Acknowledgment

The author enjoyed stimulating interaction with T. Y. Kong on the subject of this paper and received valuable suggestions from the anonymous reviewers.

References

[1] C. Arcelli, L. Cordella, and S. Levialdi, "Parallel thinning of binary pictures," *Electron. Lett.*, vol. 11, pp. 148–149, 1975.

[2] R. T. Chin, H. -K. Wan, D. L. Stover, and R. D. Iverson, "A one-pass thinning algorithm and its parallel implementation," *Comput. Vision Graphics Image Processing*, vol. 40, pp. 30–40, 1987.

[3] M. J. B. Duff and T. J. Fountain, *Cellular Logic Image Processing*. New York: Academic, 1986.

[4] T. J. Fountain and M. J. Shute, Eds., *Multiprocessor Computer Architectures*. Amsterdam: North-Holland, 1990.

[5] M. Gökmen and R. W. Hall, "Parallel shrinking algorithms using 2-subfields approaches," *Computer Vision Graphics Image Processing*, vol. 52, pp. 191–209, 1990.

[6] Z. Guo and R. W. Hall, "Parallel thinning with two-subiteration algorithms,"*Commun. ACM*, vol. 32, pp. 359–373, 1989.

[7] ——, "Fast fully parallel thinning algorithms," *CVGIP: Image Understanding*, vol. 55, pp. 317–328, 1992.

[8] R. W. Hall, "Fast parallel thinning algorithms: Parallel speed and connectivity preservation," *Commun. ACM*, vol. 32, pp. 124–131, 1989.

[9] ——, "Tests for connectivity preservation for parallel reduction operators," *Topol. Applications*, vol. 46, pp. 199–217, 1992.

[10] ——, "Optimally small operator supports for fully parallel thinning algorithms," Techn. Rep. TR-SP-91-01, Dept. of Elect. Eng., Univ. of Pittsburgh.

[11] C. M. Holt, A. Stewart, M. Clint, and R. H. Perrott, "An improved parallel thinning algorithm," *Commun. ACM*, vol. 30, pp. 156–160, 1987.

[12] T. Y. Kong and A. Rosenfeld, "Digital topology: Introduction and survey," *Comput. Vision Graphics Image Processing*, vol. 48 pp. 357–393, 1989.

[13] K. Preston and M. J. B. Duff, *Modern Cellular Automata*. New York: Plenum, 1984.

[14] C. Ronse, "A topological characterization of thinning," *Theoret. Comput. Sci.*, vol. 43, pp. 31–41, 1986.

[15] ——, "Minimal test patterns for connectivity preservation in parallel thinning algorithms for binary digital images," *Discrete Applied Mathe.*, vol. 21, pp. 67–79, 1988.

[16] A. Rosenfeld, "A characterization of parallel thinning algorithms," *Inform. Contr.*, vol. 29, pp. 286–291, 1975.

[17] A. Rosenfeld and A. C. Kak, *Digital Picture Processing*. New York: Academic, 1982, vol. 2.

[18] R. Stefanelli and A. Rosenfeld, "Some parallel thinning algorithms for digital pictures," *J. ACM*, vol. 18, pp. 255–264, 1971.

On Learning to Recognize 3-D Objects from Examples

Shimon Edelman

***Abstract*—Previous results on nonlearnability of visual concepts relied on the assumption that such concepts are represented as sets of pixels [1]. This correspondence uses an approach developed by Haussler [2] to show that under an alternative, feature-based representation, recognition is PAC learnable from a feasible number of examples in a distribution-free manner.**

***Index Terms*—Complexity, learning from examples, object recognition, representation, vision.**

I. Introduction

A. Background

Whatever innate mechanisms may be available to the human visual system for distinguishing between important and unimportant features of the outside world, there is little doubt that descriptions of *objects* built from these features are learned from examples. Since vision is the primary source of data for category formation, the study of visual learning can lead to important insights into the structure of cognition.

Any theoretical investigation of learning must start with the selection of a class from which the concepts to be learned will be drawn. This selection poses the difficult problem of achieving a compromise between the conflicting requirements of description and generalization. On one hand, the concepts are required to be sufficiently expressive to describe faithfully the target patterns and to capture any fine distinctions that may be present among them. On the other hand, concepts whose descriptions must be learned from examples and, at the same time, support generalization to novel situations should be kept as simple as possible.

The present note addresses this dilemma in the context of learning object recognition. First, it provides a background for the discussion by restating a general formulation of the notion of learnability due to Valiant and Haussler [3], [2]. Next, it mentions previous approaches to the analysis of learnability of recognition [1], [4] and attributes their negative results to a particular choice of concept class that appears to favor description capability at the expense of generalization properties. An alternative, feature-based approach to the learning of visual recognition is then formulated and analyzed. Finally, computer simulations whose results are compatible with the proposed theoretical approach [5], [6] are described and discussed.

Learnability of visual concepts can be formalized as follows (see [2]). Let $\mathcal{P}$ be the problem of learning functions belonging to a hypothesis space $\mathcal{F}$, with *domain* X and *range* Y. A pair $(x, y) \in S = X \times Y$ is called an *example*; a sequence of examples is called a *sample*. Let $L : Y \times Y \to [0, M]$ for some real $M > 0$ be the *loss function*. Finally, let $\mathcal{D}$ be a family of probability measures on S. One can now define what it means to solve the learning problem $\mathcal{P}$.

Definition *(page 6 of Haussler [2]):* Let $\mathcal{P}$ be a learning problem defined by $X, Y, \mathcal{F}, L$, and $\mathcal{D}$. Let $\mathcal{L}$ be a learning function from the

Manuscript received May 31, 1991; revised November 3, 1991. Recommended for acceptance by Associate Editor M. Nagao.

The author is with the Department of Applied Mathematics and Computer Science, Weizmann Institute of Science, Rehovot, Israel.

IEEE Log Number 9209990.

set of all samples over S into $\mathcal{F}$, that is, $\mathcal{L} : \cup_{m \geq 1} S^m \rightarrow \mathcal{F}$. $\mathcal{L}$ is said to solve $\mathcal{P}$ if for all $\nu > 0$, $0 < \alpha < 1$ and $0 < \delta < 1$, there exists a finite sample size $m = m(\nu.\alpha.\delta)$ such that for all $D \in \mathcal{D}$, if $\bar{\xi}$ consists of m examples drawn independently at random according to D with probability at least $1 - \delta$, then

$$d_\nu\big(\boldsymbol{er}_D\big(\mathcal{L}(\bar{\xi})\big).\, optimum(D.\mathcal{F})\big) \leq \alpha \qquad (1)$$

where the error $\boldsymbol{er}_D(\mathcal{L}(\bar{\xi}))$ is defined as the expectation of $\mathcal{L}(\bar{\xi})$ with respect to the distribution D, $optimum(D.\mathcal{F})$ is the infimum of $\boldsymbol{er}_D(f)$ over all $f \in \mathcal{F}$, and the metric d_ν is $d_\nu(r.s) = \frac{|r-s|}{\nu+r+s}$. In other words, the function $\mathcal{L}$ solves the learning problem if it produces with high probability a hypothesis that is acceptably close to the optimal hypothesis in $\mathcal{F}$ (this parallels Valiant's definition of probably approximately correct (PAC) learning [3]).

B. Previous Work on Learnable and Nonlearnable Visual Concepts

The issue of *representation* corresponding to the choice of the hypothesis space $\mathcal{F}$ in the above definition is of utmost importance in vision [7]. The choice of representation is not entirely free. At the very least, it is constrained by the fact that input transducers in both biological and artificial visual systems provide signal that is spatially discrete (sampled). To date, complexity analyses of visual learning have taken this constraint at face value in assuming that the basic unit of representation is the pixel [1], [4]. As a result, counterintuitive conclusions regarding the nonlearnability of the recognition of visual concepts defined by templates [1] have been obtained. Specifically, the number of examples m required to achieve PAC learning of Boolean template representations was shown to be unfeasible.

Fortunately, it appears that a somewhat misleading definition of visual complexity may have been responsible for these results. As far as the human visual system is concerned, the definition of complexity in terms of image resolution is inappropriate for two reasons. First, although spatial resolution in human vision is limited by the discrete sampling of the retinal image at the photoreceptor level, the optics of the eye act as a low-pass spatial filter, cutting off frequencies above approximately 60 cycles per degree [8]. Thus, increasing the resolution of a picture above a certain level cannot possibly affect the percept it evokes. Second, experimental findings suggest that varying the amount of detail in a picture has little effect on recognition, even when the variations themselves are easily noticeable [9]. The important factor in recognition seems to be closeness to a prototype; a faithful line drawing rendition of an object is recognized just as easily as its photograph.

II. Recognition of 3-D Objects: Theoretical Learnability

A. An Alternative Formulation of the Complexity of Recognition

In contradistinction to pixel-based definitions of visual complexity, the author proposes a resolution-independent complexity measure whose primitives are inspired by the notion of *primal sketch*, which was introduced by Marr [7]. According to Marr, primal sketch is an intermediate representation of the visual input that is formed by the first stage of bottom-up visual processing in which simple spatial properties of the input (e.g., localization of intensity or texture gradients, spatial aggregation) are made explicit. The independence of the resulting representation on input resolution is a major consequence of this process of abstraction. In fact, algorithms for low-level visual tasks such as edge detection and binocular stereo work best when applied at several levels of resolution simultaneously; see, e.g., [10].

To assess visual complexity of a class of objects for the purpose of recognition, it is sufficient to consider sets of object features that leave out irrelevant details. A feature is defined as a function from the set of all objects into $\mathcal{R}^n$. If the range of a feature bears no relationship to the space in which the object's geometry is described, that feature may be referred to as abstract (a typical example of such a feature is color). Another possibility is to consider the location of a certain relatively compact part of the object (say, an eye in face recognition) as a *spatial* feature.

In pure shape-based recognition, which is the main concern of this note, it is assumed that objects can be adequately described by their spatial features. It turns out that for such objects, any collection of at least three noncollinear features forms a *diagnostic* set, that is, allows the object to be recognized (distinguished from other objects of the same category) in a 2-D image [11], [12]. The author shall define, therefore, the complexity of a class of objects as the size k of their diagnostic feature set.

In the simplest case of 3-D point sets, the features are merely the locations of the points themselves. Clearly, the method of comparing two such feature sets must allow for some location uncertainty; otherwise, any small perturbation would render an object unrecognizable. Note that this is where the proposed approach differs from pixel-based methods, which raise combinatorial problems by considering explicitly the various pixel configurations caused by shape perturbation. One simple way to achieve relative insensitivity to feature location uncertainty is to blur at least one of the feature sets before comparison, e.g., by convolving it with a bell-shaped kernel such as the Gaussian. (Convolution with a Gaussian kernel has been previously proposed as a method for regularizing the estimation of an unknown continuous probability density from a set of discrete samples [13], [14]).

An object can then be represented as the set F of allowed transformations of its "canonical" $3k$-dimensional state (a concatenation of k triples of coordinates: one per each feature). In other words, the indicator function for the set F $f = f_2 \circ f_1$ is a composition of a transformation $f_1 : \mathcal{R}^{3k} \times T \rightarrow \mathcal{R}^{3k}$, where $T = \mathcal{R}^n$ is the n-dimensional parameter space of a Lie group of one's choice (see [15] and [16] for examples of application of Lie groups to perception) and a fuzzy indicator $f_2 : \mathcal{R}^{3k} \rightarrow [0.1]$. For concreteness, the subsequent analysis is limited to the general linear group in 3-D $GL(3)$, which includes rigid rotations in 3-D as a subgroup. Adopting the $GL(3)$ group while restricting the range of transformation parameters is equivalent to allowing the object to undergo certain nonrigid deformation and still requiring that it be recognized, provided that the deformation is not too severe. Note that this general representation is somewhat impractical because the visual system has no direct access to the 3-D coordinates of the features. One must consider, therefore, *projection* $p : \mathcal{R}^{3k} \rightarrow \mathcal{R}^{2k}$ as the first stage in any process of recognition (for simplicity, projection is assumed to be orthographic).

The problem of learning to recognize 3-D objects can now be given two different formulations, depending on the availability of *correspondence* [17] information. The first formulation assumes that the constituent features of an object can be singled out in its retinal projection and can be matched to corresponding features in the stored representation of the object. Although finding the correct correspondence between object and model features is, in general, a difficult computational problem, in many practical situations, it can be solved efficiently by using distinctive features [18] or an incremental approach resembling pruned search [19]. Note that the representation under the correspondence formulation is effectively $2k$-dimensional, as suggested in the previous paragraph.

The second formulation considered below does not assume knowledge of correspondence and postulates instead a "collapsed" 2-D representation of an object by a superposition of 2-D projections of all k of its features. Thus, objects are represented in this case by collections of their snapshots (this resembles Russell's definition of a visual object as the collection of all of its views [20]). The price for giving up the assumption of known correspondence under this formulation is in the increased likelihood of interference among the different features. As we shall see, this is reflected in an increase in the number of examples required for learning recognition without correspondence.

B. Learnability Analysis

A common feature of these two formulations is that they reduce the problem of learning recognition to the problem of approximating a multivariate function from a set of examples (cf. [21]). One way to analyze the learnability of recognition is to draw conclusions about the requisite number of examples from the knowledge of smoothness properties of the function to be approximated [22], [23]. However, approximation theory is, in a sense, too strict for the present specific purpose. A function $\hat{f}$ for which $sign(\hat{f} - 0.5) = sign(f - 0.5)$ everywhere would be a perfect recognizer of the object represented by the set $\{\boldsymbol{x} \mid f(\boldsymbol{x}) > 0.5\}$ but could still be considered a bad approximation for f under reasonable definitions of goodness of fit.

Combinatorial geometry seems better suited for the analysis of learnability of recognition than approximation theory. Fortunately, the necessary tools for such an analysis are already available and are exploited below. According to the definition given in Section I-A, a learning problem is solvable if the solution can be approximated from a finite number of examples. A bound on the number of examples can be obtained using the following theorem, which is due to Haussler:

Theorem *(page 26 of Haussler [2]):* Let $\mathcal{F}$ be a family of functions from X into a metric space (Y, d_Y) of diameter M such that $\boldsymbol{F} = \{\boldsymbol{L}_f : f \in \mathcal{F}\}$, where $\boldsymbol{L}_f(x, y) = d_Y(f(x), y)$ is permissible.[1] Let D be a probability measure on $S = X \times Y$. Assume $m \geq 1$, $\nu > 0$, and $0 < \alpha < 1$. Let $\bar{\xi}$ be generated by m independent draws from S according to an arbitrary but fixed distribution D. Then

$$\begin{aligned} &\boldsymbol{PR}\{\exists f \in \mathcal{F} : d_\nu(\hat{\boldsymbol{er}}_{\xi}(f), \boldsymbol{er}_D(f)) > \alpha\} \\ &\leq 4\mathcal{C}(\alpha\nu/8, \mathcal{F})e^{-\alpha^2 \nu m/8M} \end{aligned} \tag{2}$$

where $\boldsymbol{er}_D(f)$ is the true expectation of f and $\hat{er}_{\xi}(f)$—its estimate is based on the sample $\bar{\xi}$. This theorem gives a bound on the probability of having an indicator function $f \in \mathcal{F}$ for which the estimated expected loss would differ too much from its true value in terms of the *capacity* $\mathcal{C}$ of the class $\mathcal{F}$ (the capacity of a set is defined here, after [2], as the supremum on the size of its smallest ϵ-cover taken over all possible probability measures). Thus, if one can devise an algorithm for learning $f \in \mathcal{F}$ from examples in such a way that its error on the example set is small, the above bound could be used to compute the number of examples necessary to assure acceptable error rate throughout the input space. A simple greedy algorithm satisfying this requirement is described in Section III.

Now, consider the problem of learning a canonical $3k$-dimensional state of an object from a sequence of views. Assuming correspondence, each of the views is a $2k$-dimensional vector composed of k coordinate pairs: one per feature. The action of the transformation group on the canonical state $\boldsymbol{x}^0 \in \mathcal{R}^{3k}$, followed by orthographic projection, can be represented as the multiplication of $\boldsymbol{x}^0$ by a $2k \times 3k$ matrix $T = diag(L, \ldots, L)$, where L is a 2×3 matrix (cf., the appendix in [12]). Each of the components of a view is therefore of the form $x_i = \sum_{j=1}^{3k} t_{ij} x_j$ (where all but three of the t_{ij} for each given i vanish) and can be considered an element of a 6-D vector space of functions from $\mathcal{R}^{3k}$ to $\mathcal{R}$. Consequently, the combinatorial dimension of the class of functions $x_i(\boldsymbol{x}^0)$ for all possible $\boldsymbol{x}^0$ is equal to 6 (e.g., page 19 of [2]). To obtain the combinatorial dimension of the class of entire views, note that each view is a k-fold *free product* of functions.[2] The required dimension is therefore $6k$, which gives the class $\mathcal{X}$ of views of the capacity $\mathcal{C}(\epsilon, \mathcal{X}) = 2C^{6k}$, where $C = (\frac{2eM}{\epsilon} \ln \frac{2eM}{\epsilon})$, and M is a uniform upper bound on $\mathcal{X}$ (page 24ff [2]).

Note that to learn $\boldsymbol{x}^0$ in this formulation, one must be given examples of the form $(T, \boldsymbol{x})$, that is, the transformation must be known. This knowledge can come from standard visual motion algorithms, which work well when objects are defined by a sufficient number of discrete features, with frame-to-frame correspondence, as in the present case. In fact, given enough corresponding features in several frames, both the motion and the structure of an object can be recovered algorithmically (see, e.g., [24]). Nevertheless, it is interesting that a recognition algorithm for a given object can be learned from examples, provided that a general-purpose motion algorithm is available.

As an example of the no-correspondence case, consider the problem of learning to classify vectors $\boldsymbol{v} \in \mathcal{R}^n$ obtained by sampling the intensity of the projections of a k-feature object onto a 2-D "retina" at n fixed locations. The combinatorial dimension of each v_i, which is a superposition of the projections of k features, is k. Invoking again the free product theorem from [2], we obtain the following capacity for the no-correspondence representation: $\mathcal{C}(\epsilon, \mathcal{V}) = 2C^{nk}$, where $C = (\frac{2\epsilon M}{\epsilon} \ln \frac{2\epsilon M}{\epsilon})$, and M is a bound on $\mathcal{V}$.

Thus, under both formulations, recognition of 3-D objects is learnable; the number of examples necessary to ensure that with probability greater than $1 - \delta$, the true error rate for a k-feature object will lie within α (using the d_ν metric) of the error rate on the training set and is linear in k:

$$m = O\left(\frac{1}{\alpha^2 \nu}\left(\log \frac{1}{\delta} + K \log \frac{1}{\alpha\nu}\right)\right) \tag{3}$$

where $K = 6k$ in the correspondence case and $K = nk$ ($n \gg 6$) in the no-correspondence case. The value of n depends on details such as the average spacing of object features and the point spread function of the stages prior to recognition.

III. Recognition of 3-D Objects: A Practical Approach

The above results indicate that recognition is learnable. The author shall now outline a simple greedy algorithm that learns to represent an object from positive examples using the no-correspondence approach. A view of the object is represented as a 2-D distribution of "retinal" activity, which is caused by the simultaneous presence of k features at certain locations in the image. Thus, a view can be considered a conjunction of localized feature (CLF) occurrences (see [6]). An object is represented by a collection of just enough views to ensure that any new view obtained by a rigid rotation of the same object will fall close enough to one of the stored views (cf., page 281 of [25]). In a sense, then, an object is defined as a disjunction of conjunctions of feature occurrences [5].

The algorithm is incremental in that it accepts the examples one by one and is related to the well-known statistical technique of learning vector quantization. It starts with an empty set V of views and then iterates over views in the sample set S, retaining any view that is sufficiently distant (under a simple 2-D correlation metric) from each of the currently stored views. Details of the implementation of this algorithm are irrelevant to the present discussion and are omitted here. They can be found in [6].

[1] Permissibility is a measurability condition for uncountable classes of measurable functions. In the present case, $\boldsymbol{F}$ is permissible (see page 196 of [14]).

[2] A free product of k functions $f_i(x)$ is the form $(f_1(x), f_2(x), \ldots, f_k(x))$ [2].

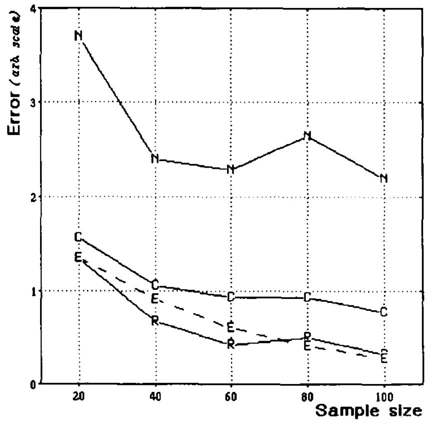

Fig. 1. Performance (measured by the ratio of the maximum distance between desired and actual output on the training set to the minimum distance on the test set; see [12]) of three recognition algorithms related to CLF versus the number m of training views (sample size). The three solid curves are data for a no-correspondence algorithm, which is a simplified version of CLF (N), a nearest-neighbor classifier with correspondence (C), and a radial basis function algorithm (R; see [5]). For comparison, a plot of the right-hand side of (2) for a particular combination of parameters ($2e^{-0.02m}$) is also shown (E, dashed curve). Note that the average performance of the no-correspondence algorithm is considerably worse than that of the other algorithms for the given range of sample sizes although its *dependence* on the sample size is the same, as expected from the results of Section II-B.

```
begin LEARN-CLF;
  1) $V \leftarrow \emptyset$;
  2) $v \leftarrow v_m \mid v_m \in S$; $S \leftarrow S - \{v_m\}$;
  3) if $min_{v_i \in V}\{d(v, v_i)\}$ is large enough then $V \leftarrow V \cup \{v\}$;
  4) if $S = \emptyset$, then return $V$; else go to 2.
end
```

The CLF algorithm satisfies trivially the requirements posed by the PAC learning paradigm; since its error rate on the training set of views can be made arbitrarily low, a similar success on a random collection of views can be assured, provided that the training set is large enough. An analysis of the structure of the problem given in the previous section indicates that the required size of the training set is feasible.

An empirical evaluation of different algorithms that learn to recognize 3-D objects from examples supports the theoretical conclusions. Results of this evaluation (but not the details of the algorithms, which are outside the scope of this correspondence) are reported below. Fig. 1 shows the dependency of an arbitrary performance measure related to error rate on sample size, for three algorithms, all of which learn by collecting and retaining specific views of objects (most of the results are from [12]; the objects are defined as collections of points in 3-D).[3] The differences between these algorithms lie in their use of the stored views.

The first of the three algorithms uses each stored view as a 2-D snapshot in which feature correspondence information is not explicitly available. The performance measure plotted for this algorithm is inversely related to the average correlation between test views and the stored representation [6]. The second algorithm is a simple nearest-neighbor classifier that represents each stored view as a $2k$-dimensional vector and, consequently, uses correspondence [12]. Finally, the third algorithm is based on view interpolation by radial basis functions (RBF's; see [5], [12]). Given a set V of views of a target object, which is obtained by the greedy procedure described above, the RBF algorithm learns to recognize that object by computing a vector of coefficients $\boldsymbol{c}$ that minimizes

$$\sum_{i=1}^{N}\left(1-\sum_{j=1}^{|V|} c_j G\left(\|\boldsymbol{v}_i-\boldsymbol{v}_j\|^2\right)\right)^2 \tag{4}$$

where N is the total number of training views $\boldsymbol{v}_i$ (which may be larger than $|V|$), $\boldsymbol{v}_j \in V$, and $G(\cdot)$ is the gaussian function. Minimizing the above expression is equivalent to finding a smooth multidimensional spline surface over $\mathcal{R}^{2k}$ that is close to 1 at the training views and falls off to 0 elsewhere. The minimization can be performed by computing the pseudoinverse of a $|V| \times N$ matrix [21]. The coefficients c_j found in this manner are used to decide whether a new view $\boldsymbol{v}$ belongs to the target object by comparing the value of the expression $\sum_j c_j G\left(\|\boldsymbol{v}-\boldsymbol{v}_j\|^2\right)$ to a threshold situated between 0 and 1.

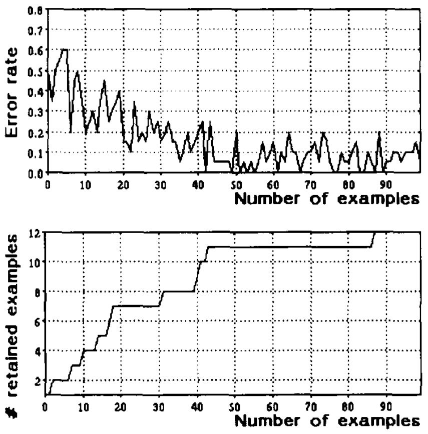

Fig. 2. Performance of a recognition module that combines Algorithm LEARN-CLF with RBF interpolation to acquire and use a multiple-view representation of an object defined by a set of six points in 3-D. *Top:* The dependence of the error rate on the number of examples presented to the module. The error rate shown in the plot is the average of miss rate (the proportion of rejected views of the target object) and false alarm rate (the proportion of accepted views of nine other objects of the same class). *Bottom:* The number of views retained by the learning algorithm asymptotes at about 12 for 100 random training views.

Fig. 2 describes an application of this algorithm to the recognition of an object defined by six randomly placed points in 3-D. The upper part of the figure shows the progress of the error rate of an RBF recognizer, which is presented with a sequence of 100 random views of the object. As shown in the lower part of the figure, 12 of these views are retained by Algorithm LEARN-CLF as a representation of the object. When the same learning algorithm was applied to objects consisting of four, five, and six 3-D points, respectively, the number of examples it took to achieve an error rate better than 25% was 21, 38, and 54 (see Fig. 3). The linear dependence of the number of examples on the complexity of the problem measured by the number of points or features defining an object should be compared with the prediction of (3).

IV. Discussion

The present note has been motivated by the recently published negative results regarding the learnability of visual recognition [1],

[3] For the restricted case of orthographic projection and rigid transformation, there exists an algorithm that can learn to recognize an object of this type from just six views obtained by Gram-Schmidt orthonormalization from a few tens of random views [11].

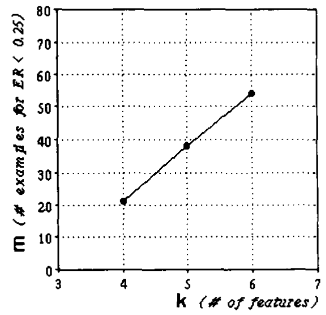

Fig. 3. Linear dependence of the number m of training examples required by the recognition algorithm on the number k of constituent points ("features") of the object is compatible with the results of Section II-B. The number m was defined as the number of the first trial for which the moving average error remained below 25% for five consecutive trials.

which appeared to contradict both intuition and facts (namely, the existence of successful learning algorithms for recognition [6], [5]). The source of the contradiction was traced back to an unrealistic definition of visual complexity in terms of image resolution. An alternative definition in terms of the number k of fuzzily positioned features was shown to lead to more plausible learnability results. Specifically, the number of examples necessary to achieve PAC learning of 3-D object recognition was found to be $O(k)$, with different coefficients of proportionality for the correspondence and no-correspondence cases. As pointed out in the introduction, the formulation of the problem of learnability is constrained by conflicting requirements of description and generalization adequacy. The formulation proposed in this note achieves a measure of compromise between these two requirements. A practical approach to the recognition of a class of 3-D objects represented by sets of spatially localized features has been outlined. At the same time, it has been shown that recognition of these objects can be learned from a feasible number of examples.

One way to draw conclusions on the learnability of recognition would be to invoke approximation theory to show that radial basis functions are adequate for 3-D object recognition and then use Haussler's proof of the learnability of RBF's [2]. Indeed, there are empirical indications that such an approach would be successful (the CLF representation outlined in Section III is related to RBF interpolation; see also [5]). The approach taken above is different in that it attempts to characterize and analyze the generic problem of learning to recognize objects while imposing as few constraints as possible on the form of the solution. The emerging picture can be summarized by observing that the most difficult part of recognition may be accumulating the relevant feature alphabet [26]. In comparison, methods for coping with viewpoint dependency of apparent shape of objects are provably learnable from examples.

Acknowledgment

The auhtor wishes to thank O. Maler, T. Poggio, and the anonymous reviewers for their comments on a draft of this paper.

References

[1] H. Shvaytser, "Learnable and nonlearnable visual concepts," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. 12, pp. 459–466, 1990.

[2] D. Haussler, "Generalizing the PAC model for neural net and other learning applications," UCSC-CRL 89-30, Univ. of California, Santa Cruz, 1989.

[3] L. G. Valiant, "A theory of the learnable," *Commun. ACM*, vol. 27, pp. 1134–1142, 1984.

[4] H. Shvaytser, "Toward a computational theory of model based vision and perception," in *Proc. 3rd Int. Conf. Comput. Vision* (Tokyo), 1990.

[5] T. Poggio and S. Edelman, "A network that learns to recognize three-dimensional objects," *Nature*, vol. 343, pp. 263–266, 1990.

[6] S. Edelman and D. Weinshall, "A self-organizing multiple-view representation of 3D objects," *Biol. Cybern.*, vol. 64, pp. 209–219, 1991.

[7] D. Marr, *Vision.* San Francisco, CA: W. H. Freeman, 1982.

[8] K. R. Boff, L. Kaufman, and J. P. Thomas, Eds., *Handbook of Perception and Human Performance.* New York: Wiley, 1986.

[9] I. Biederman and G. Ju, "Surface versus edge-based determinants of visual recognition," *Cognitive Psych.*, vol. 20, pp. 38–64, 1988.

[10] W. E. L. Grimson, *From Images to Surfaces.* Cambridge, MA: MIT Press, 1981.

[11] S. Ullman and R. Basri, "Recognition by linear combinations of models," A. I. Memo 1152, Artificial Intell. Lab., Mass. Inst. Technol., 1990; *IEEE Trans. Patt. Anal. Machine Intell.*, vol. 13, pp. 992–1005, 1991.

[12] S. Edelman and T. Poggio, "Bringing the Grandmother back into the picture: a memory-based view of object recognition," A. I. Memo 1181, Artificial Intell. Lab., Mass. Inst. of Technol., 1990; *Int. J. Patt. Recog. Artif. Intell.*, vol. 6, pp. 37–61, 1992.

[13] S. Geman and C. -R. Hwang, "Nonparametric maximum likelihood estimation by the method of sieves," *Ann. Stat.*, vol. 10, pp. 400–414, 1982.

[14] D. Pollard, *Convergence of Stochastic Processes.* New York: Springer, 1984.

[15] P. C. Dodwell, "The Lie transformation group model of visual perception," *Perception Psychophys.*, vol. 34, pp. 1–16, 1983.

[16] T. Tsao and L. Kanal, "A Lie group approach to visual perception," TR 1851, Univ. of Maryland, College Park, 1987.

[17] S. Ullman, *The Interpretation of Visual Motion.* Cambridge, MA: MIT Press, 1979.

[18] D. P. Huttenlocher and S. Ullman, "Object recognition using alignment," in *Proc. 1st Int. Conf. Comput. Vision* (London), June 1987, pp. 102–111.

[19] D. G. Lowe, "Three-dimensional object recognition from single two-dimensional images,"*Artificial Intell.*, vol. 31, pp. 355–395, 1987.

[20] B. Russell, *Analysis of Mind.* London: Allen and Unwin, 1921.

[21] T. Poggio and F. Girosi, "Regularization algorithms for learning that are equivalent to multilayer networks," *Sci.*, vol. 247, pp. 978–982, 1990.

[22] F. Girosi and T. Poggio, "Networks and the best approximation property," A. I. Memo 1164, Artificial Intell. Lab., Mass. Inst. of Technol., 1990.

[23] C. J. Stone, "Optimal global rates of convergence for nonparametric regression," *Ann. Stat.*, vol. 10, pp. 1040–1053, 1982.

[24] S. Ullman, "Computational studies in the interpretation of structure and motion: Summary and extension," in *Human and Machine Vision* (J. Beck, B. Hope, and A. Rosenfeld, Eds.). New York: Academic, 1983.

[25] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth, "Learnability and the Vapnik-Chervonenkis dimension," *J. ACM*, vol. 36, pp. 929–965, 1989.

[26] S. Edelman, "Features of recognition," CS-TR 10, Weizmann Inst. of Sci., 1991.

A Constrained Approach to Multifont Chinese Character Recognition

Xiaofei Huang, Jun Gu, *Senior Member, IEEE*, and Youshou Wu

***Abstract*—Recognizing multifont, multiple-size Chinese characters was a difficult task in the area of optical character recognition (OCR). In this correspondence, we introduce the constraint graph as a general character representation framework. Each character class is described by a constraint graph model. Sampling points on a character skeleton are taken as *nodes* in the graph. Connection constraints and position constraints are taken as *arcs* in the graph. For patterns of the same character class, this model captures both the topological invariance and the geometrical invariance in a general and uniform way. Character recognition is then formulated as a constraint-based optimization problem. A cooperative relaxation matching algorithm that solves this optimization problem is developed. A practical OCR system able to recognize multifont, multiple-size Chinese characters with a satisfactory performance was implemented.**

***Index Terms*—Constraint graph, correspondence mapping, force-driven elastic matching, optical character recognition (OCR), relaxational optimization.**

I. Introduction

Due to a very large number of character classes and higher complexity, the recognition of Chinese characters has been a very difficult task.

Traditional work in the area of character recognition mainly falls into two categories [8]: a statistical-decision approach [1], [6], [10], [13] and a structural approach [5], [11], [3], [2], [18]. A recent survey of the latest work in OCR research and development can be found in [15].

The constrained approach described in this correspondence is similar to the neural network approach to character recognition. Recently, there have been several other works that also apply artificial neural networks to OCR. Krzyzak and Suen [12] and Cun *et al.* [4] used the backpropagation model for the recognition of unconstrained handwritten digits. Gan and Lua [7] proposed an adaptive resonance network (ARN) for Chinese character classification. The ARN classifier divides 3755 Chinese characters into seven classes. For *Song* font and *Hei* font, they achieved a 90% classification rate. Leung [14] used graph matching by neural networks for character recognition.

In this correspondence, we present a general representation scheme called *constraint graph* that incorporates different types of knowledge of Chinese character patterns into a unified framework. For patterns of the same character class printed in different fonts and sizes, their structural and geometrical invariance are represented by the topological constraints and the geometrical constraints imposed on the primitives of the character patterns. The specific knowledge concerning character variabilities, as well as the general knowledge concerning character invariances, can be uniformly represented as constraints in the graph. Such a unified representation framework facilitates an automated learning process and an efficient character recognition process.

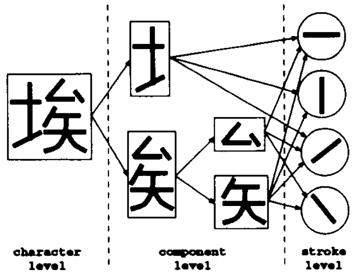

Fig. 1. Chinese character pattern can be described based on a three-level representation hierarchy.

Based on this model, character recognition is formulated as a constrained optimization problem, and a cooperative relaxation matching algorithm is developed for Chinese character recognition. We have developed a practical Chinese character recognition system on an IBM-PC with recognition rates from 95 to 99%.

The rest of the correspondence is organized as follows: In Section II, we describe a general representation framework for character recognition. We formulate the character recognition problem as a constraint-based optimization problem in Section III. In Section IV, a cooperative relaxation algorithm that finds a correspondence mapping is given. Section V provides experimental results from our character recognition system. Section VI concludes this correspodence.

II. A General Representation Framework Based on a Constraint Graph

In this section, we first describe the basic structural characteristics of Chinese character patterns. We introduce a constraint graph [9] as a general representation framework to encapsulate different types of knowledge about character patterns.

A. Structural Characteristics

A character defines a character *class* for recognition. Due to a variety of fonts, sizes, and positions, even for the same character class, there may be a great diversity of appearances and structures. Furthermore, noise and input distortions may lead to many variations of a character image. Characterizing these variations is the first critical step to character recognition.

There are several basic formation rules for Chinese characters. A Chinese character is formed by a number of components of simple structures. Some of the components are formed by basic components. Eventually, those components are disassembled into a number of strokes. Hence, a Chinese character can be described in a three-level representation hierarchy, that is, the character level, the component level, and the stroke level. An example of this representation is given in Fig. 1.

There are approximately 400 basic components and some 20 essential strokes. These numbers are much less than the total number of Chinese characters, which is over 50 000. This three-level hierarchical representation simplifies the task of character representation. It reflects the inherent characteristics that people use in the writing or in the recognition of Chinese characters.

Manuscript received October 31, 1989; revised October 4, 1992. Recommended for acceptance by Editor-in-Chief A. K. Jain.

X. Huang and J. Gu are with the Department of Electrical and Computer Engineering, University of Calgary, Calgary, Canada T2N 1N4.

Y. Wu is with the Department of Electrical Engineering, Tsinghua University, Beijing, China.

IEEE Log Number 9208436.

0162–8828/93$03.00 © 1993 IEEE