramid, whose nodes are at $\{(0,0,2), (1,0,0), (0,1,0), (-1,0,0), (0,-1,0)\}$, we analyze the aspect where 4 feature points, 3 of the basis nodes $\{(1,0,0), (0,1,0), (-1,0,0)\}$ and the top of the pyramid $(0,0,2)$, are visible. The flattest view of this aspect is shown in Fig. 6. The flattest view of a box, in one of its aspects, is also shown in Fig. 6. Fig. 5 shows the flattest view of a straight corner.

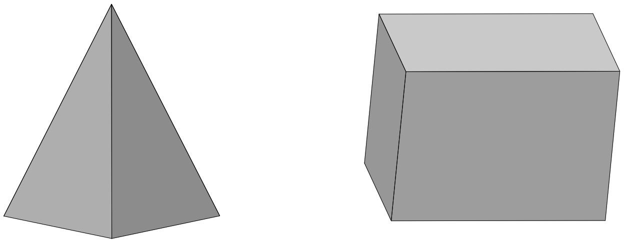

Figure 6: The flattest views of two opaque objects: a square pyramid and a box of dimensions 40 × 30 × 25.

5.2 Maximum likelihood object recognition

We return now to the example discussed in the introduction and illustrated in Fig. 1. We are given a (large) database, which includes the 3 polyhedral objects shown at the bottom row of Fig. 1. Each object in the database is represented by the coordinates of a set of feature points (8 vertices in the case of the 3 polyhedral objects). To recognize the 3 images shown at the top row of Fig. 1, we proceed as follows:

Step 1: geometrical object recognition A feature-based object recognition technique (e.g., alignment [14] or geometric hashing [16]) is first used to compute a list of candidate objects which match the given images. In order to predict which objects each system finds without simulating the actual algorithm, we perform the following meta-analysis: We compute the model-to-image distances[3] described in [2] to evaluate how well each object fits each image. The objects which achieve a small model-to-image distance are feasible matches for these geometrical object recognition methods. We therefore assume that all objects which achieve a subthreshold model-to-image distance are chosen by the object recognition method of choice. The model-to-image distances are given in Table 1.
It follows that an affine-based recognition algorithm, such as geometric hashing or linear combination [18], will produce the set of all 3 objects as feasible matches for each image. Assuming an error threshold tolerance

[3] A model-to-image distance measures the distance between the closest view of the object to the given image, up to some $2D$ image transformation.

	left image	middle image	right image
cube	0 (0)	0.3 (0)	0.006 (0)
flat box	0.01 (0)	0 (0)	0.11 (0)
elongated box	0.0005 (0)	0.009 (0)	0 (0)

Table 1: The model-to-image distances between each of 3 objects in the database: a cube, a flat box, and an elongated box, to the 3 images shown in the top row of Fig. 1. The distances up to $2D$ similarity transformation are given, whereas the distances up to $2D$ affine transformations are given in parentheses. Whenever the numbers are small, it means that there exists a viewpoint from which the object appears similar to the image up to $2D$ rotation and scale.

of about 0.05, a similarity-based algorithm, such as alignment, will produce shorter lists:

- left image $\Longrightarrow$ {cube, flat box, elongated box}
- middle image $\Longrightarrow$ {flat box, elongated box}
- right image $\Longrightarrow$ {elongated box, cube}

The first object in each list matches the image exactly, and corresponds to the matching shown with thick white arrows in Fig. 1. The second object in each list predicts the image quite well (the difference is illustrated in Fig. 7), and corresponds to the matching shown with thin black arrows in Fig. 1.

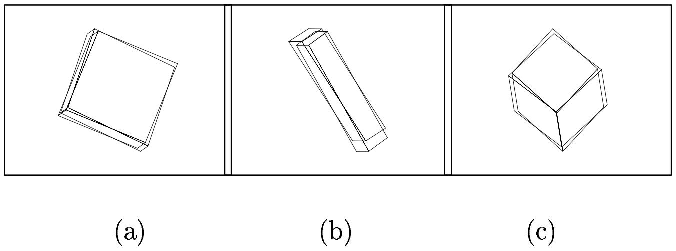

Figure 7: When the images in the top row of Fig. 1 are matched to the most likely object (illustrated by thin black arrows in Fig. 1), the matching is not precise. For each image, the difference between the closest view of the recognized object and the image is shown.

Step 2: maximum likelihood object recognition The view likelihood is used to select the best interpretation from the list produced by the geometrical object recognition algorithm. We assume a uniform prior, so that $l(\vartheta, \varphi; \mathcal{O}) = s(\vartheta, \varphi; \mathcal{O})$. $l(\vartheta, \varphi; \mathcal{O})$ is computed for each pair [object=$\mathcal{O}$, image] obtained as a possible solution in step 1: it is the likelihood of the closest view of the object $\mathcal{O}$ to the image. (The computation of $s(\vartheta, \varphi; \mathcal{O})$ for one feasible [object,image] pair is described in detail in appendix B.) Table 2 contains the *view likelihood* values.

Thus the most-likely and most-stable interpretation for each image, based on either the affine or the similarity metrics, is:

left image: the flat box (at least 6 times as likely as any other object)

middle image: the elongated box (an order of magnitude more likely than the other object suggested by the alignment method)

object/image	left image	middle image	right image
cube	2.24, 1.16	(2.6), 1.36	3.12, 1.6
flat box	13.2, 19.4	1.16, 0.36	(1.12), 1.76
elongated box	0.09, 0.48	27.6, 6.12	0.15, 0.1

Table 2: Each entry in the table corresponds to a pair [object,image]. Two numbers are given in each entry: the similarity stability defined in (15), and the affine stability defined in (14), of the closest view of the object $\mathcal{O}$ to the image. Non-feasible solutions (from the lists produced in step 1 of the algorithm) are given in parentheses.

right image: the cube according to the similarity metric; the affine metric cannot decide between the cube and the flat box, and should return both as likely solutions

Reading Table 2 by rows, we see that the likelihood of the images of the cube, the most symmetrical object, vary much less than the likelihood of the images of the flat box or the elongated box. This is because the eigenvalues of the cube differ much less (note that in the limit where $a = b = c = 1$, all the images of such an object are equally likely).

5.3 Maximum likelihood reconstruction:

When there exist cues for $3D$ structure, but many reconstructions are feasible, the view likelihood may be used to select the best among the solutions under consideration (cf. [10]). We will describe below 2 examples of such a process: one with an object with fiducial points, the other with a shaded smooth object. In these examples once again, we assume that the prior on the viewing sphere is uniform and thus view stability and view likelihood are identical and can be used interchangeably.

5.3.1 Object matching using fiducial points

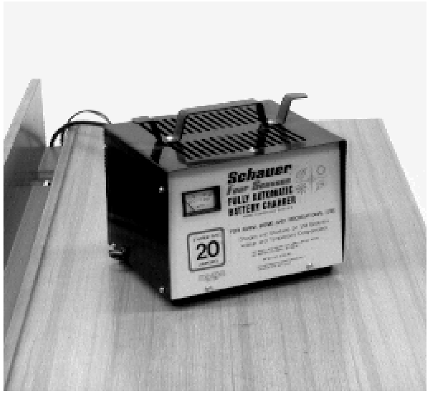

Figure 8: A battery charger. Its real dimensions are: depth - 22.5, length - 28, and height - 19 (all in cm.). When normalized by length, the maximum likelihood estimation of these dimensions gives: depth - 19.5, height - 21.

Consider the battery charger picture shown in Fig. 8. From the text on the object we know that the object is a battery charger, and we presumably know that battery chargers are box-like in shape. Our task is to compute the dimensions (up to scale) of the charger from the image coordinates of the 7 visible vertices of the charger's enveloping box. It follows from the computation described in Appendix C that the most likely interpretation of the picture is a box of dimensions 19.5 × 21 × 28, whereas the actual dimensions of the charger are 22.6 × 19.1 × 28 cm. Thus the picture is interpreted as a bit flatter and shorter than it really is.

5.3.2 Matching smooth objects using grey levels

In order to compute the stability and likelihood of gray level images, we assume a fixed lighting source and the knowledge of the reflectance map of the object. The distance between similarity normalized gray level pictures (pictures that are normalized so that their scale is, say, 1, and their main axis is aligned along the X-axis, for example) is taken to be the sum of squared differences of gray levels. We have computed the stability of this distance function numerically for Lambertian ellipsoids.

Consider the ellipsoids in Fig. 9: the length of the first two axes of the ellipsoid to the left are immediately measurable from the picture, but the height (depth) can only be derived from the gray levels. If we want to select the best interpretation for the height of the ellipsoid given that the picture is noisy, we pick the most likely height such that the distance between rendered pictures is small (less than 5 gray level values per pixel on the average). This results in choosing a flatter ellipsoid with parameters $1, 2, 3.2$ instead of $1, 2, 5$. This flattening effect is consistent with psychological evidence in humans [5]. (Note that human judgment takes into account many other priors and heuristics on the kind of shapes one is likely to encounter; thus we only expect to see correspondence between human performance and the most stable views under rather simple and impoverished conditions, such as the images of Lambertian ellipsoids.)

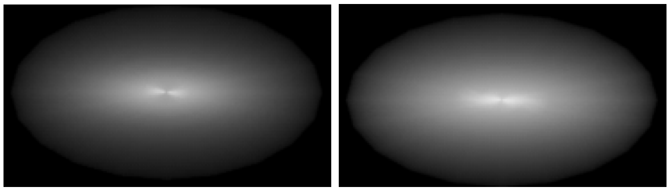

Figure 9: Left: the rendering of an ellipsoid with principal axes of length $1, 2, 5$. Right: the rendering of the most likely reconstruction of the left picture, an ellipsoid with principal axes of length $1, 2, 3.2$.

6 Summary

We described how to obtain *characteristic* views: one can use the view likelihood which measures how *typical* an image is, or the view stability which measures how *generic* an image is. Both measures are identical when the prior distribution of camera orientations is uniform. We showed how to compute these measures in the general case, given only the $2D$ metric that tells images apart. We then elaborated on two examples of feature-based image metrics. The incorporation of these measures in $3D$ object recognition and $3D$ reconstruction is likely to increase the robustness of the system.

A Affine and similarity distances

Using Maple©, we simplified $d(\vartheta, \varphi, \delta, \psi; \mathcal{O})$ for the two metrics defined in (13). For the similarity metric we get:

$$d^2_{sim}(\vartheta, \varphi, \delta, \psi; \mathcal{O}) = \frac{(1 - \cos\delta)\,(abv_1 + acv_2 + bcv_3)}{u(aw_1 + bw_2 + cw_3)} \quad (16)$$

where

$$\begin{aligned}
v_1 &= 1 - 2\,\cos^2\vartheta\cos\delta + 2\,\cos\vartheta\sin\delta\sin\vartheta\cos\psi + \cos\delta \\
v_2 &= 1 - 2\,\cos\delta\cos^2\varphi\sin^2\vartheta + \cos\delta - \\
&\quad 2\,\cos\psi\cos^2\varphi\sin\delta\sin\vartheta\cos\vartheta + \\
&\quad 2\,\sin\psi\sin\delta\sin\varphi\cos\varphi\sin\vartheta \\
v_3 &= 1 + 2\,\cos\delta\cos^2\varphi + 2\,\cos\delta\sin^2\varphi\cos^2\vartheta - \\
&\quad 2\,\sin\psi\sin\delta\sin\varphi\cos\varphi\sin\vartheta - \\
&\quad 2\,\cos\psi\sin^2\varphi\sin\delta\sin\vartheta\cos\vartheta - \cos\delta \\
u &= a\left(1 - \sin^2\varphi\sin^2\vartheta\right) + b\left(1 - \cos^2\varphi\sin^2\vartheta\right) + c\sin^2\vartheta \\
w_1 &= -2\,\cos\delta\cos\psi\sin\delta\cos\vartheta\sin\vartheta\sin^2\varphi - \\
&\quad 2\,\cos\varphi\sin\varphi\cos\vartheta\sin^2\delta\cos\psi\sin\psi - \\
&\quad \cos^2\vartheta\sin^2\delta\cos^2\psi\sin^2\varphi + 1 - \\
&\quad 2\,\cos\varphi\cos\delta\sin\psi\sin\delta\sin\vartheta\sin\varphi - \\
&\quad \sin^2\vartheta\cos^2\delta\sin^2\varphi - \cos^2\varphi\sin^2\delta\sin^2\psi \\
w_2 &= 2\,\cos\varphi\sin\varphi\cos\vartheta\sin^2\delta\cos\psi\sin\psi + \\
&\quad \sin^2\varphi\sin^2\delta\cos^2\psi + \cos^2\varphi\sin^2\delta + \\
&\quad \cos^2\vartheta\cos^2\delta\cos^2\varphi + \cos^2\delta\sin^2\varphi + \\
&\quad 2\,\cos\varphi\cos\delta\sin\psi\sin\delta\sin\vartheta\sin\varphi - \\
&\quad 2\,\cos\delta\cos\psi\sin\delta\cos\vartheta\sin\vartheta\cos^2\varphi - \\
&\quad \cos^2\vartheta\sin^2\delta\cos^2\psi\cos^2\varphi \\
w_3 &= 1 + 2\,\cos\vartheta\sin\delta\sin\vartheta\cos\delta\cos\psi - \\
&\quad \cos^2\vartheta\cos^2\delta - \sin^2\vartheta\sin^2\delta\cos^2\psi
\end{aligned}$$

For the affine metric we get:

$$d^2_{aff}(\vartheta, \varphi, \delta, \psi; \mathcal{O}) = abc\frac{\sin^2\delta(av_1 + bv_2 + cv_3)}{u(abw_1 + acw_2 + bcw_3)} \quad (17)$$

where

$$\begin{aligned}
v_1 &= (\cos\psi\cos\varphi - \sin\psi\cos\vartheta\sin\varphi)^2 \\
v_2 &= (\cos\psi\sin\varphi + \sin\psi\cos\vartheta\cos\varphi)^2 \\
v_3 &= \sin^2\psi\sin^2\vartheta \\
u &= ab\cos^2\vartheta + ac\cos^2\varphi\sin^2\vartheta + bc\sin^2\varphi\sin^2\vartheta \\
w_1 &= (\sin\delta\cos\psi\sin\vartheta - \cos\delta\cos\vartheta)^2 \\
w_2 &= (\cos\delta\sin\vartheta\cos\varphi - \sin\delta\sin\psi\sin\varphi)^2 + \\
&\quad \sin^2\delta\cos^2\psi\cos^2\vartheta\cos^2\varphi - \\
&\quad 2\,\cos\psi\sin\psi\cos\vartheta\cos\varphi\sin^2\delta\sin\varphi + \\
&\quad 2\,\cos\delta\cos\psi\cos^2\varphi\sin\delta\sin\vartheta\cos\vartheta \\
w_3 &= 2\,\cos\delta\sin\delta\sin\varphi\cos\varphi\sin\psi\sin\vartheta +
\end{aligned}$$

$$
\begin{aligned}
&2\cos\delta\cos\psi\sin^2\varphi\sin\delta\sin\vartheta\cos\vartheta + \\
&\sin^2\delta\,(\cos\psi\cos\vartheta\sin\varphi + \sin\psi\cos\varphi)^2 + \\
&\cos^2\delta\sin^2\vartheta\sin^2\varphi
\end{aligned}
$$

B Calculating view likelihood and stability:

Below we go through the details of the computation of $l(\vartheta,\varphi;\mathcal{O})$ and $s(\vartheta,\varphi;\mathcal{O})$ for one [object,image] pair: the flat box and the left image in Fig. 1.

1. We take the stored $3D$ coordinates of the set of 7 matching vertices of the flat box, and translate and rotate them (as described in Appendix 4.3) so that $V_{\vartheta=0,\varphi=0}$ is the flattest view. Let $\mathbf{P}$ denote the 3×7 matrix representing the model of the flat box, whose 3 principal eigenvalues are: $a = 8, b = 5.7, c = 0.25$. The 7 visible vertices of the flat box, before and after the transformation to a canonical system, are:

$$
\begin{bmatrix}
0.2 & -0.2 & -0.2 & 0.2 & 0.2 & 0.2 & -0.2 \\
1 & 1 & -1 & -1 & -1 & 1 & 1 \\
1 & 1 & 1 & 1 & -1 & -1 & -1
\end{bmatrix}
\Longrightarrow
$$

$$
\mathbf{P} = \begin{bmatrix}
0 & 0 & 1.4 & 1.4 & 0 & -1.4 & -1.4 \\
-1.2 & -1.2 & 0.19 & 0.21 & 1.6 & 0.21 & 0.19 \\
0.24 & -0.16 & -0.24 & 0.16 & 0.076 & 0.16 & -0.24
\end{bmatrix}
$$

2. Using the algorithm described in [2], we compute the view (up to similarity transformation) of the flat box closest to the left image. Let $\mathbf{p}$ denote the 2×7 matrix representing this image:

$$
\mathbf{p} = \begin{bmatrix}
0.35 & 0.47 & 1.4 & 1.3 & -0.59 & -1.5 & -1.4 \\
-0.99 & -1.2 & 0.70 & 0.87 & 1.5 & -0.37 & -0.54
\end{bmatrix}
$$

The difference between this view of the flat box and the left image is 0.01 (see Table 1). Fig. 7a shows the closest view of flat box superimposed on the left image.

3. We compute the angles $[\vartheta,\varphi]$ such that:

$$
\Pi s \begin{bmatrix}
\cos\mu & \sin\mu & 0 \\
-\sin\mu & \cos\mu & 0 \\
0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
1 & 0 & 0 \\
0 & \cos\vartheta & \sin\vartheta \\
0 & -\sin\vartheta & \cos\vartheta
\end{bmatrix}
$$

$$
\begin{bmatrix}
\cos\varphi & -\sin\varphi & 0 \\
\sin\varphi & \cos\varphi & 0 \\
0 & 0 & 1
\end{bmatrix}
\cdot \mathbf{P} = \mathbf{p}
$$

where Π is the orthographic projection matrix $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$, s a scalar, and $\mu \in [0, 2\pi)$ representing the rotation of the object in the image plane. We therefore need to solve the following equation:

$$
\frac{\mathbf{p}\cdot\mathbf{P}^+}{\|\mathbf{p}\cdot\mathbf{P}^+\|} = \left[\begin{matrix}
\cos\mu\cos\varphi + \sin\mu\cos\vartheta\sin\varphi \\
-\sin\mu\cos\varphi + \cos\mu\cos\vartheta\sin\varphi
\end{matrix}\right.
$$

$$
\left.\begin{matrix}
-\cos\mu\sin\varphi + \sin\mu\cos\vartheta\cos\varphi & \sin(\mu)\sin\vartheta \\
\sin\mu\sin\varphi + \cos\mu\cos\vartheta\cos\varphi & \cos\mu\sin\vartheta
\end{matrix}\right]
$$

where $\mathbf{P}^+ = \mathbf{P}^T(\mathbf{P}\mathbf{P}^T)^{-1}$ denotes the pseudo-inverse of $\mathbf{P}$. This equation has a unique solution in the range $\vartheta \in [0,\pi], \varphi \in [0, 2\pi]$, since we started with a matrix $\mathbf{p}$ which is a real projection after rotation of the matrix $\mathbf{P}$. The simplest way to compute ϑ,φ is to solve for the matrix equality above element-by-element. For the matrices $\mathbf{p}$ and $\mathbf{P}$ of the flat box we get:

$$
\vartheta = 26^o, \quad \varphi = -15^o, \quad \mu = -37^o.
$$

We now substitute $a = 8, b = 5.7, c = 0.25$ and $\vartheta = 0.46, \varphi = -0.25$ (the values in radians) into (14)-(15), to obtain the measures of *view likelihood* and *view stability* of the left image when compared to the flat box.

C Max-likelihood reconstruction:

We denote the dimensions of the charger shown in Fig. 8 by $d \times h \times 28$, where d is the depth of the charger, h its height, and 28 the length of the front face (which scales the remaining measurements). To find the best reconstruction of the charger, we search the parameter space (d, h) in 2 stages:

1. We first compute the model-to-image distance between each model and the picture. This gives us the function shown in Fig. 10. A picture of the correct model, which obtains a small image error, is shown in Fig. 11-left.

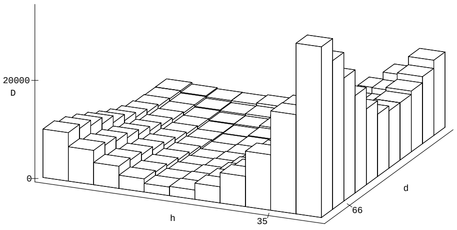

Figure 10: The model-to-image distances, as a function of the parameters d, h, of the picture shown in Fig. 8. The coordinates are drawn in log scale in the ranges $d \in [14, 86]$, $h \in [14.5 - 41]$.

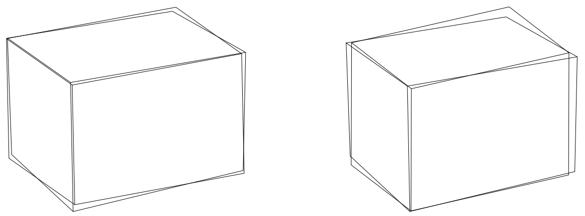

Figure 11: Left: a view of the correct model (which obtains a small model-to-image distance), super-imposed on the image of the original vertices of the charger. Right: the view of the most likely model, superimposed on the original vertices. Clearly, both interpretations match the data reasonably well.

2. We estimate an upper bound on the noise in the image to be 4 times the $3D$-affine distance between the model and the image[4]. Among all the models, whose model-to-image distance is smaller than this noise threshold, we choose the most likely one based on the view likelihood of the best view of each model.

Fig. 12 shows the likelihood of all the interpretations for which the model-to-image distance was smaller than the noise threshold; this function was computed as described in the object recognition example in Section 5.2.

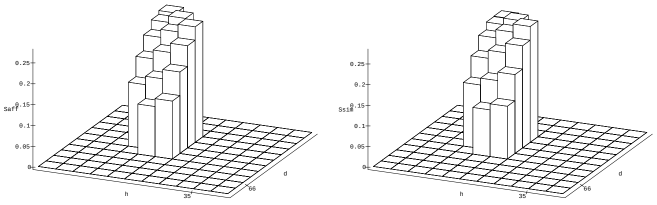

Figure 12: The likelihood of the picture shown in Fig. 8, as a function of the parameters d, h. The coordinates are drawn in log scale, in the ranges $d \in [14 - 86]$, $h \in [14.5 - 41]$. The likelihood is set to 0 for parameter values for which the model-to-image distance is larger than the noise threshold. The left image gives l_{aff}, and the right is l_{sim}.

Acknowledgments: This research was sponsored by ARPA through the U.S. Office of Naval Research under Grant N00014-93-1-1202, R&T Project Code 4424341—01. We thank Bill Freeman, Davi Geiger, Amnon Shashua, Yuval Peres, Yael Karshon, and Dror Bar-Nathan for many useful discussions and suggestions.

References

[1] F. Attneave. Some informational aspects of visual perception. *Psychol. Rev.*, pages 183–193, 1954.

[2] R. Basri and D. Weinshall. Distance metric between 3D models and 2D images for recognition and classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(4):465–470, 1996.

[3] J. Ben-Arie. The probabilistic peaking effect of viewed angles and distances with application to 3-d object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(8):760–774, 1990.

[4] T.O. Binford and T.S. Levitt. Quasi-invariants: Theory and exploitation. *Image Understanding Workshop*, pages 819–829, 1993.

[5] H. H. Bülthoff and H. A. Mallot. Interaction of different modules in depth perception. In *Proceedings of the 1st International Conference on Computer Vision*, pages 295–305, June 1987.

[6] J.B. Burns, R. Weiss, and E. Riseman. View variation of point-set and line segment features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(1):51–68, 1993.

[7] I. Chakravarty and H. Freeman. Characteristic views as a basis for three-dimensional object recognition. In *Proc. SPIE Conf. on Robot Vision*, volume 336, pages 37–45, 1982.

[8] S. J. Dickinson, A. P. Pentland, and A. Rosenfeld. 3-D shape recovery using distributed aspect matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):174–198, 1992.

[9] Attneave F and M.D. Arnoult. The quantitative study of shape and pattern perception. *Psychol. Bull.*, pages 452–471, 1956.

[10] W. T. Freeman. Exploiting the generic view assumption to estimate scene parameters. In *Proceedings of the 4th International Conference on Computer Vision*, pages 347–356, Berlin, Germany, 1993. IEEE, Washington, DC.

[11] W. T. Freeman. The generic viewpoint assumption in a framework for visual perception. *Nature*, 368(6471):542–545, April 7 1994.

[12] Y. Gdalyahu and D. Weinshall. Measures for silhouettes resemblance and the most representative silhouette of a curved object. In *Proceedings of the 4th European Conference on Computer Vision*, Cambridge, UK, 1996. Springer-Verlag.

[13] E. C. Hildreth, N. M. Grzywacz, E. H. Adelson, and V. K. Inada. The perceptual buildup of three-dimensional structure from motion. *Perception & Psychophysics*, 48(1):19–36, 1990.

[14] D. P. Huttenlocher and S. Ullman. Recognizing solid objects by alignment with an image. *International Journal of Computer Vision*, 5:195–212, 1990.

[15] K. Kanatani. *Group Theoretical Methods in Image Understanding.* Springer, Berlin, 1990.

[16] Y. Lamdan and H. Wolfson. Geometric hashing: a general and efficient recognition scheme. In *Proceedings of the 2nd International Conference on Computer Vision*, pages 238–251, Tarpon Springs, FL, 1988. IEEE, Washington, DC.

[17] I. Rigoutsos and R. Hummel. Distributed bayesian object. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, pages 180–186, 1993.

[18] S. Ullman and R. Basri. Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(10):992–1006, 1991.

[19] D. Weinshall and R. Basri. Distance metric between 3d models and 2d images for recognition and classification. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, pages 220–225, New-York City, NY, 1993. IEEE, Washington, DC.

[20] M. Werman and D. Weinshall. Similarity and affine distance between point sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):810–814, 1995.

[21] A. P. Witkin and J. M. Tenenbaum. On the role of structure in vision. In J. Beck, B. Hope, and A. Rosenfeld, editors, *Human and Machine Vision*, pages 481–544. Academic Press, New York, NY, 1983.

[4] $3D$-affine distance between a $3D$ model and a $2D$ image is the difference between the closest view of the model, transformed by any linear transformation, to the image. This distance should be 0 in our example for every model under consideration, since the image is a view of a box, which is related by an affine transformation to any other box. When this number is different from 0, it gives us an estimate of the noise in the image.