View-based Recognition: A comparison of three methods

Antonio M. Rodriguez (tonyr@uci.edu)
Dept.of Cognitive Science, University of California, Irvine

Bruce M. Bennett (bbennett@math.uci.edu)
Dept.of Mathematics, University of California, Irvine

Donald D. Hoffman (ddhoff@uci.edu)
Dept.of Cognitive Science, University of California, Irvine

Hongkai Zhao (zhao@math.uci.edu)
Dept.of Mathematics, University of California, Irvine

Abstract. We compare three different approaches to view-based rigid-object recognition, all of which use features that can be assigned point coordinates: Linear combination of views, view interpolation, and recognition polynomials. Using Monte Carlo simulations we generate ROC curves for each method for a total of 45 different viewing conditions: three different noise levels, three levels of perspective distortion, and five different amounts of prior information, in the form of different numbers of models of the object which are given prior to the recognition task. For each method and for each condition, we compute the number A', a measure of detection efficacy given by the area under the ROC curve. A' may also be interpreted as the probability of correct response to the corresponding 2-alternative forced choice task. We find that (1) the optimal choice of method varies with viewing condition, however (2) recognition polynomials provide the greatest expected efficacy over all conditions. We compare the computational complexity of each method, and discuss how recognition polynomials, being O(n) (where n is the number of features summed over all views) incur the lowest possible computational cost.

Keywords: vision, rigidity, computer vision

1. Introduction

There has been much work in recent years to develop computational theories of 3-D object recognition from just 2-D image data. These theories differ from approaches which assume that the visual system stores and manipulates 3-D object data (for example, [4, 16, 25]). In this paper we discuss and compare three theories for the recognition of isolated rigid objects, theories that use 2-D shape information alone (ignoring, for example, such visual cues as color, texture, context and expectation): "linear combination of models," "view interpolation" and "recognition polynomials." We review these theories in the sections below. Each uses 2-D information from discrete views consisting of discrete points, which are assumed to be projections onto the image plane of features on a 3-D object. The recognition polynomials considered here allow " weak" perspective, i.e., they are explicitly constructed to detect orthographically-projected rigid objects which may be scaled differently in different views. The linear-combination-of-models theory implicitly allows weak perspective. View interpolation does not take perspective into account. Thus none of the three approaches is mathematically designed to take full perspective projection into account. Nonetheless we can compare the performance of the three methods under varying degrees of perspective distortion, varying noise level, and varying numbers of models of the object which are given as prior information.

An interesting approach not considered in this paper is the "rigidity checking" algorithm of McReynolds and Lowe presented in [17] for determining when two perspective views are consistent with a rigid interpretation. From the ROC curves published in [17] it appears that the performance of this algorithm for the case of one model view and high perspective distortion is superior to any of the three methods considered in this paper, and for one model view and low perspective distortion its performance seems inferior to that of the recognition polynomial method. Data are not available to discuss its performance in the case of larger numbers of model views.

Throughout this paper we refer to "objects", "models", and "views". An "object" is an array of n points in 3-D together with the origin; n is fixed. A "model" or "rigid model" of a given object is a rotated version of it; we consider only rotations which fix the origin, i.e., in which the origin is foveated. We will also speak of "affine models." An "affine model" of an object is obtained by applying an arbitrary linear transformation of 3-D space, not necessarily a rotation. A "view" is an array of n points in 2-D. A "view of an object" is the projection onto an image plane of a model of the object. Unless otherwise specified the projections are orthographic; in this case the view consists of the array of points in 2-D given by the (x, y) coordinates of the points in a model of the object. We assume throughout that the correspondence problem is solved, i.e., we assume that we know which points in the various views correspond to each other.

Each of the three theories provides a detector for the purpose of recognizing a rigid object from one view, called a "novel" view to distinguish it from the model views of the object which have been given as prior information. (This extends in a natural way to give detectors for recognizing the object from multiple novel views, but we do not pursue this here.) The detector is, in each case, represented as a function $F(\mathbf{Q})$ which assigns a non-negative real number to the novel view Q; Q is specified by $2n$ real variables, i.e., F is a function on the space $(\mathbb{R}^2)^n$ of arrays of n points in 2-D. Ideally this number is 0 if and only if the novel view is a rigid view of the given object. In practice, however, because of noise or because of the mathematical structure of the detectors, this will not be the case. Therefore it is necessary to choose a threshold value: a view will be considered a rigid view of the given object only if the detector yields a number less than the threshold. Of course, the larger the threshold, the greater the probability of false targets (incorrect identifications of non-rigid views as rigid). According to the principles of signal detection theory, the performance of the detector over all possible choices of threshold values is expressed in the detector's ROC (receiver operating characteristic) curve. One way to view the ROC curve is as a plane curve parameterized by threshold values t; to each t is associated the point $(h(t), i(t))$, where $h(t)$ is the probability of correctly identifying a rigid view (probability of hits) if t is used as the threshold, and $i(t)$ is the probability of incorrectly identifying a non-rigid view as rigid (probability of false targets) if t is used. A discrete approximation to the ROC curve may be constructed from the lists of detector values (i.e., values of the function F) obtained from a Monte Carlo simulation as described in section 5. A main point is that the entire ROC curve as a geometric object, independent of any particular parameterization, is what contains the information about the performance of the detector. However, certain numerical invariants of the curve such as its subtended area, denoted A', are useful for making overall comparisons of detectability of the various methods ([10], see also section 6).

In particular the ROC curve is invariant under replacement of the detector function f by $g(f)$, where g is any monotone increasing function. In fact to produce the ROC curve we create two lists of numbers in order of increasing size. One list consists of the detector values obtained for a sequence of N hit trials and the other list consists of the detector values for a series of N miss trials. The ROC curve is then the set of all points in the plane of the form $(i/N, n(i)/N)$ as $i = 1, ..., N$, where $n(i)$ is the position in the hit list occupied by the same number which is in the ith position of the miss list. Therefore as long as the same monotone increasing function is applied to the numbers in both lists, the pairs $(i/N, n(i)/N)$ as $i = 1, ..., N$ remain the same. For example, in the linear combination of views method (§2), the criterion to recognize a novel view Q as a rigid view of a given object P is that Q belongs to a certain linear subspace L of the space $(\mathbb{R}^2)^n$ of all views. Thus, the Euclidean distance from Q to L in $(\mathbb{R}^2)^n$ is a natural candidate for $f(Q)$. However, the square of that distance serves equally well – it results in the same ROC curve.

For each of the three theories, we consider 45 viewing conditions including all possible combinations of three levels of noise, three levels of perspective distortion and five different numbers of models of an object given a priori (1, 2, 3, 4, and 10 models). However we take the number n of points (in addition to the origin) to be 11 throughout, i.e., our objects always consist of 12 points one of which is the origin. For each of the 45 possible viewing conditions for each detector, we generate ROC curves using 1,000-trial Monte Carlo simulations; this procedure is discussed in detail below. In this way we can compare the performance of the three detector types and, more importantly, identify the viewing conditions under which the various detectors excel. In fact, we can then define a "supremum" detector which uses each of the three detectors in that range of conditions where its performance is superior.

In the sections below we discuss the three theories and associated algorithmic and computational issues. We present in detail the methods for generating the corresponding Monte Carlo simulations. We present the ROC curves and discuss some comparisons that they make possible, and we display the supremum detector.

2. Linear Combinations of Models.

The linear combinations method has been presented in [27] and [26], and has two variants. The theory which underlies this method deserves mathematical clarification and commentary. For this reason our discussion of this method is longer than that for the other two methods. At the end of the section we discuss some issues of computation.

In the first variant, the criterion to recognize a novel view as a view of a given object is that the n-vector of the x-coordinates of the points in the novel view are linear combinations of the n-vectors of x-coordinates of the points in the views of some given models of the object; similarly the n-vector of the y coordinates of the points in the novel view are linear combinations of the n-vectors of y coordinates of the same model views. The coefficients used in the linear combination of the x-coordinate n-vectors will, in general, be different from the coefficients in the linear combination of the y-coordinate n-vectors. In the second variant, the criterion is that the vector consisting of all the n points in the novel view (i.e., the list of the n points in the novel view considered as a single $2n$-vector) is a linear combination of the vectors of the n image points of the views of given models of the object. We will see below that this is mathematically equivalent to using the first variant where twice the number of model views are used.

We now discuss this in detail. Let P be an object in 3-D consisting of n points, together with the origin. We will call another such object P_1 a *model* of P if there exists a rotation R of 3-D space such that $P_1 = RP$.

Now let R, S, T, U be distinct rotations of 3-D space (which fix the origin), and let $P_1 = RP, P_2 = SP, P_3 = TP, P' = UP$ be the objects which results from applying R, S, T, U to P.

We think of P_1, P_2, P_3 as given models of P, and we think of P' as a novel rotated version of P; each of these objects, like P, consists of n points in 3-D together with the origin.

We now consider the first variant of the criterion. Let $P_{1,X}, P_{2,X}, P_{3,X}, P'_X$ denote the vectors of X-coordinates of the points in P_1, P_2, P_3, P'; each of these vectors is an n-vector. Similarly let $P_{1,Y}, P_{2,Y}, P_{3,Y}, P'_Y$ denote the n-vectors of Y-coordinates of the points in P_1, P_2, P_3, P'. Let R_X, S_X, T_X, U_X denote the first rows of the matrix representations of R, S, T, U, and let R_Y, S_Y, T_Y, U_Y denote the second rows. For generic choices of P_1, P_2, P_3, i.e., for generic choices of R, S, T, both sets of vectors $\{R_X, S_X, T_X,\}$ and $\{R_Y, S_Y, T_Y\}$ are linearly independent. Then U_X is a linear combination of R_X, S_X, T_X, and U_Y is a linear combination of R_Y, S_Y, T_Y. It follows that P'_X is linear combination of $P_{1,X}, P_{2,X}, P_{3,X}$, and P'_Y is a linear combination of $P_{1,Y}, P_{2,Y}, P_{3,Y}$. (These two linear combinations will, in general, involve different coefficients.) To summarize:

Proposition 1. If $P_1, ..., P_m$ are generic models of P for $m \geq 3$, and if P' is any rotated version of P, then P'_X is a linear combination of $P_{1,X}, ..., P_{m,X}$, and P'_Y is a linear combination of $P_{1,Y}, ..., P_{m,Y}$.

Ullman and Basri propose to use the converse of Proposition 1 as a criterion for recognition:

Linear Combination of Models Recognition Criterion - First Variant: Let $P_1, ..., P_m$ be given generic models of P, in the sense that $P_i = R_i P$ for generic rotations R_i of 3-D space. Then we will 'recognize' a novel view P' to be a view of P if P'_X is a linear combination of $\{P_{1,X}, ..., P_{m,X}\}$, and P'_Y is a linear combination of $\{P_{1,Y}, ..., P_{m,Y}\}$, where $P_{i,X}$ denotes the vector of x-coordinates of P_i, and $P_{i,Y}$ denotes the vector of y-coordinates of P_i. Note: to apply the criterion it is only necessary to know the model *views* $P_{j,X,Y}$, not the actual 3-D model P_j.

Remark. If $m < 3$ the proposition fails, so that generic views of P will not satisfy the criterion. We therefore expect poor detection in this case due to failure to recognize generic views of P.

Let E_P denote the set of all rotations $P' = UP$ of P in 3-D. We have a map $\pi_X : E_P \to \mathbb{R}^n$ which sends a model P' of P to the n-vector of x-coordinates of its points; similarly we have $\pi_Y : E_P \to \mathbb{R}^n$. We let $\pi = (\pi_X, \pi_Y) : E_P \to \mathbb{R}^n \times \mathbb{R}^n = \mathbb{R}^{2n}$. Let $I_{P_1,...,P_m}$ denote the set of image data satisfying the above criterion ($m \geq 3$). It is equivalent to say that if $[P_{1,X}, ..., P_{m,X}]$ denotes the subspace of $\mathbb{R}^n$ spanned by $P_{1,X}, ..., P_{m,X}$, and if $[P_{1,Y}, ..., P_{m,Y}]$ denotes the subspace of $\mathbb{R}^n$ spanned by $P_{1,Y}, ..., P_{m,Y}$, then $I_{P_1,...,P_m} = [P_{1,X}, ..., P_{m,X}] \times [P_{1,Y}, ..., P_{m,Y}] \subset \mathbb{R}^n \times \mathbb{R}^n$. Proposition 1 states that the image of π is contained in $I_{P_1,...,P_m}$. Moreover, the argument leading to the Proposition shows that $[P_{1,X}, ..., P_{m,X}] = [P_{1,X}, P_{2,X}, P_{3,X}]$, and $[P_{1,Y}, ..., P_{m,Y}] = [P_{1,Y}, P_{2,Y}, P_{3,Y}]$. Hence $I_{P_1,...,P_m} = [P_{1,X}, P_{2,X}, P_{3,X}] \times [P_{1,Y}, P_{2,Y}, P_{3,Y}]$, so that $I_{P_1,...,P_m}$ has dimension 6 regardless of the number of points n or the number of models m, as long as $m \geq 3$. Moreover, the image of π generates $I_{P_1,...,P_m}$. The same argument shows that for $m \geq 3$ the space $I_{P_1,...,P_m}$ is independent of the particular choice of models $\{P_1, ..., P_m\}$, provided that they are generic in the sense that some choice of three of them arise by applying to P rotation matrices which have linearly independent first rows, and linearly independent second rows,

Remark. Starting with at least three models, if the number of points n is held constant, we do not expect improvement in detection as additional models are added to the prior information. In fact, as we have just noted, for a generic choice of $\{P_1, ..., P_m\}$, as m increases beyond 3 the linear subspace $I_{P_1,...,P_m}$ remains constant. Since membership in this subspace is the recognition criterion, increasing m beyond 3 will have no effect from a purely theoretical point of view. However, in practice, the redundancy entailed in such an increase may serve to reduce the effect of noise. In fact, we found that this was the case: we noted a marked increase in detection going from 3 to 4 model views in the presence of medium and high perspective distortion (see Figure 1).

Now E_P can be identified with the set of rotations of 3-D, where the rotation R corresponds to RP in E_P. Thus E_P has dimension 3; in fact it is a nonlinear sub-manifold of the vector space $\mathbb{R}^9$ of all 3×3 matrices. Indeed, we can view the map $\pi : E_P \to \mathbb{R}^{2n}$ as the restriction to E_P of the linear map $\Pi : \mathbb{R}^9 \to \mathbb{R}^{2n}$ which sends any matrix M to the pair of n-vectors consisting of the x-coordinates and the y-coordinates of the points in MP. Let K denote the kernel of the linear map π. We claim that the dimension of K is 3, provided that $n \geq 3$ and the n points of P are in general position, i.e., the points of P span $\mathbb{R}^3$. In fact, K is the set of those MP which consist of points whose x- and y-coordinates are all 0. Let the n points of P be denoted $P^1, ..., P^n$. Then the condition for MP to be in K is that $(M_X P^1, ..., M_X P^n)$ and $(M_Y P^1, ..., M_Y P^n)$ are the 0-vector in n-space (where as usual M_X and M_Y denote the first and second rows, respectively, of the matrix M). Since $P^1, ..., P^n$ span $\mathbb{R}^3$, this implies that M_X and M_Y are $(0, 0, 0)$. Thus, the only free parameters for M in K are the entries in its third row M_Z. This shows that K has dimension 3 as claimed.

By the dimension formula for linear maps, we infer that the image of Π in $\mathbb{R}^{2n}$ is a linear subspace of dimension 6. Since the domain of Π contains the domain of π, the linear subspace of $\mathbb{R}^{2n}$ which is the image of Π contains the linear subspace generated by the image of π, i.e., it contains $I_{P_1,...,P_m}$, which also has dimension 6. We conclude that these two subspaces coincide. This means that the linear combination of models recognition criterion is justified on mathematical grounds for the recognition of arbitrary affine models of P (i.e., objects of the form MP where M is an arbitrary matrix), even more than for the recognition of rigid models of P (i.e., objects of the form RP where R is an rotation matrix): The fact that the space of all affine models of P, identified with $I\!R^9$, projects *onto* $I_{P_1,\ldots,P_m}$ means that (assuming no noise and perfect resolution) a view satisfies the criterion if and only if it is the view of an affine model of P. It follows that, de facto, the linear combination approach incorporates weak perspective recognition, since uniform scaling is an affine transformation.

Thus the use of the criterion to infer rigid structures is – from a mathematical standpoint– an enormous inductive leap, as is usually the case in perception. In fact, it embodies a bias in favor of rigid interpretations. This bias is presumably justified because of a matching bias in the environment: Motion which is non-rigid but affine is rare. In the Monte Carlo simulations, we study the performance of the criterion to distinguish rigid views of P from randomly generated views of P. The rarity of non-rigid affine motion in the environment is reflected in the fact that non-rigid affine views will occur with zero probability among the randomly generated non-rigid views in the simulation. However as noise is increased, the probability increases that a randomly generated object will lie within detector threshold distance of an affine – but not rigid –view of P, thereby producing a false target. Thus we expect that for higher noise levels, the linear combinations method will not perform as well as the recognition polynomial method whose detector theoretically registers a hit only if the view has a rigid interpretation.

Linear Combination of Models Recognition Criterion -Second Variant: Let $P_1, \ldots, P_m$ be given generic models of P, in the sense that $P_i = R_i P$ for generic rotations R_i of 3-D space. Then we will 'recognize' a novel view P' to be a view of P if P' is a linear combination of $\{P_{1,X,Y}, \ldots, P_{m,X,Y}\}$, where $P_{i,X,Y}$ denotes the $2n$-vector of (X, Y)- coordinates (i.e., the 2-D images) of the n points in the ith model.

Let $J_{P_1,...,P_m}$ denote the linear subspace of $I\!R^{2n}$ generated by $\{P_{1,X,Y},...,P_{m,X,Y}\}$. Assume $m \geq 6$. By adding suitable linear combinations of $R_{1,X}, R_{2,X}, R_{3,X}$ to $R_{i,X}$ for each $i \neq 1,2,3$ we can produce a matrix whose first row is zero. Applying this matrix to P yields a member of $J_{P_1,...,P_m}$ which is a list of n points in the (X,Y)-plane whose X-coordinates are all 0. In this manner, since $m \geq 6$, we get at least three generic members of the set $[P_{1,Y},...,P_{m,Y}]$. Similarly, by adding suitable linear combinations of $R_{4,X}, R_{5,X}, R_{6,X}$ to $R_{i,X}$ for each $i \neq 4,5,6$ we can produce a matrix whose second row is zero. Applying this matrix to P yields a member of $J_{P_1,...,P_m}$ which is a list of n points in the (X,Y)-plane whose Y-coordinates are all 0. In this manner, since $m \geq 6$, we get at least three generic members of the set $[P_{1,X},...,P_{m,X}]$. This shows that $I_{P_1,...,P_m} \subset J_{P_1,...,P_m}$. On the other hand, a linear combination of vectors each of which is a list of (X,Y)-coordinates of n points is a special case of a linear combination of the lists of X- and Y-coordinates separately. Thus $J_{P_1,...,P_m} \subset I_{P_1,...,P_m}$. We obtain

Proposition 2. The second variant of the linear combination of models method using at least six models $m \geq 6$ coincides with the first variant of the method using at least three models.

We use the first variant of the criterion in the sequel.

Some remarks on computation. We briefly discuss the computational procedure for the linear combination detector. If model views $P_{1,X,Y},...,P_{m,X,Y}$ are given, then the criterion for a novel view V to be rigid is that V lies in the linear subspace $I_{P_1,...,P_m}$ of $I\!R^{2n}$. Thus we take our detector function $F(V)$ to be the square of the Euclidean distance from Q to $I_{P_1,...,P_m}$ in $I\!R^{2n}$. In the papers [27] and [26] a gradient descent search method is mentioned for this purpose: the idea is to search for a point Q in $I_{P_1,...,P_m}$ whose distance to V is minimum, and then use this distance for $F(V)$. Gradient descent searches can, in general, run the risk of getting trapped in minima which are local but not global. In this case, however, given the linearity of the manifold $I_{P_1,...,P_m}$, the squared distance function to V has a unique supremum which is the desired minimum, so local entrapment is not a problem. However the search procedure is computationally expensive, especially given the high dimensionality of the ambient space $I\!R^{2n}$.

In our Monte Carlo simulations we used the Nelder-Mead simplex method of gradient descent in MATLAB to calculate F. However, we describe below a purely algebraic procedure to calculate the distance directly, which is several orders of magnitude less costly than the search procedure.

As usual we let n denote the number of points and m the number of views; we assume $n > m$. Let V be a novel view consisting of n points in 2-D, and let V_X, V_Y denote the n-vectors of X-coordinates and Y-coordinates respectively of the points in V. Given views $P_1, ..., P_m$ of the 3-D object P, we want to find the square of the Euclidean distance in $I\!R^{2n}$ from V to $I_{P_1,...,P_m}$. Since $I_{P_1,...,P_m} = [P_{1,X}, ..., P_{m,X}] \times [P_{1,Y}, ..., P_{m,Y}]$, this is the same as the sum of the squares of the Euclidean distances from V_X to $[P_{1,X}, ..., P_{m,X}]$ and from V_Y to $[P_{1,Y}, ..., P_{m,Y}]$. We compute these squared distances as the squares of the length of the projections of V_X and V_Y on the orthogonal complements of the subspaces $[P_{1,X}, ..., P_{m,X}]$ and $[P_{1,Y}, ..., P_{m,Y}]$ respectively. In fact we will directly obtain orthonormal bases $\{U_{X,1}, ..., U_{X,p}\}$ and $\{U_{Y,1}, ..., U_{Y,p}\}$ for these orthogonal complements, and then our detector function F applied to V will be

$$F(V) = \sum_{k=1}^{p} (V_X \cdot U_{X,k})^2 + \sum_{k=1}^{p} (V_Y \cdot U_{Y,k})^2$$

.

We now describe the method we used to produce the orthonormal bases $\{U_{X,1}, ..., U_{X,p}\}$ and $\{U_{Y,1}, ..., U_{Y,p}\}$. The procedure rests on the following fact, easily verified as a corollary of Cramer's Rule:

Fact. Let $A = ((a_{ij}))$ be an $(m+1) \times m$ matrix. Let U denote the $m+1$ - dimensional column vector whose ith component is $(-1)^i detA(i)$, where $A(i)$ denotes the $m \times m$ matrix obtained by deleting the ith row of A. Then U is orthogonal to every column of A.

Now let A be an $n \times m$ matrix, $n > m$. For every choice of $n - m - 1$ rows, we can delete them to obtain an $(m + 1) \times m$ matrix, A', and use the method above to produce a $m + 1$-vector U' orthogonal to every column of A'. Now expand U' to an n-vector U by inserting 0's as additional coordinates in places corresponding to the rows of A that we deleted. It is clear that this U will be orthogonal to every column of the original A. In this way we generate a family of vectors $\{U_k\}$, each of which is orthogonal to every column of A. Finally, we can extract an orthonormal basis for the subspace generated by the $\{U_k\}$. If we apply this procedure to the $n \times m$ matrix whose columns are the n-vectors $P_{1,X}, ..., P_{m,X}$, we obtain the desired orthonormal basis $\{U_{X,1}, ..., U_{X,p}\}$ for the orthogonal complement of $[P_{1,X}, ..., P_{m,X}]$. Similarly we obtain the orthonormal basis $\{U_{Y,1}, ..., U_{Y,p}\}$. We then define $f(V)$ by the formula above.

We recall that the dimension of $[P_{1,X}, ..., P_{m,X}]$ and $[P_{1,Y}, ..., P_{m,Y}]$ will in general be 3 for all $m \geq 3$. Thus for all $m \geq 3$ we will have $p = n - 3$ in general.

3. View Interpolation Model

The view interpolation model [19] attempts to achieve recognition of an object by the method of Generalized Radial Basis Functions (GRBFs). This method is based on the principle of Cover's Theorem [6]: a classification problem is more likely to be linearly separable when it is cast in a higher-dimensional space. It is interesting that this approach is a generalization of the linear combination of views method described earlier [19], and is also equivalent to standard regularization [21, 24] and generalized splines [5, 20].

This method yields a detector function of the form:

$$F(x) = \sum_{j=1}^{N} w_j G(x - c_j). \tag{1}$$

Here G is an appropriate basis function on $(I\!R^2)^n$, such as the gaussian function; recall n is the number of points in our object. The centers c_j, $j = 1, ..., n$, are chosen randomly as points in $(I\!R^2)^n$, i.e., each center is an array of n points in $I\!R^2$. The more centers that are used, the higher the dimension of the space in which the problem is cast and so the better will be the approximation to the desired output [7]. The weights w_j, $j = 1, ..., n$, are real numbers found during a learning stage to be described below; the weights are specific to a given rigid object consisting of n points in 3D together with a given set of $2m$ training views. Once the function F is found it will be used as a detector of views of the given rigid object in the following way. Suppose x_0 is an initial view of the object. Ideally, then, the function $F = 1$ if x is a rigid view of x_0 and 0 otherwise. The problem is to choose w_j for which F will achieve this ideal.

The w_j's are found during the training stage: training views x_i as $i = 1, ..., 2m$ are chosen randomly subject to the condition that $x_1, ..., x_m$ are model (rigid) views of x_0 and $x_{m+1}, ..., x_{2m}$ are not rigidly related to x_0. A $(2m \times N)$-dimensional matrix $\mathbf{G} = ((G_{ij}))$ is defined by $G_{ij} = G(||x_i - c_j||), i = 1, ..., 2m, j = 1, ..., N$, where $||x_i - c_j||$ denotes Euclidean distance in $(I\!R^2)^4$. Let $\mathbf{W}$ denote the unknown weight vector $(w_1, ..., w_N)^T$, i.e., $\mathbf{W}$ is a column vector. We want $\mathbf{W}$ to satisfy

$$\mathbf{GW} = \mathbf{D}, \tag{2}$$

where $\mathbf{D} = (\mathbf{1}, ..., \mathbf{1}, \mathbf{0}, ..., \mathbf{0})^{\mathbf{T}}$, i.e., $\mathbf{D}$ is the $2m$-dimensional column vector whose first m entries are 1 and the last m are 0. The idea here is train the detector to return the value 1 in the case of a hit and 0 in the case of a miss; the training consists in finding $\mathbf{W} = (w_j)$ so that 2 is satisfied. Once W is chosen, letting $F(x)$ be as in 1 we can think of the equation $F(x) = 1$ as interpolating a surface on the points $x_1, ..., x_m$, i.e., on the m model views (hits) of the original object. Our detector measures distance from this surface. Note that if the matrix $\mathbf{G}$ is square (i.e., if the number N of centers equals number $2m$ of training views) it will be invertible by Michelli's Theorem. In this case we can simply solve for $\mathbf{W} = \mathbf{G}^{-\mathbf{1}}\mathbf{D}$. In our case however we always used $N = 100$ centers, but the number m varied with the experimental conditions. Recall that we considered cases $m = 1, 2, 3, 4$ and 10 of model views. Since $\mathbf{G}$ is not square the method of "pseudo-inverses" may be used to obtain a solution for $\mathbf{W}$ from the equation 2. The pseudo-inverse of $\mathbf{G}$ is defined to be the matrix $\mathbf{P} = (\mathbf{G}^{\mathbf{T}}\mathbf{G})^{-\mathbf{1}}\mathbf{G}^{\mathbf{T}}$. Note that if $\mathbf{G}$ is square $\mathbf{P} = \mathbf{G}^{-\mathbf{1}}$. In any case, we let $\mathbf{W} = \mathbf{P}\mathbf{D}$.

$$\mathbf{W} = \mathbf{P}\mathbf{D}, \qquad \mathbf{P} = (\mathbf{G}^{\mathbf{T}}\mathbf{G})^{-\mathbf{1}}\mathbf{G}^{\mathbf{T}}, \tag{3}$$

In our Monte Carlo simulation we used a Gaussian $\mathbf{G}$ so that in our case the equation 1 is of the form

$$F(x) = \sum_{j=1}^{N} w_j e^{-||x-c_j||^2}. \tag{4}$$

4. Recognition Polynomials

By a "k-views of n points rigid structure recognition polynomial" we mean (1) a polynomial in the image data, i.e., in the coordinates of the n points in the k given views, which (2) evaluates to zero when the image data are consistent with a rigid 3D structure, and which (3) almost surely (Lebesgue) does not evaluate to zero otherwise [2, 3]. Thus, by this definition, the recognition polynomial detects rigidity. But it can also be used for recognition purposes: If we plug into a k -view polynomial the coordinates of the points in $k-1$ given model views of an 3-D object P, there results a polynomial in the variables corresponding to one remaining view. This polynomial evaluates to 0 on the coordinates of the points in a novel view if and only if the novel view is consistent with an interpretation as a rigid view of P; in this sense the polynomial recognizes P. For example, polynomials for detection of rigid motion from 2 orthographic views were developed in [2]; the application to the recognition problem was discussed in [3].

In this paper we consider weak-perspective recognition polynomials, which recognize rigidity up to uniform scaling. Theorems about the existence and uniqueness of weak-perspective solutions for 3D structure were indicated in [11, 13]: for two weak-perspective views there is a one-parameter family of solutions, and for three views the solutions are unique up to reflection. In addition, there is a large literature exploring the computation of 3D structure from weak-perspective views [1, 8, 9, 14, 15, 18, 22, 23, 28, 29]. However from the approach taken in those papers it seems not easy to develop recognition polynomials for the weak-perspective detection of rigidity.

Kontsevich [12] developed a mathematical approach to the weak-perspective detection of rigidity, using some of the same geometric ideas in [11], notably the decomposition of rigid rotation into a rotation about the viewing direction and a rotation about an axis in the image plane. However, using this approach Kontsevich developed a two-view detection criterion which is particularly simple and elegant, and which permits the development of recognition polynomials in an effective manner. A similar approach, in the two-view case, was later presented by [22].

While a recognition polynomial for the two-view case was not explicitly stated in [12], a recognition criterion which easily leads to such a polynomial was clearly presented. An approach to the three-view case was also described in [12], but in this paper we use the two-view polynomial in a more direct fashion to construct polynomials to recognize a single novel view as a view of an object P, given various numbers of model views of P. From the ROC curves derived from our Monte Carlo simulations, we will see that the weak-perspective recognition polynomials perform well: Over all viewing conditions (varying the number of model views, the amount of perspective distortion and the amount of noise) the weak perspective recognition polynomials out-perform the other methods, and they are the least computationally expensive.

We now present Kontsevich's derivation of the polynomial constraint to detect rigidity from two views in weak perspective. The fundamental case of this theory is where the views consist of 4 points together with the origin. The recognition criterion developed for this case may then be adapted to the general case of 2 views of n points, as we discuss below. Recall that a rigid motion in weak perspective projection is equivalent to a rigid motion composed with a uniform scaling, projected orthographically. We ignore translations parallel to and reflections through the viewing plane: the motions we consider are thus rigid rotations in 3-space which fix the origin or translations along the viewing direction, followed by uniform scaling.

Suppose the object has four distinguished feature points, labelled "1, 2, 3 and 4," together with the origin, labelled "0." Denote by r_j the edge in $\mathbb{R}^3$ between point 0 and point j in the *first* view (with $j = 1, 2, 3, 4$). r'_j is then the corresponding edge, in $\mathbb{R}^3$, in the second view. If we let π denote orthographic projection to the image plane, denote the projected edges by

$$p_j = \pi(r_j); \qquad p'_j = \pi(r'_j). \tag{5}$$

Kontsevich's geometrical insight is to employ the fact that, given a fixed (image) plane, any rigid rotation in $\mathbb{R}^3$ can be thought of as a composition of two rotations: first, a rotation about a *unit* axis vector v parallel to that image plane and second, a rotation about an axis vector perpendicular to the image plane. The second rotation takes the first axis vector v to a new unit vector v^1 in the image plane. Finally, if the uniform scaling factor is s, we let

$$v' = \frac{1}{s} v^1. \tag{6}$$

The norm $\|v'\|$ of v' thus need not be 1.

Consider the first of the decomposed rotations, around the axis v. It is clear that the respective projections of r_j and p_j onto v are equal. Now consider the second rotation, which takes v to v^1. If we let r^1_j be the edge between 0 and j after that rotation, and denote $\pi(r^1_j) = p^1_j$, it is clear that the respective projections of r^1_j and p^1_j onto v^1 are equal, and that the lengths of these projections are the same as those of r_j and p_j onto v. Thus $p_j \cdot v = p^1_j \cdot v^1$. Finally, consider the scaling. Since π is orthographic, the scaling factor of s results in $p_j = s\, p^1_j$ and, in virtue of equation 6, we arrive at the *linear* Kontsevich equations:

$$p_j \cdot v - p'_j \cdot v' = 0, \tag{7}$$

with

$$\|v\| = 1; \qquad \|v'\| = \|v\|/s; \qquad \|p'\| = s\|p\|. \tag{8}$$

Now 7 is a homogeneous system of four linear equations in the four unknown coordinates c_1, c_2, c_1' and c_2', where

$$v = (c_1, c_2), \qquad v' = (c_1', c_2'). \tag{9}$$

If we denote the data by

$$p_j = (x_j, y_j), \qquad p_j' = (x_j', y_j'), \tag{10}$$

the condition for there to be a nontrivial solution to 7 is that the coefficient matrix have rank less than 4, i.e., that the determinant in equation 11 is zero. Thus v and v' can be known only up to an overall scale factor (at best); choosing a solution for which $\|v\| = 1$ allows us to compute the scale factor s from the middle equation in 8. This completes the exposition of Kontsevich's two-view derivation.

$$\det \begin{vmatrix} x_1 & y_1 & x_1' & y_1' \\ x_2 & y_2 & x_2' & y_2' \\ x_3 & y_3 & x_3' & y_3' \\ x_4 & y_4 & x_4' & y_4' \end{vmatrix} = 0. \tag{11}$$

Note, however, that it is not enough to require merely a nontrivial solution: *both* vectors v and v' need to be nonzero, and this requires that the 2×2 diagonal minors be nonsingular too. Moreover, an unambiguous computation of the scale factor s requires that the matrix have rank 3 (i.e., there is a 1-parameter set of solutions). However, among those matrices whose rank is less than or equal to 3, the ones with rank equal to 3 are generic.

We can now see how a *recognition polynomial* arises for the two-view weak perspective recognition of four points (plus one foveated point which projects to the origin) in rigid motion. This is a polynomial in the 16 data values $\{x_j, y_j, x'_j, y'_j\}_{j=1,\ldots,4}$, which must evaluate to zero for there to be a rigid interpretation. The condition that 7 has a nontrivial solution is then that the determinant of the coefficient matrix vanish. Ignoring the negative signs in the last two columns of this matrix, we get Equation 11. Therefore the two-view recognition polynomial is the determinant in 11.

For a generic system that satisfies 11 is there is a one-parameter family of rigid interpretations. This means that there is a one-parameter family of pairs P, P' of objects in 3-D which project to the first and second views respectively, where P and P' are related by a rotation in 3-D. Each such rotation may be decomposed into a rotation about the same axis v in the image plane, followed by a rotation about the z -axis (as in the derivation of the recognition polynomial above). However the angles of these rotations vary in the family. In fact, let us choose any angle α. There is a pair P, P' in the family such that the rotation from P to P', when decomposed as above, involves a rotation through α about v. Letting z_j denote the depth coordinate of the jth point of this P , we have

$$z_j = \frac{-c_2 x_j + c_1 y_j}{(c_1^2 + c_2^2)^{1/2}} \cot\alpha - \frac{-c'_2 x'_j + c'_1 y'_j}{(c_1^2 + c_2^2)^{1/2}} \csc\alpha \tag{12}$$

Here c_1, c_2, c'_1, c'_2 are the unique solution to the equation 7 above which define v subject to the condition that $\|v\| = 1$; recall that the vanishing of the recognition polynomial guarantees the existence of this solution. (The derivation of this formula follows from the approach to depth reconstruction presented in [12]). Note that the different P's in the family (i.e., P's whose depth coordinates are given by 12 for different α's) are *not* rigidly related.

Since the depth coordinate z_j in 12 depends only on (x_j, y_j), (x'_j, y'_j) and the choice of α, increasing the number of points does not remove the ambiguity: no matter how many points are in the views, the one-parameter ambiguity of 3-D interpretations remains. Thus, the recognition polynomial is actually recognizing rigid views P' of *any* of the P's in the family. What, then, is the justification for using the polynomial to recognize a *given* 3-D object P? Suppose that the image of a given 3-D object P is acquired, followed shortly thereafter by the image of some other object. If these two successive images admit a rigid motion interpretation, the real-world probability is high that the second object is rigidly related to P (rather than to a different object in the one-parameter family).

Suppose P is an object with four points plus one foveated point, and let models $P_1, ..., P_m$ of P be given. Let f_j denote the two-view recognition polynomial for P_j. f_j is a polynomial in variables $(x_1, y_1), ..., (x_4, y_4)$, which vanishes on an array of four points in 2-D if and only if that array can be interpreted as a rigid view of P_j, i.e., as a rigid view of P. Thus

$$F_{P_1,...,P_m} = \sum_{j=1}^{m} f_j^2 \tag{25}$$

is a recognition polynomial for the recognition of a single novel view as having a weak perspective interpretation as a rigid view of P, *given the models* $P_1, ..., P_m$ *of* P. Now suppose P (and hence each P_j) consists of n points, $n > 4$. For each choice of four indices $i_1, ..., i_4$ out of n, we can consider the sub object $P_j^{i_1,...,i_4}$ of P_j consisting of the four points with those indices. (Since we assume the correspondence problem is solved, we may view each P_j as consisting of an *ordered* set of n points.) Let $f_j^{i_1,\ldots,i_4}$ be the two-view recognition polynomial for $P_j^{i_1,\ldots,i_4}$, and let

$$F_{P_1,..,P_m}^{i_1,..,i_4} = \sum_{j=1}^{m} (f_j^{i_1,\ldots,i_4})^2$$

be the recognition polynomial given the models $P_1^{i_1,\ldots,i_4}, \ldots, P_m^{i_1,\ldots,i_4}$. To obtain a recognition polynomial for the entire n-point object P given the models $P_1, \ldots, P_m$ of P, one could simply take

$$\sum_{i_1,\ldots,i_4} F_{P_1,\ldots,P_m}^{i_1,\ldots,i_4}$$

where the sum is taken over all choices of four indices $i_1, \ldots, i_4$ out of n. But it is unnecessary to use all these summands.

In fact it suffices to take the sum over all sets of 4 consecutive indices begining with the set $\{1, 2, 3, 4\}$ and wrapping around to end with the set $\{n, 1, 2, 3\}$. Thus we will let

$$F = F_{P_1,\ldots,P_m}^{1,2,3,4} + F_{P_1,\ldots,P_m}^{2,3,4,5} + \ldots + F_{P_1,\ldots,P_m}^{n,1,2,3}. \tag{13}$$

This is the polynomial that we use in our Monte Carlo simulations for $m = 1, 2, 3, 4$ and 10. One sees from the resulting ROC curves that, while the performance is good even for $m = 1$, it improves as m increases, i.e., as the a priori knowledge about P increases. This is evident from the increase in the statistic A' (the area under the ROC curve) as m increases, even though the ROC curves themselves are not easily distinguishable in the small display format.

5. Monte Carlo Simulations and Generation of the ROC Curves

To determine the performance of each of the three methods under varying conditions of different numbers of model views, noise levels, and distortion levels, a Monte Carlo simulation was conducted. In this simulation each method was trained with $m = 1, 2, 3, 4$ or 10 rigid model views of randomly-generated test objects consisting of 10 points together with the origin in 3D, to produce a detection function F. This F was then tested with a novel view which was either a rigid view of the object (hit case) or consisted of random dots (false alarm case). For each value of m the various experimental conditions were achieved by subjecting both the model views and the novel view to three levels of noise: none, low (1% Gaussian), or high (5% Gaussian) and three levels of perspective distortion: none, low, and high. Thus each method was tested with 45 conditions: 5 different values of m (1, 2, 3, 4, or 10 model views), 3 levels of noise, and 3 levels of distortion.

The test objects were created by first randomly generating 11 points in $\mathbb{R}^3$ all within the interval $[0, 1]^3$. These 11 points were then uniformly scaled by a factor of either 1, 2, or 4, corresponding to no, low, and high perspective distortion. The distortion factor increases the size of the object and thus enhances any distortions due to perspective. Once the test object was constructed a set of m model views of this test object was created by rotating the test object about a randomly-selected, unit-normalized vector and projecting to a view in $\mathbb{R}^2$. In the case of no distortion the projection was orthographic, otherwise the projection was performed as weak-perspective with the following equation used for projection of the point $p = (p_x, p_y, p_z)$,

$$d = f/(f + p_z), \tag{14}$$

$$p'_x = p_x * d, \tag{15}$$

$$p'_y = p_y * d, \tag{16}$$

with f being the focal length which is always set to 10, and p'_x, p'_y being the x, y-coordinates of the projection of the point p. Gaussian noise was applied additively to each model view (after the perspective distortion). Three levels of noise–low, medium and high–were used corresponding to σ^2 (variance) values of $.01, .05$, and $.1$. In a given experimental condition, the same degree of noise and perspective distortion that were applied to the model views were also applied to the novel rigid views.

For each method and each condition (specified by a number m of model views, a level of perspective distortion, and a level of noise) 1,000 iterations were performed for both the hit case and the false alarm case. Each hit iteration was obtained, as described above, by choosing at random a test object together with m model (rigid) views of it which were then modified by the given levels of distortion and noise. A novel rigid view of the test object was also generated at random and subjected to modification by the same levels of distortion and noise. Each of the three detection methods, given the m model views, provides a measuring device in the form of a function F which when applied to a novel object, yields a non-negative number. This number is ideally zero when the novel object is a rigid view of the test object and increases in size as the novel object deviates from a rigid view of the test object. In this manner, each of the thousand hit iterations yielded a positive real number and these were listed in order of increasing size. Each false alarm iteration was obtained by generating test objects and model views as above, but now the novel view consisted of randomly chosen points. As in the hit case, the detector produced a non-negative real number for each false-alarm iteration, and these numbers were also listed in increasing order of size. Thus, for each condition we get two lists of non-negative real numbers in increasing order: a hit list and a false-alarm list.

From these two lists an ROC curve was generated. The ROC curve consists of $1,000$ points $\{(a_i, b_i)\}_{i=1,\ldots,1000}$ in a plane whose horizontal axis corresponds to false alarm rate and whose vertical axis corresponds to hit rate. Here $a_i = i/1000$ and $b_i = h(i)/1000$ where $h(i)$ is the position in the hit list of the same non-negative real number which occupies the ith position in the false alarm list. In other words, if the non-negative real number c appears in the ith position in the false alarm list and in the jth position in the hit list then, $j = h(i)$, and $(a_i, b_i) = (i/1000, j/1000)$. Note that the pairs $(i, h(i))$ match the positions in the respective lists where the same numbers occur; the pairs do not depend on the values of the numbers themselves. Thus, if the same monotone increasing function is applied to numbers in both lists the pairs $(i, h(i))$ will not change and hence the ROC curve will not be effected.

6. Evaluation of the ROC curves; the detectability measure A'.

To compare ROC curves for the different models and simulation conditions the A' statistic [10] was computed. A trapezoidal summation was used to compute the area under each ROC curve and thus A'.

Analysis of variance was conducted on the A' values for each method. All analyses were performed with $\alpha = .05$. Tukey-HSD was used as the post-hoc test in all calculations.

Insert Fig 1 about here.

Insert Fig 2 about here.

Insert Fig 3 about here.

6.1. Model

There was a significant main effect of models, $F(2,135) = 4.753$, $p = .011$. The linear combination and the view interpolation methods achieved similar performance across all conditions of the simulation, $\mu = 77.54$, and $\mu = 77.97$, respectively, $p = .769$. The recognition polynomial achieved the best overall performance with $\mu = 79.37$, $p = .011$.

6.2. Number of Model Views

Overall, the performance of correct detection significantly increased, as expected, with increasing number of model views, $F(4,135) = 1299$, $p < .001$. A' for $1, 2, 3, 4$, and 10 views were $\mu = 64.86$, $\mu = 72.72$, $\mu = 81.36$, $\mu = 85.34$, $\mu = 87.16$, respectively. All means were significantly different, except the 4 and 10 model view case, $t(x) = y$, $p = .160$.

6.3. Amount of perspective distortion

Performance decreased significantly with increasing amounts of perspective distortion, $F(2,135) = 2370$, $p < .001$. For no distortion, $\mu = 86.04$, medium distortion, $\mu = 81.53$, and high distortion, $\mu = 67.31$.

6.4. Noise Level

Overall detection ability decreased significantly, as expected, with increasing levels of noise, $F(4,135) = 25.06$, $p < .001$. Low noise $\mu = 79.69$, medium level of noise $\mu = 78.87$, and high level of noise $\mu = 76.33$.

6.5. Model and Number of Model Views

The interaction between model and number of model views was significant $F(8,135) = 1118$, $p < .001$. Both the linear combination and the view interpolation models achieved better performance with an increase in the number of views. However, it is interesting to note that the recognition polynomial decreased in performance as more model views were employed. The performance of each model for each view condition is shown in Figure 4 and Table I.

Insert Fig 4 about here.

Insert Table I about here.

6.6. Perspective Distortion

There was a significant interaction between model and level of perspective distortion, $F(4,135) = 2930$, $p < .001$. The linear combination and the recognition polynomial approaches tended to decrease in performance as perspective distortion was increased. However, the view interpolation model performed better with high distortion case than with low distortion. In fact, the linear combination and recognition polynomial methods are designed explicitly to assume no perspective distortion, so they do worse with more distortion. By contrast the view interpolation method does not incorporate any notion of perspective in its design; it simply learns ad hoc relationships between input and output patterns. While it is not suprising that for this reason the method might perform better in the presence of high perspective distortion, we have not identified a mathematical reason why this should be the case. The performance of each method is depicted in Figure 5 and Table II.

Insert Fig 5 about here.

Insert Table II about here.

6.7. Level of Noise

The performance with respect to increasing levels of noise was similar for all three models. No model degraded significantly as the noise level increased. The performance of each method for each noise level is shown in Figure 6 and Table III.

6.8. Summary of Analysis

Three interesting results from the above analysis are: 1) the recognition polynomial performance decreased with more model views, 2) the view interpolation method improved with more perspective distortion, and 3) the linear combination was the least sensitive to increasing noise. Keeping these differences in mind we will now propose a new method which combines the best performance of each of these models.

Insert Fig 6 about here.

Insert Table III about here.

7. The Supremum Method

One way to construct a computer vision system would be to combine different modules which instantiate detectors with different known domains of optimal performance. A module is brought into play only when the optimal environmental conditions for its detector are present. Three types of variation of environmental conditions were considered in this paper: 1) number of model views, 2) amount of noise, and 3) amount of perspective distortion. The domain of optimal performance of each detector is a range of conditions specified by values of the parameters corresponding to the environmental variations. Thus, for the purpose of deciding when a given detector should be activated, the visual system must employ a priori techniques for measuring these parameters. For example, the number of model views is proportional to the amount of time a given stimulus is presented. Noise level could be estimated using a measure of spatial frequency in the source image. Amount of perspective distortion might be ascertained by accommodation and convergence cues. Once these measurements are obtained for a given environmental condition the results are used to select the appropriate module. We refer to this procedure as the *supremum method*.

In order to carry out the supremum method to combine the three detectors considered in this paper, we use the results of our ROC analysis to determine the domains of optimal performance. During each condition given a specific false alarm rate the method which yields the highest hit rate is found. It is the hit rate from the best performing method which becomes the hit rate in the ROC curve of the supremum method. This calculation was performed on all 15 conditions of the simulation. The results of this method are shown in Figure 7.

Insert Fig 7 about here.

8. Discussion

Overall, the recognition polynomial achieved the best performance across all conditions $A' = 79.37$, $F(2, 135) = 4.753$, $p = .011$. The linear combination and view interpolation methods yielded slightly (but significantly) lower overall performance, $A' = 77.54$ and $A' = 77.97$, respectively. All models achieved better performance with an increasing number of model views, $F(4, 135) = 1299$, $p < .001$. Performance decreased significantly for the recognition polynomial with an increase in perspective distortion, $F(2, 135) = 2370$, $p < .001$. However, a marked increase in performance was observed for the view interpolation method in the case of high perspective distortion for all numbers of model views and for the linear combination method for large numbers of model views. For all methods performance decreased with an increase in the amount of noise, $F(4, 135) = 25.06$, $p < .001$. No model performed better than any other with respect to noise.

Each model exhibited superior performance within a certain range of parameters. The linear combination model exhibited superior performance with high numbers of model views. The recognition polynomial method was superior for low numbers of model views and low to medium distortion. The view interpolation method was superior in the case of high perspective distortion for all numbers of model views (except in the case of 10 model views when the linear combination method was superior half the time). The supremum method is a viable and high-performance approach for practical machine vision recognition of 3D objects from 2D views.

References

1. Alter, T. D.: 1994, '3-D pose from 3 points using weak-perspective'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **16**, 802–808.
2. Bennett, B. M., D. D. Hoffman, J. E. Nicola, and C. Prakash: 1989, 'Structure from two orthographic views of rigid motion'. *Journal of the Optical Society of America A* **6**, 1052–1069.
3. Bennett, B. M., D. D. Hoffman, and C. Prakash: 1993, 'Recognition polynomials'. *Journal of the Optical Society of America A* **10**, 759–764.
4. Biederman, I.: 1985, 'Human Image Understanding: Recent Research and a Theory'. *Computer Vision, Graphics, and Image Processing* **32**(1), 29–73.
5. Broomhead, D. S. and D. Lowe: 1988, 'Multivariable Functional Interpolation and Adaptive Networks'. *Complex Systems* **2**, 321–355.
6. Cover, T. M.: 1965, 'Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition'. *IEEE Transactions on Electronic Computers* **14**, 326–334.
7. Haykin, S.: 1999, *Neural Networks; A comprehensive Foundation*. New York: Macmillan, 2nd edition.
8. Horaud, R., F. Dornaika, B. Lamiroy, and S. Christy: 1997, 'Object pose: The link between weak perspective, paraperspective, and full perspective'. *International Journal of Computer Vision* **22**, 173–189.
9. Huang, T. S., A. M. Bruckstein, R. J. Holt, and A. N. Netravali: 1996, 'Uniqueness of 3D pose under weak perspective - a geometrical proof'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **17**, 1220–1221.
10. Iverson, G. and D. Bamber: 1997, 'The Generalized Area Theorem in Signal Detection Theory'. In: A. A. J. Marley (ed.): *Choice, Decision, and Measurement*. Mahwah, pp. 302–318.
11. Koenderink, J. J. and A. J. van Doorn: 1991, 'Affine structure from motion'. *Journal of the Optical Society of America A* **8**, 377–385.
12. Konstevich, L. L.: 1993, 'Pairwise comparison technique: A simple solution for depth reconstruction'. *Journal of the Optical Society of America A* **10**, 1129–1135.

13. Lee, C. H. and T. Huang: 1990, 'Finding point correspondences and determining motion of a rigid object from 2 weak perspective views'. *Computer Vision Graphics and Image Processing* **52**, 309–327.

14. Lei, Y. W. and K. C. Wong: 1999, 'Detection and localisation of reflectional and rotational symmetry under weak perspective projection'. *Pattern Recognition* **32**, 167–180.

15. Lind, M.: 1996, 'Perceiving motion and rigid structure from optic flow - a combined weak-perspective and polar-perspective approach'. *Perception and Psychophysics* **58**, 1085–1102.

16. Lowe, D. G. and T. O. Binford: 1985, 'The Recovery of Three-Dimensional Structure From Image Curves'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **7**, 320–326.

17. McReynolds, D. P. and D. G. Lowe: 1996, 'Rigidity checking of 3D point correspondences under perspective projection'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **18**(12), 759–764.

18. Ostuni, J. and S. Dunn: 1996, 'Motion from three weak perspective images using image rotation'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **18**, 64–69.

19. Poggio, T. and S. Edelman: 1990, 'A Network that Learns to Recognize 3D Objects'. *Nature* **343**(6255), 263–266.

20. Poggio, T. and F. Girosi: 1989, 'A Theory of Networks for Approximation and Learning'. Technical Report A. I. Memo 1140, Massachusetts Institute of Technology, Artificial Intelligence Laboratory and Center for Biological Information Processing, Whitaker College, Massachusetts.

21. Poggio, T., V. Torre, and C. Koch: 1985, 'Computational Vision and Regularization Theory'. *Nature* **317**, 314–319.

22. Shapiro, L. S., A. Zisserman, and M. Brady: 1995, '3D motion recovery via affine epipolar geometry'. *International Journal of Computer Vision* **16**, 147–182.

23. Shimshoni, I., R. Basri, and E. Rivlin: 1999, 'A geometric interpretation of weak-perspective motion'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **21**, 252–257.

24. Tikhonov, A. N. and V. I. Arsenin: 1977, *Solutions of ill-posed problems [Metody resheniia nekorrektnykh zadach]*. New York: Halsted Press.

25. Ullman, S.: 1989, 'Aligning Pictorial Descriptions: An Approach to Object Recognition'. *Cognition* **32**(3), 193–254.

26. Ullman, S.: 1998, 'Three-dimensional object recognition based on the combination of views'. *Cognition* **67**, 21–44.

27. Ullman, S. and R. Basri: 1991, 'Recognition by linear combination of models'. *IEEE Transaction on Pattern Analysis and Machine Intelligence* **13**(10), 992–1005.

28. Weber, J. and J. Malik: 1997, 'Rigid body segmentation and shape description from dense optical flow under weak perspective'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **19**, 139–143.

29. Xu, G. and N. Sugimoto: 1999, 'A linear algorithm for motion from three weak perspective images using Euler angles'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **21**, 54–57.

List of Tables

List of Figures

Table I. Performance of Models vs. Number of Model Views.

Models	Model Views 1	2	3	4	10
Lin. Comb.	38.88	63.73	87.05	98.09	99.96
View Interp.	71.69	74.50	77.90	80.21	85.56
Rec. Poly.	84.03	79.94	79.16	77.73	75.99

Table II. Performance of Models vs. Distortion Level.

Models	Distortion Low	Medium	High
Lin. Comb.	92.56	77.24	62.83
View Interp.	66.36	79.04	88.51
Rec. Poly.	99.21	88.31	50.59

Table III. Performance of Models vs. Nose Level.

Models	Noise		
	Low	Medium	High
Lin. Comb.	77.62	77.91	77.10
View Interp.	79.24	78.02	76.66
Rec. Poly.	82.21	80.67	75.23

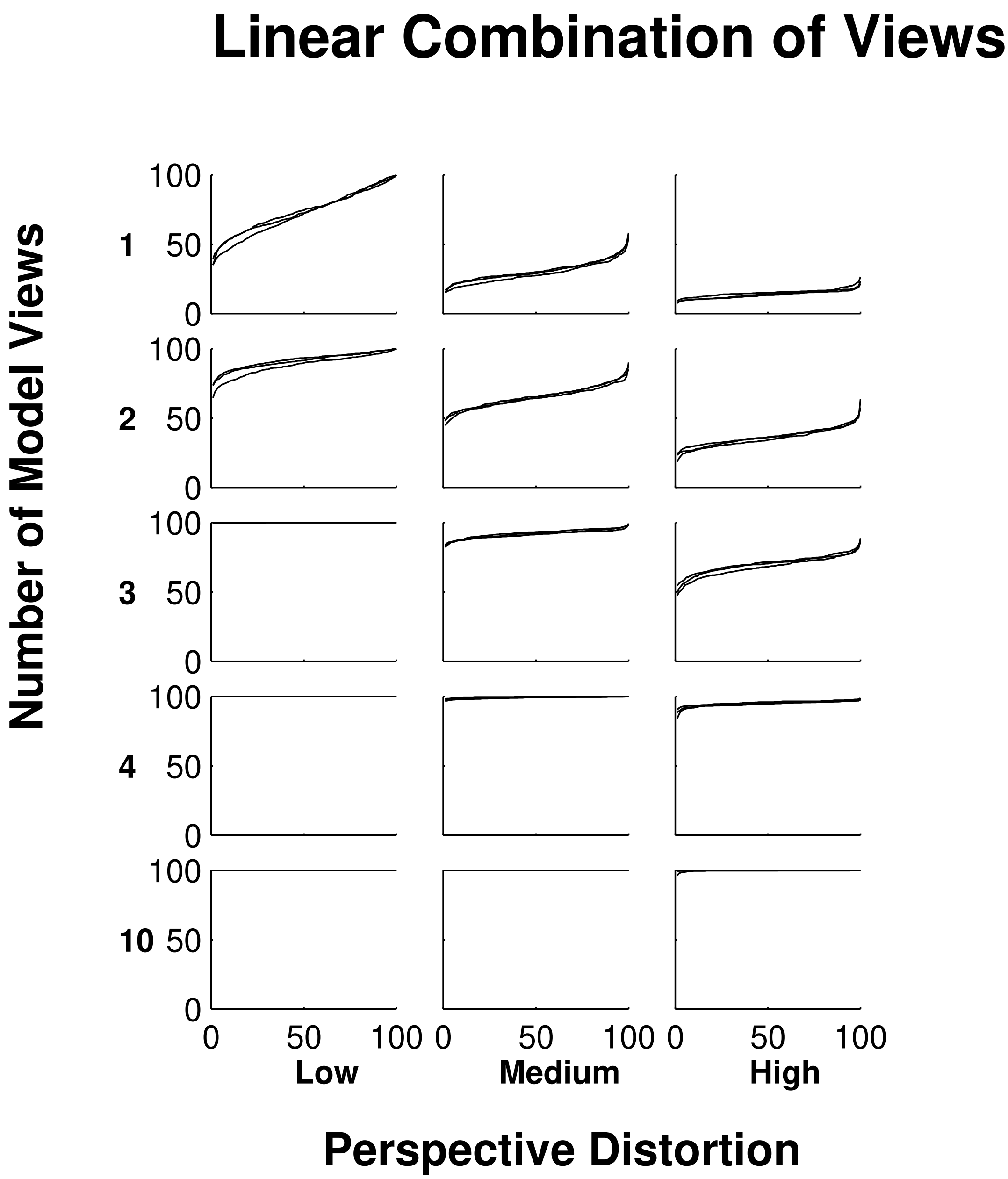

Figure 1. Linear combination of views ROC.

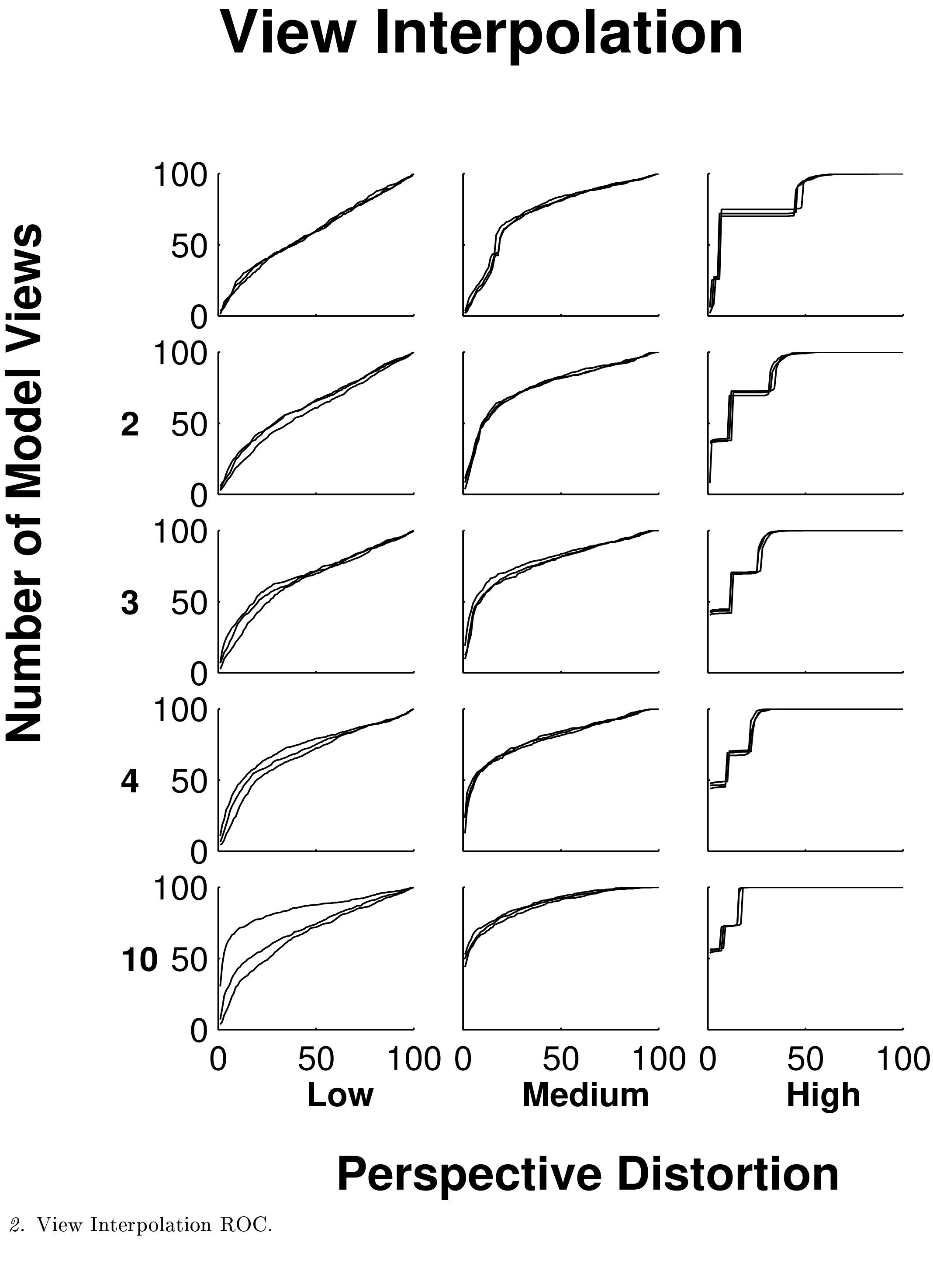

Figure 2. View Interpolation ROC.

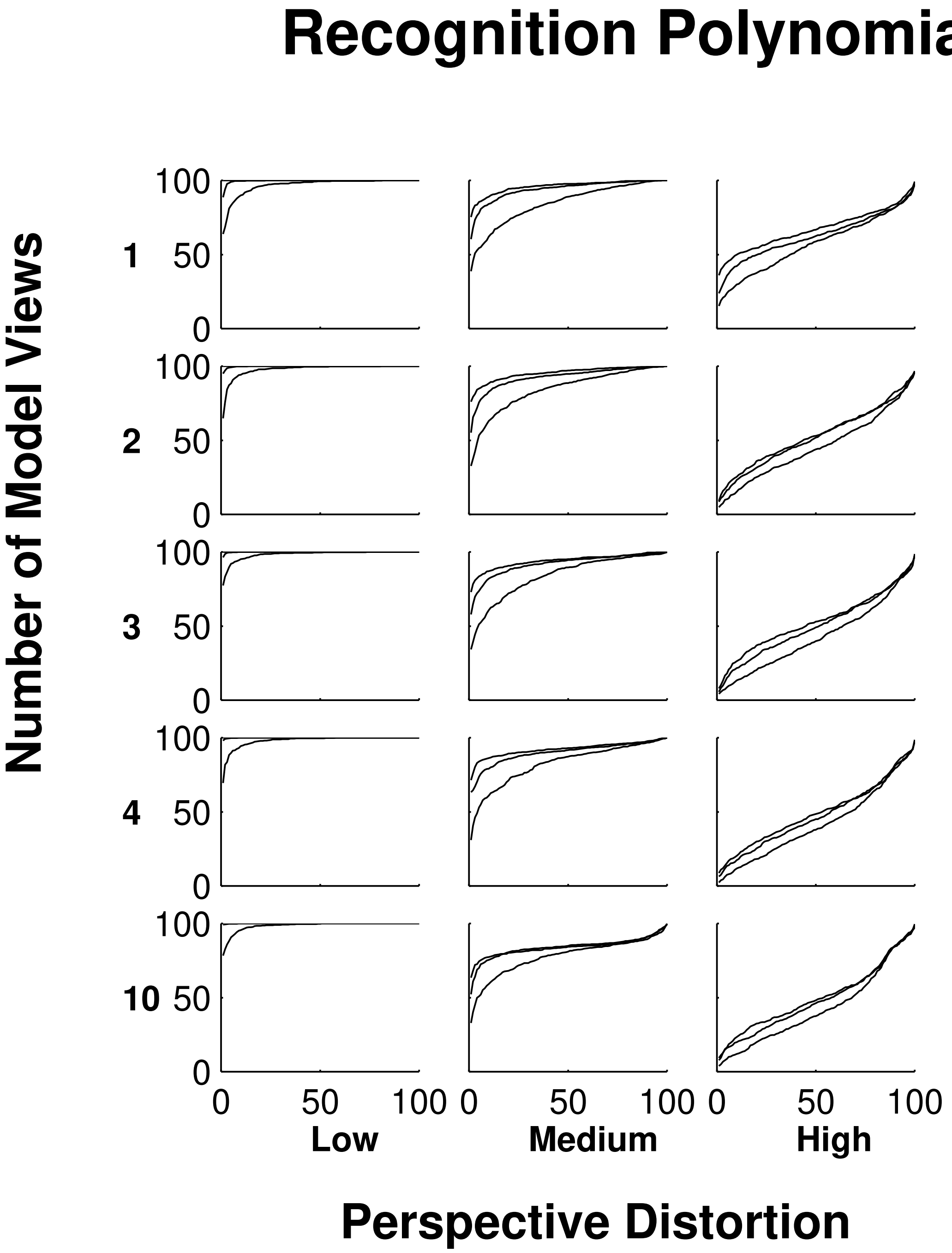

Figure 3. Recognition Polynomial ROC.

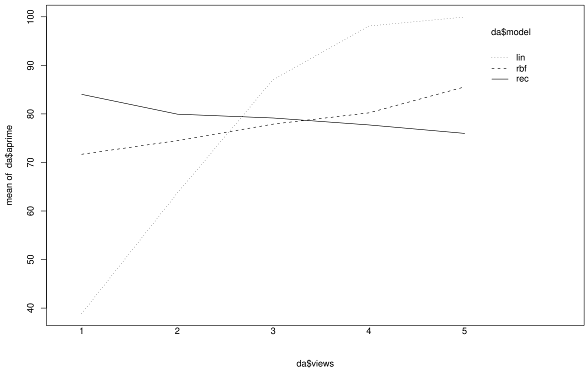

Figure 4. Model and views interaction.

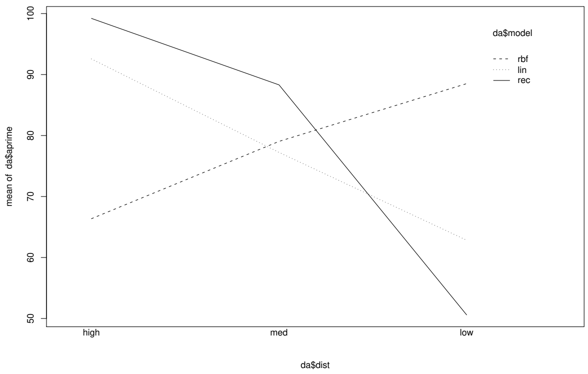

Figure 5. Model and distortion interaction.

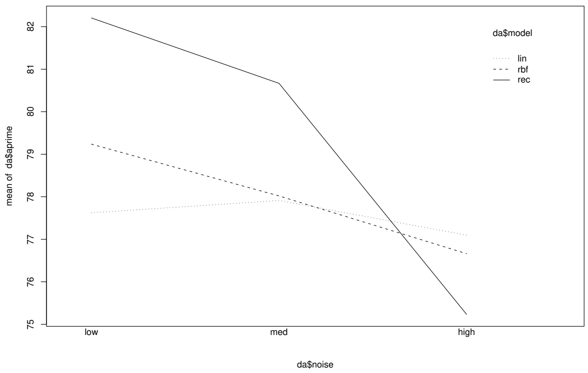

Figure 6. Model and noise interaction.

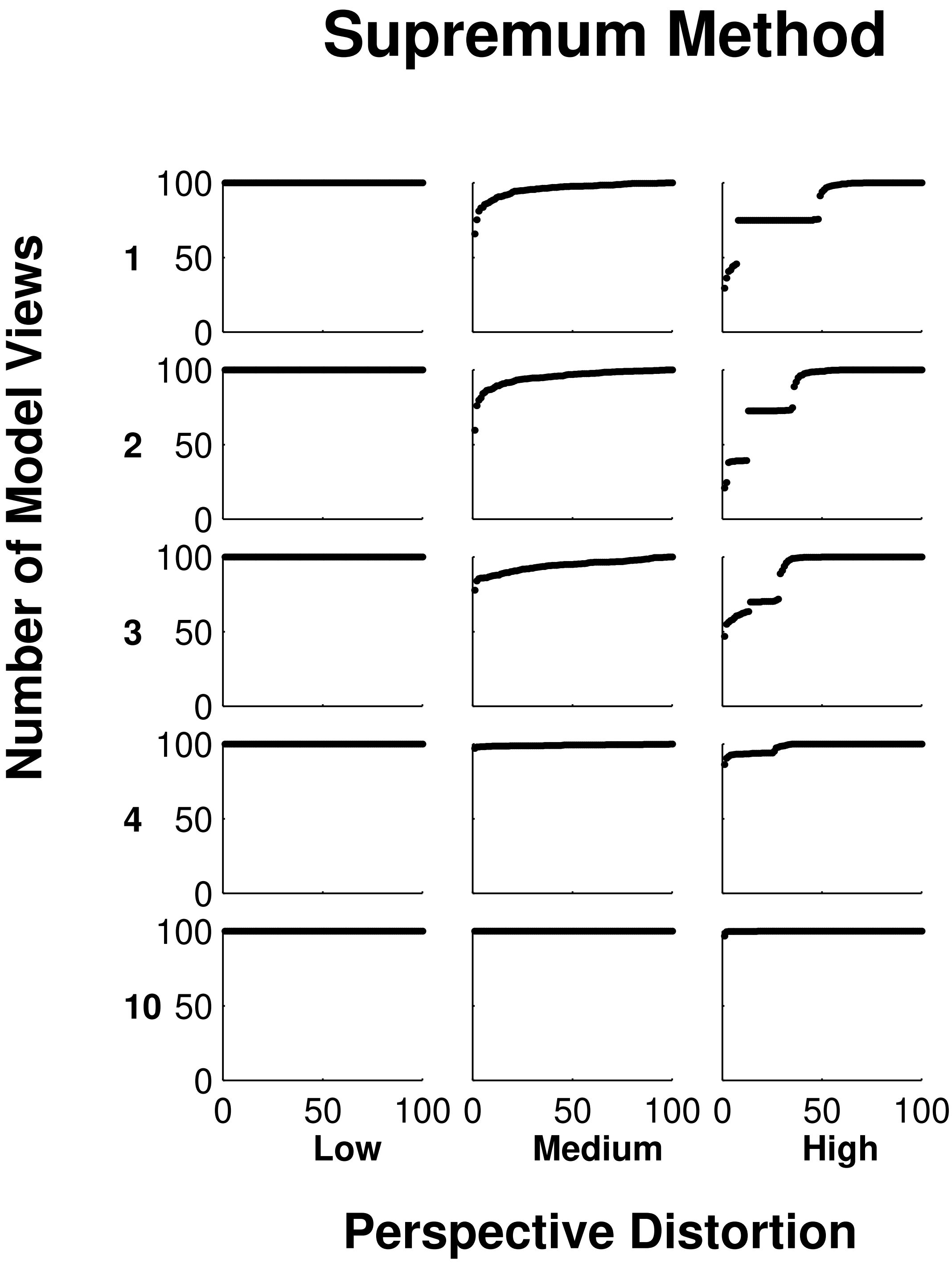

Figure 7. Supremum Method ROC.