Distance learning for similarity estimation

Yu, J.; Amores, J.; Sebe, N.; Radeva, P.; Tian, Q.

DOI
10.1109/TPAMI.2007.70714

Publication date
2008

Published in
IEEE Transactions on Pattern Analysis and Machine Intelligence

Distance Learning for Similarity Estimation

Jie Yu, *Member*, *IEEE*, Jaume Amores, Nicu Sebe, *Member*, *IEEE*, Petia Radeva, and
Qi Tian, *Senior Member*, *IEEE*

Abstract—In this paper, we present a general guideline to find a better distance measure for similarity estimation based on statistical analysis of distribution models and distance functions. A new set of distance measures are derived from the harmonic distance, the geometric distance, and their generalized variants according to the Maximum Likelihood theory. These measures can provide a more accurate feature model than the classical euclidean and Manhattan distances. We also find that the feature elements are often from heterogeneous sources that may have different influence on similarity estimation. Therefore, the assumption of single isotropic distribution model is often inappropriate. To alleviate this problem, we use a boosted distance measure framework that finds multiple distance measures, which fit the distribution of selected feature elements best for accurate similarity estimation. The new distance measures for similarity estimation are tested on two applications: stereo matching and motion tracking in video sequences. The performance of boosted distance measure is further evaluated on several benchmark data sets from the UCI repository and two image retrieval applications. In all the experiments, robust results are obtained based on the proposed methods.

Index Terms—Image classification, information retrieval, pattern recognition, artificial intelligence, algorithms.

1 Introduction

Similarity has been a research topic in the field of psychology for decades (see Wallach [1] and Tversky and Krantz [2]), but recently, there has been a huge resurgence in the topic. Similarity judgments are considered to be a valuable tool in the study of human perception and cognition and play a central role in the theories of human knowledge representation, behavior, and problem solving. Tversky [3] describes the similarity concept as "an organizing principle by which individuals classify objects, form concepts, and make generalizations."

1.1 Similarity Estimation in Image Retrieval

Retrieval of images by similarity, that is, retrieving images that are similar to an already retrieved image (retrieval by example) or to a model or schema, is a relatively old idea. Some might date it to antiquity, but more seriously, it appeared in specialized geographic information systems databases around 1980, in particular, in the "Query by Pictorial Example" system of IMAID [4]. From the start, it was clear that retrieval by similarity called for specific definitions of what it means to be similar. In the mapping system, a satellite image was matched to existing map images from the point of view of similarity of road and river networks, easily extracted from images by edge detection. Apart from theoretical models [5], it was only in the beginning of the 1990s that researchers started to look at retrieval by similarity in large sets of heterogeneous images with no specific model of their semantic contents. The prototype systems of Kato [6], followed by the availability of the QBIC commercial system using several types of similarities [7], contributed to making this idea more and more popular.

Typically, a system for retrieval by similarity rests on three components:

- Extraction of features or image signatures from the images and an efficient representation and storage strategy for this precomputed data.
- A set of similarity measures, each of which captures some perceptively meaningful definition of similarity and which should be efficiently computable when matching an example with the whole database.
- A user interface for the choice of which definition of similarity should be applied for retrieval, presentation of retrieved images, and supporting relevance feedback.

The research in the area has made the following evident:

- A large number of meaningful types of similarity can be defined. Only part of these definitions is associated with efficient feature extraction mechanisms and (dis)similarity measures.
- Since there are many definitions of similarity and the discriminating power of each of the measures is likely to degrade significantly for large image databases, the user interaction and the feature storage strategy components of the systems will play an important role.
- Visual content-based retrieval is best used when combined with the traditional search, both at the user interface and at the system level. The basic reason for this is that content-based retrieval is not seen as a

- *J. Yu is with the Intelligent Systems Group, Kodak Research Labs, 1999 Lake Ave., Mail Code: 02103, Rochester, NY 14615.*
E-mail: Jerry.J.Yu@gmail.com.
- *J. Amores is with the IMEDIA Research Group, Institut National de Recherche en Informatique et en Automatique (INRIA), France.*
E-mail: Jaume.Amores@inria.fr.
- *N. Sebe is with the Faculty of Science, University of Amsterdam, Room F 0.06, Kruislaan 403, 1098 SJ Amsterdam, The Netherlands.*
E-mail: nicu@science.uva.nl.
- *P. Radeva is with the Universitat Autònoma de Barcelona, Department of Informàtica, Computer Vision Center, Edifici O, 08193 Bellaterra (Barcelona), Catalunya, Spain. E-mail: petia@cvc.uab.es.*
- *Q. Tian is with the Department of Computer Science, University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249.*
E-mail: qitian@gmail.com.

Manuscript received 6 July 2006; revised 4 Jan. 2007; accepted 2 May 2007; published online 7 June 2007.
Recommended for acceptance by J. Luo.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0503-0706.
Digital Object Identifier no. 10.1109/TPAMI.2007.70714.

0162-8828/08/$25.00 © 2008 IEEE Published by the IEEE Computer Society

replacement of parametric (SQL), text, and keywords search. The key is to apply content-based retrieval where appropriate, which is typically where the use of text and keywords is suboptimal. Examples of such applications are where visual appearance (for example, color, texture, and motion) is the primary attribute as in stock photo/video, art, and so forth.

Gudivada and Raghavan [8] listed different possible types of similarity for retrieval: color similarity, texture similarity, shape similarity, spatial similarity, and so forth. Some of these types can be considered in all or only part of one image, can be considered independently of scale or angle or not, depending on whether one is interested in the scene represented by the image or in the image itself. Representation of features of images, such as color, texture, shape, motion, is a fundamental problem in visual information retrieval. Image analysis and pattern recognition algorithms provide the means to extract numeric descriptors that give a quantitative measure of these features. Computer vision enables object and motion identification by comparing extracted patterns with predefined models.

1.2 Distance Measure for Similarity Estimation

In many science and engineering fields, the similarity between two features is determined by computing the distance between them using a certain distance measure. In computer vision, as well as some other applications, the euclidean distance or sum of the squared differences (L_2—SSD) is one of the most widely used measures. However, it has been suggested that it is not appropriate for many problems [9]. From a maximum likelihood (ML) perspective, it is well known that the SSD is justified when the feature data distribution is Gaussian [10], whereas the Manhattan distance or sum of the absolute differences (L_1—SAD), another commonly used measure, is justified when the feature data distribution is Exponential (double or two-sided exponential). Therefore, which measure to use can be determined if the underlying data distribution is known or well estimated. The common assumption is that the real distribution should fit either the Gaussian or the Exponential. However, in many applications, this assumption is invalid. Finding a suitable distance measure becomes a challenging problem when the underlying distribution is unknown and could be neither Gaussian nor Exponential [10].

In content-based image retrieval [11], feature elements are extracted for different statistical properties associated with the entire digital images or perhaps with a specific region of interest. The heterogeneous sources suggest that the elements may be from different distributions. In previous work, most of the attention focused on extracting low-level feature elements such as color-histogram [12], wavelet-based texture [13], [14], and shape [15] with little or no consideration of their distributions. The most commonly used method for calculating the similarity between two feature vectors is still to compare the euclidean distance between them.

Although some work has been done to utilize the data model in similarity image retrieval [16], [17], [18], [10], the relation between the distribution model and the distance measure has not been fully studied yet. It has been justified that Gaussian, Exponential, and Cauchy distribution result in L_2, L_1, and Cauchy metrics, respectively. However, distance measures that fit other distribution models have not been studied yet. The similarity estimation based on feature elements from unknown distributions is an even more difficult problem.

In this paper, based on previous work [16], [17], [18], we propose a guideline to learn a robust distance measure for accurate similarity estimation. The novelty and contribution of our work lie on two folds. First, we study the relation of data distribution, distance function, and similarity estimation. We prove that the well-known euclidean and Manhattan distances are not the optimal choices when the data distribution is neither Gaussian nor Exponential. Further, our study on the relation between mean estimation and data distribution found a new set of distance measures. They correspond to a set of distributions that cannot be mathematically formulated and have not been reported in literature before. Our experiments show that these new distance measures perform better than traditional distance measures, which implies that the new distributions model the data better than the well-known Gaussian and Exponential distributions. Second, a boosted distance measure framework is used to automatically find the best distance functions from a set of given measures and choose the feature elements that are most useful for similarity estimation. It is especially robust to small sample set problem because the best measures are learned on each selected feature elements. Experimental results show the superior performance of the proposed method. It is also worth mentioning that arbitrary distance functions can be plugged into the boosting framework, which may provide more accurate similarity estimation.

The rest of this paper is organized as follows: Section 2 presents a distance measure analysis using the ML approach. Section 3 describes the boosted distance measure. In Section 4, we apply the new distance measures to estimate the similarity in a stereo matching application, motion tracking in a video sequence, and content-based image retrieval. Discussion and conclusions are given in Section 5.

2 Distance Measure Analysis

2.1 Maximum Likelihood Approach

The additive model is a widely used model in computer vision regarding ML estimation. Haralick and Shapiro [19] consider this model in defining the M-estimate: "Any estimate μ defined by a minimization problem of the form $\min \sum_i f(x_i - \mu)$ is called an M-estimate." Note that the operation "–" between the estimate and the real data implies an additive model. The variable μ is either the estimated mean of a distribution or, for simplicity, one of the samples from that distribution.

Maximum Likelihood theory [19] allows us to relate a data distribution to a distance measure. From the mathematical-statistical point of view, the problem of finding the right measure for the distance comes down to the maximization of the similarity probability.

We use image retrieval as an example for illustration. Consider first, the two subsets of N images from the database $(D) : \boldsymbol{X} \subset \mathrm{D}, \boldsymbol{Y} \subset \mathrm{D}$, which according to the ground truth are similar

$$\boldsymbol{X} \equiv \boldsymbol{Y} \text{ or } \boldsymbol{x}_i \equiv \boldsymbol{y}_i, i = 1, \ldots, N, \tag{1}$$

where $\boldsymbol{x}_i \in \boldsymbol{X}$ and $\boldsymbol{y}_i \in \boldsymbol{Y}$ represent the images from the corresponding subsets.

Equation (1) can be rewritten as

$$\boldsymbol{x}_i = \boldsymbol{y}_i + \boldsymbol{d}_i, i = 1, \ldots, N, \tag{2}$$

where $\boldsymbol{d}_i$ represents the "distance" image obtained as the difference between image $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$.

In this context, the similarity probability between two sets of images $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be defined

$$P(\boldsymbol{X},\boldsymbol{Y}) = \prod_{i=1}^{N} p(\boldsymbol{x}_i, \boldsymbol{y}_i), \tag{3}$$

where $p(\boldsymbol{x}, \boldsymbol{y})$ describes the similarity between images $\boldsymbol{x}$ and $\boldsymbol{y}$ (measured by the probability density function of the difference between $\boldsymbol{x}$ and $\boldsymbol{y}$). Independence across images is assumed. We define

$$f(\boldsymbol{x}_i, \boldsymbol{y}_i) = -\log p(\boldsymbol{x}_i, \boldsymbol{y}_i). \tag{4}$$

Equation (3) becomes

$$P(\boldsymbol{X},\boldsymbol{Y}) = \prod_{i=1}^{N} \{\exp[-f(\boldsymbol{x}_i, \boldsymbol{y}_i)]\}, \tag{5}$$

where the function f is the negative logarithm of the probability density function of $\boldsymbol{x}$ and $\boldsymbol{y}$.

According to (5), we have to find the function f that maximizes the similarity probability. This is the *Maximum Likelihood* estimator for $\boldsymbol{X}$, given $\boldsymbol{Y}$ [19].

Taking the logarithm of (5), we find that we have to minimize the expression

$$\sum_{i=1}^{N} f(\boldsymbol{x}_i, \boldsymbol{y}_i). \tag{6}$$

In our case, according to (2) the function f does not depend individually on its two arguments, query image $\boldsymbol{x}_i$, and the predicated one $\boldsymbol{y}_i$ but only on their difference. We have thus a local estimator, and we can use $f(\boldsymbol{d}_i)$ instead of $f(\boldsymbol{x}_i, \boldsymbol{y}_i)$, where $\boldsymbol{d}_i = \boldsymbol{x}_i - \boldsymbol{y}_i$ and the operation "–" denotes pixel-by-pixel difference between the images or an equivalent operation in feature space. Therefore, minimizing (6) is equivalent to minimizing

$$\sum_{i=1}^{N} f(\boldsymbol{d}_i). \tag{7}$$

Maximum Likelihood estimation shows a direct relation between the data distribution and the comparison measure. One can note that the Gaussian model is related to L_2 metric, whereas the Exponential model is related to L_1 metric and so is Cauchy metric, respectively [16], [10]. Note that although we consider images as an example, this notion can be extended to feature vectors associated with the images when we are working with image features or, even, can be extended to pixel values in the images.

2.2 Distance Measure Analysis

The Gaussian, Exponential, and Cauchy distribution models result in the L_2 metric, L_1 metric, and Cauchy metric, respectively [10]. It is reasonable to assume that there may be other distance measures that fit the unknown real distribution better. More accurate similarity estimation is expected if the measure could reflect the real distribution. We call this problem of finding the best distance measure *distance measure analysis*. It can be mathematically formulated as follows.

Suppose we have observations $x_i{}^1$ from a certain distribution

TABLE 1
Distance Measures and the Mean Estimation for Different Distributions

	Distance Measure	Mean Estimation
Arithmetic	$\varepsilon = \sum_{i=1}^{N}(x_i - \hat{\mu})^2$	$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$
Median	$\varepsilon = \sum_{i=1}^{N} \vert x_i - \hat{\mu} \vert$	$\hat{\mu} = med(x_1, \cdots, x_N)$
Harmonic	$\varepsilon = \sum_{i=1}^{N} x_i(\frac{\hat{\mu}}{x_i} - 1)^2$	$\hat{\mu} = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}}$
Geometric	$\varepsilon = \sum_{i=1}^{N} [\log(\frac{x_i}{\hat{\mu}})]^2$	$\hat{\mu} = (\prod_{i=1}^{N} x_i)^{\frac{1}{N}}$

$$x_i = \mu + d_i, \tag{8}$$

where d_i, $i = 1, \cdots, N$ are data components and μ is the distribution mean or a sample from the same class if it is considered as center of a subclass from a locality point of view. In most cases, μ is unknown and may be approximated for similarity estimation. For a distance function

$$f(x, \mu) \geq 0, \tag{9}$$

which satisfies the condition $f(\mu, \mu) = 0$, μ can be estimated by $\hat{\mu}$ which minimizes

$$\varepsilon = \sum_{i=1}^{N} f(x, \hat{\mu}). \tag{10}$$

It is equivalent to satisfy

$$\sum_{i=1}^{N} \frac{d}{d\hat{\mu}} f(x_i, \hat{\mu}) = 0. \tag{11}$$

For some specific distributions, the estimated mean $\hat{\mu} = g(x_1, x_2, \cdots, x_N)$ has a closed form solution. The arithmetic mean, median, harmonic mean, and geometric mean in Table 1 are in this category.

It is well known that the L_2 metric (or SSD) corresponds to the arithmetic mean, whereas the L_1 metric (or SAD) corresponds to the median. However, no literature has discussed the distance measures associated with the distribution models that imply the harmonic mean or the geometric mean. Those measures in Table 1 are inferred using (11). Fig. 1a illustrates the difference among the distance functions $f(x, \hat{\mu})$ for the arithmetic mean, median, harmonic mean, and geometric mean. For fair comparison, the value of μ is set to be 10 for all distributions. We found that in distributions associated with the harmonic and geometric estimations, the observations that are far from the correct estimate (μ) will contribute less in producing μ, as distinct from the arithmetic mean. In that case, the estimated values will be less sensitive to the bad observations (that is, observations with large variances), and they are therefore more robust.

2.3 Generalized Distance Measure Analysis

The robust property of harmonic and geometric distance measures motivates us to generalize them and come up with new measures that may fit the distribution better. Three families of distance measures in Table 2 are derived from the

[1] The observations can be in scalar or vector form. For simplicity, our derivations are based on the scalar form, but the extension to vector form is straightforward. Each vector element can be treated as a scalar.

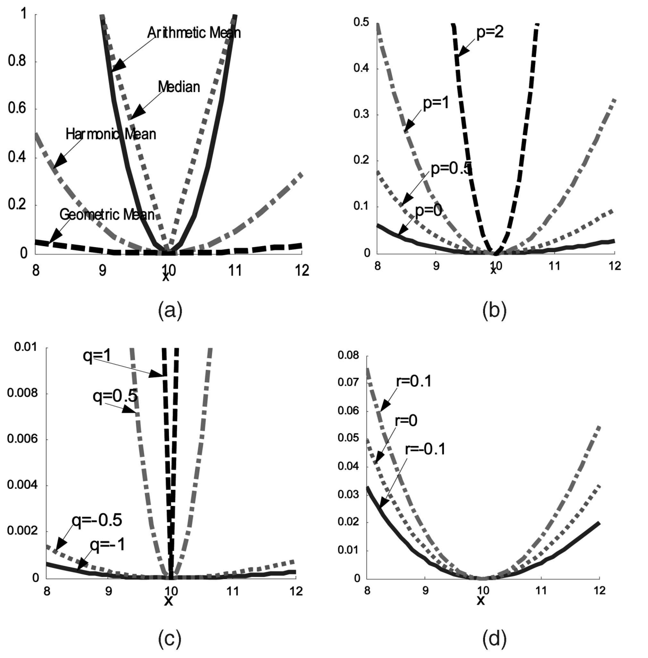


Fig. 1. The distance function $f(x, \mu)$ of (a) the arithmetic mean, median, harmonic mean, and geometric mean, (b) first-type, (c) second-type generalized harmonic mean, and (d) the generalized geometric mean (μ is fixed and set to 10).

TABLE 2
Generalized Distance Measures

Distance Family	Distance Measure	Mean Estimation
Generalized harmonic mean (1st type)	$\varepsilon = \sum_{i=1}^{N} (x_i)^p (\frac{\hat{\mu}}{x_i} - 1)^2$	$\hat{\mu} = \frac{\sum_{i=1}^{N} (x_i)^{p-1}}{\sum_{i=1}^{N} (x_i)^{p-2}}$
Generalized harmonic mean (2nd type)	$\varepsilon = \sum_{i=1}^{N} [(x_i)^q - (\hat{\mu})^q]^2$	$\hat{\mu} = [\frac{N}{\sum_{i=1}^{N} (x_i)^q}]^{-\frac{1}{q}}$
Generalized geometric mean	$\varepsilon = \sum_{i=1}^{N} [(x_i)^r \log(\frac{x_i}{\hat{\mu}})]^2$	$\hat{\mu} = [\prod_{i=1}^{N} (x_i)^{(x_i)^{2r}}]^{\frac{1}{\sum_{i=1}^{N} (x_i)}}$

generalized mean estimation using (10). The parameters p, q, and r define the specific distance measures and describe the corresponding distribution models that may not be explicitly formulated as Gaussian and Exponential. We found that in the generalized harmonic mean estimation, the first type is generalized based on the distance measure representation, whereas the second type is generalized based on the estimation representation. However, if $p = 1$ and $q = -1$, both types will become ordinary harmonic mean, and if $p = 2$ and $q = 1$, both types will become an arithmetic mean. As for the generalized geometric mean estimation, if $r = 0$, it will become an ordinary geometric mean. It is obvious that the generalized measures correspond to a wide range of mean estimations and distribution models. Figs. 1b, 1c, and 1d show the distance measure function $f(x, \hat{\mu})$ corresponding to the first type and second type generalized harmonic mean and the generalized geometric mean estimation, respectively. It should be noted that not all mean estimations have a closed-form solution, as in Tables 1 and 2. In that case, $\hat{\mu}$ can be estimated by numerical analysis, for example, greedy search of $\hat{\mu}$ to minimize ε.

3 Boosting Distance Measures for Similarity Estimation

3.1 Motivation

As mentioned in Section 1, the most commonly used distance measure is the euclidean distance that assumes that the data have a Gaussian isotropic distribution. When the feature space has a large number of dimensions, an isotropic assumption is often inappropriate. Besides, the feature elements are often extracted by different statistical approaches, and their distributions may not be the same and different distance measures may better reflect the distributions. Thus, an anisotropic and heterogeneous distance measure may be more suitable for estimating the similarity between features.

Mahalanobis distance $(\boldsymbol{x}_i - \boldsymbol{y}_i)^T W (\boldsymbol{x}_i - \boldsymbol{y}_i)$ is one of the traditional anisotropic distances. It tries to find the optimal estimation of the weight matrix W. It is worth noting that it assumes that the underlying distribution is Gaussian, which is often not true. Furthermore, if k is the number of dimensions, the matrix W contains k^2 parameters to be estimated, which may not be robust when the training set is small compared to the number of dimensions. Classical techniques such as Principal Component Analysis (PCA) [20] or Linear Discriminant Analysis (LDA) [21] may be applied to reduce the dimensions. However, these methods cannot solve the problems of a small training set, and they also assume Gaussian distribution.

3.2 Boosted Distance Measures

Based on the analysis in Section 3.1, we use a boosted distance measure for similarity estimation, where the similarity function can be estimated by a generalization of different distance measures on selected feature elements. In particular, we use AdaBoost with decision stumps [22] and our distance measure analysis to estimate the similarity. Given a training set with feature vectors $\boldsymbol{x}_i$, the similarity estimation is done by training AdaBoost with differences $\boldsymbol{d}$ between vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, where each difference vector $\boldsymbol{d}$ has an associated label l_d:[2]

$$l_d = \begin{cases} 1 \text{ if } \boldsymbol{x}_\text{i} \text{ and } \boldsymbol{x}_\text{j} \text{ are from same class} \\ 0 \text{ otherwise.} \end{cases} \tag{12}$$

A weak classifier is defined by a distance measure m on a feature element f with estimated parameter(s) θ, which could be as simple as the mean and/or a threshold. The label prediction of the weak classifier on feature difference $\boldsymbol{d}$ is $h_{\text{m,f},\theta}(\boldsymbol{d}) \in \{0, 1\}$.

In this paper, for simplicity, the weak classifier learns two parameters as θ: $\text{p} \in \{-1, 1\}$ and *threshold*. The prediction is defined as follows:

[2] The elements of difference vector $\boldsymbol{d}$ between vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ can be measured by different metrics, for example, euclidean distance $\boldsymbol{d} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2$ or Manhattan distance $\boldsymbol{d} = |\boldsymbol{x}_i - \boldsymbol{x}_j|$.

$$h(\boldsymbol{d}) = \begin{cases} 1 \text{ if } d_f \cdot p < threshold \cdot p \\ 0 \text{ otherwise.} \end{cases} \tag{13}$$

The boosted distance measure $H(\boldsymbol{d})$ is learned by weighted training with different distance measures on each feature element and by selecting the most important feature elements for similarity estimation iteratively. Consequently, we derive a predicted similarity $S(\boldsymbol{x}, \boldsymbol{y}) = H(\boldsymbol{d})$ that is optimal in a classification context. The brief algorithm is listed below.

Please note that the resulting similarity $S(\boldsymbol{x}, \boldsymbol{y})$ may not be a traditional metric. It does not have the metric properties such as symmetry and triangular inequality. However, it is not necessarily a disadvantage because the proposed application of similarity estimation does not rely on the metric properties. Indeed, nonmetric distances measure can be more accurate for comparing complex objects, as have been studied recently in [23].

```
Boosting Distance Measure Algorithm
  Given: Pairwise difference vector set $D$ and the
         corresponding label $L$
         Number of iterations $T$
         Weak classifiers based on each distance measure $m$
         for each feature element $FE$
  Initialization: weight $w_{i,t=1} = 1/|D|$, $i = 1, \dots, D$
  Boosting:
  For $t = 1, \dots, T$
      • Train the weak classifier on the weighted sample set.
      • Select the best weak classifier giving the smallest error
        rate:
            $\varepsilon_t = \min_{m,FE,\theta} \sum_i w_{i,t} |h_{m,FE,\theta}(\boldsymbol{d}_i) - l_i|$.
      • Let $h_t = h_{m_t,FE_t,\theta_t}$ with $m_t$, $FE_t$, $\theta_t$ minimizing error
        rate.
      • Compute the weights of classifiers $(\alpha_t)$ based on its
        classification error rate:
      Let $\beta_t = \frac{\varepsilon_t}{1-\varepsilon_t}$, $\alpha_t = \frac{1}{\log(\beta_t).}$
      • Update and normalize the weight for each sample:
            $w_{i,t+1} = w_{i,t} \beta_t^{1-|h_{t,i}-l_i|}$
            $w_{i,t+1} = w_{i,t+1} / \sum_i w_{i,t+1}$.
      end for $t$
  Final prediction $H(\boldsymbol{d}) = \sum_t \alpha_t h_t(\boldsymbol{d})$.
```

The method has three main advantages: 1) the similarity estimation uses only a small set of elements that is most useful for similarity estimation, 2) for each element, the distance measure that best fits its distribution is learned, and 3) it adds effectiveness and robustness to the classifier when we have a small training set compared to the number of dimensions.

Because the feature elements may be from different sources, they may be modeled as different distributions. Actually, the correlation of distribution is very difficult to be mathematically modeled even if we assume the same distribution for different features as in Relevant Component Analysis (RCA) and [24]. The boosting scheme alleviates that problem because the feature elements selected have complimentary properties in similarity estimation and, consequently, the correlation among the selected feature elements should be low. Furthermore, since the training iteration T is usually much less than the original data dimension, the boosted distance measure works as a nonlinear dimension reduction technique similar to Viola and Jones [25], which keeps the most important elements to similarity judgment. It could be very helpful to overcome the small sample set problem [26]. It is worth mentioning that the proposed method is general and can be plugged into many similarity estimation techniques, such as widely used K-NN [27].

Compared with other distance measures proposed for K-NN, the boosted similarity is especially suitable when the training set is small. Two factors contribute to this. First, if N is the size of the original training set, this is augmented by using a new training set with $O(N^2)$ relations between vectors. This makes AdaBoost more robust against overfitting. Second, AdaBoost complements K-NN by providing an optimal similarity. Increasing the effectiveness for small training sets is necessary in many real classification problems, and in particular, it is necessary in applications such as retrieval where the user provides a small training set online.

3.3 Related work

We notice that there have been several works on estimating the distance to solve certain pattern recognition problems. Domeniconi et al. [28] and Peng et al. [29] propose specific estimations designed for the K-NN classifier. They obtain an anisotropic distance based on local neighborhoods that are narrower along relevant dimensions and more elongated along nonrelevant ones. Xing et al. [30] propose estimating the matrix W of a Mahalanobis distance by solving a convex optimization problem. They apply the resulting distance to improve the K-means behavior. Bar-Hillel et al. [31] also use a weight matrix W to estimate the distance by RCA. They improve the Gaussian Mixture EM algorithm by applying the estimated distance along with equivalence constraints.

The work by Athitsos et al. [32] and Hertz et al. [33] resemble the boosting part of our method, although their approach is conceptually different. Athitsos et al. [32] proposed a method called BoostMap to estimate a distance that approximates a certain distance, for example, EMD distance by Ruber et al. [46] but with a much smaller computational cost. Our method does not approximate or emulate any given distance. What we want to do is to learn a new distance function that is accurate for our problem. Hertz's work [33] uses AdaBoost to estimate a distance function in a product space (with pairs of vectors), whereas the weak classifier minimizes an error in the original feature space. Therefore, the weak classifier minimizes a different error than the one minimized by the strong classifier AdaBoost. In contrast, our framework utilizes AdaBoost with weak classifiers that minimize the same error as AdaBoost and in the same space. Apart from this conceptual difference, Hertz et al. [33] use Expectation-Maximization of Gaussian Mixture as a weak classifier, where they assume that the data have a Gaussian Mixture distribution and estimate several covariance matrices, which may not work well when the real distribution is not Gaussian or the training set is small compared to the dimensionality of the data.

4 Experiments and Analysis

4.1 Distance Measure Analysis in Stereo Matching

Stereo matching is the process of finding correspondences between entities in images with overlapping scene content. The images are typically taken from cameras at different viewpoints, which imply that the intensity of corresponding pixels may not be the same.

In stereo data sets, the ground truth for matching corresponding points may be provided by the laboratory where these images were taken. This ground truth is a result of

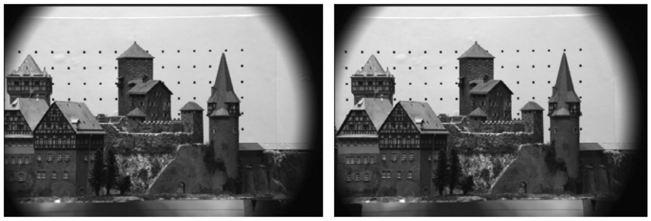


Fig. 2. A stereo image pair from the Castle data set.

mapping the world coordinates, in which the camera is moving to the image coordinates, using the three-dimensional geometry relations of the scene. In this case, an automatic stereo matcher, which is able to detect the corresponding point pairs registered in the stereo images of the test set scenes can be tested. For this stereo matcher, it is possible to determine the best measure when comparing different image regions to find the similar ones. The optimum measure in this case will give the most accurate stereo matcher.

We use two standard stereo data sets (Castle set and Tower set) provided by Carnegie Mellon University. These data sets contain multiple images of static scenes with accurate information about object locations in 3D. The images are taken with a scientific camera in an indoor setting, the Calibrated Imaging Laboratory at CMU. The 3D locations are given in $X - Y - Z$ coordinates with a simple text description (at best accurate to 0.3 mm), and the corresponding image coordinates (the ground truth) are provided for all 11 images taken for each scene. For each image, there are 28 points provided as ground truth in the Castle set and 18 points in the Tower set. An example of two stereo images from the Castle data set is given in Fig. 2.

In each of the images, we consider the points, which are given by the ground truth, and we want to find the proper similarity estimation, which will ensure the best accuracy in finding the corresponding points according to the ground truth.

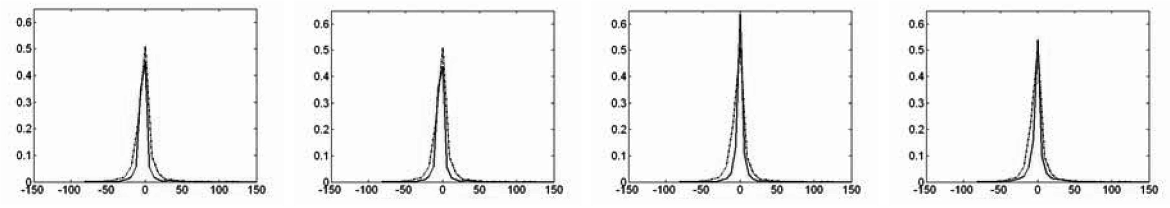
We cannot use a single pixel information but have to use a region around it, so we will perform template matching. Our automatic stereo matcher will match a template defined around one point from an image with the templates around points in the other images to find similar ones. If the resulting points are equivalent to those provided by the ground truth, we consider that we have a *hit*; otherwise, we have a *miss*. The accuracy is given by the number of the hits divided by the number of possible hits (number of corresponding point pairs). Because the ground truth is provided with subpixel accuracy, we consider that we have a hit when the corresponding point lies in the neighborhood of one pixel around the point provided by the ground truth. Our intention is to try distance measures other than SSD, that is, L_2, (which is used in the original algorithms) in calculating the disparity map. The algorithm is described in the following:

1. Obtain the ground truth similarity distance distribution A template of size 5×5 is applied around each ground truth point (that is, 28 points for each image), and the real distance is obtained by calculating the difference of pixel intensities within the template between sequential frames, which is the difference between frame 2 and frame 1, frame 3 and frame 2, and so on.
2. Obtain the estimated similarity using distance measure analysis:
    a. Given the 28 ground truth points in one frame, say, frame k, the template matching centered at a ground-truth point is applied to find its corresponding point in frame $k+1$.
        - To find the corresponding point in frame $k+1$, we search a band centered at the row coordinate of the pixel provided by the test frame k with a height of seven pixels and width equal to the image dimension. The template size is 5×5.
        - The corresponding point is determined to minimize the quantity of *distance*, which is defined by distance measures. For example, the distance under L_1 metric is the summed as the absolute difference between the intensity of each pixel in the template and that in the searching area, that is, $\sum_{i=1}^{25} |x_{i,k+1} - x_{i,k}|$, and the distance under L_2 is the summed squared difference between each pixel intensity in the template and that in the searching area, that is, $\sum_{i=1}^{25} (x_{i,k+1} - x_{i,k})^2$. For other distance measures, ε, see Tables 1 and 2.
    b. Apply the template centered at the ground-truth point in frame k and its tracked point in frame $k+1$ to calculate pixel intensity difference as the estimated similarity measurement.
3. Apply the Chi-square test [34].

The estimated distance and the real distance are compared using Chi-square test.

DistBoost [33] and the Boosted Distance are also tested for comparison. Note that Chi-square test cannot be applied on these two techniques. For the parameterized measures, we should choose the parameter value that minimizes the Chi-square test. As our first attempt, the parameters of p, q, and r are tested in the range of -5 to 5 with step size 0.1. Two thirds of the reference points pairs are randomly selected for training, and the rest are used for testing.

Fig. 3 shows the real distance distribution and the estimated distance distribution for the distance measures on the Castle data set. Both the solid and dashed curves are sampled with 233 points at equal intervals. The Chi-square test value is shown for each measure in Table 3. The smaller the Chi-square test value, the closer the estimation is to the real distribution. The generalized geometric mean measure has the best fit to the measured distance distribution. Therefore, the accuracy should be the greatest when using the generalized geometric mean measure (Table 3). In all cases, the hit rate for the generalized geometric mean $(r = 1.5)$ is 80.4 percent, and the hit rate for the Cauchy measure is 78.9 percent. The hit rates obtained with L_1 and L_2 are both 78.2 percent. The Cauchy measure performs better than both L_1 and L_2. It should be noted here that the Chi-square test score is not exactly in the same order of the hit rate though the winner is consistent in both cases. This is because the ground truth is provided with subpixel accuracy during the data collection process, and we consider that it is a hit when the corresponding point lies in the neighborhood of one pixel around the point provided by the ground truth. The inconsistency introduced by this rounding distance may explain the observation (not in the exact order for both measures). The boosted measure outperformed the



(a) (b) (c) (d)

Fig. 3. The real distance distribution (dashed line) versus the estimated distance distribution (solid line) for the Castle data set. (a) L_1 (0.0366). (b) L_2 (0.0378). (c) Cauchy $a = 17$ (0.0295). (d) Generalized geometric $r = 1.5$ (0.0239).

DistBoost. Similar results were obtained for the Tower set, and they are not shown here.

To evaluate the performance of the stereo matching algorithm under difficult matching conditions, we also use the ROBOTS stereo pair [35]. This stereo pair is more difficult due to varying levels of depth and occlusions (Fig. 4). For this stereo pair, the ground truth consists of 1,276 points pairs with one pixel accuracy. Two thirds of the reference points pairs are randomly selected for training, and the rest are used for testing.

Consider a point in the left image given by the ground truth. The disparity map gives the displacement of the corresponding point position in the right image. The accuracy is given by the percentage of pixels in the test set, which the algorithm matches correctly. Table 4 shows the accuracy of the algorithms when different distance measures are used. Note that the accuracy is lower using the ROBOTS stereo pair, showing that, in this case, the matching conditions are more difficult. However, still, the second-type generalized harmonic mean with $q = 4.1$ gives the best result. The Cauchy measure still performs better than L_1 and L_2, and this observation is consistent with [10]. Our best single distance measure even outperformed the learning-based DistBoost, whereas the Boosted Distance measure outperformed all other distance measures. It is worth mentioning that our improvement for the stereo matching experiments is relatively small. We believe that the cost of searching for the better measure is small, and our approach could give an even larger improvement on other test sets.

TABLE 3
The Accuracy (Percent) of the Stereo Matcher on the Castle Set (Best Parameter Is Shown)

Distance Measure	Chi-square test	Hit rate (%)
L_1	0.0366	78.2
L_2	0.0378	78.2
Cauchy [11]	0.0295 ($\partial = 17$)	78.9 ($\partial = 17$)
Harmonic mean	0.0273	78.2
Geometric mean	0.0378	77.1
1st-type generalized harmonic mean (1st-gh)	0.0328 ($p = 1.5$)	78.2 ($p = 1.5$)
2nd-type generalized harmonic mean (2nd-gh)	0.0272 ($q = 1.6$)	78.6 ($q = 1.6$)
Generalized geometric mean (gg)	0.0239 ($r = 1.5$)	80.4 ($r = 1.5$)
DistBoost [30]	N/A	84.3
Boosted Distance	N/A	**89.7**

4.2 Distance Measure Analysis in Motion Tracking

In this experiment, the distance measure analysis is tested on a motion tracking application. We use a video sequence containing 19 images on a moving head in a static background [36]. For each image in this video sequence, there are 14 points given as a ground truth. The motion tracking algorithm between the test frame and another frame performs template matching to find the best match in a 5×5 template around a central pixel. In searching for the corresponding pixel, we examine a region of width and the height of seven pixels centered at the position of the pixel in the test frame. The idea of this experiment is to trace moving facial expressions. Therefore, the ground truth points are provided around the lips and the eyes, which are moving through the sequences.

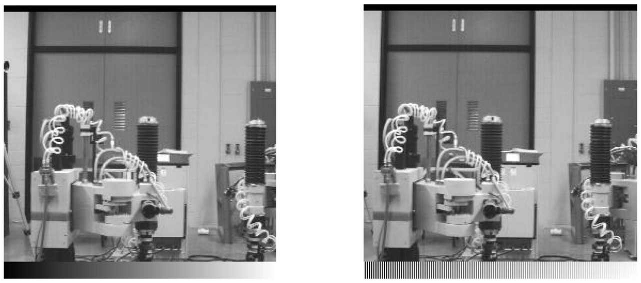
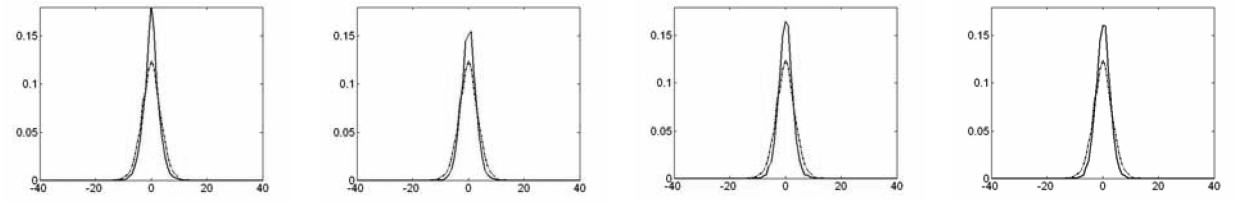


Fig. 4. ROBOTS stereo pair.

TABLE 4
The Accuracy (Percent) of the Stereo Matcher on the Robots Stereo Pair (Best Parameter Is Shown)

Distance Measure	Chi-square test	Hit rate (%)
L_1	0.0399	61.20
L_2	0.0481	59.60
Cauchy	0.0392 ($a = 1.3$)	62.80 ($a = 1.3$)
Harmonic mean	0.0782	58.40
Geometric mean	0.0319	54.50
1st-type generalized harmonic mean (1st-gh)	0.0340 ($p = 4.7$)	60.40 ($p = 4.7$)
2nd-type generalized harmonic mean (2nd-gh)	0.0201 ($q = 4.1$)	65.60 ($q = 4.1$)
Generalized geometric mean (gg)	0.0511 ($r = -4.3$)	58.00 ($r = -4.3$)
DistBoost [30]	N/A	64.32
Boosted Distance	N/A	**71.31**



(a) (b) (c) (d)

Fig. 5. The real data distribution (dashed line) versus the estimated data distribution (solid line) for motion tracking. (a) L_1 (0.0997). (b) L_2 (0.0765). (c) Cauchy $a = 7.1$ (0.0790). (d) Generalized geometric mean with $r = 7.0$ (0.0712).

In Fig. 5, we display the fit between the real data distribution and the four distance measures. The real data distribution is calculated using the template around points in the ground truth data set considering sequential frames. The best fit is the generalized geometric mean measure with $r = 7.0$.

Between the first frame and a later frame, the tracking distance represents the average template matching results. Fig. 6 shows the average tracking distance of the different distance measures. The generalized geometric mean measure with $r = 7.0$ performs best, whereas Cauchy measure outperforms both L_1 and L_2.

4.3 Boosted Distance Measure in Image Retrieval

As we discussed in Section 3.2, the boosted distance measure performs an element selection that is highly discriminant for similarity estimation, and it does not suffer from the small sample set problem as LDA and other dimension reduction techniques. To evaluate the performance, we tested the boosted distance measure on image classification against some state-of-the-art dimension reduction techniques: PCA, LDA, Nonparametric Discriminant Analysis (NDA) [37], and plain euclidean distance in the original feature space.

The two data sets we used are 1) a subset of the MNIST data set [38], containing similar handwritten 1s and 7s (Fig. 7a) and a gender recognition database containing facial images from the AR database [39] and the XM2TVS database [40] (Fig. 7b). Raw pixel intensity is used as feature elements. Using raw pixels is just a simple form of representation and is considered valid in this case because the object appears aligned in the image. Similar data representation has been used in other research work, for example, face retrieval work by Moghaddam et al. [26]. It is noted that our method can be applied to arbitrary features used in CBIR applications. We could use any other form of representation because the method does not depend on this particular choice. The dimension of the feature for both databases is 784, whereas the size of the training set is fixed at 200, which is small compared to the dimensionality of the feature. In such a circumstance, selecting an appropriate feature element is very important. In our previous study on face recognition, we found that it is difficult for classic techniques such as PCA, LDA, and Fisherface (PCA + LDA) [41]. In this experiment, the difference measure m is fixed as L_1, that is, $\boldsymbol{d} = \boldsymbol{x}_i - \boldsymbol{x}_j$ for simplicity. It will be easily extended to different measures by feeding difference d obtained with different measures such as euclidean distance in the next experiment. Nearest-neighbor classifier is used in the reduced dimension space.

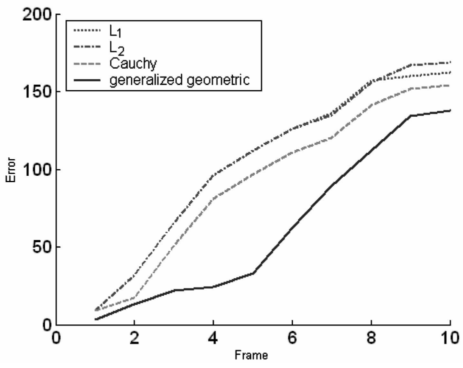
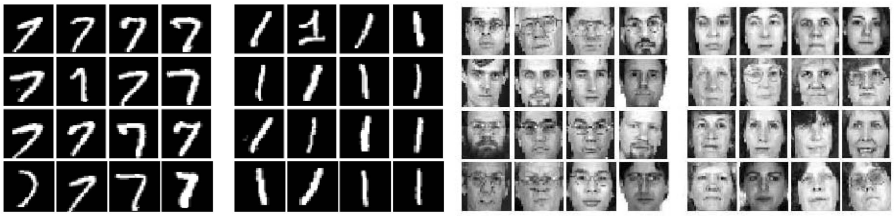


Fig. 6. Average tracking distance of the corresponding points in successive frames; for Cauchy, $a = 7.1$, and for generalized geometric mean, $r = 7.0$.

Fig. 8 shows the classification accuracy versus the projected dimension, which, for our boosted distance measure, is the number of iterations T. Because of the small sample problem, the accuracy of LDA is poor, at 50 percent and 49.9 percent, and is not shown in the figure. A simple regularization scheme can improve its performance but LDA still remains much worse than other techniques. It is clear that the traditional methods perform poorly due to the fact that we use a very small training set compared to the dimensionality of the data. Note that all traditional methods rely on estimating a covariance or scatter matrix with k^2 elements, where k is the number of dimensions. Empirical experience suggests that we need a training set of size greater than $3k^2$ to obtain a robust estimation of k^2 parameters. However, our boosted distance measure needs to estimate only a very few parameters on each dimension, which provides a robust performance on the small training set and makes it outperform the well-known techniques.

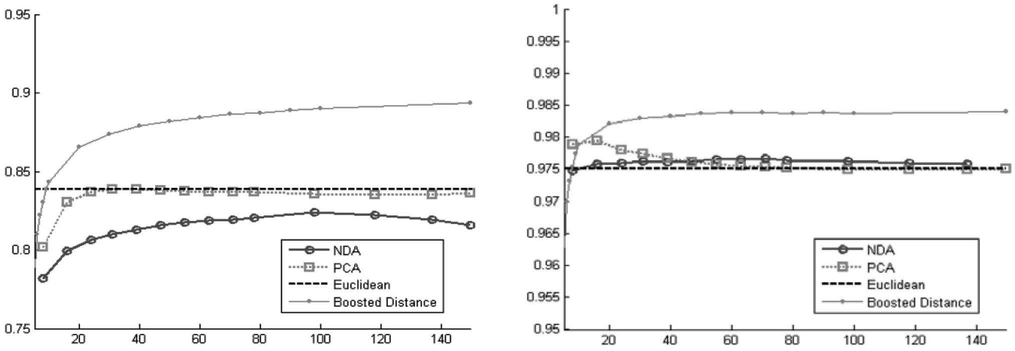
4.4 Boosted Distance Measure on Benchmark Data Set

In this section, we compare the performance of our approach by boosting multiple distance measures with boosting single measure and several well-known traditional approaches. The experiment is conducted on 13 benchmark data set from the University of California, Irvine (UCI) [42] and two data sets used in the third experiment: gender and written digits recognition. For the benchmark data set tests, we used 20 percent of the data as training set and 80 percent as testing set. The traditional distance measures we tested are euclidean Distance, Manhatan Distance, RCA distance [31], Mahalanobis distance with the same covariance matrix for all the classes *(Mah)*, and Mahalanobis with a different covariance matrix for every class *(Mah-C)*. The last three measures are sensitive to small sample set problem. A diagonal matrix D could be estimated instead of original weight matrix W to simplify that problem, and consequently, we can obtain three measures *RCA-D*, *Mah-D*, and *Mah-CD*, respectively. To make the comparison complete, we also test original AdaBoost with



(a) (b)

Fig. 7. Example images from (a) handwritten digits and (b) gender recognition.



(a) (b)

Fig. 8. Accuracy of classification on (a) gender recognition (b) and written digits.

TABLE 5
Comparison to Traditional Distance Measure and AdaBoost on UCI Data Sets

Error Rate (%)	Traditional Measure	AdaBoost +d.s.	AdaBoost +C4.5	*B. L_1*	*B. L_2*	Boosted Multiple Measures
Ad	17.31 (L_1)	12	11.42	9.35	10.23	**8.92**
gender	15.38 (L_1)	12.27	11.89	10.67	11.54	**10.57**
Mnist	2.34 (RCA-D)	2.22	2.14	1.82	1.63	**1.32**
arrhythmia	37.02 (RCA-D)	31.39	29.94	26.68	25.83	**25.62**
Splice	10.55 (Mah-D)	5.94	4.84	5.02	5.87	**4.61**
sonar	26.1 (Mah-CD)	25.95	25.81	25.74	27.72	**25.35**
spectf	31.16 (Mah-D)	28.65	27.18	27.03	**25.93**	26.2
ionosphere	**10.78** (RCA)	19.92	19.92	17.83	18.78	17.35
Wdbc	6.83 (Mah-CD)	5.81	5.37	4.96	4.49	**4.32**
german	38.74 (Mah-D)	34.31	33.18	32.21	33.76	**31.6**
vote1	9.07 (L_1)	6.37	6.37	6.98	6.91	**6.18**
credit	19.18 (Mah-CD)	17.97	**17.21**	18.39	17.86	17.63
Wbc	5.25 (RCA)	5.7	5.34	4.89	5.78	**4.23**
Pima	34.55 (Mah-CD)	31.02	29.96	29.58	31.24	**28.12**
Liver	41.11 (Mah)	35.51	35.43	33.26	35.69	**32.77**

decision stump *(d.s.)*, C4.5 [43], boosted $L_1(B.L_1)$ and $L_2(B.L_2)$. The AdaBoost C4.5 decision tree is implemented in the Matlab Classification Toolbox [44]. Due to the space limitation, only the traditional distance measure that gives the best performance in each data set is shown in Table 5. The smallest error rates are highlighted in bold.

From the results in Table 5, we can find that the boosted multiple distance measures performs the best in 12 out of 15 data sets. It provides comparable results to the best performance on two data sets (spectf and credit). Only in one data set (ionosphere), our method is outperformed by the traditional distance measure. It proves that the method could discover the best distance measure that reflects the distribution and selects the feature elements that are discriminant in similarity estimation. It is worth mentioning that our framework does not consider correlation between feature elements explicitly as other distance measures such as the Mahalanobis distance do. However, the boosting process will rarely select features that are strongly correlated to each other. This is because, at each round, boosting selects one feature that provides information not included in the already selected ones. Therefore, the estimated distance is based on features complementary to each other. On the other hand, traditional methods such as the Mahalanobis, used for considering feature correlation are not robust in spaces of high dimensionality and small number of training objects, as shown in the results.

5 Discussions and Conclusions

This paper presents a comprehensive analysis on distance measure and boosting heterogeneous measure for similarity estimation. Our study shows that learning the similarity measure is an important step (mostly ignored by the existing literature) for many computer vision applications. The main contribution of our work is to provide a general guideline for designing a robust distance estimation that could adapt data distributions automatically. Novel distance measures deriving from harmonic, geometric mean, and their generalized forms are presented and discussed. We examined the new measures for several applications in computer vision, and the estimation of similarity can be significantly improved by the proposed distance measure analysis.

The relationships between probabilistic data models, distance measures, and ML estimators have been widely studied. The creative component of our work is to start from an estimator and perform reverse engineering to obtain a measure. In this context, the fact that some of the proposed measures cannot be translated into a known probabilistic model is both a curse and a blessing. A curse, because it is really not clear what the underlying probabilistic models are (they certainly do not come from any canonical family), and this is usually the point at which one starts. After all, the connection between the three quantities (metric, data model, and ML estimator) is probabilistic. It is a bit unsettling to have no idea of what these models are. It is a blessing because this is probably the reason why these measures have not been previously proposed. However, they seem to work very well according to the experimental result in this paper.

In similarity, estimation of the feature elements are often from heterogeneous sources. The assumption that the feature has a unified isotropic distribution is invalid. Unlike a traditional anisotropic distance measure, our proposed method does not make any assumptions on the feature distribution. Instead, it learns the distance measure for each element to capture the underlying feature structure. Because the distance measure is trained on the observations of each element, the boosted distance does not suffer from the small sample set problem. Considering that not all feature elements are related to the similarity estimation, the boosting process in the proposed method provides a good generalization of the feature elements that are most important in a classification context. It also has a dimension reduction effect, which may be very useful when the original feature dimension is high. The automatic measure adaptation and element selection in the boosted distance measure bridge the gap between the high-level similarity concept and low-level features. With this approach, we guarantee that the measure factor is filtered out (it is optimized), and the user can concentrate on getting better features for improving the matching. Another nice feature of the approach is that it can be applied to a wide variety of algorithms, and it is not dependent on them. We tested the method on image retrieval, stereo matching, and motion tracking applications to show that the approach can be applied to a wide variety of computer vision algorithms, which pose different challenges. The experimental results show that the boosted measure is more effective than traditional distance measures.

In the future, we would like to continue this research work in the following directions:

1. studying the correlation between feature elements and formulating them mathematically,
2. incorporating our new measure into state-of-the-art classification techniques,
3. comparing with different feature selection schemes and evaluating the performance improvement (comparison with two feature selection schemes are reported in Appendix A), and
4. using a larger set of distance measures [47] to further enhance the estimation accuracy.

Appendix A

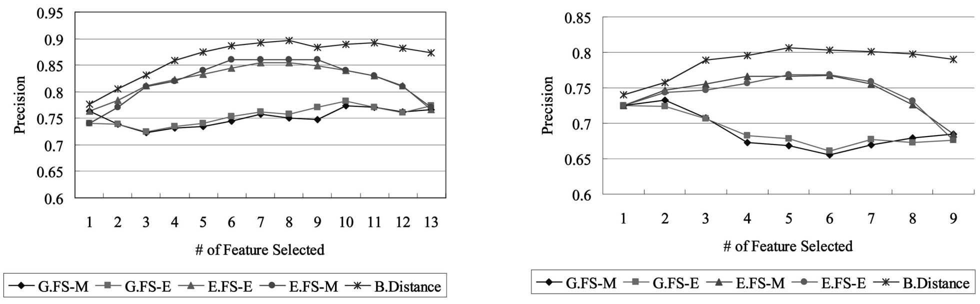
Comparison with Two Feature Selection Schemes

In our framework, feature selection is performed by boosting and decision stumps, which are classical methods. The main purpose of the paper is not to introduce a new feature selection algorithm. Nevertheless, we compare our boosted distance with the following two feature selection schemes to show the strength of our method: Greedy Feature Selection (G.FS) first ranks the individual performance of each element. Then, it selects the best K feature elements, where K is the feature set size. We construct two classifiers for G.FS method using euclidean (G.FS-E) and Manhattan (G.FS-M) distance, respectively. Exhaustive Feature Selection (E.FS) compares all the combination of feature elements with a fixed feature set size and chooses the element combination that gives the best performance. Similarly, we use euclidean and Manhattan metrics to construct two classifiers, E.FS-E and E.FS-M, respectively. Although E.FS is theoretically better than or equivalent to G.FS, its computational cost is high due to the exhaustive testing process, which makes it infeasible for data set with large dimensionality. The data sets we used are Heart and Breast-Cancer from UCI repository [42]. The dimensionalities data are 13 and 9, respectively. Two thirds of the data are used for training and the rest for testing.

The results in Fig. 9 suggest that the boosted distance outperforms both feature selection schemes. The reasons lie in twofold: First, the new distance measures fit the data better than traditional euclidean and Manhattan distances. Second, the weighted feature combination makes the performance of boosted distance less sensitive to the size of the feature set. Note that using boosting and decision stumps as feature selection has been reported in the literature [25]. Our contribution is to integrate this approach into a distance estimation framework.

Acknowledgments

A short version of this manuscript was published in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, New York, 2006. This work was supported in part by the US Army Research Office (ARO) grant under W911NF-05-1-0404, and by the Center of Infrastructure Assurance and Security (CIAS), the University of Texas at San Antonio. J. Yu's work was done while he was with the University of Texas at San Antonio.



(a) (b)

Fig. 9. Comparison of boosted distance with two feature selection schemes. (a) Heart data set. (b) Breast-cancer data set.

REFERENCES

[1] M. Wallach, "On Psychological Similarity," *Psychological Rev.,* vol. 65, no. 2, pp. 103-116, 1958.

[2] A. Tversky and D.H. Krantz, "The Dimensional Representation and the Metric Structure of Similarity Data," *J. Math. Psychology,* vol. 7, pp. 572-597, 1977.

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

[4] N.S. Chang and K.S. Fu, "Query by Pictorial Example," *IEEE Trans. Software Eng.,* vol. 6, no. 6, pp. 519-524, June 1980.

[5] P. Aigrain, "Organizing Image Banks for Visual Access: Model and Techniques," *Proc. Int'l Meeting for Optical Publishing and Storage,* pp. 257-270, 1987.

[6] K. Kato, "Database Architecture for Content-Based Image Retrieval," *Proc. SPIE Conf. Image Storage and Retrieval Systems,* vol. 1662, pp. 112-123, 1992.

[7] M. Flicker, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by Image and Video Content: The QBIC System," *Computer,* vol. 28, no. 9, pp. 23-32, Sept. 1995.

[8] V.N. Gudivada and V. Raghavan, "Design and Evaluation of Algorithms for Image Retrieval by Spatial Similarity," *ACM Trans. Information Systems,* vol. 13, no. 2, pp. 115-144, 1995.

[9] M. Zakai, "General Distance Criteria," *IEEE Trans. Information Theory,* pp. 94-95, Jan. 1964.

[10] N. Sebe, M.S. Lew, and D.P. Huijsmans, "Toward Improved Ranking Metrics," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 10, pp. 1132-1143, Oct. 2000.

[11] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-Based Image Retrieval at the End of the Early Years," *IEEE Trans. Pattern Analysis Machine Intelligence,* vol. 22, no. 12, pp. 1349-1380, Dec. 2000.

[12] M. Swain and D. Ballard, "Color Indexing," *Int'l J. Computer Vision,* vol. 7, no. 1, pp. 11-32, 1991.

[13] R.M. Haralick, K. Shanmugam, and I. Dinstein, "Texture Features for Image Classification," *IEEE Trans. Systems, Man, and Cybernetics,* vol. 3, no. 6, pp. 610-621, 1973.

[14] J.R. Smith and S.F. Chang, "Transform Features for Texture Classification and Discrimination in Large Image Database," *Proc. IEEE Int'l Conf. Image Processing,* 1994.

[15] B.M. Mehtre, M. Kankanhalli, and W.F. Lee, "Shape Measures for Content Based Image Retrieval: A Comparison," *Information Processing Management,* vol. 33, no. 3, pp. 319-337, 1997.

[16] Q. Tian, Q. Xue, J. Yu, N. Sebe, and T.S. Huang, "Toward an Improved Error Metric," *Proc. IEEE Int'l Conf. Image Processing,* Oct. 2004.

[17] J. Amores, N. Sebe, and P. Radeva, "Boosting the Distance Estimation: Application to the K-Nearest Neighbor Classifier," *Pattern Recognition Letters,* vol. 27, no. 3, pp. 201-209, Feb. 2006.

[18] J. Yu, J. Amores, N. Sebe, and Q. Tian, "Toward Robust Distance Metric Analysis for Similarity Estimation," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* June 2006.

[19] R. Haralick and L. Shapiro, *Computer and Robot Vision II.* Addison-Wesley, 1993.

[20] I.T. Jolliffe, *Principal Component Analysis,* second ed. Springer, 2002.

[21] R. Duda, P. Hart, and D. Stork, *Pattern Classification,* second ed. John Wiley & Sons, 2001.

[22] R.E. Schapire and Y. Singer, "Improved Boosting Using Confidence-Rated Predictions," *Machine Learning,* vol. 37, no. 3, pp. 297-336, 1999.

[23] D.W. Jacobs, D. Weinshall, and Y. Gdalyahu, "Classification with Nonmetric Distances: Image Retrieval and Class Representation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 6, pp. 583-600, June 2000.

[24] P.J. Phillips, "Support Vector Machines Applied to Face Recognition," *Proc. Advances in Neural Information Processing Systems,* vol. 11, 1998.

[25] P. Viola and M.J. Jones, "Robust-Real Time Face Detection," *Int'l J. Computer Vision,* vol. 57, no. 2, pp. 137-154, 2004.

[26] B. Moghaddam, T. Jebara, and A. Pentland, "Bayesian Face Recognition," *Pattern Recognition,* 2000.

[27] T.M. Cover and P.E. Hart, "Nearest Neighbor Pattern Classification," *IEEE Trans. Information Theory,* vol. 13, pp. 21-27, Jan. 1968.

[28] C. Domeniconi, J. Peng, and D. Gunopulos, "Locally Adaptive Metric Nearest Neighbor Classification," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 9, pp. 1281-1285, Sept. 2002.

[29] J. Peng, D. Heisterkamp, and H.K. Dai, "LDA/SVM Driven Nearest Neighbor Classification," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 940-942, 2001.

[30] E.P. Xing, A.Y. Ng, M.I. Jordan, and S. Russell, "Distance Metric Learning, with Application to Clustering with Side-Information," *Proc. Neural Information Processing Systems,* pp. 505-512, 2003.

[31] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, "Learning Distance Functions Using Equivalence Relations," *Proc. Int'l Conf. Machine Learning,* pp. 11-18, 2003.

[32] V. Athitsos, J. Alon, S. Sclaroff, and G. Kollios, "BoostMap: A Method for Efficient Approximate Similarity Rankings," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2004.

[33] T. Hertz, A. Bar-Hillel, and D. Weinshall, "Learning Distance Functions for Image Retrieval," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 570-577, 2004.

[34] P.J. Huber, *Robust Statistics.* John Wiley & Sons, 1981.

[35] M.S. Lew, T.S. Huang, and K. Wong, "Learning and Feature Selection in Stereo Matching," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 16, no. 9, pp. 869-882, Sept. 1994.

[36] L. Tang, Y. Kong, L.S. Chen, C.R. Lansing, and T.S. Huang, "Performance Evaluation of a Facial Feature Tracking Algorithm," *Proc. NSF/ARPA Workshop: Performance versus Methodology in Computer Vision,* 1994.

[37] K. Fukunaga and J. Mantock, "Nonparametric Discriminant Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 2, pp. 671-678, Feb. 1997.

[38] Y. LeCun and C. Cortes, "MNIST Database," http://yann.lecun.com/exdb/mnist/, 1998.

[39] A. Martinez and R. Benavente, *The AR Face Database,* technical report, vol. 24, Computer Vision Center, 1998.

[40] J. Matas, M. Hamouz, K. Jonsson, J. Kittler, C. Kotropoulos, A. Tefas, I. Pitas, T. Tan, H. Yan, F. Smeraldi, J. Bigun, N. Capdevielle, W. Gerstner, S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz, "Comparison of Face Verification Results on the Xm2vts Database," *Proc. Int'l Conf. Pattern Recognition,* pp. 858-863, 1999.

[41] J. Yu and Q. Tian, "Constructing Discriminant and Descriptive Features for Face Classification," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing,* May 2006.

[42] C. Merz and P. Murphy, "UCI Repository of Machine Learning Databases," http://www.ics.uci.edu/mlearn/MLRepository.html, 1998.

[43] J.R. Quinlan, "Bagging, Boosting, and C4.5," *Proc. Nat'l Conf. Artificial Intelligence,* pp. 725-730, 1996.

[44] D.G. Stork and E. Yom-Tov, "Computer Manual in MATLAB to Accompany," *Pattern Classification,* John Wiley & Sons, 2004.

[45] T. Hertz, A. Hillel, and D. Weinshall, "Learning a Kernel Function for Classification with Small Training Samples," *Proc. ACM Int'l Conf. Machine Learning,* 2006.

[46] Y. Rubner, C. Tomasi, and L.J. Guibas, "The Earth Mover's Distance as a Metric for Image Retrieval," *Int'l J. Computer Vision,* 2000.

[47] J. Lafferty, V. Pietra, and S. Pietra, "Statistical Learning Algorithms Based on Bregman Distances," *Proc. Canadian Workshop Information Theory,* 1997.

Jie Yu received the BE degree in electrical engineering from Dong Hua University, Shanghai, China, in 2000 and the PhD degree in computer science at the University of Texas at San Antonio (UTSA), San Antonio, in 2007. He is a research scientist in the Intelligence Systems Group at Kodak Research Labs. His research interests include multimedia information retrieval, machine learning, computer vision, and pattern recognition. He has published 18 journal articles, conference papers, and book chapters in these fields. He is the recipient of the Student Paper Contest Winner Award of IEEE ICASSP 2006 and the Presidential Dissertation Award of UTSA in 2006. He is a member of the IEEE and the ACM.

Jaumes Amores received the BSc degree in computer science from the University of Valencia, Spain, in 2000 and the MSc degree in computer science and the PhD degree from the Computer Vision Center, Autonomous University of Barcelona (UAB), Spain, in 2003 and 2006, respectively. Recently, he has moved to the Institut National de Recherche enInformatique et en Automatique (INRIA), France, where he holds a postdoctoral position in the IMEDIA Research Group under the direction of Dr. N. Boujemaa. His research interests include statistical learning, content-based image retrieval, object recognition, and medical image registration and retrieval.

Nicu Sebe is with the Faculty of Science, University of Amsterdam, the Netherlands. His research interests include multimedia information retrieval and human-computer interaction in computer vision applications. He is the author of two books and has published more than 100 technical papers in the areas of computer vision, content-based retrieval, pattern recognition, and human-computer interaction. He has organized several conferences and workshops in these areas, including the International Conference Image and Video Retrieval (CIVR '07) and was a guest editor of special issues in *IEEE Computer*, *Computer Vision and Image Understanding*, *ACM Transactions on Multimedia Computing, Communication, and Applications*, and *ACM Multimedia Systems and Image and Vision Computing*. He is an associated editor of the *Machine Vision and Application Journal* and is the general cochair of the IEEE International Conference on Automatic Face and Gesture Recognition 2008. He is a member of the IEEE.

Petia Reveda received the bachelor's degree in applied mathematics and computer science at the University of Sofia, Bulgaria, in 1989, the MS degree in 1993, and the PhD degree at the Universitat Autònoma de Barcelona (UAB) in 1998, working on the development of physics-based models applied to image analysis. She joined the Computer Science Department, UAB in 1991, as a teaching professor. Currently, she is a research project manager at the Computer Vision Center (CVC), an R&D institute founded by the UAB and the Generalitat de Catalunya. She has been and is the principal researcher or coordinator of several European and national research and industrial projects related to computer vision technologies. She has more than 120 international publications in international journals and proceedings in the field of medical imaging, image segmentation, pattern recognition, and computer vision. Her present research interests include development of physics-based approaches (in particular, statistics methods and deformable models) for medical image analysis, industrial vision, and remote sensing.

Qi Tian received the BE degree in electronic engineering from Tsinghua University, China, in 1992 and the PhD degree in electrical and computer engineering from the University of Illinois, Urbana–Champaign in 2002. He is an assistant professor in the Department of Computer Science, University of Texas at San Antonio (UTSA) and an adjunct assistant professor in the Department of Radiation Oncology, the University of Texas Health Science Center at San Antonio. He was a summer intern (2000, 2001) and a visiting researcher (2001) at the Mitsubishi Electric Research Laboratories (MERL), Cambridge, Massachusetts. He was a visiting professor in the Multimodal Information Access and Synthesis (MIAS) center (May–June 2007), University of Illinois, Urbana–Champaign (UIUC), a visiting researcher in the Web Mining and Search Group (Summer 2007), Microsoft Research Asia (MSRA), and a visiting professor in the Video Media Understanding Group (Summer 2003), NEC Laboratories America, Inc., Cupertino, California. His research interests include multimedia information retrieval, computer vision, and pattern recognition. He has published more than 70 refereed book chapters, journals, and conference papers in these fields. His research projects are funded by the US Army Research Office (ARO), Department of Homeland Security (DHS), San Antonio Life Science Institute (SALSI), Center of Infrastructure Assurance and Security (CIAS), and UTSA. He was the coauthor of the Best Student Paper with Jie Yu in IEEE ICASSP 2006. He has been in the International Steering Committee for ACM Workshop Multimedia Information Retrieval (MIR) (2006-2009), ICME 2006 Best Paper Committee member, conference chair for Visual Information Processing (VIP '07), Fifth International Conference on Intelligent Multimedia and Ambient Intelligence (IMAI 2007), ACM Workshop Multimedia Information Retrieval (2005), SPIE Internet Multimedia Management Systems (2005), and Eighth Multimedia Systems and Applications, SPIE's International Symposium on Optics East (2006), publicity cochairs of ACM Multimedia (2006) and ACM International Conference of Image and Video Retrieval (2007), special session chair of the Pacific-rim Conference on Multimedia (2007), track chair of multimedia content analysis at the IEEE International Conference on Multimedia and Expo (2006). He also served as the session/special session chair and TPC members in more than 60 IEEE and ACM conferences including ACM Multimedia, ICME, ICPR, ICASSP, PCM, CIVR, MIR, HCI, and VCIP. He is the guest coeditor of *Computer Vision and Image Understanding* for a special issue on similarity matching in computer vision and multimedia is in the editorial board of the *Journal of Multimedia*. He is a senior member of the IEEE and a member of the ACM.