

Therefore, the quantifier free part of the prenex formula (13) depends on $2k$ polynomial sets with a maximal degree of $8m$. By applying well-known algorithm of Collins (Collins 75), the three coordinates of \bar{s} can be eliminated leaving a quantifier free logic formula with $k_p = (2k + 3)^8(2 \cdot 8m)^{81}$ polynomial sets of maximal degree $m_p = 0.5(2 \cdot 8m)^8$. (The three quadratic constraints relating the sines and cosines of the Euler angles are also imposed.) \square

The VC-dimension of the class $C_{3D}^{project}(V)$ is given by the following theorem.

theorem 2 *For every semi algebraic set V of degree (k, m) in \mathbb{R}^3 ,*

$$VCdim(C_{3D}^{project}(V)) = O(\log km)$$

Proof: [sketch] The proof relies on results developed in previous papers. Let $S = \{x_1, \dots, x_N\}$ be a subset of \mathbb{R}^2 that is shattered by the class $C_{3D}^{project}(V)$. The union of boundaries $B_S = \bigcup_{i=1}^N \partial K_{x_i}^V$ of the semi algebraic parameter sets $\{K_{x_i}^V\}$ divides the parameter space \mathbb{R}^9 into connected components.

Milnor's classical theorem (Milnor 64) states that any partition of \mathbb{R}^n , that obeys a set of k polynomial inequalities, has at most $\frac{1}{2}(2 + d)^n$ connected components. (d is the total degree $\sum_{i=1}^k deg(f_i)$.)

Recall that, by Collins decomposition, each of the parameter sets $K_{x_i}^V$ is specified by $k_p = (2k + 3)^8(2 \cdot 8m)^{81}$ -many polynomial sets of the form $\{\bar{t}|f_j(\bar{t}, x_i) > 0\}$ each of degree $m_p = 0.5(2 \cdot 8m)^8$ or lower. Note that at least one of the functions $f_{ij}(\bar{t}) = f_j(\bar{t}, x_i)$ vanishes on each point of the boundary of $K_{x_i}^V$.

Consider now the product function $G(\bar{t}) = \prod_{i,j} f_{ij}(\bar{t})$. Any connected component of $\mathbb{R}^9 \setminus B_S$ corresponds to a union of one or more connected components of $\{t : G(t) > 0\}$ or of $\{t : G(t) < 0\}$. $G(\bar{t})$ is a $(k_p m_p N)$ -degree polynomial in 9 real variables, and, by our modification to Milnor theorem, the number of connected components of its positive set $\{\bar{t}|G(\bar{t}) > 0\}$ (as well as of its negative set, which is the pos-set of $-G$) is not higher than $(2 + k_p m_p N)^9$. Therefore any cardinality N of a point set that is shattered must satisfy the following relation

$$2^N \leq 2(2 + k_p m_p N)^9 \quad (14)$$

The theorem, as well the asymptotic lower bound (6) follows by a straightforward calculation. \square

A straightforward application of the bounds given in (Blumer et al. 89) may now give the number of data features which guarantees that the hypothesized instance is not more than ϵ_0 different from the true instance. More concrete assertions, such as that the localization result is "good enough" follow by specifying the required localization precision, and using (3) to specify ϵ_0 .

Conclusion

We analyzed the amount of data required to localize a 3D object from its 2D perspective projection, and obtained a rigorous upper bound on the number of data features required to draw a reliable hypothesis. The analysis was carried independently of the recognition method used, and in a certain sense, independently of the particular objects considered.

The same approach was used to derive the number of data features required to localize instances of 2D objects, associated with Euclidean, Similarity, Affine and Perspective transformation classes. It was also generalized to analyse the general model-based recognition task. Specifically, it was shown that the number of data features required for recognition grows at most logarithmically with the library size (Lindenbaum and Ben-David 94).

References

- Blumer, A., A. Ehrenfeucht, D. Haussler and M.K. Warmuth, 1989, "Learnability and The Vapnik-Chervonenkis Dimension", *JACM*, **36**(4), 929-965.
- S. Ben-David and M. Lindenbaum, 1993, "Localization vs. Identification of Semi-Algebraic Sets", Proceedings of the 6th ACM Conference on Computational Learning Theory, pp. 327-336.
- Collins, G.E., 1975, "Quantifier Elimination for Real Closed Fields by Cylindrical Algebraic Decomposition", Proceedings of the 2nd GI Conf. On Automata Theory and Formal Languages, *Springer Lec. Notes Comp. Sci.* **33**, pp. 515-532.
- Goldberg P. and M. Jerrum, 1993, "Bounding the Vapnik-Chervonenkis Dimension of Concept Classes Parametrized by Real Numbers", Proceedings of the 6th ACM Conference on Computational Learning Theory, pp. 361-368.
- Grimson, W.E.L., and D.P. Huttenlocher, 1991, "On the Verification of Hypothesized Matches in Model-Based Recognition", *IEEE Trans. on Pattern Analysis and Mach. Intel.*, **PAMI-13**(12), pp. 1201-1213.
- Kriegman, D.J. and J. Ponce, 1990, "On Recognizing and Positioning Curved 3D objects from Image Contours", *IEEE Trans. on Pattern Analysis and Mach. Intel.*, **PAMI-12**, pp. 1127-1137.
- Lindenbaum, M., 1993, "Bounds on Shape Recognition Performance", submitted.
- Lindenbaum, M. and S. Ben-David, 1994 "Applying VC-dimension Analysis to Object Recognition", 3rd European conference on Comp. Vision (to appear).
- Milnor, J., 1964, "On the Betti Numbers of Real Varieties", *Proc. Amer. Math. Soc.* **15**, pp. 275-280.
- Taubin, G., and D.B. Cooper, 1992, "2D and 3D Object Recognition and Positioning with Algebraic Invariants and Covariants", in *Symbolic and Numerical Computation for Artificial Intelligence*, B.R. Donald, D. Kapur, and J.L. Mundy, eds.

gradient, implying that the following degree- m polynomial constraint, ($m = \text{deg}(f_i)$),

$$[\bar{s} - \bar{f}] \cdot \nabla f_i(\bar{s}) = 0. \quad (9)$$

In addition, the projected point \bar{s} is included in the polynomial surface itself, and thus satisfies

$$f_i(\bar{s}) = 0 \quad (10)$$

- The other source for visible contours is the intersection of two polynomial boundaries which create normal discontinuities and are thus visible due to shading, texture, etc. The points on the intersection of the polynomial surfaces $\{f_i(\bar{s}) = 0\}$ and $\{f_j(\bar{s}) = 0\}$ are simply specified by requiring them to satisfy both polynomials. Note that such visible curves may lie, in the projection, within the outline but also on it.

An Algebraic expression for the extended boundary.

By definition, the extended boundary contains all points that are close enough to the perspective projection of some point in the contour generator. Formally, let $G(V_i)$ be the contour generator of the transformed three dimensional object, and $(G(V_i))_p$ be its perspective projection. The extended boundary of this projection, denoted $[(G(V_i))_p]^\Delta$, is given by

$$[(G(V_i))_p]^\Delta = \{\bar{q} = (q_x, q_y, 0) \mid \exists \bar{s} \in G(V_i) \text{ s.t. } \|\bar{q} - \text{proj}(\bar{s})\| < \Delta\} \quad (11)$$

We would like to know what is the number of random measurements needed, to guarantee with confidence $1 - \delta$ that the distance between the true instance of the object and any hypothesized instance that is consistent with the measurements is smaller than some value. The distance between instances is measured between the corresponding observable objects, that is, as the normalized area difference between the extended boundaries $[(G(V_i))_p]^\Delta$. To find a sufficient number of measurements, we proceed now to bounding the VC dimension of the associated concept class

$$C_{3D}^{\text{project}}(V) = \{[(G(V_i))_p]^\Delta \mid t \in T\}. \quad (12)$$

We apply the following technique:

- We assume that some set of points S of cardinality N is shattered by the concept class.
- We observe that every point in S corresponds to a partition of the parameter space into two parts: one of parameters for which the corresponding extended boundary of transformed set includes that point, and another that includes the parameters for which the corresponding extended boundary does not include that point.
- We observe that the N points in S partition the parameter space into connected components, such that all parameter in the same connected component correspond to extended boundaries that contain the same subset of S .

- We prove that the number of these connected components is polynomial in N implying that the number of subsets $A \subseteq S$ that may be written in the form $A = [(G(V_t))_p]^\Delta \cap S$ is also polynomial.

- In order to shatter the set S , every one of its 2^N subsets should be expressed as $[(G(V_t))_p]^\Delta \cap S$ for some t . Since only polynomial number of subsets can be written in this form, we conclude that this class of extended boundaries cannot shutter arbitrarily large point sets.

Partitioning the parameter space.

The first step in this direction is to find the structure of the parameters space:

Lemma 1 *For any semi algebraic set $V \subseteq \mathbb{R}^3$ of degree (k, m) ($m \geq 2$), transformed by a 3D rigid transformation, and projected using perspective projection on the image plane, and for every \bar{s} in that image plane, the set of transformation parameters*

$$K_{x_i}^V = \{\bar{t} \mid x_i \in [(G(V_i))_p]^\Delta\}$$

is also a semi-algebraic set of degree $(k_p = (2k+3)^8(2 \cdot 8m)^{81}, m_p = 0.5(2 \cdot 8m)^8)$ (in the parameter space \mathbb{R}^9).

Proof: The proof is based on the theory of quantifier elimination from Logic theory. The parameter set $K_{x_i}^V$ may be written as the truth set of a prenex formula in the coordinates of \bar{s} and the 9 parameters t_1, \dots, t_9 as variables. Recall that t_1, \dots, t_9 are the parameters of the inverse transformation, which transform every point on V_i into a point on V .

$$K_{x_i}^V = \{\bar{t} \mid \exists \bar{s} \text{ s.t. } \bar{s} \in G(V_i) \wedge \|\bar{x}_i - \text{proj}(\bar{s})\| < \Delta\} \quad (13)$$

Now, the second condition, $\|\bar{x}_i - \text{proj}(\bar{s})\| < \Delta$, does not depend on the transformation and can be easily transformed to a polynomial inequality of second degree in the coordinates of \bar{s} . The first condition is more complicated: a point \bar{s} in the contour generator $G(V_i)$ of the transformed object V_i must be either in the transformed intersection of two polynomial surfaces or on the occluding boundary of one transformed polynomial surface.

- To satisfy the first option it suffice that will satisfy two polynomial constraints, such as $f_j(\mathbf{R}'\bar{s} + \mathbf{t}') > 0$ and $f_{j'}(\mathbf{R}'\bar{s} + \mathbf{t}') > 0$, (or \geq or $=$), where f_j and $f_{j'}$ are two of the polynomials that specify V . Considering both the coordinates of \bar{s} and the transformation parameters as variables, these polynomials are of maximal degree of $4m$.
- For the point \bar{s} to be on a smooth occluding contour, the gradient of the transformed polynomial must be orthogonal to the viewing vector $[\bar{s} - \bar{f}]$. The orthogonality is preserved if the coordinate system is changed and therefore we can write this condition as

$$\nabla [f_j(\mathbf{R}'\bar{s} + \mathbf{t}')] \cdot [\mathbf{R}'(\bar{s} - \bar{f}) + \mathbf{t}'] = 0$$

This constraint is polynomial with maximal degree of $8m$.

we may now calculate the number of data features sufficient to guarantee that every consistent hypothesis is ϵ_0 -accurate with confidence $1 - \delta$. In the rest of the paper we bound the VC-dimension of one particular class: extended boundaries of perspective projections of 3D objects. We do not refer to particular objects, but just assume that the object belongs to the extremely large class of objects, defined in the next section.

The class of objects considered - Semi-algebraic sets

We shall focus on well behaved geometrical objects - the Semi-Algebraic subsets of \mathbb{R}^2 and \mathbb{R}^3 .

Definition 2: A semi-algebraic open set of degree (k, m) in \mathbb{R}^n is a set that can be represented as a boolean combination of k sets of the form $\{\bar{x} \in \mathbb{R}^n : f_j(\bar{x}) Q 0\}$ where the functions f_j are real polynomials of maximal degree m , and Q is one of the relations $\leq, =, <$.

Polynomial objects of modest degrees (e.g. 4) suffice to describe complicated objects and thus provide high representation power (see, e.g. (Taubin and Cooper 92)). The class we consider here is even richer: besides polynomial objects it also contains combinations of them which include, e.g., polygonal objects (which, for k being the number of polygon sides, are semi algebraic sets of degree $(k, 1)$). The family of Semi-Algebraic sets is parametrized, meaning that the class of objects considered is actually not limited.

Localization - The VC-dimension of transformed Semi-Algebraic sets

Our general approach treats both two dimensional and three dimensional semi-algebraic objects and a wide class of transformation. Here, we focus on three dimensional semi-algebraic objects and perspective projection of them, and analyse the class of concepts which are the extended boundary of these projections. We show that the VC-dimension of this class is logarithmic in the complexity of the object, and obeys the asymptotic upper bound

$$B_{3D\Delta}^{project}(V) = 712 \log(km), \quad (6)$$

thereby providing the parameter needed to determine the number of two dimensional data features (taken from the projected image), required to localize the object with the required precision and confidence. The bound does not depend on the particular object chosen but only on its complexity, as expressed by the number of polynomials that define it, k , and by their degree, m .

The VC-dimension of projected 3D semi-algebraic objects.

We consider the common imaging procedure, which involves projecting the object on an image plane and

getting the information from the projection. We assume here that the imaging process is done by a pin-hole camera, which implements a perspective projection and, for our purposes, is a good approximation to common realistic cameras. Furthermore, we follow Kriegman and Ponce approach (Kriegman and Ponce 90) and assume that only sharp edges in the projected image are observable. Such sharp edges in the image may come either from the outline of the object, or from discontinuities of its surface normal that are usually the result of two intersecting polynomial surfaces.

The object instance class

Considering the model-based localization problem, we assume that the object present in the scene is an instance V_t of a known object model V , associated with some unknown but general rigid transformation $t = (\mathbf{R}, \bar{\mathbf{t}})$

$$V_t = \{\bar{s}' = \mathbf{R} \bar{s} + \bar{\mathbf{t}} | \bar{s} \in V\} \quad (7)$$

where both \bar{s} and \bar{s}' are 3D coordinate vectors that describe points in the 3D space, $\bar{\mathbf{t}}$ is a 3D translation vector and \mathbf{R} is a rotation matrix. (Note the following small change in notation: Unlike the description of the general framework, V_t does not describe the object after the full transformation but denotes the object before the projection. Consequently, the extended boundary will be redefined.) To parametrize this transformation, we use the parameter vector $\bar{t} = \{t_1, \dots, t_9\}$, which includes the translation components and the sines and cosines of the Euler rotation angles of the inverse transformation. The class of 9-tuples which are valid parameters of the rigid transformation is constrained by some equalities between the parameters and is denoted T .

The perspective projection process

Let the optical axis of the pin-hole camera coincide with the z -axis, the image plane be on the $z = 0$ plane, and the focal point be at $\bar{f} = (0, 0, -f)$. One line passes between every point $\bar{s} = (s_x, s_y, s_z)$ in the 3D space and the focal point, and specifies the projection of \bar{s} as its intersection with the image plane. This implies the simple expression for perspective projection of \bar{s} : $proj(\bar{s}) = \bar{r} = (r_x, r_y, 0)$.

$$r_x = \frac{f}{s_z + f} s_x \quad r_y = \frac{f}{s_z + f} s_y. \quad (8)$$

The contour generators

Clearly, not all points of the object V_t are projected to the visible curves in the image. Points that are projected belong either to the occluding contour or to discontinuities of the surface normal and thus must obey some constraints:

- The projected point may be on the occluding contour but only on one polynomial surface $f_i(s_x, s_y, s_z) = 0$. In this case the viewing direction vector $\bar{s} - \bar{f} = (s_x, s_y, s_z) - (0, 0, -f)$ is tangent to the polynomial surface and perpendicular to the

seems very difficult to model. The simple model, suggested in the following lines, is not claimed to cover all situations in computer vision. It addresses, however, the uncertainty on the observed part of the object and the inaccuracy of the measurements.

We model the uncertainty in the data features available by assuming that the data features are randomly drawn in the neighborhood of the object boundary. Let ∂V_t be the boundary of the instance of the object V , after a transformation t . In the simple case, where only boundary points associated with inaccuracy Δ are available, we assume that they are independently sampled according to a uniform distribution, inside

$$V_t^\Delta = \{ r \mid \exists s \in \partial V_t \text{ s.t. } \|s - r\| < \Delta \}, \quad (1)$$

to which we refer as either “extended boundary” or “observable object”. More complicated data collection models, which include arbitrary but bounded sampling distributions and data features which include boundary slope measurements, are considered in the full version.

The interpretation stage

We refrain from referring to any particular method for inferring the hypothesis. The only assumption taken is that the interpretation stage may draw any hypothesis that is consistent with the data. Let H be the set of possible hypotheses, which, in the model based setting, may contain instances of different objects under different transformations. Then, for M being the data set, the algorithm may draw any hypothesis in $\{h \mid M \subset h; h \in H\}$.

An error measure

We treat all recognition tasks uniformly and consider them successful if a special error measure, defined below, between the true object and the hypothesized one, is guaranteed to be lower a threshold value. For V_t being the true object that is present in the scene and $W_{t'}$ being some hypothesized instance, the error associated with this hypothesis is defined as the normalized difference between the volumes of the corresponding observable objects.

$$E(V_t, W_{t'}) = \frac{Vol(V_t^\Delta \setminus W_{t'}^\Delta)}{Vol(V_t^\Delta)} \quad (2)$$

This error measure agrees with the intuitive meaning of recognition and localization. High localization accuracy, for example, implies that the boundaries of the true object and the hypothesis are very close, and leads to a small difference between the corresponding extended boundaries. Low localization accuracy, on the other hand, allows larger error.

The uniform recognition accuracy measure may be used to specify recognition success in the more familiar forms, by setting the maximal error, for which the hypotheses is still considered successful. For example, regarding the *localization task*, one may consider any

distance measure $D(\cdot, \cdot)$ (say, Hausdorff distance,) between two object instance, and denote a localization procedure successful if, for the hypotheses drawn, the distance between the true object V_t and the hypothesis $V_{t'}$ is d_0 or smaller. (The value d_0 may be adjusted arbitrarily according to the localization precision required.) Requiring a recognition accuracy better than

$$e_0 = \max_{t, t' \in T; D(V_t, V_{t'}) > d_0} E(V_t, V_{t'}). \quad (3)$$

guarantees that no instance of V which is d_0 -far from the true instance is drawn as an hypothesis.

Therefore, we are interested in the following question:

How many measurements are needed to guarantee, with a certain confidence $1 - \delta$, that all hypotheses that are at least e_0 -far from the true object instance are rejected ?

Learning and recognition

Now, the equivalence between the localization task and PAC learning should be apparent: let $\{V_t \mid t \in T\}$ be a set of instances associated with one object V and a class of instances T . To every instance from this set, associate a concept identical to the extended boundary.

$$V_t \longleftrightarrow V_t^\Delta \quad (4)$$

$$\{V_t \mid t \in T\} \longleftrightarrow C_{T^\Delta}(V) = \{V_t^\Delta \mid t \in T\} \quad (5)$$

Every data feature extracted from the object boundary provides a (positive) example to the corresponding concept. Learning a concept in $C_{T^\Delta}(V)$ with an accuracy better than e_0 means that all concepts in the class, associated with a symmetric difference greater than e_0 , are not consistent with the examples. Note however, that according to our data collection model, the density is zero everywhere except inside the concept itself. Assuming further that the distribution is uniform within the extended boundary, implies that the recognition error (2) is also smaller than e_0 , and that the recognition task is successful.

While the PAC learnability results usually holds for arbitrary distribution, we will assume that the data features are placed according to a uniform distributions densities. The reason is the need to establish a relation between the recognition accuracy measure $E(V_t, W_{t'})$ and the symmetric difference $V_t^\Delta \Delta W_{t'}^\Delta$, induced by the sampling density. This cannot be achieved by all distributions: Consider for example a distribution that is concentrated in a single point. The learning performance in this case will be excellent as the density weighted symmetric difference and the associated prediction error will be null after one example. The knowledge about the location of the object will, however, be poor because completely different hypotheses can be consistent if they share one point with the true object (either inside or outside).

Inserting the VC-dimension of the concept class $C_{T^\Delta}(V) = \{V_t^\Delta \mid t \in T\}$ into the bound in theorem (1),

give here, was already considered in (Ben-David and Lindenbaum 93) and (Goldberg and Jerrum 93).

The paper is divided into two major parts: explaining the relation between learning and recognition, and calculating the VC-dimension associated with the task of localizing a 3D object from its 2D perspective image.

Learnability and the VC-Dimension

Given a collection, \mathcal{K} , of subsets of some base set, X , and a measure of difference between the members of \mathcal{K} , a set of points $\{x_1, \dots, x_n\} \subset X$ is said to ϵ -pin down \mathcal{K} , if, for every pair of sets $A, B \in \mathcal{K}$, if $A \cap \{x_1, \dots, x_n\} = B \cap \{x_1, \dots, x_n\}$ then the difference between these members of \mathcal{K} is at most ϵ .

It is evident that the size of such 'pinning down' sets, as well as their number, depends upon the family \mathcal{K} of sets. The theory of computational learnability formalizes this issue within the framework of Valiant's PAC learning model. In that model the family of sets \mathcal{K} is usually called a 'concept class' and its members are 'concepts'. The model assumes the existence of some probability distribution P over X . This probability plays a double role: First, the difference between concepts is specified as the P -probability of hitting their symmetric difference. Second, the 'fraction' of pinning-down n -tuples (among all n -tuples of points of X) is measured by the probability of picking such a tuple by i.i.d. sampling n -many times according to P .

A class \mathcal{K} is called *PAC-learnable* (or just 'learnable') if, for every positive ϵ, δ , there exists a finite number m (depending upon these parameters) such that for every probability distribution P over X , the P^m -probability of picking an m -tuple that ϵ -pins down \mathcal{K} exceeds $(1 - \delta)$. It turns out that a concept class is learnable **iff** a purely combinatorial parameter – the Vapnik-Chervonenkis dimension of this class, is finite (Blumer et al. 89).

Definition 1: [Vapnik-Chervonenkis Dimension] Let X be some set and \mathcal{K} a collection of its subsets.

- We say that \mathcal{K} shatters a set $A \subseteq X$, if, for every $B \subseteq A$, there exists some $C \in \mathcal{K}$ such that $C \cap A = B$.
- The Vapnik-Chervonenkis Dimension (in short, *VC-dim*) of \mathcal{K} is the maximum number d such that \mathcal{K} shatters a set of size d . (If \mathcal{K} shatters sets of unbounded size, we say that its *VC-dim* is ∞).

Example: Let X be the unit interval and \mathcal{K} be the collections of all its subintervals whose length is 0.1. I.e., $\mathcal{K} = \{[a, a + 0.1] : 0 \leq a \leq (1 - 0.1)\}$. It is not hard to realize that \mathcal{K} shatters every pair of points in $[0.1, 0.9]$ which are at most 0.1 apart. On the other hand, \mathcal{K} shatters no subset A of the interval whose cardinality exceeds 2. It follows that $\text{VC-dim}(\mathcal{K}) = 2$.

We can now state the result of Blumer et. al. (Blumer et al. 89) showing how the VC-dim of a class determines its learnability.

theorem 1 [(Blumer et al. 89)]

- A class \mathcal{K} is PAC-learnable **iff** it has a finite VC-dimension.
- If $\text{VC-dim}(\mathcal{K}) = d$ then, for every positive ϵ and δ ,
 1. if

$$m \geq \max \left(\frac{4}{\epsilon} \log \frac{2}{\delta}, \frac{8d}{\epsilon} \log \frac{13}{\epsilon} \right)$$

then, for every probability distribution P over X , the P^m -probability of picking an m -tuple that ϵ -pins down \mathcal{K} exceeds $(1 - \delta)$.

2. On the other hand, if

$$m < \max \left(\frac{1 - \epsilon}{\epsilon} \ln \frac{1}{\delta}, d(1 - 2(\epsilon(1 - \delta) + \delta)) \right)$$

then, there exists a probability distribution P over X , such that the P^m -probability of picking an m -tuple that ϵ -pins down \mathcal{K} is less than $(1 - \delta)$.

Note that the upper bound of this theorem guarantees the existing of many ϵ -pinning-down tuples of size linear in the VC-dim of a class and in $\frac{1}{\epsilon}$, for every underlying probability distribution. The lower bound, on the other hand, only states the *existence* of a 'difficult' distribution and does not rule out the possibility that, for some specific distribution, the task of pinning down a class may require fewer sample points. In (Ben-David and Lindenbaum 93) some evidence is provided to show that, for classes of algebraically-defined objects in the Euclidean space, the lower bound above is indeed a close estimate of the minimal size of pinning-down sets relative to the uniform distribution.

Learning and recognition

This section discusses the relation between learning and recognition, and shows that in a proper setting, recognition tasks are equivalent to learning tasks in the sense that an object is recognized (or localized) if some related concept class is PAC learned with a certain prediction power.

We consider recognition processes that are composed of a data collection stage followed by an interpretation stage. In the first stage data features are collected in random locations, independently, and according to fixed distribution. In the second stage the data collected is combined with prior knowledge, and is interpreted, to yield an hypothesis on the identity and pose of the object in the scene. These stages are described in the next two sections.

The data collection stage

In Vision scenarios, information is usually obtained from the observed object's edges in an image, and is usually associated with some location error. Data extraction from images involves many factors including illumination, occlusion, the effect of edge detectors and

Applying VC-dimension Analysis To 3D Object Recognition from Perspective Projections *

Michael Lindenbaum and Shai Ben-David

Computer Science Department, Technion

Haifa 32000, ISRAEL

(mic, shai) @cs.technion.ac.il

Abstract

We analyze the amount of information needed to carry out model-based recognition tasks, in the context of a probabilistic data collection model, and independently of the recognition method employed. We consider the very rich class of semi-algebraic 3D objects, and derive an upper bound on the number of data features that (provably) suffice for localizing the object with some pre-specified precision. Our bound is based on analysing the combinatorial complexity of the hypotheses class that one has to choose from, and quantifying it using a VC-dimension parameter. Once this parameter is found, the bounds are obtained by drawing relations between recognition and learning, and using well-known results from computational learning theory. It turns out that this bounds grow logarithmically in the algebraic complexity of the objects.

Introduction

We present here a quantitative analysis of the amount of information required for Model-based object recognition. Taking a statistical approach, we consider a random data collection model and analyse the number of measurements that guarantees recognition success within a certain confidence. Intuitively, more data is needed if the recognition procedure is required to discriminate between object instances that are visually similar, and if more alternatives are allowed by the possible instance specification. In this paper these intuitive observations are quantified by deriving a rigorous upper bound on the number of features required to succeed. Our approach is very general and applies to a very large class of objects, and to several transformation classes. It is based on a combinatorial analysis, which provides the VC-dimension of concept classes associated with this objects and the transformations. In this note we concentrate on localizing 3D objects from their perspective projections.

The bounds are derived relying on the observation that the recognition task is related to a learning task,

*This work was supported by the Technion fund for the promotion of research and by the Smoler research fund

in which one tries to learn a subset of some space, by observing samples of this space. We consider the *Probably Approximately Correct* (PAC) learning model, which assumes that the samples available are randomly drawn, and requires that the hypothesis provided is a good approximation to the true subset, within a certain prespecified confidence. In this setting, the elegant PAC learning theory guarantees that the number of samples required to learn is not higher than a certain threshold, which grows with the accuracy of the hypothesis, the required confidence, and a certain parameter, associated with the of allowed hypotheses, and known as the VC-dimension. The mathematical heart of our result is therefore an analysis of the VC-dimension of a certain concept class, related to the localization task. Interestingly, the analysis and its results are independent of the particular object considered, and the derived VC-dimension parameter depends only on the object's complexity and the class of transformations.

The results we provide, besides quantifying the *fundamental difficulty* of recognition tasks, should be useful for analyzing reported results by comparing them to the theoretical bounds, and to designing recognition procedures. Many recognition paradigms use a consistent data subset as a sufficient evidence to the presence of an object in the scene. Our results, together with other considerations described latter, may be used to set the sufficient size of such subsets that guarantees the reliability of such a procedure.

The *fundamental difficulty* of recognition tasks was already considered before in several papers: Lindenbaum used a different approach to set upper and lower bounds on the amount of data required to succeed in recognition and localization tasks (Lindenbaum 93). Grimson and Huttenlocher considered a complementary aspect of the recognition *fundamental difficulty* (Grimson and Huttenlocher 91). While we basically assume that all the data features belong to an object, they examine the possibility that a subset of "noise data features" will give a false evidence for the presence of an object in the scene. Some of the abstract mathematical treatment, without the interpretation we