MEMO #93–09
July 1993

View-Based Recognition[1]

Thomas M. Breuel
IDIAP, C.P. 609, 1920 Martigny, Switzerland
tmb@idiap.ch

ABSTRACT

In this paper, we propose view-based recognition, a method for 3D object recognition based on multi-view representations. We analyze view-based recognition and compare its performance theoretically and empirically with one of the most commonly used method for 3D object recognition, 3D bounded error recognition. In particular, we show that the probability of false positive or false negative matches in a view-based recognition system is not substantially different from the probability of similar errors in other commonly used recognition systems. Furthermore, we derive an upper bound on the number of views needed to be stored by a view-based recognition system in order to achieve zero probability of false negative matches. Simulations and experiments on real images suggest that these estimates are conservative and that view-based recognition is a robust and simple alternative to the more traditional 3D shape based recognition methods.

Introduction

In this paper, I describe and analyze a view-based recognition (VBR) system for the recognition of 3D objects in 2D images.

Unlike previous 3D recognition systems, which have generally combined both view-based and 3D model-based approaches,[4,7] this system uses a *strictly view-based* approach to the representation of 3D objects. That is, a model of a 3D object consists simply of a collection of 2D views of the 3D object. In order to recognize objects in images, all the views representing each 3D object are compared against the image using a 2D matching algorithm.

View-based approaches to 3D object recognition have several important advantages over 3D model-based approaches. VBR greatly simplifies model acquisition problem, the representation of partial object models, the representation of smooth surfaces, and the modeling of effects such as lighting and shadows. In practice, VBR turns out to be very robust and easy to implement. And, VBR allows us to address questions such as similarity measures and recognition by parts in a simpler 2D (rather than 3D) framework.

[1]A version of this paper has appeared in MVA '92 (Breuel, 1992).

Despite these obvious advantages, the acceptance of view-based methods has been hindered by concerns about the space- and time-requirements of such methods ("how many views are needed?"), and by the approximate and seemingly heuristic nature of view-based approach. To address these concerns, I present a number of theoretical and empirical results.

Bounded Error Recognition

The formalization of the 3D recognition problem that we chose here is that of *bounded error recognition*. Bounded error recognition has been studied extensively in the computer vision literature and forms the basis of many different recognition systems (Grimson[5] gives an extensive review and references).

The idea behind 3D recognition under bounded error is the following. First, we assume that objects have visual characteristics (*features*) that can be localized in images and transform as if they were rigidly attached to the object as the object undergoes 3D rigid body transformations.

In order to account for variability in object shape, limited sensor resolution, and sensitivity of the feature extraction process to lighting, we do not require that features occur in the image exactly in the positions predicted by the mathematical model of the object. Instead, we allow them to be displaced by a small, bounded amount from their true locations.

Mathematically, we can formalize the bounded error recognition model as follows. Assume the object model consists of a collection $\{m_i\}$ of points in $\mathbb{R}^3$. An image consists of a collection $\{b_j\}$ of points in $\mathbb{R}^2$. A bounded error match consists of a set of correspondences j_i (usually, 1-1 but not onto) between model features and image features together with a 3D rigid body transformation T_3 such that $\|PT_3m_i - b_{j_i}\| \leq \epsilon$, where P is the camera model—usually orthographic or perspective projection—and ϵ is an error bound.

View-Based Recognition

As in the 3D bounded error recognition case, we assume that images consist of features. However, rather than using object models that represent objects as collections of 3D points, we use object models that represent objects as collections of views, where each view is a collection of feature locations in $\mathbb{R}^2$. We declare a match between a the model and the image if for any view of the object, we can find a bounded error match under 2D equiform transformations (translation, rotation, and scale).

More formally, we write the view-based model as $\{m_j^r\}$, where r identifies the view. A match under the view-based approximation then consists of a view r, a set of correspondences j_i between features in the view and image features, and a 2D transformation T_2 (translation, rotation, and scale) satisfying: $\|T_2m_i^r - b_{j_i}\| \leq \delta$.

The motivation behind this approach is the following. Consider a set of points in $\mathbb{R}^3$ that can undergo rigid body transformations and scaling. Such a transformation is given by 7 parameters: 3 parameters to specify a translation, 3 parameters to specify a rotation, and one parameter to specify scale.

Let us assume an orthographic projection model. Then it is easy to see that translation along the projection axis does not affect the projected image of the points. Furthermore, by symmetry, translations, rotations, and scale in the image plane can compensate for 4 of the remaining 6 parameters describing the 3D pose of the set of points.

This leaves us with two parameters (e.g., identifiable with slant and tilt or the points on the surface of a sphere, the *viewing sphere*) determining the actual location of points in the projection of the set of 3D points, up to 2D translation, rotation, and scale.

The changes induced in the image of a 3D object by varying these remaining two parameters appear like non-rigid deformations of a 2D model. Therefore, instead of modeling them exactly, we can simply attempt to model them as "2D error" or "noise" on the location of features.

Probability of Error

View-based recognition is only an approximation to 3D bounded error recognition, in the sense that the possibility of false positive or false negative matches exists (i.e., that a VBR system incorrectly declares an object as present or absent in a scene). Intuitively, the probability of such errors depends on the number of views used by the VBR system and on the parameter δ. In this paper, we will assume that δ has been chosen and sufficient number of views has been stored such that the probability of false negative matches is zero (it can be shown that this is always possible). It remains then to estimate the probability of false positive matches.

The basic idea is the following. We can represent the complete set of views of an object consisting of k features as a subset of $\mathbb{R}^{2k}$ by concatenating the $2k$ coordinates of the feature locations into a single $2k$-dimensional vector. The shape of this set (the *view set*) will be determined by two components: the shape of the object and the error model we use. Different error models give rise to different kinds of view sets. Let us denote the view set for some given object under a bounded error model as S_{BE} and the view set for the same object using some alternative error model as S_a. Then $S_{\text{BE}} - S_a$ represents the set of views that are recognized by the bounded error model but not by the alternative model (false negative matches), and $S_a - S_{\text{BE}}$ represents the set of views that are recognized by the alternative model but not by the bounded error model (false positive matches). The volume of the set $S_a - S_{\text{BE}}$ is related (via a probability distribution on all possible inputs to the recognition system) to the probability of a false positive matches.

Space does not permit us to present a complete analysis here, but it can be shown[2] that for a number of commonly used recognition methods, including least-square error recognition and alignment, the volumes of $S_a - S_{\text{BE}}$ are within a fixed constant factor of each other.

For example, for comparing recognition under bounded error with recognition under least square error, we can observe that the corresponding view sets S_{BE} and S_{LSQ} are dilations of a single manifold under different but similar metrics in the space of all views. This lets us relate the volume of the difference $S_{\text{LSQ}} - S_{\text{BE}}$ to the constant in the definition of similarity between the two metrics in view space. Analogous analyses can be made for methods like alignment.

But the same can be found to be true for the view-based approximation, if we choose $\delta = c\epsilon$ for some constant c. However, unlike the effect of choices like bounded error recognition vs. least-square error recognition, which is of fixed magnitude, the view-based approximation to bounded error recognition actually lets us approximate bounded error recognition arbitrarily well by choosing a smaller constant c. That is, by decreasing c, we can make the volume of $S_{\mathrm{VBR}} - S_{\mathrm{BE}}$ (and hence the probability of false positive matches under most probability distributions) arbitrarily small. Of course, as we will see below, we have to pay in terms of storage and computation: we need to store and match $\Theta(\delta^{-2})$ views.

Number of Views

We noted above that the appearance of a 3D object in a 2D image is determined by two parameters (e.g., identified with points on the viewing sphere) after accounting for 2D equiform transformations.

Now, if we assume that the relationship between these two parameters of the viewing transformation and the location of features in the image is piecewise smooth (an assumption that is certainly satisfied for features that are "attached" to the object), then it will be true that small changes in slant and/or tilt will give rise to only small changes in the location of features in the image.

Generalization from a Single View We can formalize the notion of smoothness of the viewing transformation as requiring piecewise uniform continuity over the viewing sphere. To do this, we make use of the modulus of continuity:

$$w_f(\delta) = \sup_{|x_1 - x_2| \leq \delta} |f(x_1 - x_2)| \tag{1}$$

It is not difficult to see that if the maximum modulus of continuity of the viewing transformation is bounded as $\mu \geq \frac{w_f(\delta)}{\delta}$, then each individual view of an object will match a solid angle of approximately $\frac{\delta}{\mu}$ on the viewing sphere.

Since the viewing sphere is a 2D surface, and since we are covering it with patches of linear dimension $\frac{\delta}{\mu}$, we expect that we need $O(\left(\frac{\delta}{\mu}\right)^{-2})$ different patches (and hence, views) of an object.

We can derive more concrete bounds by actually bounding the modulus of continuity. Let us assume that translations have already been accounted for.

Then, a viewing transformation consists of a rotation R, a change of scale S, and a projection P: $b_{j_i} = P\,S\,R(m_i)$.

We know that for small rotations (say, of size ϵ) around an axis given by a unit vector r, the displacement of a vector v is given by:

$$\Delta v = \epsilon\, r \times v \tag{2}$$

Since for a unit vector r, it is true that $\|r \times v\| \leq \|v\|$ we know that $\|\Delta v\| \leq \|v\|$. Furthermore, since $P = \mathrm{diag}(1,1,0)$, $\|P\,v\| \leq \|v\|$.

Therefore, we see that for any axis of rotation, scale factor s, and small angles of rotation ϵ, the projection of an attached feature v does not move by more than $\epsilon\|v\|$ (this bound actually also works for large ϵ).

Hence, the modulus of continuity of the viewing transformation with respect to any rotation is bounded as $\mu = s\|v\|$. Now, because images are formed on a sensor of finite diameter (retina, CCD array) $s\|v\|$ is bounded by a constant determined by the sensor hardware. So, if we assume that the sensor is bounded by a circle of radius D, then μ is simply D.

Covering the Viewing Sphere Above, we have seen that for an individual view, for smooth viewing transformations, changes in slant/tilt of order ϵ will move the location of features in the image by less than $\epsilon\mu$. Since we require that $\delta \geq \epsilon\mu$, this means that for a given δ, ϵ is at least as large as $\frac{\delta}{\mu}$.

Now, allowing changes in slant/tilt by an amount of ϵ corresponds to an area of α on the viewing sphere:

$$\alpha = 2\pi(1 - \cos\epsilon) \geq \frac{23}{24}\pi\,\epsilon^2 \tag{3}$$

(the last inequality comes from the Taylor series expansion of cos).

The viewing sphere has total area 4π. The total number V of circular patches required to cover the viewing sphere, if we could choose their placement, is then bounded (including a factor of 2 to account for the fact that we cannot cover the viewing sphere without overlap using circular tiles):

$$V \leq \left\lceil 2\,\frac{4\pi}{\alpha}\right\rceil = \left\lceil 2\frac{4\pi}{\frac{23}{24}\pi\,\epsilon^2}\right\rceil \tag{4}$$

$$\leq \left\lceil 9\epsilon^2\right\rceil \leq \left\lceil 9\left(\frac{\delta}{\mu}\right)^2\right\rceil = \left\lceil 9\left(\frac{\delta}{D}\right)^2\right\rceil \tag{5}$$

This is the bound on the number of views of a 3D object under a bounded error recognition model and allowing the view based recognition algorithm to choose the individual views.

The ratio $\frac{\delta}{D}$ is the error that is tolerated by the recognition system relative to the size to the image of the object. In practice, this ratio will be somewhere around 5%. If we choose δ in the view-based approximation such that $\frac{\delta}{D} = 0.05$ this results in an *upper bound* on the number of views of 3600.

Note that the resulting bounds on the number of views of an object are independent of the complexity (number of features) of the object. Object complexity does have an influence on the number of different views in the *presence* of occlusions: objects with more features tend to have a larger number of aspects (a bound on the number of aspects in terms of the complexity of an object is given in[6]).

We will see below that the number of views required in an actual view-based system can be much smaller. One reason for this is the frequent occurrence of approximate invariants and the presence of characteristic non-metric information (topology, non-geometric information) in images.

Efficiency

View-based recognition lets us replace matching of a single 3D model with matching of a larger number, say R, of 2D models. This may not appear to be a good tradeoff from an efficiency point of view. However, upon closer examination, it appears that view-based recognition may actually be faster than direct 3D recognition. The reason is the following.

The complexity of bounded error recognition algorithms is dominated by the minimum number of correspondences between image and model features that determine an alignment (among other things, because of the potential size of the output of the algorithm). Let us consider the case in which no additional "grouping" or "segmentation" information is available, and in which there are N image features and M model features. Then, a 3D recognition algorithm will have complexity of approximately $\Omega(V(N,M)\,N^3M^3)$, where $V(N,M)$ is the time required for "verifying" a match, while a 2D recognition algorithm will have complexity of approximately $\Omega(V(N,M)\,N^2M^2)$.

If anything the constants in these asymptotic complexities will be better for the 2D algorithm due to the simpler geometric computations involved, and, hence, 2D recognition can be carried out faster than 3D recognition by a factor of NM. Now, as long as $NM > R$ (recall that R is the number of 2D models in the view-based approximation), the view-based approach to recognition will be faster than the direct 3D approach; this inequality is satisfied for commonly used error bounds and all but very simple scenes and objects.

Simulations

Above, we have seen theoretical analyses that support the idea that view-based recognition does not differ significantly from 3D methods in terms of the probability of false positive errors, and that view-based recognition does not require "too many" views in order to work.

Since large data bases of images and object models for testing 3D recognition systems are not available, we had to rely on simulations in order to compare the performance of different 3D recognition methods (3D alignment, least-square matching, linear combination of views) with view-based recognition on large model bases. The simulations used data bases consisting of 1000 differently bent "paper clips" (these object were chosen because they have also been used in a variety of other simulations and psychophysical experiments. In some representative experiments, each paper clip consisted of 20 line segments, and the location of features (bends) in the image was uncertain by approx. 5% of the total projected size of the clip. The simulations were more difficult for the view-based recognition algorithm than the case analyzed above, since it was given a collection of random views of the object as input, from which a model had to be built. In contrast, the 3D recognition algorithms received as input the (perfect) 3D model used to generate the images in the simulation.

Under these conditions, we found that 300 views needed to be stored for each view-based model in order to achieve an error rate smaller than that of optimal 3D matching algorithms.

The predictions about robustness of view-based recognition were confirmed. For exam-

Figure 1: Example of an airplane recognized by the view-based recognition system.

ple, in a different set of experiments, intersections between projections of segments of the paperclips were used as features, rather than the locations of bends. 3D based methods designed for attached features (not surprisingly) failed completely on this problem, while view-based methods only required 2–3 times as many training examples to achieve the same error rates as in the case of attached features.

Real Images

We have implemented a prototype view-based recognition system for 3D objects that builds object models automatically from examples. The system can reliably recognize and distinguish model airplanes in scenes of 3D objects in the presence of significant clutter and occlusion.

An example of the optimal match and initial pose estimate returned by the system is shown in Figure 1. Input to the recognition module consisted of the raw output of a Canny edge detector. The 2D matching algorithm did not require (or take advantage of) any grouping or segmentation information, nor did it require the extraction of "point features". Furthermore, there was no attempt to tune any of the system's parameters: the Canny edge detector was used with its standard settings, 2D error bounds of 10 pixels were used.

The internal view-based model was built from 32 different views of the airplane at different elevations and orientations. Models were acquired automatically by the system. These views were matched against an input image using a modified version of the RAST algorithm.[1]

The system was tested on 22 scenes containing the model and other objects (toy airplanes, cars, blocks, etc.). Only in one of the 22 scenes was the first choice of the system incorrect (in that case, the second choice gave the right match).

An example of a match is shown in Figure 1. Note that the model view includes a shadow of the airplane, a useful and salient feature for recognizing this kind of object.

Discussion

As we noted in the introduction, the idea of view-based recognition itself is not new. However, up to now, it has been used apologetically and as a heuristic. In the analysis and empirical results presented above, we have established clearly the relationship between view-based recognition and one of the most commonly used approaches to 3D recognition—3D bounded error recognition. Based on such results, the author hopes that view-based recognition will be viewed as a well-founded, simple, and robust approach to 3D object recognition, rather than as a heuristic.

From the theoretical considerations, we can infer that view-based recognition is particularly well-suited to recognition tasks in which scenes are cluttered, but in which very precise pose estimates are not needed. But even in cases where precise pose estimates are needed, view-based recognition is still a useful pre-processing step—the initial match and approximate pose estimate returned by a view-based system can be refined using other techniques.

From a practical point of view, we believe that view-based methods are currently the only feasible methods for general-purpose, robust, integrated 3D recognition systems, i.e., systems that address both the model acquisition and the recognition problem for complex scenes.

References

[1] Thomas M. Breuel. Fast Recognition using Adaptive Subdivisions of Transformation Space. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, 1992.

[2] Thomas M. Breuel. *Geometric Aspects of Visual Object Recognition.* PhD thesis, Massachusetts Institute of Technology, 1992.

[3] Thomas M. Breuel. View-Based Recognition. In *MVA '92, IAPR Workshop on Machine Vision Applications*, December 1992.

[4] R. T. Chin and C. R. Dyer. Model-based Recognition in Robot Vision. *ACM Computing Surveys*, 18(1):67–108, March 1986.

[5] Eric Grimson. *Object Recognition by Computer.* MIT Press, Cambridge, MA, 1990.

[6] K. Ikeuchi and T. Kanade. Applying sensor models to automatic generation of object recognition programs. In *Proceedings of the International Conference on Computer Vision*, pages 228–237, Tarpon Springs, FL, 1988.

[7] Matthew R. Korn and Charles R. Dyer. 3d multiview object representations for model-based object recognition. *Pattern Recognition*, 20(1):91–103, 1987.