Pattern Recognition 47 (2014) 2721–2731

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Good recognition is non-metric

Walter J. Scheirer [a,e,*], Michael J. Wilber [b], Michael Eckmann [c], Terrance E. Boult [d,e]

[a] Harvard University, 52 Oxford St. NWL 209, Cambridge, MA 02138, United States
[b] Cornell University, 402 Gates Hall, Ithaca, NY 14853, United States
[c] Skidmore College, 815 North Broadway, Saratoga Springs, NY 12866, United States
[d] Securics, Inc, 1867 Austin Bluffs Parkway, Suite 100, Colorado Springs, CO 80918, United States
[e] University of Colorado Colorado Springs, 1420 Austin Bluffs Parkway P.O. Box 7150, Colorado Springs, CO 80933-7150, United States

ARTICLE INFO

Article history:
Received 11 May 2013
Received in revised form
5 December 2013
Accepted 27 February 2014
Available online 11 March 2014

Keywords:
Machine learning
Metric learning
Recognition
Computer vision
Face recognition
Object recognition

ABSTRACT

Recognition is the fundamental task of visual cognition, yet how to formalize the general recognition problem for computer vision remains an open issue. The problem is sometimes reduced to the simplest case of recognizing matching pairs, often structured to allow for metric constraints. However, visual recognition is broader than just pair-matching: what we learn and how we learn it has important implications for effective algorithms. In this review paper, we reconsider the assumption of recognition as a pair-matching test, and introduce a new formal definition that captures the broader context of the problem. Through a meta-analysis and an experimental assessment of the top algorithms on popular data sets, we gain a sense of how often metric properties are violated by recognition algorithms. By studying these violations, useful insights come to light: we make the case for local distances and systems that leverage outside information to solve the general recognition problem.

1. Introduction

Recognition is a term everyone in computer vision and machine learning understands – or at least we think we do. Despite multiple decades of research, it may be somewhat surprising to learn that a very basic question remains unresolved: *is recognition metric*? Familiar distance metrics used in computer vision include Euclidean distance and Mahalanobis distance, both computed in feature space. Given one of these metrics, the task of recognizing an unknown object can be approached by finding the class label of its nearest neighbor under that distance metric in a set of training samples. Such an approach provides a recognition function, thus some level of recognition can be accomplished with a metric. However, at a more fundamental level, we would like to know if distance truly captures all that is meant by the term recognition, and if metrics are good approaches to solving complex recognition tasks in computer vision. In this review paper, we adopt the convention that a problem is metric if the best solutions to that problem can be achieved by directly applying a distance metric to compute the answer.

An important observation with implications for recognition is that in separable metric space, using a distance metric and the nearest neighbor (NN) algorithm provides useful consistency. As the number of i.i.d. samples from the classes approaches infinity, the NN algorithm will converge to an error rate no worse than twice the Bayes error rate, *i.e.* no worse than twice the minimum achievable error rate given the distribution of the data [3]. To many, this convergence theorem suggests that recognition can always be formulated as NN matching with an appropriate distance metric. However, having to double the error of the optimal algorithm over the same data often does not lead to a particularly good algorithm. This becomes apparent when actual error rates are considered during experimentation.

With the recent popularity of metric learning [4–13] for various recognition tasks, where a metric is learned over given pairs of images that are similar or dissimilar, one might infer that recognition is always a metric process. We note that the NN convergence theorem [3] is true for *any* metric – hence any improvements from the choice of metric, or metric learning, are not about the asymptotic error, but something else such as the error for finite samples and/or the rate of convergence. We will show that while metric learning can produce reasonable results, enforcing metric properties leaves out information, often limiting the quality of recognition with finite data. This is consistent with supporting prior work [14] in pattern recognition that shows increasing discriminative power for non-metric distance measures over visual data.

If the convergence theorem itself is about recognition, then the recognition problem is *assumed* to be formulated in an asymptotic

* Corresponding author at: Harvard University, 52 Oxford St. NWL 209, Cambridge, MA 02138, United States. Tel.: +1 610 657 1538; fax: +1 617 496 5424.
E-mail addresses: wscheirer@fas.harvard.edu (W.J. Scheirer), mjw285@cornell.edu (M.J. Wilber), meckmann@skidmore.edu (M. Eckmann), tboult@securics.com (T.E. Boult).

http://dx.doi.org/10.1016/j.patcog.2014.02.018
0031-3203/

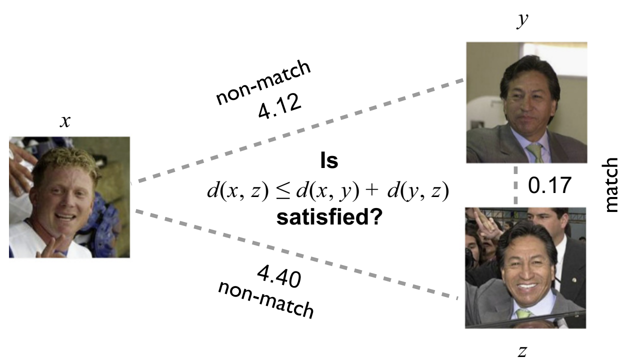

Fig. 1. Assumptions are often made about the underlying nature of recognition in computer vision that do not hold true in practice. A common constraint placed upon recognition algorithms is that they must be *metric*, meaning their distance scores adhere to the properties of non-negativity, identity, symmetry and the triangle inequality. At first glance, the scores from many recognition algorithms appear to satisfy these constraints. However, violations can be subtle. For example, the distance scores produced by the top-performing Tom-vs-Pete algorithm [1] for these images from LFW [2] violate the triangle inequality.

sense with infinite i.i.d. samples. We argue that visual recognition does not rely on either of those assumptions, but rather focuses on maximizing the accuracy for finite, and, unfortunately, opportunistic and hence potentially biased sampling.

A metric function is defined as follows:

Definition 1 (*Distance Metric*). A function $d: X \times X \rightarrow \mathbb{R}$ is metric over a set X if it satisfies four properties for $\{x, y, z\} \subseteq X$:

1. $d(x, y) \geq 0$ (non-negativity).
2. $d(x, y) = 0 \Leftrightarrow x = y$ (identity).
3. $d(x, y) = d(y, x)$ (symmetry).
4. $d(x, z) \leq d(x, y) + d(y, z)$ (the triangle inequality).

Metric functions have useful properties that allow one to show that a particular problem can be formulated as a convex minimization problem, or, as we have stated, that various types of sequences converge in the limit. There are also several cases where one of the properties is excluded. Functions that do not satisfy the triangle inequality are called semimetrics, those that violate symmetry are called quasimetrics, and those missing one or both halves of the identity requirement are called pseudometrics.[1] While the term "distance measure" is sometimes used to mean a distance metric, it is more appropriate to use this term to mean a measurement that provides information about dissimilarity, but may be formally non-metric (our use of the term follows this convention).

Is it reasonable to assume that a distance metric d maps pairs of elements from X into $\mathbb{R}$ during recognition? When a person recognizes an object, do they refer to an actual image of the object of interest? A more likely alternative is a comparison to a stored model with a more complex internal representation, not a direct copy of some prior trained input. This view is consistent with prototype theory [16] in cognitive psychology. Thus, at a structural level, recognition in this mode takes an input $x \in X$, and a model M, and hence cannot be metric because it is not even of the proper functional form. It is possible to build a model using just x, and then consider the distance between models in a nearest neighbor fashion. Many instance learning algorithms do just that. However, for many other commonly used recognition algorithms, one cannot induce a proper model from a *single* input.[2] Thus, the general problem of recognition cannot be restricted to just metrics, even though it must include them.

In the core pattern recognition literature, this issue has been raised specifically in the context of Euclidean distance. Pȩkalska et al. [17] observe that "Non-metric dissimilarity measures may arise in practice, *e.g.* when objects represented by sensory measurements or by structural descriptions are compared." Experiments to confirm this have included: comparing distance measures before and after Euclidean transforms are applied [17,18]; an examination of the parameter space of data for metricity [14]; and an evaluation of dissimilarity representations for classification [18–22]. In all cases, an enforcement of Euclidean constraints does not help classification performance [23], and non-Euclidean measures are often shown to be better, leading Pȩkalska et al. to conclude "that non-Euclidean and/or non-metric distances can be informative and useful in statistical learning" [14].

However, even in light of this finding, the research area of metric learning for computer vision remains quite active. A key difference from earlier work in metrics for statistical learning is that recent work in visual learning, with its strong need for data normalization, eschews Euclidean distance in favor of Mahalanobis distance [4]. In our review of the literature, we take a broader look at the many non-Euclidean metric learning approaches that have been proposed since the above studies were conducted.

Beyond statistical learning, it is natural to ask if the human mind, a most successful recognition system operates in a way that satisfies the key metric properties of symmetry and the triangle inequality. The consensus in the cognitive psychology community is a definitive "no". In seminal work, Tversky [24] showed that human analysis of "similarity" is non-symmetric and is context dependent. One of the visual experiments conducted by Tversky was a simple pair-matching task, where subjects were asked if two block letters were the same or not. A similarity function $\mathcal{S}(p, q)$ indicated the frequency at which subjects noted letter p to be the same as q. The experiment showed that the order of presentation of the letters mattered in a statistically significant way: $\mathcal{S}(p, q) \neq \mathcal{S}(q, p)$. This result, along with others for matching faces, abstract symbols, and the names of countries led Tversky to conclude that "similarity is not necessarily a symmetric relation."

In the subsequent work, Tversky and Gati [25] examined if the triangle inequality (Fig. 1) is satisfied by humans when assessing similarity. Because the triangle inequality can always be satisfied by adding a large constant to the distances between individual points when measuring dissimilarity on an ordinal scale, Tversky and Gati proposed a test that assumes segmental additivity: $d(x, z) = d(x, y) + d(y, z)$. Over numerous pair-matching trials across stimuli, human similarity judgments were found to violate the triangle inequality in a statistically significant manner. Even without the triangle inequality for additive functions, it is still possible to induce metric models with subadditive metrics. However, in experiments where subjects provided subjective probability estimates instead of ordinal numbers, Tversky and Koehler [26] were only able to show that the reported scores are often, *but not always*, subadditive.[3]

Linking these findings back to pattern recognition, Duin [28,29] finds a similar effect for the problem of judging difference between real world objects, and highlights the need for a reconsideration of

[1] Note that without the property of identity, the theorem of NN convergence [3] does not hold. It has also been shown [15] that the optimal distance measure, in the sense of minimal Bayes risk, always violates the identity property and therefore is not metric.

[2] For example, consider support vector machines (SVM): one cannot draw a conceptual decision boundary without both positive and negative samples.

[3] It is possible to work around the constraint of segmental additivity using a subadditive metric based on Shepard's universal law of generalization to induce a metric from finite sets of data [27], but the result is still not consistent with the human perception findings of Tversky and Koehler [26].