Classification with Nonmetric Distances: Image Retrieval and Class Representation

David W. Jacobs, *Member*, *IEEE Computer Society*,
Daphna Weinshall, *Member*, *IEEE Computer Society*, and Yoram Gdalyahu

Abstract—One of the key problems in appearance-based vision is understanding how to use a set of labeled images to classify new images. Classification systems that can model human performance, or that use robust image matching methods, often make use of similarity judgments that are nonmetric; but when the triangle inequality is not obeyed, most existing pattern recognition techniques are not applicable. We note that exemplar-based (or nearest-neighbor) methods can be applied naturally when using a wide class of nonmetric similarity functions. The key issue, however, is to find methods for choosing good representatives of a class that accurately characterize it. We show that existing condensing techniques for finding class representatives are ill-suited to deal with nonmetric dataspaces. We then focus on developing techniques for solving this problem, emphasizing two points: First, we show that the distance between two images is not a good measure of how well one image can represent another in nonmetric spaces. Instead, we use the vector correlation between the distances from each image to other previously seen images. Second, we show that in nonmetric spaces, boundary points are less significant for capturing the structure of a class than they are in Euclidean spaces. We suggest that atypical points may be more important in describing classes. We demonstrate the importance of these ideas to learning that generalizes from experience by improving performance using both synthetic and real images. In addition, we suggest ways of applying parametric techniques to supervised learning problems that involve a specific nonmetric distance functions, showing in particular how to generalize the idea of linear discriminant functions in a way that may be more useful in nonmetric spaces.

Index Terms— Nonmetric, image retrieval, classification, supervised learning, median, condensing, nearest-neighbor, triangle inequality, robust distance, representation.

1 Introduction

Two fundamental issues in pattern recognition and cognitive science include the problems of *clustering*—the partitioning of data into parts or classes, and *classification*—the representation of classes and the association of a new data item (query) with a certain class. In this paper, we are interested in the *classification* of images, including class representation and image retrieval.

Approaches to classification can be characterized by the type of data used and its representation. Two cases are typically considered: either the data items are mapped to some real normed vector space (called *feature space*), or they are mapped to the nodes of a weighted graph, with edge weights representing similarity or dissimilarity relations (henceforth, we will call the dissimilarity value between 2 images "distance"). The second form, called "pairwise representation," lacks geometrical notions such as interpoint Euclidean distance and its size is $O(N^2)$ for N datapoints. However, it has the advantage that no feature selection is required. Moreover, pairwise relations may violate metric properties such as the triangular inequality, a situation that cannot be modeled when data is embedded in a normed vector space. The same is true with respect to symmetry violation, which can be represented by a directed graph.

We assume here the second type of data representation as a weighted graph of image dissimilarities, since feature selection is often an elusive task (especially as it concerns images). In principle, it may be possible to embed this representation in a vector space, obtaining the first type of data representation, as assumed in most of the work on image classification. If the distances are metric, such embedding is feasible as discussed below. However, most recent work in computer vision compares images using measures of similarity that are complex and nonmetric, in that they do not obey the triangle inequality ([1], [7], [9], [11], [22], [23], [25], [32], [33], [35]). This can occur because the triangle inequality is difficult to enforce in complex matching algorithms that are statistically robust (see discussion in Section 2.1). Also, when matching is conceptualized as the comparison between two probability distributions, there may be strong reasons for using distances such as the Kullback-Leibler measure of cross-entropy, which is asymmetric and does not obey the triangle inequality (eg., see [28]). Moreover, much

- *This paper is based on previously published research, please see the Acknowledgement for details.*
- *D.W. Jacobs is with the NEC Research Institute, 4 Independence Way, Princeton, NJ 08540. E-mail: dwj@research.nj.nec.com.*
- *D. Weinshall is with the Institute of Computer Science, The Hebrew University, 91904 Jerusalem, Israel. E-mail: daphna@cs.huji.ac.il.*
- *Y. Gdalyahu is with IBM Research, Haifa Laboratory, MATAM Haifa, Israel. E-mail: yoramg@il.ibm.com.*

Manuscript received 4 Feb. 1999; revised 29 Dec. 1999; accepted 17 Feb. 2000. Recommended for acceptance by K. Bowyer.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 109058.

research in psychology suggests that human similarity judgments are also not metric. This raises serious questions about the extent to which existing work on classification can be applied using complex models of similarity, either in computer vision or as cognitive models.[1]

1.1 Vectorial Embedding

Most work on supervised classification assumes vectorial representation, typically with the ℓ_2 (Euclidean) norm. For example, in linear discriminant analysis and in many multilayer neural network models, hyperplanes are found that separate labeled points as well as possible. Similarly, the mean point of a set is used in methods such as k-means clustering, and minimal linear spanning bases are used in methods such as Radial Basis Functions [34] or Principal Component Analysis (PCA). For some applications, a vectorial representation can describe the data in a very natural way.

In many other applications, however, there is no obvious vectorial representation, although algorithms do exist to compare different objects. To use the techniques mentioned above, in this case, *metric* dissimilarity relations d_{ij} between datapoints may be transformed to vectorial representation. The embedding problem is to map each node (datapoint) i to a vector v_i in some real normed space, such that $\|v_i - v_j\| = d_{ij}$. It can be shown that the mapping that assigns $v_i \leftarrow [d_{i1}, \ldots, d_{in}]$ embeds n points in $\mathcal{R}^n$ with the max norm ℓ_∞. Recently, for example, [15] have applied support vector machines to classification after embedding datapoints in a Euclidean, or pseudo-Euclidean space. Alternately, they have shown improvements over nearest-neighbor methods by representing each object by its distance to all others, and then applying support vector machines to the resulting vectors. However, the runtime of this latter approach is at least as great as a brute force approach to nearest neighbor, since the distance from a test object to all training objects must be computed.

However, n is usually very large and dimensionality reduction is often required, where possible methods are PCA, random projection [26], principal curves [19], and others, thus introducing distortion into the low-dimensional representation. A low-dimensional graph embedding with controlled distortion is proposed in [30], using the transformation described above for metric dissimilarity relations. An alternative method is Multidimensional Scaling (MDS) [27], where the embedding may preserve the ranking of the pairwise distances, but not necessarily their ratios. We note that dimensionality reduction is another name for the problem of feature selection, namely it involves the assumption that a small number of features can be found that describe the datapoints with sufficient accuracy.

1. We should note that, in contrast to supervised learning, there has been a good deal of work on clustering, or *unsupervised* learning that is applicable to nonmetric spaces. An early example of such work, which stresses the importance of nonmetric distances, is [33]. Brand [5] has proposed clustering based on E-M in nonmetric spaces. In addition, clustering methods based on graph partition or physical models (e.g., [4], [21]) are suitable for nonmetric clustering. These works do not directly address the issue of using these clusters for classification of new data (see [33] for some comments).

Our work is motivated by work on image similarity that suggests that practical and psychologically valid measures of similarity are nonmetric. Note that some authors use the phrase *nonmetric* to imply a qualitative, or rank ordering of distances. We use *nonmetric* in the standard mathematical sense. A distance function, $D(i_1, i_2)$, between pairs of images, is metric when:

1. $D(i_1, i_2) \geq 0$
2. $D(i_1, i_2) = 0$ if and only if $i_1 = i_2$.
3. $D(i_1, i_2) = D(i_2, i_1)$ (symmetry).
4. $D(i_1, i_3) \leq D(i_1, i_2) + D(i_2, i_3)$ (the triangle inequality).

We are interested in spaces in which the last condition fails. Failure of symmetry is also of interest, but this is beyond the scope of the present paper. We will also consider some robust distances in which Condition 2 does not hold because some limited deviation between objects may be ignored.

Nonmetric distances turn up in many application domains, such as string (DNA) matching, collaborative filtering (where customers are matched with stored "prototypical" customers), and retrieval from image databases. Fig. 11 shows one example, the output of an algorithm for judging the similarity of the silhouettes of different objects [11]. Given a series of labeled pictures of common objects (cars and cows, Fig. 11a and Fig. 11b, we may wish to identify new silhouettes (Fig. 11d) based on their similarity to the previously seen ones. In [1], many such algorithms are reviewed, showing why their use will typically lead to nonmetric distances (see discussion in Section 2.1).

Unfortunately, the methods described above cannot guarantee low distortion vectorial embedding of nonmetric dissimilarity relations. We will give a concrete example below. Consequently the vectorial methods described above may not be suitable for many applications. We therefore focus our attention on classification methods that use pairwise representation.

One other basic tenet of pattern recognition is that all the information is in the distribution of the data. One should note, however, that most methods to estimate such distributions based on a small sample, including nonparametric methods such as Parzen windows and interpolation methods such as Radial Basis Functions (RBF), implicitly assume Euclidean structure on the data by the very essence of interpolation. Once again, this assumption makes these algorithms unsuitable for our domain; instead, we directly estimate the distribution of the data in term of its pairwise dissimilarity representation.

1.2 Dissimilarity-Based Methods

In classification methods that use pairwise representation, retrieval is based on the distance from query to class, typically measured by the k-nearest neighbor of a query. Brute force implementation of nearest neighbors is of high complexity in both space and time—linear in the size of the data where sublinearity is desired with large databases. In particular, time complexity may be high because many similarity functions used in computer vision are complex, and require considerable computation. Many techniques therefore have been devised to speed up nearest-neighbor

Fig. 1. The Voronoi diagram for two points using, from left to right, p-distances with $p = 2$ (Euclidean distance), $p = 1$ (Manhattan distance, which is still metric), the nonmetric distances arising from $p = .5$, $p = .2$, and the min (1-median) distance. The p-distance between two points (x_1, y_1) and (x_2, y_2) is: $(|x_1 - x_2|^p + |y_1 - y_2|^p)^{\frac{1}{p}}$; the min distance is $min(|x_1 - x_2|, |y_1 - y_2|)$. Min distance in 2D is illustrative of the behavior of the other median distances in higher dimensions. The region of the plane closer to one point is shown in dark gray, and closer to the other in light gray. Note that only the first is invariant to our choice of axis direction.

classification. Some of these methods build tree structures that rely on the points lying in a Euclidean [10] or at least in a metric space [43]. So once again, these methods are not appropriate for our domain.

A more heuristic approach to decrease the time and space complexity of nearest-neighbor classification—the use of *condensing* algorithms—has been developed in the context of metric spaces but is potentially applicable in nonmetric spaces as well. These algorithms select subsets of the original training classes that are as small as possible, and that will correctly classify all the remainder of the training set, which is then discarded. Condensing algorithms do not explicitly rely on distances that obey the triangle inequality. However, we will show below that the performance of existing condensing methods can be severely degraded with nonmetric distances. In particular, the selection of representative data items in the condensing process should take into account the nature of the dissimilarity between items; for example, you may only need to compare a query with boundary points in the Euclidean space, but the nature of boundaries changes tremendously with the change of distance function. This problem is illustrated in Fig. 1, showing the different decision boundaries that arise when using three nonmetric distance functions, in comparison with the metric Manhattan and Euclidean distances.

1.3 Condensing Methods—Previous Work

Condensing algorithms select small subsets of the original data such that they will correctly classify all the remainder of the data using nearest-neighbor classification; such methods are essential to decrease the complexity of nearest-neighbor classification and make it practical. Many condensing methods have been developed: Hart [18] proposed an iterative algorithm that initializes a representative set and then continues to add elements from the training set that cannot be correctly classified by the representative set, until all remaining training elements can be correctly classified. Gowda and Krishna [14] propose an algorithm like Hart's, but which explicitly attempts to add first to the representative set the points nearest the boundary. Fukunaga and Mantock [16] propose an iterative algorithm, that attempts to optimize a measure of how well distances to the representative set approximate distances to the full training set.

Dasarathy [8] more explicitly searches for the smallest representative set that will correctly classify the training set. We will discuss this algorithm in more detail and show experiments with it in Section 4. Dasarathy notes that once an element is added to the representative set, it is guaranteed that other elements of the same class will be correctly classified if they are closer to this representative element than to any element of any other class. A greedy algorithm is used, in which the representative set is augmented at each step by the training element that guarantees correct classification of the largest number of elements not yet guaranteed to be correctly classified. This basic step is followed by subsequent steps that refine the representative set, but in practice these are not found to be important.

1.4 Our Contribution

In this paper, we make two contributions: First, we discuss the use of nonmetric distances in applications and show that existing classification methods can indeed encounter considerable difficulties when applied to nonmetric similarity functions. Second, we propose a new condensing method, demonstrate analytically some of the conditions under which this strategy is preferable, and show experimentally that it is effective.

The rest of this paper is organized as follows: In Section 2, we discuss why nonmetric dissimilarity-based representations of data are commonly used (and needed) in applications. In Section 3, we describe our approach to the organization (or condensing) of a nonmetric image database. In Section 4, we describe a concrete condensing algorithm and compare it to other algorithms using both simulations and real data. Finally, in Section 5, we discuss the possibility of relaxing the assumptions in this paper and building parametric supervised learning algorithms suited to specific nonmetric distances.

2 Nonmetric Distances—Why and When

In the introduction, we described the difficulties with using existing classification techniques when given nonmetric dissimilarity-based representations of data. In this section, we show that such representations are common in applications and discuss some of the reasons why they are needed.

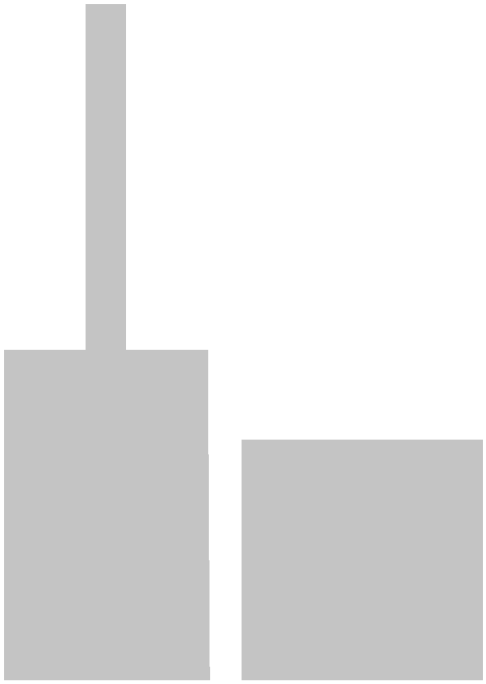

Fig. 2. Object 1 (left) and object 2 (right) used in our discussion of Hausdorff matching.

2.1 Nonmetric Distances in Applications

Distance functions that are robust to outliers or to extremely noisy data will typically violate the triangle inequality. One group of such functions is the family of image comparison methods that match subsets of the images and ignore the most dissimilar parts (see [1], [11], [2], [32]). As one example, Huttenlocher et al. [22], [23] perform recognition and motion tracking by comparing point sets using the Hausdorff distance. They consider only a fixed fraction of the points for which this distance is minimized. By not considering the most dissimilar parts of the images, these methods become both robust to image points that are outliers, and nonmetric. We call these nonmetric methods median distances. A k-median distance between two vectors (or images, represented as vectors) $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ is defined as:

$$d(\mathbf{x}, \mathbf{y}) = k - median\{|x_1 - y_1|, \ldots, |x_n - y_n|\},$$

where the k-median operator returns the kth value of the ordered difference vector. Related robust techniques, such as M-estimation which identify outliers and weigh them less heavily, also lead to nonmetric distances [17].

We will now describe an example of matching with a robust Hausdorff distance. This illustrates robustness considerations that are very common in machine vision and will also allow us to be more concrete about the difficulties raised by nonmetric distances. Fig. 2 shows two shapes. We will compare these shapes to each other and to other shapes by comparing points sampled along their boundaries. Specifically, the two shapes are compared by translating one with respect to the other and choosing the translation that minimizes the k-median Hausdorff distance between the two. Hausdorff distance is the maximum distance between any point in one shape and the point that is closest to it in the other. That is, for point sets $\mathcal{I}, \mathcal{J}$ it is:

$$\max(\max_{i \in \mathcal{I}} \min_{j \in \mathcal{J}} ||i - j||, \max_{j \in \mathcal{J}} \min_{i \in \mathcal{I}} ||i - j||).$$

By k-median Hausdorff distance we mean (following Huttenlocher et al. [22]) that instead of considering the maximum distance over all points in a set to their nearest neighbor, we consider the distance at some percentile. In the following example, we consider 60 percent of the points in each set that are closest to the other set.

In Fig. 3, we show a picture of one of these objects partially occluded. It is such cases that motivate the use of robust matching techniques—when we can extract points along part of the boundary of the object, but should only expect a partial match between the image and the objects in the data base. In more complex problems, our database itself may consist of images of noisy or partially occluded objects, rather than idealized perfect appearance-based representations.

In Fig. 3, we also show the translation that minimizes the robust Hausdorff distance between each object and the partially occluded object, and also between the two objects. In this example, the distance from the occluded version of object 1 and the unoccluded version is zero, because 60 percent of the object was visible. The distance from object 1 to object 2 is 2 pixels, and the distance from the occluded version of object 1 to object 2 is $4\frac{1}{4}$. These distances clearly violate the triangle inequality.

This violation of the triangle inequality is not an artifact of a poor choice of features. It is inherent in the idea of robust matching, which allows one portion of an object to be matched to one image, and a different portion to match a different image. These objects cannot be mapped into a metric feature space without large distortions in the distances between them. Moreover, we know of no work that suggests how one can extract features from these shapes that enable robust matching using the Euclidean norm in a vector space.

We can contrast this with statistical methods for handling missing data, which can be robust without introducing nonmetric distances. These rely on identifying outliers explicitly using domain specific knowledge or parametric models. For example, the values of some variables obtained in a survey may be identified by hand as clearly incorrect; or values that deviate too far from the mean value may be discarded. In vision, missing data can arise in problems such as structure-from-motion when some point features disappear from view during the course of a motion sequence, or some values clearly disagree with a linear motion model. When some values are known to be outliers, they can then be replaced by other reasonable values. In statistics, sometimes missing values are replaced by the mean value; or linear regression may allow one to infer the value of missing data. [31] provides an overview of these methods, while [39] and [24] describe solutions to missing data problems in vision. However, in the example above, the shape of the points that are missing due to occlusion could only be inferred after the object has been recognized. For this reason, missing data techniques are not applicable in problems of interest to us here.

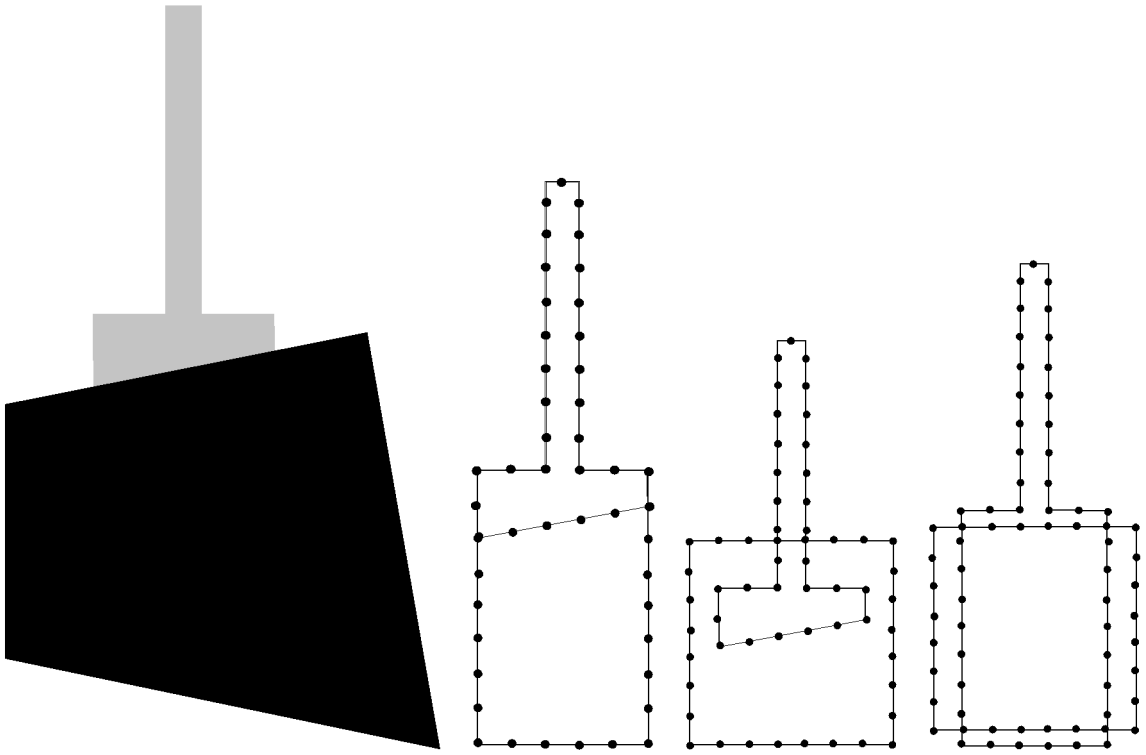

Fig. 3. On the left, object 1 is shown partially occluded. The middle two figures show alignments between points sampled along its visible boundary, and along the boundaries of objects 1 and 2. These alignments minimize the robust Hausdorff distance between the objects and the occluded version of object 1. On the right, we show the best alignment between object 1 and 2 (a range of alignments produce the same Hausdorff distance between these shapes).

As another example, the use of nonmetric ℓ_p distances, with $0 < p < 1$ has been suggested for robust image matching [9]. ℓ_p distance (or *p-distance*) between two vectors $\mathbf{x}, \mathbf{y}$, is defined as:

$$d(\mathbf{x}, \mathbf{y}) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}.$$

Thus, $p = 2$ gives the Euclidean distance, and $p = 1$ gives the Manhattan distance. These distances are nonmetric for $p < 1$ (see Royden [36]). As shown in Fig. 4, they are less affected by extreme differences than is Euclidean distance, and can therefore be more robust to outliers. Related robust distance functions have become widespread in many aspects of vision in recent years, especially for problems such as boundary detection, motion estimation, and parametric object detection (eg., [29], [13], [3], [2], [12], [32]). In this paper, we will focus more on the use of robust distances in supervised learning, while noting that their significance in vision seems quite broad.

One interesting property of robust distances such as the median and nonmetric p-distances is that although they apply to feature vectors, they treat the space these vectors occupy in a way that is not rotationally symmetric. The distance between two points doesn't just depend on the length of the vector connecting one to another, it also depends on how this vector is aligned relative to the coordinate axes (this is true for all p-distances except Euclidean distance). This is appropriate when different coordinates denote independent features. For example, when the pixels of an image are used as coordinates, the axes of the space denote separate pixels, while an arbitrary

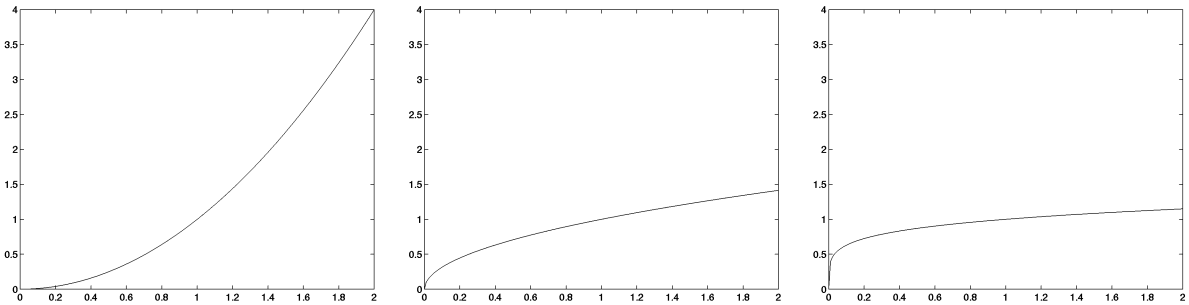

Fig. 4. Graphs of x^2 (left), $x^{0.5}$ (middle) and $x^{0.2}$ (right). We can see that when properly normalized, for p-distances less than 1 the cost function rises quickly at first, then more slowly, so that extreme values do not dominate the total cost.

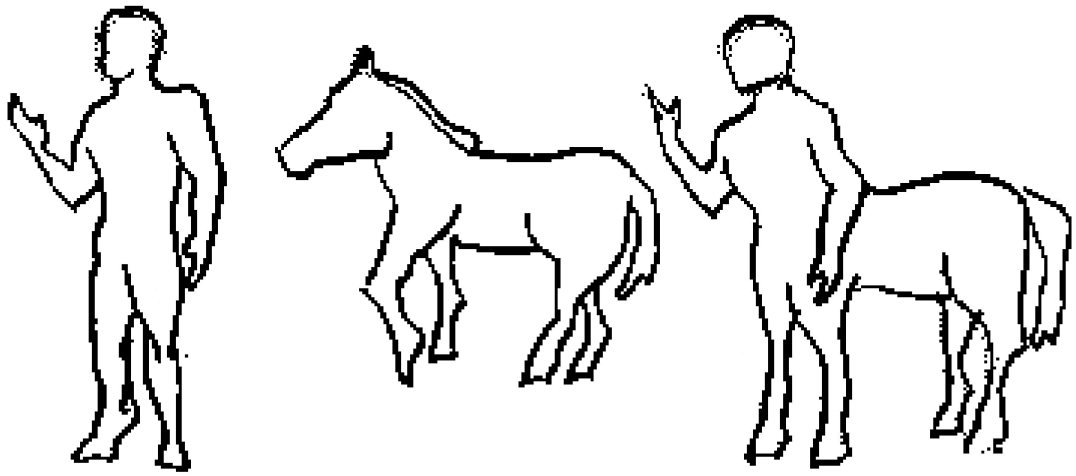

Fig. 5. Judgments of visual similarity between these three images may violate the triangle inequality.

direction represents some linear combination of all pixels. When we wish to be robust to the possibility that some pixels have spurious values, it is important to treat the directions of the axes quite differently from other directions.

Besides a desire for robustness, another reason for nonmetric distances to arise is that image distances may be the output of a complex algorithm, which has no obvious way of ensuring that the triangle inequality holds. Jain et al. [25], for example, perform character recognition and other image comparisons using a deformable template matching scheme that yields distances that are not symmetric and do not obey the triangle inequality. Related elastic matching methods have been widely used (e.g., [1], [11], [20], [40], [44], [38]) often in ways that do not appear to lead to metric distances.

In fact, Basri et al. [1] show that contradictions can exist between forcing such methods to obey the triangle inequality and other goals that are desirable in deformable template matching. More specifically, they consider the class of algorithms that match closed contours elastically, using dynamic programming to find correspondences between the contours that minimize the sum of a local cost function based on comparisons between small portions of the contours. They articulate a set of desirable properties for such a cost function and show that they cannot all simultaneously be met. Specifically, they consider three criteria: 1) that the more one deforms a contour by bending it, the more dissimilar it becomes from its original shape; 2) that a series of small deformations should affect similarity less than one large deformation equal to the sum of the small ones in magnitude; and 3) the triangle inequality. They show that these criteria cannot all be achieved with such an elastic matching approach, explaining why such approaches may adopt nonmetric distances.

Finally, we are interested in image comparison methods that model human vision. This may also be desirable in many applications. However, there is much work in psychology that suggests that human similarity judgments are nonmetric. Most notably, Tversky et al. (e.g., [41]) showed in a series of studies that human similarity judgments often violate metric axioms: in particular, the judgment need not be symmetric (one would say "North Korea is similar to Red China," but not "Red China is similar to North Korea"), and the triangle inequality rarely holds (transitivity should follow from the triangle inequality, but the similarity between Jamaica and Cuba and between Cuba and Russia does not imply similarity between Jamaica and Russia). This occurs because different features can be attended to when different comparisons are made. In some cases, this may allow objects to be represented in a small number of different spaces in each of which we may use metric distances. However, in visual comparisons this is not possible.

Fig. 5 provides a visual analog of this example that demonstrates this point. Many observers will find that the centaur is quite similar to the person and to the horse. However, the person and the horse are quite different from each other. For observers who determine that the dissimilarity between the horse and person is more than twice as great as the dissimilarity between either shape and the centaur, there is a violation of the triangle inequality. Intuitively, this occurs because when comparing two images we focus on the portions that are very similar, and are willing to pay less attention to regions of great dissimilarity. However, one cannot necessarily divide objects into a few discrete components that are individually compared using metric distances. Consequently, any computer vision system that attempts to faithfully reflect human judgments of similarity is apt to devise nonmetric image distance functions (see also [37] for discussion of this issue).

2.2 Problems that Are Inherently Nonmetric

Are there problems that are inherently nonmetric, or is nonmetricity only induced by one's choice of representation and distance function? Our answer is tentative: We consider what happens when researchers approach a problem by first designing the best possible distance function or matching algorithm. Issues related to dealing with missing data, noise, and robustness should all be taken into account. When these conditions can be modeled, one can even design an optimal distance function. For example, if one knows that point sets have been generated by adding noise to some points of a model, and adding some random outlying points, a robust, nonmetric distance will be

optimal. If the optimal distance is nonmetric, and if the data cannot be embedded in a metric space without greatly distorting this distance function, then our answer is that the problem is inherently nonmetric.

As described above, methods such as PCA and MDS can be used to embed any data in a metric space by accepting distortions of the distance function. Moreover, one can always trivially transform any set of distances using a monotonic function so that they form a metric by, for example, scaling all distances to lie within the range from one to two. While such embeddings are possible, they are not useful. For example, neither a trivial nor a very noisy embedding can be used to assist in finding nearest neighbors quickly using a tree structure. If large deviations from the triangle inequality are a necessary consequence of the image comparisons that are desired, accepting distortions of the distance function as a form of noise will lead to poor performance. Moreover, the view that violations of such metric properties as the triangle equality are "Euclidean inconsistencies," evidence for a poor design of the distance function, is not true in the domains we consider. We therefore would avoid data embedding with high distortion cost.

Under certain conditions, data can be mapped with low distortion to a metric space. For example:

- We may assume that there are "hidden variables" such that in a higher-dimensional space (with more features than the original feature space) the appropriate distance function is metric, and possibly even Euclidean. An assumption to this effect underlies many methods in pattern recognition and statistical estimation, which then proceed in the quest to find the hidden variables. The good performance of support vector machines in many applications may indicate that this assumption is often reasonable and useful.
- Since axes in the feature space are chosen somewhat arbitrarily, we may assume that there exist "objective" axes where the distance between datapoints is invariant with respect to rotation and translation of these axes. (Typically, however, features are not completely arbitrary but have a meaning inherent to the data and the measurement procedures.)

In this paper, we address problems for which such assumptions cannot be made. Since there is no inherent reason (other than computational convenience) to assume that one of these assumptions would always hold, and since nonmetric distances are so common in applications, techniques to address such problems should be of practical interest. This commits us to nearest-neighbor methods in preference of many better-behaved pattern recognition techniques. By relying directly on the distance function, we exploit whatever "Euclidean inconsistencies" there are in the data, rather than try to remove them.

3 Our Approach

Our goal is to develop condensing methods for selecting a subset of the training set, and then to classify a new image according to its nearest neighbor among this subset. We seek a subset of the training set that minimizes errors in the classification of new datapoints: a representative subset of the training data whose nearest distance to most new data items approximates well the nearest distance to all the training set. Thus, we emphasize that the representative set maintains the same generalization function as the whole dataset.

3.1 How to Determine Data Redundancy

In designing a condensing method, one needs to understand *when is one image a good substitute for another?* For example, when are two images a redundant set? Our answer to this question is what most distinguishes our approach from previous work on nearest neighbors in metric spaces; specifically, our answer depends on a relevant statistical measure which we call redundancy, and not directly on distances.

3.1.1 Redundancy

We begin by considering the answer to this question that is implicit in previous work. In what follows, let i_1 and i_2 denote two arbitrary elements of the training set, and let i denote a new image that we wish to classify. Let $d(i_1, i_2)$ denote the distance between the two elements i_1, i_2.

In a metric space, symmetry and the triangle inequality guarantee that the distance between i_1 and i_2 bounds the difference between $d(i_1, i)$ and $d(i_2, i)$, that is, $|d(i_1, i) - d(i_2, i)| \leq d(i_1, i_2)$. Thus, when $d(i_1, i_2)$ is small, we know that one of these images can substitute for the other. However, in a nonmetric space, the value of $d(i_1, i_2)$ need not provide us with any information about the relationship between $d(i_1, i)$ and $d(i_2, i)$. Our first observation is that we must use more information than just the distance between two datapoints to decide whether the presence of one in the training set makes the other superfluous.

More specifically, our desired condensing algorithm will discard an image, in favor of another one which is already stored, only when the two images are expected to have similar distances to other images, yet unseen; i.e., when they have high redundancy. We define this property—the **redundancy** relation between two images—as follows:

Definition 1. *The* **redundancy** $R(i_1, i_2)$ *between two images* i_1, i_2 *is the probability that images* i_1, i_2 *have similar distances to other images, i.e.,*

$$R(i_1, i_2) = \int_{|d(i_1,i)-d(i_2,i)|<\epsilon} f(i)di$$

for some small ϵ *and image sampling distribution* $f(i)$.

3.1.2 Atypical Points

Existing condensing methods focus on choosing representative points from the boundaries between classes. Boundary points are especially important to retain as representatives in metric spaces because they influence the decision boundary between two classes in the region that is near the classes. Our second main intuition is that it is important to retain atypical examples in the training set rather than just boundary points. An "atypical" image is any image that is dissimilar from most other images in the class, and especially from the already chosen representatives of the class. Atypical points can be important to retain as

Fig. 6. The Voronoi diagram for two sets of points, each containing five points (shown as squares and crosses). The distance functions used are, from left to right, p-distances with $p = 2$ (Euclidean distance), $p = 1$ (Manhattan distance), $p = 0.5$, $p = 0.2$, and median distance. The region of points closer to the class of crosses is shown in black, and the region closer to the class of squares is shown in white.

representatives because points that are far from the other representatives can have a big effect on the decision boundary between the two classes. This can also be true in metric spaces, but it is especially true in nonmetric spaces because there can be points in nonmetric spaces that are close to elements in each class even though these elements are not boundary points. We give an example of this below.

3.1.3 Example: A 2D Domain

Let us illustrate these points in a simple 2D domain. Thus, in this section, all "images" are 2D vectors. In this domain, we will use the nonmetric 1-median distance, i.e., the min distance. Min distance may not be a good robust estimator; our goal is only to demonstrate ideas that apply to other median distances in higher dimensions. The relevant geometrical structure is the Voronoi diagram: a division of the plane into regions in which the points all have the same nearest neighbor.

Our first example is Fig. 1, showing that nonmetric distances (Fig. 1, three right-most pictures) can produce much more complex Voronoi diagrams than do metric distances (Fig. 1, two left-most pictures). Fig. 6 further illustrates the complex structures that classes can take with nonmetric distances. These examples illustrate our second point: the difficulty in relying on decision boundaries when dealing with nonmetric data.

Next, Fig. 7 shows a simple example illustrating our first point: the potential value in looking at pairwise redundancy in determining image interchangeability as representatives of a class. In the top right, we show the Voronoi diagram, using median distance for two clusters of points (i.e., each point in the plane is labeled according to the cluster to which its nearest neighbor belongs). One cluster, P, consists of four points (labeled p_1, p_2, p_3, p_4 in the upper left and shown as black squares) all close together, both according to median distance, and to Euclidean distance. The second cluster, Q, (labeled $q_1, \ldots q_5$ in the upper left and shown as black crosses), all have the same x coordinate, and so all are separated by zero distance when using the median distance; however, the points are divided into two subgroups in feature space: q_1 and q_2 on top, and q_3, q_4, q_5 on the bottom.

To characterize Q well, it is important to pick representatives from both the bottom group and the top group. To illustrate this, the bottom lefthand figure shows the Voronoi

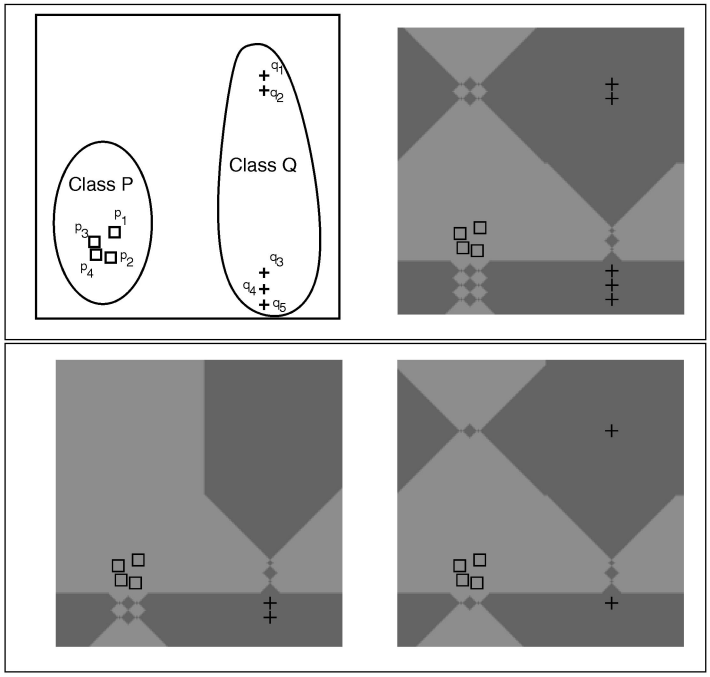

Fig. 7. Top: two clusters of labeled points (left) and their Voronoi diagram (right) using the median (min) distance. Bottom: the Voronoi diagram resulting when only q_3 and q_4 represent class Q (left) and when q_2 and q_3 are chosen as representatives (right).

diagram produced when we represent Q using q_3 and q_4, while the bottom right figure shows the Voronoi diagram resulting when we choose q_2 and q_3. Clearly, the latter choice produces a Voronoi diagram much more faithful to the true one.

Existing condensing algorithms cannot handle this situation. First, every element of Q will correctly classify all the other elements; no point will be preferable in this regard. Second, q_3, q_4, and q_5 are nearest to class P, and would be judged boundary points. So, although cluster Q is really split into two distinct groups, this is not directly reflected in their distances to each other. However, one can detect the two subgroups of cluster Q in a natural way by looking at the pairwise redundancy of elements of Q; for example, although $d(q_1, q_5) = 0$, $R(q_1, q_5)$ is small because the distances from q_1 and q_5 to the elements of P are quite different. Notice in this example that q_2 has a big effect on the Voronoi diagram even though it is not a boundary point in the sense of being close to P. Such mishaps happen more frequently in nonmetric spaces; in our example, even though the distance from q_2 to P is fairly large, there are points in the space that have zero distance to both q_2 and to elements of P.

In summary, by considering a 2D domain, we first see that robust, nonmetric distances lead to rather odd, nonintuitive decision boundaries between clusters. Second, we see that for the median distance, the extent to which two images have similar distances to other images can be a good predictor of their interchangeability as class representatives, much better than just the distance between them. We wish to emphasize that while these results are illustrated in a 2D domain, it should be clear that they extend to higher dimensions as well. In fact, in higher-dimensional spaces one can show that on average the median distance will be less correlated with Euclidean distance.

3.2 Measuring Redundancy: Correlation vs. Distance

We compare two methods of estimating redundancy: 1) the distance between the two images; and 2) the correlation between their *distance vectors*: $\{d(i_1, k) | \forall k \in T, k \neq i_1, i_2\}$ and $\{d(i_2, k) | \forall k \in T, k \neq i_1, i_2\}$, where T is the training set. Starting with (2), we first discuss in Section 3.2.1 when correlation is a good estimator of redundancy. Moving to (1), we analyze the relation between distance and redundancy for two specific examples: in Section 3.2.2, we show that in Euclidean space, pairwise image distance alone gives an excellent indication of redundancy; in Section 3.2.3, we show a robust space where more reliable results are obtained by using correlation to estimate redundancy.

3.2.1 Correlation and Redundancy

Correlation predicts redundancy well when the distances from one datapoint P to other datapoints in the training set are sampled from the same distribution as the distances from P to other datapoints in the test set. Thus, the use of correlation as a measure of redundancy does not entail any assumption about the nature of the distance function, but requires that the statistical estimation is proper in that the training set is large and representative enough. In other words, correlation is a good estimator of redundancy for distance functions whose corresponding graphs, or "pairwise representation," are smooth enough to enable estimation using the small available sample.

3.2.2 Euclidean Space: Distance Predicts Redundancy

From the definition of redundancy in Section 3.1.1, it can be readily shown that the Euclidean distance between points in the Euclidean plane is monotonically decreasing with redundancy (i.e., closer points are more redundant). Thus, in Euclidean space, distance is a good estimator of redundancy. In fact, since this estimator is independent of the sample training set, it is preferable over correlation.

More specifically, we can show the following:

Lemma 1. *For three points i_1, i_2, i_3 in Euclidean space, and assuming data is sampled from the uniform distribution over a compact subspace of Euclidean space,*

$$d(i_1, i_2) < d(i_1, i_3) \implies R(i_1, i_2) > R(i_1, i_3).$$

Proof. From the definition of redundancy in Section 3.1.1 and assuming a uniform prior, the redundancy between two points i_1, i_2—$R(i_1, i_2)$—equals the volume of space which includes all images i, where $|d(i_1, i) - d(i_2, i)| < \epsilon$. (We neglect discrepancies due to boundary effects at the boundaries of our compact subspace.) It immediately follows (in Euclidean space) that redundancy is the volume between the two surfaces of a hyperboloid sheet, defined by $d(i_1, i) - d(i_2, i) = \pm\epsilon$. All that remains to be shown is that this volume monotonically decreases with the distance between the points $d_{12} = d(i_1, i_2)$.

Fig. 8 illustrates the situation in $\mathcal{R}^2$, clearly showing that the delineating surfaces where $|d(i_1, i) - d(i_2, i)| = \epsilon$ are getting closer together—meaning decreased redundancy—as the distance between the points increases. We next outline a proof that this is always the case, which completes the proof of the lemma. We work in $\mathcal{R}^2$ for simplicity of notations, although the proof readily extends to any dimension $\mathcal{R}^d$.

We first translate and rotate the plane so that $i_1 = [0, \frac{d_{12}}{2}]$ and $i_2 = [0, -\frac{d_{12}}{2}]$. When $d_{12} > \epsilon$,[2] $|d(i_2, [x, y(x)]) - d(i_1[x, y(x)])| = \epsilon$ defines a hyperbola $y(x)$, whose two vertices are at $[0, \frac{\epsilon}{2}]$, and $[0, -\frac{\epsilon}{2}]$. Denoting the two hyperbolic curves passing respectively through the two vertices by $y_1(x)$ and $y_2(x)$, we can show (by differentiation with respect to d_{12}) that $y_1(x) - y_2(x)$ is monotonically decreasing with d_{12} at every point $x \neq 0$. At $x = 0$ the distance is ϵ, independent of d_{12}. □

In fact, we can show another interesting lemma (the proof is omitted) which sheds more light on the connection between distance and redundancy in the Euclidean space:

Lemma 2. *For three points i_1, i_2, i_3 in the Euclidean space, and a random point j sampled from a "reasonable" prior,*

$$d(i_1, i_2) < d(i_1, i_3) \implies$$
$$\Pr\left(|d(i_1, j) - d(i_2, j)| < |d(i_1, j) - d(i_3, j)|\right) > \frac{1}{2}.$$

That is, it is more likely that two points that are closer together will have similar distances to a new point than two points that are further apart.

2. When $d_{12} < \epsilon$, redundancy is 1—all points have similar distances to i_1, i_2.

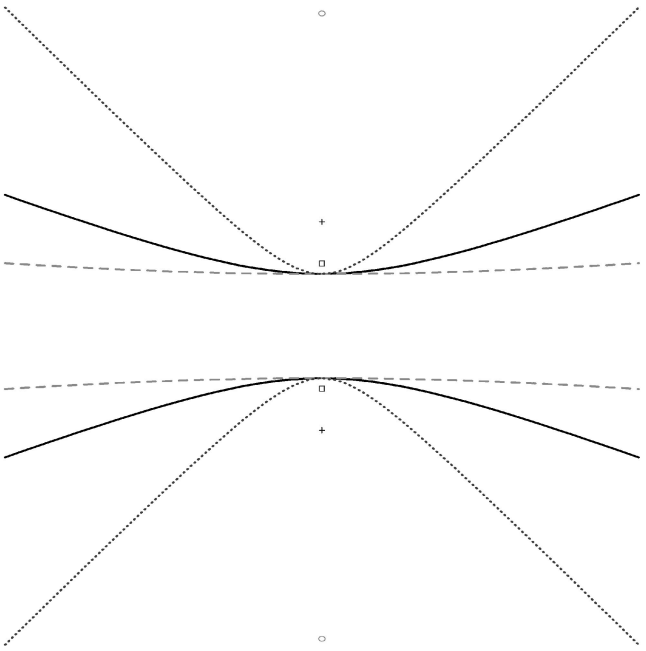

Fig. 8. The two delineating curves, where $|d(i_1, i) - d(i_2, i)| = \epsilon$, are shown for three pairs of points i_1, i_2 in $\mathcal{R}^2$ located on the vertical axis symmetrically around the origin. The most distant pair of points (in light gray) are separated from each other by 6ϵ: the points are marked with circles, and their corresponding delineating curves are hyperbolas shown as dashed lines. The second pair (in black) is separated by 2ϵ: the points are marked with crosses, and the delineating curves are shown as solid lines. The closest pair (in dark gray) is separated by only 1.2ϵ: the points are marked with boxes, and the delineating curves are shown as dotted lines.

3.2.3 Nonmetric Space: Distance and Redundancy

We have shown that in Euclidean space, when points are closer together they must have higher redundancy. This need not be true in nonmetric spaces as can be readily shown by example, since almost anything can happen in arbitrary nonmetric spaces. We will demonstrate what typically happens when using one nonmetric distance—the 2D min distance (a member of the family of median distances). We demonstrate three characteristics of the relation between distance and redundancy in this space: 1) pairs of points separated by the same distance can have very different redundancy (which is **not possible** in Euclidean space); 2) nevertheless, other factors being equal, smaller distances typically lead to equal or greater redundancy (as in Euclidean space); 3) empirically measuring the redundancy using the distance between the two points and other points is typically more reliable than estimating redundancy solely from the distance separating two points. We therefore conclude that in the absence of any additional information, the k-median distance between two points provides a useful clue to their redundancy; but when other points are available they will provide valuable additional information about redundancy.

Fig. 9 shows redundancy between different pairs of points. Let i_1 and i_2 denote two points, shown with a square and a cross in the pictures; then each point i in the plane is shaded dark gray if $|d(i_1, i) - d(i_2, i)| < \epsilon$, and redundancy is measured by the area of points shaded dark. We show eight cases, arranged in four pairs where the min distance between i_1, i_2 is the same; but in the top picture the difference in the x and y coordinates of i_1 and i_2 is also the same, while in the bottom picture the difference in the y coordinate has been enlarged. This does not change the min distance between i_1 and i_2, which is based on only the closest coordinates. In the four pairs of pictures we gradually increase the distance between i_1 and i_2, from less than ϵ (left) to greater than ϵ (right).

First, by comparing the top and bottom pictures, especially on the left, we can see that the redundancy between i_1 and i_2 can vary dramatically, even when the distance between them is held constant. At the same time, in some cases, points separated by a larger distance can

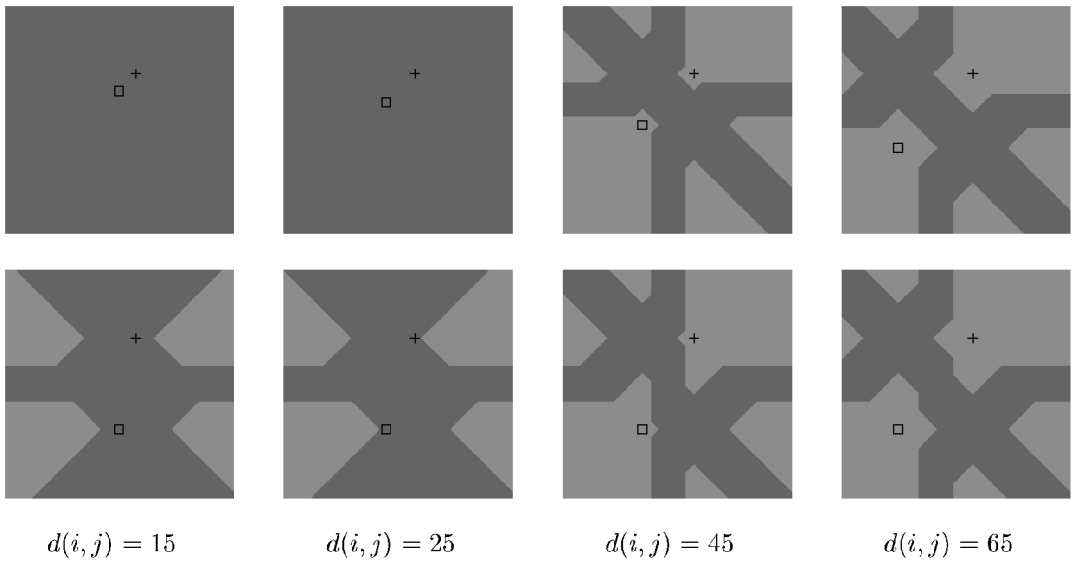

Fig. 9. In each picture, two points are shown as a square and cross. The darkly shaded region shows all points such that the difference in the distance from this point to the square and cross is less than $\epsilon = 30$ pixels (image is 200×200). The min distance is used here. The min distance between the cross and square varies between 15 (left) to 65 (right). For each distance, we show two pairs of points: The min distance between these pairs is the same, but the difference in the other coordinate varies considerably—same in the first row, different in the second row.

have smaller redundancy (compare the top picture with distance 25 to the bottom picture with distance 15). This effect is common with median distances since the distance between two objects tells us only about one part of the object and ignores other parts that can still affect the distance to other objects. Thus, we have demonstrated point 1 above—pairs of points separated by the same distance can have very different redundancy.

Next, we see that increasing the distance between i_1, i_2 generally increases redundancy: in each row, as the distance increases from left to right, the dark shaded area decreases at the same time. In particular, increasing the distance between points from less than ϵ to greater than ϵ can dramatically affect the redundancy. In other cases, increasing distance may lead to no increase in redundancy, especially when distances are large relative to ϵ. Thus, we have demonstrated point 2) above—other factors being equal, smaller distances typically lead to equal or greater redundancy. Finally, we conclude point 3) from elementary probability theory—if we have access to the distances from i_1 and i_2 to points randomly sampled in the plane, we can estimate redundancy accurately.

4 Comparison of Algorithms

We now draw on these insights to produce concrete methods for representing classes in nonmetric spaces, for nearest-neighbor classification. In the following, we will first describe four different condensing algorithms (Section 4.1). In Section 4.2, we will describe a series of simulations comparing the different algorithms under various conditions. In Section 4.3, we will describe a comparison of the different algorithms given a real data set of classes of silhouettes.

4.1 Description of Algorithms

We compared the following four algorithms: the first two algorithms, random selection and boundary enhancement, represent old condensing ideas; the last two algorithms, atypical selection and correlation cover, apply new ideas discussed above for class representation in nonmetric spaces. Each algorithm selects a representative set of examples $\mathcal{S}$ of size Q_c.

In describing these algorithms, we use the word *cover* as follows: Let $\mathcal{P}$ denote a class of images. For $p_1, p_2 \in \mathcal{P}$, p_1 *covers* p_2 if and only if $d(p_1, p_2) < d(p_2, q), \forall q \notin \mathcal{P}$. That is, choosing p_1 as a representative guarantees correct classification of p_2. We say that a subset of $\mathcal{P}$ covers $\mathcal{P}$ when every point in $\mathcal{P}$ is covered by some point in this subset. Note that $\mathcal{P}$ always has a subset that covers it, since $\mathcal{P}$ trivially covers itself. However, a useful cover only exists when the distance function succeeds in ensuring that many points within the same set are nearby relative to points in the other set. When a compact covering set does not exist, that is strong evidence that nearest-neighbor classification will be ineffective. For example, if the full set $\mathcal{P}$ is the smallest set that covers $\mathcal{P}$ then we would have 0 percent accuracy in using nearest neighbors to classify each point in $\mathcal{P}$ using all the remaining points.

Random selection. For every class $\mathcal{C}$, compute $\mathcal{S}$ by randomly (but without repetitions) choosing Q_c examples from $\mathcal{C}$.

An algorithm for the selection of class representation is potentially useful only if it outperforms random selection.

Boundary enhancement. For every class $\mathcal{C}$, compute $\mathcal{S}$ as follows:

1. compute an approximation to the minimal cover of size $\leq Q_c$ using a greedy algorithm and
2. until size of $\mathcal{S}$ is Q_c, add boundary points which are furthest from $\mathcal{S}$.

This algorithm is described in detail in Appendix A.1.

Part 1 of this algorithm resembles the first iteration of Dasarathy's algorithm [8]; subsequent iterations were not found by [8] to significantly affect the results. While capturing the essence of most condensing algorithms which look for class boundaries, our implementation ignores important differences which address the issue of computational efficiency, since such differences are not relevant for our purpose here.

Atypical selection. For every class $\mathcal{C}$, compute $\mathcal{S}$ as follows:

1. compute an approximation to the minimal cover of size $\leq Q_c$ using a greedy algorithm and
2. until the size of $\mathcal{S}$ is Q_c, add atypical points (not necessarily on the boundary) which are furthest from $\mathcal{S}$.

This algorithm is described in detail in Appendix A.2.

This algorithm tests our second observation, that class boundaries fail to capture important class structure and that "atypical" points—which are far from the representative set—should also be included in the class representation.

Selection based on correlation. Following the ideas discussed in Section 3, we compare datapoints by measuring how well their vectors of distances are correlated. More specifically, given two datapoints X, Y and their corresponding distance vectors $\mathbf{x}, \mathbf{y} \in \mathcal{R}^n$, where $\mathbf{x}$ is the vector of distances from X to all the other training points and $\mathbf{y}$ is the vector of distances from Y to all the other training points, we measure the correlation between the datapoints using the statistical correlation coefficient between $\mathbf{x}, \mathbf{y}$:

$$corr(X, Y) = corr(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} - \mu_x}{\sigma_x} \cdot \frac{\mathbf{y} - \mu_y}{\sigma_y}.$$

Above μ_x, μ_y denote the mean of $\mathbf{x}, \mathbf{y}$ respectively, and σ_x, σ_y denote the standard deviation of $\mathbf{x}, \mathbf{y}$, respectively.

We explored two ways of using correlation between datapoints. Version 1 below is significantly less efficient than the previous algorithms, since it requires the computation of correlation between any two datapoints. Version 2 only requires the computation of correlation between any two datapoints within the same class. Since both versions performed roughly the same in all our tests on both simulated and real data, version 2 is our preferred algorithm; in subsequent discussions, only results with version 2 are shown.

Version 1. Repeat the atypical selection algorithm as described above, but whenever the distance between two datapoints is used—use instead the correlation between them.

Version 2. For every class $\mathcal{C}$, compute $\mathcal{S}$ as follows: using a greedy algorithm, choose Q_c points that maximize V_X—a combined measure of incremental cover and correlation at each datapoint X.

Define V_X as $V_X = (N_X + 1) \cdot \frac{1-Corr_X}{2}$, where: $Corr_X$ is the maximal correlation of X with points in $\mathcal{S}$, and N_X is the number of not-yet-covered points which get covered by X. While ad hoc, this measure combines two useful criteria by favoring points that cover previously uncovered points, while also favoring points that are different from the currently selected points.

This algorithm is described in detail in Appendix A.3.

Note that this algorithm chooses representatives by combining N_X, which measures how well each point covers previously uncovered points in terms of distances, with $1 - Corr_X$, which indicates how redundant the point is with respect to previously chosen points.

4.2 Simulations

To compare the four algorithms described above, we simulated data representing various conditions:

1. For simplicity, each datapoint was chosen to be a vector in $\mathcal{R}^7$ or $\mathcal{R}^{25}$. Thirty clusters were randomly chosen, each with 30 datapoints.
2. To study how the structure of the data affect the performance of each algorithm, we simulated three cases:
   a. **"Vanilla."** The points in each class form a small cluster in the feature space. Specifically, the center of each cluster is chosen randomly in space, and the class members are chosen from a spherical normal distribution around the chosen center.
   b. **One outlier.** The points in each class cluster around a prototype, but many class members vary widely in one dimension (which may be different for the different class members). Specifically, the center of each cluster is chosen randomly in space, and the class members are chosen from a spherical normal distribution; for about half the points, however, one coordinate (randomly chosen) takes an arbitrary value totally different from the center value.
   c. **Irrelevant features.** The points in each class cluster around a prototype, but many class members vary widely in a small number of dimensions (less than half, and fixed for the different class members). Specifically, the center of each cluster is chosen randomly in space, and the class members are chosen from one of two normal distributions spread around the chosen center: half the points are chosen from a spherical normal distribution, and half the points are chosen from an elongated elliptical normal distribution.

   In Case 2a above, robust distances are not really required, whereas in Cases 2b and 2c robust methods improve performance.
3. We simulated four distance functions: Euclidean (ℓ_2), $\ell_{0.5}$, $\ell_{0.2}$, and median. Note that the middle two are nonmetric but bounded, i.e., the violation of the triangle inequality is bounded by a constant scaling factor independent of the size of the data (see Appendix B), while the median distance can arbitrarily violate the triangle inequality.

During the simulations, 1,000 test datapoints were randomly chosen from a uniform distribution in $\mathcal{R}^7$ and $\mathcal{R}^{25}$. Each test datapoint was classified based on: 1) all the data, 2) the representatives computed by each of the four algorithms described in Section 4.1. For each algorithm, the test is successful if the two methods (classification based on all the data and based on the chosen representatives) give the same results. Thus for each algorithm, we attach a score of percent correct: the percentage of test datapoints that scored success in this simulation block. We repeated each block 20 times (each with a different training set), to gather statistics on the variability in the percent correct value (mean and standard deviation).

Fig. 10 summarizes representative results of some of our simulations. Note that our test data comes from a different distribution than the training set. By using uniformly distributed test data, we estimate the volume of the difference in the voronoi diagrams produced by the complete set and the representative subset.

4.3 Test with Real Data

To test our method with real images, we used the local curve matching algorithm described in [11]. This curve matching algorithm was designed to compare curves which may be quite different, and return the distance between them. The steps of the algorithm include: 1) the automatic extraction of feature points with relatively high curvature and 2) feature matching using dynamic programming and efficient alignment; the final distance is the median of the interfeature distances. Thus, this algorithm is nonmetric, due to both the somewhat arbitrary selection of features and the final median step.

In our test, we used two classes with 30 images each. The first set included images of two similar cars from different points of view. The second set included images of a cow from different points of view, including partially occluded images. Two of the original images are shown in Fig. 11d. A few contours from the cow class, which were automatically extracted, are shown in Fig. 11a, while contours from the car class are shown in Fig. 11b. 30 images were used as test images; contours extracted from these images are shown in Fig. 11c. These test images are also cow contours: some were obtained from different viewpoints of the same cow, and some from the same viewpoints with more occlusion.

We used the four algorithms described in Section 4.1 to compute representative sets of five and seven contours for each class. We then compared the classification of the test data based on these representatives, to the classification obtained using all the data. Results (percent correct scores) are shown in Fig. 12.

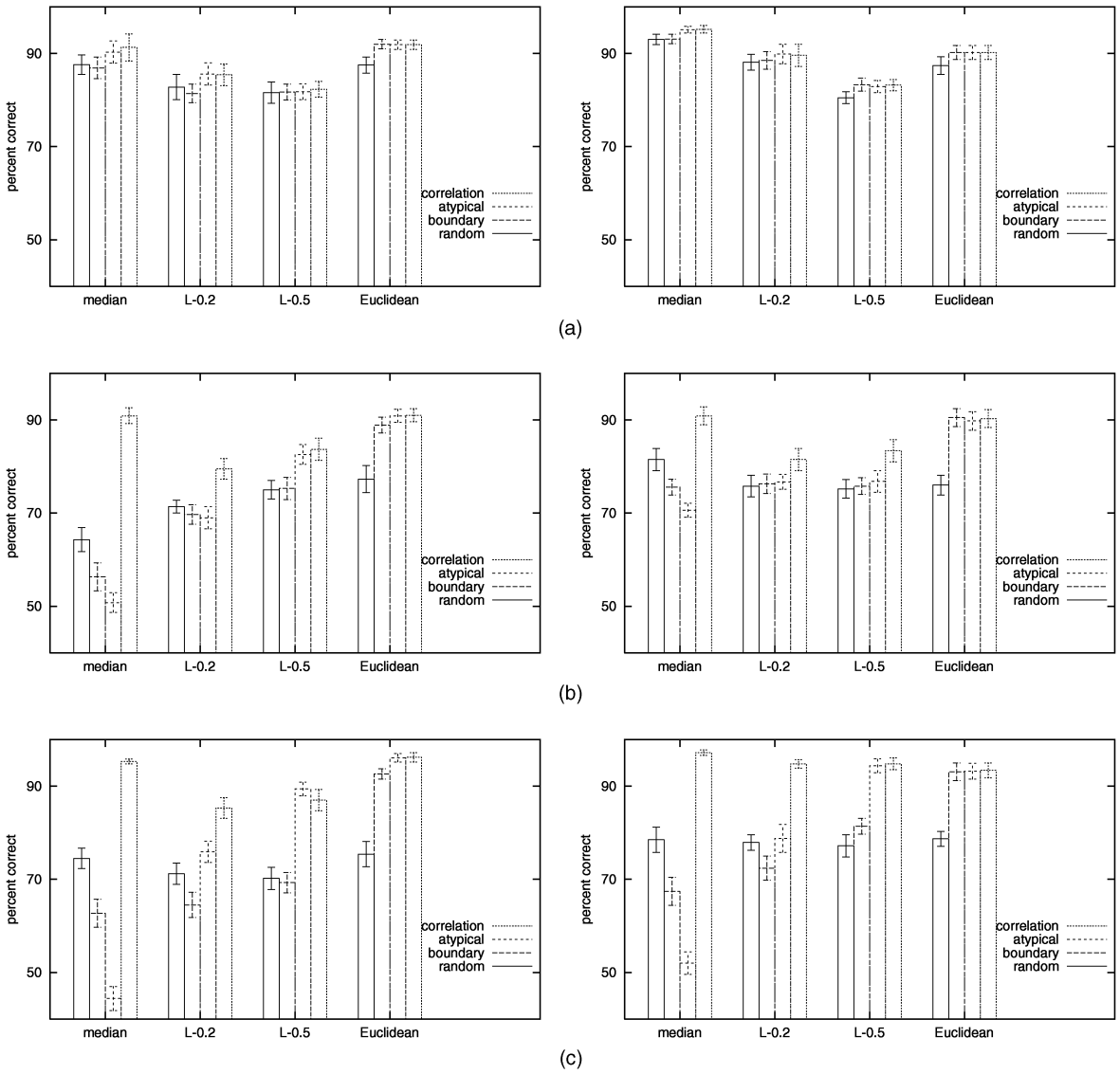

Fig. 10. Simulation results, with data chosen from $\mathcal{R}^7$ (left column) and $\mathcal{R}^{25}$ (right column). Plotted are the values of percent correct scores, as well as error bars giving the standard deviation calculated over 20 repetitions of each experiment. Each graph contains a histogram composed of four groups of four bars: each group gives the percent correct score for each of the different distance functions used here: median, $\ell_{0.2}$, $\ell_{0.5}$, and ℓ_2; for each group (distance function) we plot four bars for each of the four algorithms described in Section 4.1 (from left to right): **random** selection, **boundary** enhancement, atypical selection, and selection based on correlation. (a) "Vanilla" case, (b) one outlier, and (c) irrelevant features (see text).

4.4 Discussion

In all our experiments using simulated and real data (Figs. 10b, 10c, and 12), the correlation-based algorithm performs significantly better than any other algorithm with any of the nonmetric distance functions; given the Euclidean distance, its performance is similar (or slightly, but significantly, better) as compared to the boundary and atypicals methods. Interestingly, the boundary method performs **significantly worse** than a random selection of representatives with the median distance and the $\ell_{0.2}$ distance.

In Fig. 10a, which represents the "vanilla" case where the data lacks any "interesting" structure and where a class is just a clump of entities which are truly close to each other and well-separated from other classes, all the methods are comparable in performance. In this case, there is no need to use a nonmetric distance. Occasionally, especially with the Euclidean distance, the random selection performs significantly worse than the other methods, although its score is not much lower (and the difference may not be worth the additional effort).

4.5 Summary

Our results clearly demonstrate an advantage to our method over existing methods in the classification of data in nonmetric spaces. Almost as important, in metric spaces (fourth column in Figs. 10a-c) or when the classes lack any "interesting" structure (Fig. 10a), our method is not worse than any existing method. Thus, it should be generally used when the nature of the data and the distance function is not known a priori. Note that although the random method sometimes performs as well as the other methods, it does not provide any criteria as to how many points should be selected for the representative set. This is a rather important

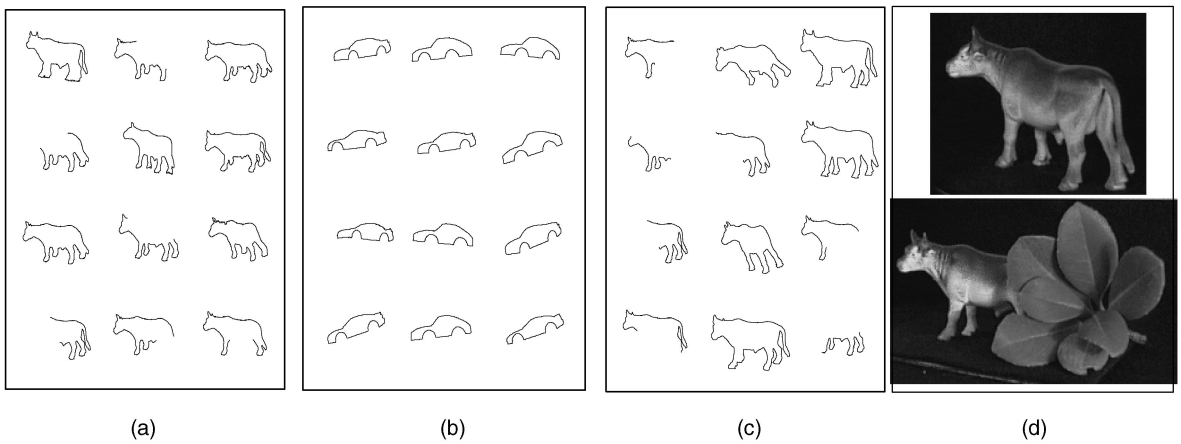

Fig. 11. Real data used to test the four algorithms: (a) Twelve examples from the first class of 30 cow contours, (b) 12 examples from the second class of 30 car contours, (c) 12 examples from the set of 30 test contours, and (d) 2 examples of the real images from which the contours in (a) were obtained.

hidden advantage of the three principled methods over the random method.

5 Extension: Parametric Methods

We now consider the extension of parametric methods to vectorial data described by nonmetric distances. Some of the simplest and most popular parametric methods seek a hyperplane that best separates two classes; yet Fig. 1 tells us that linear discriminants and hyperplane separators are problematic and ill-defined in robust nonmetric vector spaces. We note, however, that using a linear discriminant in Euclidean space is equivalent to performing nearest-neighbor classification with respect to two points (i.e., the two prototypical examples representing each class), where the separating hyperplane is perpendicular to the line connecting these two points and located midway between them. Thus, the most obvious generalization of linear discrimination to nonmetric spaces is via the use of prototypical examples and nearest-neighbor classification (or networks of Radial Basis Functions).

More specifically, we seek a prototype instance of each class and perform nearest-neighbor classification using these prototypes, and using the distance function that is appropriate to the space. Optimal prototypes are defined as points in the vector space such that nearest-neighbor classification according to the given distance produces correct classification of the training set. Unfortunately, finding the optimal prototypes for most distance functions, including the median, is prohibitively computationally demanding. In other words, an exponential search may be needed to guarantee prototype optimality. It may be fruitful in the future to approach this problem by using heuristic methods to find good, though suboptimal, prototypes.

A simpler approach is to represent each class by its generalized "centroid." In the Euclidean space, the centroid (or mean) is the point $\bar{q}$ whose sum of squared distances to all the class members $\{q_i\}_{i=1}^n$, measured by

$$\sqrt{\sum_{i=1}^{n} d(\bar{q}, q_i)^2},$$

is minimized. Suppose now that our data comes from a vector space where the correct distance is the ℓ_p distance from (1). With the natural extension of the above definition, we use the following lemma to generalize the concept of centroid to nonmetric ℓ_p spaces:

Definition 2. *For a set of points in ℓ_p, the point $\bar{q}$ which obtains the minimal sum of distances to all the set members*

$$E = \frac{1}{p}\sqrt{\sum_{i=1}^{n} d(\bar{q}, q_i)^p} \tag{3}$$

is the generalized centroid of the set.

Lemma 3. *For $p < 1$ (the nonmetric cases), the exact value of every feature of the generalized centroid $\bar{q}$ must have already appeared in at least one element in the class, i.e., $\forall\, j\ \exists\, i$ such that $\bar{q}^j = q_i^j$ (see [9] for a related result).*

Proof.

1. Since the function $f(x) = x^p$ is monotonic, the minimum of E in (3) is obtained for the same $\bar{q}$ as the minimum of E^p:

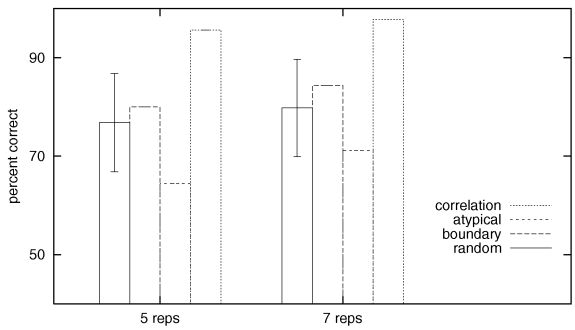

Fig. 12. Results with real data: Two histograms of four bars show the percent correct score for each of the four algorithms described in Section 4.1: **random** selection (where standard deviation of performance is also plotted), **boundary** enhancement, **atypical** selection, and selection based on **correlation**. The number of representatives chosen by the algorithm was limited to five (left histogram) and seven (right histogram).

TABLE 1
Results Showing How Often Members of a Class (from Five Arbitrary Classes in the Corel Database) Are Closer to Their Own Class Prototype Than to Other Prototypes

	Euclidean prototype (mean)	$\ell_{0.5}$ prototype (sparse)
$\ell_{0.5}$ distance	68%	80%
Euclidean distance	72%	73%

Each image is represented by a vector of 11 features (following[7]). Each row corresponds to a different distance function used to match images with prototypes: the metric Euclidean and the nonmetric $\ell_{0.5}$ (used in [7]). Each column corresponds to a different method used to compute class prototypes: the features' mean, which is the Euclidean prototype, and the $\ell_{0.5}$ sparse prototype computed as described in Lemma 3.

$$E^p = \sum_{i=1}^{n} d(\bar{q}, q_i)^p = \sum_{i=1}^{n}\sum_{j=1}^{d} |\bar{q}^j - q_i^j|^p = \sum_{j=1}^{d}\sum_{i=1}^{n} |\bar{q}^j - q_i^j|^p. \quad (4)$$

2. Since E^p in (4) is separable in the features, the minimum can be computed for each feature $\bar{q}^j$ separately by minimizing E_j^p:

$$E_j^p(\bar{q}^j) = \sum_{i=1}^{n} |\bar{q}^j - q_i^j|^p. \quad (5)$$

3. $E_j^p(\bar{q}^j)$ is a function of one variable, continuous and piecewise differential. Assume w.l.o.g. that the features are ordered so that $q_1^j < q_2^j < \ldots < q_n^j$; then $E_j^p(\bar{q}^j)$ is continuous and differential in every segment (q_i^j, q_{i+1}^j). Let us compute its second derivative inside the ith segment:

$$\frac{d^2 E_j^p}{d(\bar{q}^j)^2} = \sum_{i=1}^{n} \frac{p(p-1)}{|\bar{q}^j - q_i^j|^{2-p}} < 0,$$

where the sum is negative for $p < 1$ because each of its components is negative. This shows that $E_j^p(\bar{q}^j)$ is concave within every segment for $0 < p < 1$; thus it can only obtain a local maximum within the segment, and local minima can only be obtained at the boundary points q_i^j, q_{i+1}^j.

4. Every point $\bar{q}^j = q_i^j$ for some i is indeed a local minimum. To see this, note that the jth component of the first derivative of the sum in (5) goes to infinity as $\bar{q}^j$ approaches q_i^j from above, and to negative infinity when $\bar{q}^j$ approaches q_i^j from below. Therefore, this term dominates the derivative, so pushing $\bar{q}^j$ towards q_i^j decreases the sum.

Thus, in order to find the vector of feature values $\bar{q}$ which globally minimizes (3), we need only search among the set of existing feature values: $\{q_i^j\}_{i=1}^n$ for every j separately. □

Corollary 1. *The values of the elements of the generalized centroid can be determined separately with complexity $O(n^2)$ or less (less for $p = 1$, where the median feature obtains the minimum), and total complexity of $O(dn^2)$ given d features. $\bar{q}$ is therefore determined by a mixture of up to d exemplars, where d is the dimension of the vector space.*

The corollary implies that there are efficient algorithms for finding the generalized "centroid" of a class, even using certain nonmetric distances. Moreover, the point which replaces the centroid in ℓ_p spaces for $p < 1$ contains values from at most d datapoints, which means that the representation is "sparse"—a desirable property for data compression.

We now use this result with a concrete example to compute prototypes for the corel database, a large commercial database of images prelabeled by different categories (such as "lions"), where nonmetric distance functions have proven effective in determining the similarity of images [7]. The corel database is very large, making the use of prototypes desirable.

We represent each image using a vector of 11 features (thus, $d = 11$) describing general image properties, such as color histograms, as described in [7]. We consider the Euclidean and $\ell_{0.5}$ distances, and their corresponding prototypes: the Euclidean mean and the $\ell_{0.5}$-prototype computed according to Lemma 3. Given the first five classes, each containing 100 images, we found their corresponding prototypes; we then computed the percentage of images in each class which are closest to their own prototype, using either the Euclidean or the $\ell_{0.5}$ distance and one of the two prototypes. The results are given in Table 1, showing that indeed best performance is obtained with the nonmetric $\ell_{0.5}$ distance and the corresponding $\ell_{0.5}$ generalized centroid used as class prototype (computed as described in Lemma 3).

Another important distance function is the generalized Hamming distance: given two vectors of features, their distance is the number of features which are different in the two vectors. This distance was assumed to underlie human perception in psychophysical experiments which used artificial objects (Fribbles) to investigate human categorization and object recognition [42]. In agreement with experimental results, the prototype $\bar{q}$ for this distance computed according to the definition above is the vector of "modal" features—the most common feature value computed independently at each feature.

6 Discussion

It is an interesting question as to how such nonmetric parametric classifiers compare to other parametric methods, such as Support Vector Machines [6]. This is a problem of large scope, so we will make only some preliminary comments. We focus in this paper on classification using

data items that contain outliers and have missing elements, where robust distances like the k-median are appropriate. In this case, a single object can produce an infinite number of images, each containing some core number of elements in common, and other elements that can vary arbitrarily. A limited number of images does not accurately delineate the entire set of possible images. It is therefore difficult to expect standard parametric classification methods to fit any surface (such as a linear separator) to the existing data and use this surface to extrapolate to new data. This is because any new image may be very far (in Euclidean distance) from every previously seen image. However, when we use a robust, nonmetric distance, we are making explicit important domain specific knowledge. For example, using a median distance can allow us to trivially extrapolate to many new images, correctly ignoring outliers and spurious data. From very few samples, we may build a classifier that correctly classifies test items that are very different in Euclidean space, but very similar with the appropriate nonmetric distance. However, the general effectiveness of any of these methods will depend greatly on the specific domain and task.

7 Conclusions

We feel that our paper makes two main contributions. First, we reassessed the relevance of existing supervised classification techniques to application-based and human-like classification. We argued that classification systems that can model human performance, or use robust matching methods that are typically used in successful applications, make use of similarity judgments that are nonmetric, and in particular, do not obey the triangle inequality. In this case, most existing pattern recognition techniques are not relevant. Exemplar-based methods, however, can be applied naturally when using a wide class of nonmetric similarity functions. The key issue, however, in applying exemplar-based methods in such settings is to find methods for choosing good representatives of a class, that accurately characterize it.

We then focused on developing techniques for solving this problem, emphasizing two points: First, we showed that the distance between two images is not a good measure of how well one image can represent another in nonmetric spaces. Instead, we suggested considering the correlation between the distances from each image to other previously seen images. Second, we showed that in nonmetric spaces, boundary points are less significant for capturing the structure of a class than they are in the Euclidean space. We suggested that atypical points may be more important in describing classes. We demonstrated the importance of these ideas in greatly improving classification results using both synthetic and real images.

Finally, we have suggested ways of applying parametric techniques to supervised learning problems that involve a specific, nonmetric distance function. We have shown how to generalize the idea of linear discriminant functions in a way that may be more useful in nonmetric spaces. And we have shown a "proof-of-concept" for classification using the centroid of a class as determined by the specific distance, with ℓ_p distances for $p < 1$.

Appendix A

Detailed Description of Condensing Algorithms

A.1 Boundary Enhancement

For every class $\mathcal{C}$, compute its representative set of examples $\mathcal{S}$ of size Q_c, using the following algorithm:

1. compute cover
   - **(0)** Initialization:
     set of representative examples $\mathcal{S}$ is empty
     set of examples which are incorrectly classified $\mathcal{A} \leftarrow \mathcal{C}$
   - **(1)** For every datapoint c in $\mathcal{C} \setminus \mathcal{S}$, compute N_c—the number of points in $\mathcal{A}$ closer to c than to any point in any other class (i.e., the number of points covered by c).
   - **(2)** Find $\bar{c}$ with the largest value of N_c over all $c \in \mathcal{C} \setminus \mathcal{S}$. Let $T_{\bar{c}}$ denote the group of points in $\mathcal{A}$ closer to $\bar{c}$ than to other classes (i.e., the points covered by $\bar{c}$).
   - **(3)** update:
     $\mathcal{A} \leftarrow \mathcal{A} \setminus T_{\bar{c}}$
     $\mathcal{S} \leftarrow \mathcal{S} \cup \{\bar{c}\}$
   - $\Longleftarrow$ while size of $\mathcal{S}$ is smaller than Q_c and $\mathcal{A}$ not empty, return to (1)

2. Add boundary points: while size of $\mathcal{S}$ is smaller than Q_c, repeat:
   - **(1)** $\forall c \in \mathcal{C}$, compute
     D_c—the distance of c to the nearest datapoint in *another* class
     d_c—the distance of c to the nearest datapoint in the representative set $\mathcal{S}$
     $\Delta_c = d_c - D_c$
   - **(2)** Find $\bar{c}$ which produces the largest Δ_c over all $c \in \mathcal{C}$.
   - **(3)** update:
     $\mathcal{S} \leftarrow \mathcal{S} \cup \{\bar{c}\}$

A.2 Atypical Selection

For every class $\mathcal{C}$, compute its representative set of examples $\mathcal{S}$ of size Q_c, using the following algorithm:

1. compute cover, same as described in A.1 for the boundary algorithm.
2. Add atypical points: while size of $\mathcal{S}$ is smaller than Q_c, repeat:
   - **(1)** $\forall c \in \mathcal{C} \setminus \mathcal{S}$, compute
     d_c—the distance of c to the nearest datapoint in the representative set $\mathcal{S}$
   - **(2)** Find $\bar{c}$ which obtains the largest d_c over all $c \in \mathcal{C} \setminus \mathcal{S}$.
   - **(3)** update:
     $\mathcal{S} \leftarrow \mathcal{S} \cup \{\bar{c}\}$

A.3 Selection Based on Correlation

For every class $\mathcal{C}$, compute its representative set of examples $\mathcal{S}$ of size Q_c, using the following algorithm:

(0) Initialization:
set of representative examples $\mathcal{S}$ is empty
set of examples which are incorrectly classified $\mathcal{A} \leftarrow \mathcal{C}$
(1) For every datapoint c in $\mathcal{C} \setminus \mathcal{S}$, compute
N_c—the number of points in $\mathcal{A}$ closer to c than to any point in any other class
$Corr_c$—the maximal correlation of c with points in $\mathcal{S}$
$V_c = (N_c + 1) \cdot \frac{1 - Corr_c}{2}$
(2) Find $\bar{c}$ which obtains the largest V_c over all $c \in \mathcal{C} \setminus \mathcal{S}$. Let $T_{\bar{c}}$ denote the group of points in $\mathcal{A}$ closer to $\bar{c}$ than to other classes (i.e., the points covered by $\bar{c}$).
(3) update:
$\mathcal{A} \leftarrow \mathcal{A} \setminus T_{\bar{c}}$
$\mathcal{S} \leftarrow \mathcal{S} \cup \{\bar{c}\}$
$\Longleftarrow$ while size of $\mathcal{S}$ is smaller than Q_c and $V_c > t$ for some threshold t, return to (1).

APPENDIX B

A BOUNDED TRIANGLE INEQUALITY FOR NONMETRIC ℓ_p DISTANCES

It is known that for $p < 1$, the p-distance defined in (1) does not satisfy the triangle inequality: given three points $q_1, q_2, q_3 \in \mathcal{R}^n$, $d(q_1, q_3) > d(q_1, q_2) + d(q_2, q_3)$. More specifically, let $\mathbf{x}$ denote the vector connecting q_1, q_2, and $\mathbf{y}$ denote the vector connecting q_2, q_3; then

$$\left(\sum_1^n |x_i + y_i|^p\right)^{\frac{1}{p}} > \left(\sum_1^n |x_i|^p\right)^{\frac{1}{p}} + \left(\sum_1^n |y_i|^p\right)^{\frac{1}{p}}.$$

We will now show that there exists $\kappa > 1$ such that the κ-bounded triangle inequality is satisfied:

$$d(q_1, q_3) \leq \kappa(d(q_1, q_2) + d(q_2, q_3));$$

more specifically,

$$\left(\sum_1^n |x_i + y_i|^p\right)^{\frac{1}{p}} \leq \kappa\left(\left(\sum_1^n |x_i|^p\right)^{\frac{1}{p}} + \left(\sum_1^n |y_i|^p\right)^{\frac{1}{p}}\right). \quad (6)$$

First, from the concavity of $f(t) = t^p$ for $p < 1$, and for $a, b \geq 0$:

$$\frac{1}{2^{1-p}}(a^p + b^p) \leq (a + b)^p \leq a^p + b^p,$$

where the first inequality follows from $\frac{(a^p + b^p)}{2} \leq (\frac{a+b}{2})^p$. Let

$$a = (\sum_1^n |x_i|^p)^{\frac{1}{p}},$$

$$b = (\sum_1^n |y_i|^p)^{\frac{1}{p}};$$

then

$$\left(\left(\sum_1^n |x_i|^p\right)^{\frac{1}{p}} + \left(\sum_1^n |y_i|^p\right)^{\frac{1}{p}}\right)^p = (a + b)^p$$

$$\geq \frac{1}{2^{1-p}}(a^p + b^p) = \frac{1}{2^{1-p}}\left(\sum_1^n |x_i|^p + \sum_1^n |y_i|^p\right)$$

$$= \frac{1}{2^{1-p}}\sum_1^n \left(|x_i|^p + |y_i|^p\right)$$

$$\geq \frac{1}{2^{1-p}}\sum_1^n \left(|x_i| + |y_i|\right)^p$$

$$\geq \frac{1}{2^{1-p}}\sum_1^n |x_i + y_i|^p.$$

Thus, (6) holds with $\kappa = 2^{\frac{1-p}{p}}$. Furthermore, this inequality cannot be improved; a worst case, for which equality is obtained, is the case where n is assumed even, $\mathbf{x}$ is the vector whose even components are 1 and the odd ones are 0, and $\mathbf{y}$ is the vector whose odd components are 0 and the even ones are 1.

Note that the smaller p is the larger the bound κ is, and the further from metric the corresponding p-distance is.

ACKNOWLEDGMENTS

The authors would like to thank Liz Edlind for Fig. 5 and thank Shimon Edelman for the MDS matlab code. This paper is based on "Condensing Image Databases when Retrieval is Based on Nonmetric Distances," by David Jacobs, Daphna Weinshall, and Yoram Gdalyahu, which appeared in the Proceedings of the Sixth IEEE International Conference on Computer Vision, January 1998, and on "Supervised Learning on Nonmetric Spaces," by Daphna Weinshall, David Jacobs, and Yoram Gdalyahu, 1998, NIPS. This research was conducted while Daphna Weinshall was on sabbatical from NEC Research Instititute and, also, while Yoram Gdalyahu was with the Hebrew University.

REFERENCES

[1] R. Basri, L. Costa, D. Geiger, and D. Jacobs, "Determining the Similarity of Deformable Objects," *Vision Research*, vol. 38, no. 15-16, pp. 2,365-2,385, 1998.
[2] M. Black and P. Anandan, "The Robust Estimation of Multiple Motions: Parametric and Piecewise-Smooth Flow Fields," *Computer Vision and Image Understanding*, vol. 63, no. 1, pp. 75-104, 1996.
[3] A. Blake and A. Zisserman, *Visual Reconstruction*. Cambridge, Mass.: MIT Press, 1987.
[4] M. Blatt, S. Wiseman, and E. Domany, "Clustering Data through an Analogy to the Potts Model," *Advances in Neural Information Processing Systems*, vol. 8, pp. 416-422, 1996.
[5] M. Brand, "A Fast Greedy Pairwise Distance Clustering Algorithm and Its Use in Discovering Thematic Structures in Large Data Sets," Technical Report 406, MIT Media Lab, 1996.
[6] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, pp. 273-297, 1995.
[7] I. Cox, M. Miller, S. Omohundro, and P. Yianilos, "PicHunter: Bayesian Relevance Feedback for Image Retrieval," *Proc. Int'l Conf. Pattern Recognition*, vol. C, pp. 361-369, 1996.
[8] B.V. Dasarathy, "Minimal Consistent Set (MCS) Identification for Optimal Nearest Neighbor Decision Systems Design," *IEEE Trans. Systems, Man, and Cybernetics*, vol. 24, no. 3, pp. 511-517, 1994.
[9] M. Donahue, D. Geiger, R. Hummel, and T. Liu, "Sparse Representations for Image Decompositions with Occlusions," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 7-12, 1996.
[10] J. Friedman, J. Bently, R. Finkel, "An Algorithm for Finding Best Matches in Logarithmic Expected Time," *ACM Trans. Math. Software*, vol. 3, no. 3, pp. 209-226, 1977.

[11] Y. Gdalyahu and D. Weinshall, "Flexible Syntactic Matching of Curves and Its Application to Automatic Hierarchical Classification of Silhouettes," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 21, no. 12, pp. 1,312-1,328, Dec. 1999.

[12] D. Geiger and F. Girosi, "Parallel and Deterministic Algorithms MRFs: Surface Reconstruction," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 13, no. 5, pp. 401-412, May 1991.

[13] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 6, no. 7, pp. 721-741, July 1984.

[14] K. Gowda and G. Krishna, "The Condensed Nearest Neighbor Rule Using the Concept of Mutual Nearest Neighbor," *IEEE Trans. Information Theory,* vol. 25, no. 4, pp. 488-490, 1979.

[15] T. Graepel, R. Herbrich, P. Bollmann-Sdorra, and K. Obermayer, "Classification on Pairwise Proximity Data," *Proc. Neural Information Processing Systems,* pp. 438-444, 1999.

[16] K. Fukunaga and J. Mantock, "Nonparametric Data Reduction," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 6, no. 1, pp. 115-118, Jan. 1984.

[17] R. Haralick and L. Shapiro, *Computer and Robot Vision,* vol. 2, Addison-Wesley, 1993.

[18] P. Hart, "The Condensed Nearest Neighbor Rule," *IEEE Trans. Information Theory,* vol. 14, no. 3, pp. 515-516, 1968.

[19] T. Hastie and W. Stuetzle, "Principal Curves," *J. Am. Statistical Assoc.,* vol. 84, 502-516, 1989.

[20] G. Hinton, C. Williams, and M. Revow, "Adaptive Elastic Models for Hand-Printed Character Recognition," *Neural Information Processing Systems,* vol. 4, pp. 512-519, 1992.

[21] T. Hofman and J.M. Buhman, "Pairwise Data Clustering by Deterministic Annealing" *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 1, pp. 1-14, Jan. 1997.

[22] D. Huttenlocher, G. Klanderman, and W. Rucklidge, "Comparing Images Using the Hausdorff Distance," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 15, no. 9, pp. 850-863, Sept. 1993.

[23] D. Huttenlocher, D.J. Noh, and W. Rucklidge, "Tracking Non-Rigid Objects in Complex Scenes," *Proc. Fourth Int'l Conf. Computer Vision,* pp. 93-101, 1993.

[24] D. Jacobs, "Linear Fitting with Missing Data: Applications to Structure-from-Motion and to Characterizing Intensity Images," *IEEE Conf. Computer Vision and Pattern Recognition,* pp. 206-212, 1997.

[25] A. Jain and D. Zongker, "Representation of Handwritten Digits Using Deformable Templates," *Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 2, pp. 153-158, Feb. 1997.

[26] W. Johnson and J. Lindenstrauss, "Extension of Lipshitz Mapping to Hilbert Space," *Contemporary Math.,* vol. 26, pp. 189-206, 1984.

[27] H. Klock and J. Buhmann, "Multidimensional Scaling by Deterministic Annealing," *Proc. Int'l Workshop Energy Minimization Methods in Computer Vision and Pattern Recognition,* pp. 245-260, 1997.

[28] J. Kapur and H. Kesavan, *Entropy Optimization Principles with Applications.* Academic Press, 1992.

[29] S. Li, "On Discontinuity-Adaptive Smoothness Priors in Computer Vision," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 17, no. 6, pp. 576-586, June 1995.

[30] N. Linial, E. London, and Y. Rabinovich, "The Geometry of Graphs and Some of Its Algorithmic Applications," *Combinatorica,* vol. 15, pp. 215-245, 1995.

[31] R. Little and D. Rubin, *Statistical Analysis with Missing Data.* John Wiley and Sons, 1987.

[32] P. Meer, D. Mintz, D. Kim, and A. Rosenfeld, "Robust Regression Methods for Computer Vision: A Review," *Int'l J. Computer Vision,* vol. 6, no. 1, pp. 59-70, 1991.

[33] L. Ornstein, "Computer Learning and the Scientific Method: A Proposed Solution to the Information Theoretical Problem of Meaning," *J. Mount Sinai Hospital,* vol. 32, no. 4, pp. 437-494, 1965.

[34] T. Poggio and F. Girosi, "Regularization Algorithms for Learning That are Equivalent to Multilayer Networks," *Science,* vol. 247, pp. 978-982, 1990.

[35] J. Puzicha, J. Buhmann, Y. Rubner, and C. Tomasi, "Empirical Evaluation of Dissimilarity Measures for Color and Texture," *Proc. Int'l Conf. Computer Vision,* pp. 1,165-1,172, 1999.

[36] H. Royden, *Real Analysis.* New York: MacMillan Publishing, 1968.

[37] S. Santini and R. Jain, "Similarity Queries in Image Databases," *Proc. Conf. Computer Vision and Pattern Recognition,* pp. 646-651, 1996.

[38] C. Tappert, "Cursive Script Recognition by Elastic Matching," *IBM J. Res. Development,* vol. 26, no. 6 pp. 765-771, 1982.

[39] C. Tomasi and T. Kanade, "Shape and Motion from Image Streams under Orthography: a Factorization Method," *Int'l J. of Computer Vision,* vol. 9, no. 2, pp.137-154, 1992.

[40] W. Tsai and S. Yu, "Attributed String Matching with Merging for Shape Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 4, pp. 453-462, July 1985.

[41] A. Tversky, "Features of Similarity," *Psychological Rev.,* vol. 84, no. 4, pp. 327-352, 1977.

[42] P. Williams, "Prototypes, Exemplars, and Object Recognition," PhD thesis, Dept. of Psychology, Yale Univ., 1997.

[43] P. Yianilos, "Data Structures and Algorithms for Nearest Neighbor Search in General Metric Spaces," *Proc. Fourth Ann. ACM-SIAM Symp. Discrete Algorithms,* pp. 311-321, 1993.

[44] K. Yoshida and H. Sakoe, "Online Handwritten Character Recognition for a Personal Computer System," *IEEE Trans. Consumer Electronics,* vol. 28, no. 3, pp. 202-209, 1982.

David W. Jacobs received the BA degree from Yale University in 1982. He graduated magna cum laude, with distinction in mathematics. From 1982 to 1985 he worked for Control Data Corporation on the development of database management systems and attended graduate school in computer science at New York University. From 1985 to 1992 he attended Massachusetts Institute of Technology, where he received the MS and PhD degrees in computer science. Since then, he has been a research scientist at NEC Research Institute. In 1998, he was on sabbatical at the Royal Institute of Technology (KTH) in Stockholm. Dr. Jacobs has also been a visiting member of the Courant Institute at New York University, where he has taught classes in computer vision and learning. His research has focused on human and computer vision, especially in the areas of object recognition and perceptual organization. He has also published work in the areas of motion understanding, memory and learning, and computational geometry.

Daphna Weinshall received the BSc degree in mathematics and computer science from Tel-Aviv University, Tel-Aviv, Israel, in 1982. She received the MSc and PhD degrees in mathematics and statistics from Tel-Aviv University in 1985 and 1986, respectively, working on models of evolution and population genetics. Between 1987 and 1992, she visited the center for biological information processing at Massachusetts Institute of Technology, and IBM T.J. Watson Research Center. In 1993, she joined the Institute of Computer Science at the Hebrew University of Jerusalem, where she is now an associate professor. Her research interests include computer and biological vision, as well as machine and human learning. She has published papers on learning in machine and human vision, qualitative vision, visual psychophysics, Bayesian vision, invariants, multipoint and multiframe geometry, image and model point-based metrics, motion, and structure from motion. Selected publications can be obtained from http://www.cs.huji.ac.il/~daphna.

Yoram Gdalyahu received the BSc degree in physics and mathematics from the Hebrew University, Jerusalem, Israel. He received the MSc degree in physics from the Weizmann Institute of Science, Rehovot, Israel for his work on resonant tunneling and inelastic scattering in gallium arsenide heterostructures. His PhD thesis in computer vision was submitted to the senate of the Hebrew University in October 1999. His research combines computer vision and machine learning. Upon graduation, he received the Eshkol scholarship given by the Israeli Ministry of Science to the best PhD students. He is currently with the IBM Research Laboratory, Haifa, Israel.