# A NEW STUDY ON DISTANCE METRICS AS SIMILARITY MEASUREMENT

*Jie Yu*
Department of Computer Science
University of Texas at San Antonio

*Jaume Amores*,
IMEDIA Research Group,
INRIA,
Rocquencourt, France

*Nicu Sebe*
Faculty of Science
University of Amsterdam

*Qi Tian*
Department of Computer Science
University of Texas at San Antonio

## ABSTRACT

Distance metric is widely used in similarity estimation. In this paper we find that the most popular Euclidean and Manhattan distance may not be suitable for all data distributions. A general guideline to establish the relation between a distribution model and its corresponding similarity estimation is proposed. Based on Maximum Likelihood theory, we propose new distance metrics, such as harmonic distance and geometric distance. Because the feature elements may be from heterogeneous sources and usually have different influence on similarity estimation, it is inappropriate to model the distribution as isotropic. We propose a novel boosted distance metric that not only finds the best distance metric that fits the distribution of the underlying elements but also selects the most important feature elements with respect to similarity. The boosted distance metric is tested on fifteen 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.

## 1. INTRODUCTION

The most straightforward way to measure the similarity between two features is to compute the distance between them using a certain distance metric, which is also one of the most popular methods. In many fields such as image retrieval, the Euclidean distance, or SSD (sum of the squared differences or $L_2$), is widely used. However it has been suggested that this metric may not be suitable for all applications [1]. We find that from a Maximum Likelihood perspective the SSD metric is justified when the feature data is from Gaussian distribution [2]. Another popular distance metric, Manhattan distance or SAD (sum of the absolute differences or $L_1$), corresponds to the situation where the feature data distribution is Exponential. If the underlying data distribution is known or can be well modeled, it is possible to find the best distance function that matches the distribution. In most research work it is assumed that the real distribution is either the Gaussian or the Exponential. However such assumption is invalid for many applications. When the underlying distribution is unknown and could be neither Gaussian nor Exponential, finding a suitable distance metric becomes a challenge.

Similarity measurement could be used in content-based image retrieval where feature elements are extracted for different statistical properties for entire digital images, or perhaps with specific region of interest. Because of the heterogeneous sources those features may be from different distributions. Most of the attention in previous research work focused on extracting low-level feature elements such as color, texture, and shape with little consideration on their distributions. To compute the Euclidean distance between two feature vectors is still the most commonly used method for measuring the similarity between them.

Although we have done some research on utilizing the data distribution information for image retrieval based on similarity measurement [2, 3], the relation of the distribution model and the distance metric is still not clear. It has been proven that Gaussian, Exponential, and Cauchy distribution correspond to $L_2$, $L_1$, and Cauchy metrics, respectively [2]. However there are many other distribution models whose corresponding distance metrics are unknown. Besides, the similarity estimation based on feature elements from unknown distributions is an even more difficult problem. In this paper we extend our previous work in [2, 3] in proposing a guideline to learn a robust distance metric as accurate similarity measurement.

The rest of the paper is organized in the following way. Section 2 presents general analysis on distance metrics based on the maximum likelihood criterion. Section 3 introduces our novel distance metrics and the boosted version. In Sections 4 we use the new distance metrics as similarity measurement in experiments of motion tracking in a video sequence and content-based image retrieval. Discussions and conclusions are given in Section 5.

## 2. ANALYSIS ON DISTANCE METRICS

### 2.1 Distance Metric and Data Distribution

Based on Maximum Likelihood criteria, the $L_2$ metric, $L_1$ metric, and Cauchy metric are proven to be the optimal distance measure for the Gaussian, Exponential, and Cauchy distribution models respectively [2]. Since there are many other distribution models, it is reasonable to assume that there may be a certain model that fits the unknown data structure better. Consequently more accurate similarity estimation is expected if the metric could reflect the real distribution. We model this problem of finding the best distance metric through *distance metric analysis*. Mathematically it can be formulated in the following way.

Suppose we have some observed data from a certain distribution

$$x_i = \mu + d_i \tag{1}$$

where $d_i$, $i = 1, \cdots, N$ are data components and $\mu$ is the distribution mean or a sample from the same class. In most cases $\mu$ is unknown and may be approximated for similarity estimation. For some function

$$f(x, \mu) \geq 0 \tag{2}$$

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

$$\varepsilon=\sum_{i=1}^{N} f(x,\hat{\mu}) \tag{3}$$

It is equivalent to satisfy

$$\sum_{i=1}^{N}\frac{d}{d\hat{\mu}}f(x_i,\hat{\mu})=0 \tag{4}$$

We can find a closed-form solution of the estimated mean $\hat{\mu}=g(x_1,x_2,\cdots,x_N)$ for certain distributions. The arithmetic mean, median, harmonic mean, and geometric mean in Table 1 are in that category. It's proven that the $L_2$ metric (SSD) corresponds to the arithmetic mean while the $L_1$ metric (SAD) corresponds to the median. However, the distance metrics associated with the distribution models that imply the harmonic mean or the geometric mean haven't been introduced in literature before. Those metrics in Table 1 are inferred using equation (4). It is obvious that in distribution associated with the harmonic and geometric estimations, the observations which are far away from the correct estimate ($\mu$) will make less contribution in producing $\mu$, as distinct from the arithmetic mean. In that case the estimated values will be less sensitive to the outliers and they are therefore more robust.

**Table 1. Distance metrics and mean estimation for different distributions**

| | Distance Metric | 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 | $\varepsilon=\sum_{i=1}^{N}\lvert x_i-\hat{\mu}\rvert$ | $\hat{\mu}=med(x_1,\cdots,x_N)$ |
| 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 | $\varepsilon=\sum_{i=1}^{N}[\log(\frac{x_i}{\hat{\mu}})]^2$ | $\hat{\mu}=(\prod_{i=1}^{N}x_i)^{\frac{1}{N}}$ |
| Generalized harmonic (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 (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 | $\varepsilon=\sum_{i=1}^{N}[(x_i)^r\log(\frac{x_i}{\hat{\mu}})]^2$ | $\hat{\mu}=[\prod_{i=1}^{N}(x_i)^{(x_i)^{2r}}]^{\frac{1}{\sum_{i=1}^{N}(x_i)^{2r}}}$ |

## 2.3 Generalized Distance Metric Analysis

Inspired by the robust property of harmonic and geometric distance metrics we try to generalize those metrics and find new metrics that may fit the distribution better. In Table 1 we list three families of distance metrics which are derived from the generalized mean estimation using equation (4). Three parameters *p*, *q*, *r* are used to define the specific distance metrics and describe the corresponding distribution models which may not be explicitly formulated as Gaussian and Exponential. It is obvious that in the generalized harmonic mean estimation the 1st type is generalized based on the distance metric representation, while the 2nd 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 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 metrics correspond to a wide range of mean estimations and distribution models. It should be noted that not all mean estimations have closed-form solutions as in Tables 1. In that case $\hat{\mu}$ can be estimated by numerical analysis, e.g., greedy search of $\hat{\mu}$ to minimize $\varepsilon$.

# 3. BOOSTING DISTANCE METRICS FOR SIMILARITY ESTIMATION

## 3.1 The Problem

We found that the most widely applied distance metric is the $L_2$ distance. It assumes the data has a Gaussian isotropic distribution. However if the dimensionality of the feature space is high, the assumption of isotropic distribution is often inappropriate. Furthermore, since the feature elements are often extracted for different statistical properties, their distributions may not be the same and different distance metrics may better reflect the distributions for each feature element. Thus, an anisotropic and heterogeneous distance metric may be more desirable for estimating the similarity between features.

## 3.2 Our Approach

Inspired by the discussion in Section 3.1 we propose a novel boosted distance metric as similarity measurement where a similarity function for certain classes of samples can be estimated by a generalization of different distance metrics on selected feature elements. Specifically we adopt the idea of AdaBoost with decision stumps [4] and our novel distance metrics to measure the similarity.

Given a training set with feature vectors $\boldsymbol{x}_i$, the similarity estimation is done by training AdaBoost with difference vector $d$ obtained by different distance metrics between vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, *e.g.*, $d=\boldsymbol{x}_i-\boldsymbol{x}_j$ for $L_1$ metric, where each difference vector $d$ has an associated label $l_d$

$$l_d=\begin{cases}1 \text{ if } \boldsymbol{x}_i \text{ and } \boldsymbol{x}_j \text{ are from same class}\\ 0 \text{ otherwise}\end{cases} \tag{5}$$

A weak classifier $h_t$ is defined by a distance metric $m$ on a single feature element $f$ with estimated parameter(s) $\theta$, which could be as simple as the mean and/or a threshold. The label prediction of the weak classifier on feature difference $d$ is $h_{m,f,\theta}(d)\in\{0,1\}$. The boosted distance metric $h_t(d)$ is learned iteratively by weighted training samples with different distance metrics on each feature element and selecting the most important feature elements for similarity estimation, where $t=1,\ldots,T$. Consequently we derive a predicted similarity $S(\boldsymbol{x}_i,\boldsymbol{x}_j)=\sum_{t=1}^{T}\alpha_t h_t(d)$ that is optimal in a classification context, where $\alpha_t$ is the weight of classifier at the $t^{\text{th}}$ iteration based its classification error [4].

Three major advantages could be found in the proposed method: i) the similarity measure could only relate to the elements that are most useful for classification; ii) best distributions that fits

each element are found; iii) the proposed method is effective and robust for the classification when we have a small training set compared to the number of dimensions. The boosted distance metric could be considered as a non-linear dimension reduction technique. This retains the most important elements to similarity judgment, because the training iteration $T$ is usually much less than the original data dimension. This would be very helpful for overcoming the small sample set problem. Another good property of the boosted metric is that it is general and can be plugged into many similarity estimation techniques, such as the widely used $K$-NN.

The boosted similarity is more suitable for $K$-NN than other popular metrics when the training set is small. It could be explained in the following way: i) 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 over-fitting; ii) 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 on-line.

## 4. EXPERIMENTS AND ANALYSIS

### 4.1. Test on New Distance Metrics

In this experiment, we use our new distance metrics to estimate similarity that is mainly introduced by noise. 300 images from the Corel database are randomly selected. For each selected image, a copy is printed and digitized by a scanner, and resized. The whole process introduces noise due to the dithering patterns of the printer and scanner. We repeat the process 10 times for each selected image. According to this ground truth, we determine the real distribution of the similarity noise considering two different spaces: image space (or intensity space) and feature space. In image space, the intensity of pixel is used. In feature space, we have used two visual features: wavelet-based texture [5], and edge-based structure feature [6]. We compare our new metrics with the conventional $L_2$ and $L_1$ metrics, and the Cauchy metric.

**Table 2. The Chi-square test values for distance metrics**

| Error Metric | Structure | Texture | Intensity |
|---|---|---|---|
| $L_1$ | 0.2022 | 0.5269 | 2.6562 |
| $L_2$ | 0.2425 | 0.5017 | 4.1066 |
| Cauchy | 0.3365 ($a$=5) | 0.4953 ($a$=15.1) | 3.0607 ($a$=24.8) |
| Harmonic | 0.3087 | 0.2701 | 3.0970 |
| Geometric | 0.3138 | 0.4834 | 3.2940 |
| 1st type generalized harmonic (gh) | **0.1648** ($p$=1.8) | 0.1679 ($p$=-4.9) | 3.0970 ($p$=1.0) |
| 2nd type generalized harmonic (gh) | 0.2082 ($q$=-0.1) | **0.1331** ($q$=-0.7) | 2.2317 ($q$=-3.0) |
| Generalized geometric (gg) | 0.3086 ($r$=4.9) | 0.2959 ($r$=4.7) | **2.0374** ($r$=-4.2) |
| best metric | 1st gh ($p$=1.8) | 2nd gh ($q$=-0.7) | gg ($r$=-4.2) |

Chi-square test is first used to evaluate how well the distance metric fit the data distribution. There are several conclusions from Table 2: (i) the $L_1$ metric and Cauchy metric are more suitable than $L_2$ metric, e.g., in the intensity (image) space. This observation agrees with the results in [2]. (ii) Better estimations can be obtained by a large set of error metrics other than $L_1$, $L_2$, and Cauchy metrics, e.g., for the structure feature. This shows the effectiveness of our new metrics. Specifically, nonlinear estimations based on the generalized harmonic mean and generalized geometric mean are more robust than those based on the $L_1$ metric and $L_2$ metric.

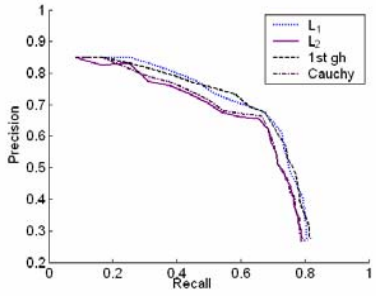


**Figure 1. The retrieval accuracy of four metrics on image database; for Cauchy metric, $a = 5$ ; and for 1st-type generalized harmonic mean, $p = 1.8$ .**

To compare the image retrieval results, we query the image database using different distance metrics and inspect how they affect the retrieval results. The precision-recall plot is used as the performance measure. Recall is a measure of the completeness of the retrieved set, i.e., the percentage of retrieved objects in the correct answer set. Precision, on the other hand, measures the purity of the retrieved set, i.e., the percentage of relevant objects among those retrieved.

Figure 1 shows the precision-recall graph. The curve corresponding to the 1st-type generalized harmonic mean ( $p = 1.8$ ) is above the others (comparable to $L_1$ for some range), showing that the method using the 1st-type generalized harmonic mean and $L_1$ are more effective than other metrics.

### 4.2. Test in Motion Tracking

In this experiment we test our novel distance metrics along with the traditional ones on motion tracking application. We use a video sequence containing 19 images on a moving head in a static background [7]. For each image in this video sequence, there are 14 points given as ground truth.

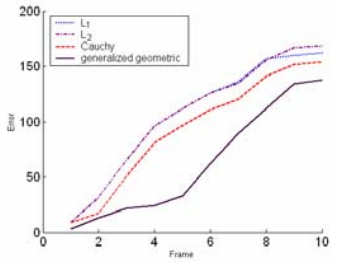


**Figure 2. Average tracking distance of the corresponding points in successive frames; for Cauchy, $\alpha = 7.1$ , and for generalized geometric mean, $r = 7.0$ .**

The motion tracking algorithm between the test frame and another frame performs template matching to find the best match in a $5 \times 5$ template around a central pixel. In searching for the corresponding pixel, we examine a region of width and the height of 7 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.

The tracking distance represents the average template matching results between the first frame and a later frame. Figure 2 shows the average tracking distance of the different distance metrics. The generalized geometric mean metric with $r = 7.0$ performs best, while Cauchy metric outperforms both $L_1$ and $L_2$.

### 4.3. Comparison to the State-of-the-Art on Benchmark Dataset

To evaluate the performance of our boosted distance metric, we compare it with several well-known traditional approaches. Thirteen benchmark datasets from UCI machine learning repository are used for training and testing.

The traditional distance metrics we tested are: Euclidean Distance, Manhatan Distance, RCA [8] distance, 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 metrics are sensitive to the small sample set problem. So a diagonal matrix *D* could be estimated instead of original weight matrix *W* to simplify that problem and consequently we can obtain three metrics *RCA-D*, *Mah-D* and *Mah-CD*. To make the comparison complete, we also test the original AdaBoost with decision stump *(d.s.)* and C4.5.

**Table 3. Comparison to traditional distance metric and AdaBoost on UCI datasets**

| Error Rate (%) | Traditional Metric | AdaBoost +d.s. | AdaBoost +C4.5 | Boosted Metric |
|---|---|---|---|---|
| ad | 17.31 ($L_1$) | 12 | 11.42 | **8.92** |
| arrhythmia | 37.02 (RCA-D) | 31.39 | 29.94 | **25.62** |
| splice | 10.55 (Mah-D) | 5.94 | **4.84** | 4.89 |
| sonar | 26.1 (Mah-CD) | 25.95 | 25.81 | **25.35** |
| spectf | 31.16 (Mah-D) | 28.65 | 27.18 | **26.2** |
| Ionosphere | **10.78** (RCA) | 19.92 | 19.92 | 17.35 |
| wdbc | 6.83 (Mah-CD) | 5.81 | 5.37 | **4.32** |
| german | 38.74 (Mah-D) | 34.31 | 33.18 | **31.6** |
| Vote1 | 9.07 ($L_1$) | 6.37 | 6.37 | **6.18** |
| credit | 19.18 (Mah-CD) | 17.97 | **17.21** | 17.63 |
| Wbc | 5.25 (RCA) | 5.7 | 5.34 | **4.23** |
| pima | 34.55 (Mah-CD) | 31.02 | 29.96 | **28.12** |
| liver | 41.11 (Mah) | 35.51 | 35.43 | **32.77** |

Due to the space limitation, only the traditional distance metric that gives the best performance in each data set is shown. The smallest error rates are highlighted in bold. Note that this experiment is different from that of [3] in that multiple distance metrics are boosted for optimal performance. From the results in Table 3 we find that our boosted distance metric performs the best in 10 out of 13 datasets. It provides comparable results to the best performance on 2 datasets. Only in Ionosphere dataset is our method outperformed by the traditional distance metric. It proves that our method could discover the best distance metric that reflects the distribution and selects the feature elements that are discriminant in similarity estimation.

## 5. DISCUSSIONS AND CONCLUSIONS

In this paper we first analyze the relation between distance metric and data distribution. New distance metrics are derived from harmonic, geometric mean and their generalized forms are presented and discussed. Those new metrics are tested on several applications in computer vision and we found the estimation of similarity can be significantly improved.

Since the feature elements used in similarity estimation are often from heterogeneous sources, we find the assumption that the feature has a unified isotropic distribution is inappropriate. To substitute traditional anisotropic distance metric, we proposed a boosted metric that does not make any assumption on the feature distribution. It can find optimal distance metrics on each element to capture the underlying feature structure. Since the learned distance metric only associates with selected elements, the boosted distance, it is more robust to small sample set problem. It also has the dimension reduction effect which may be very useful to alleviate high-dimensionality problem. We find that the automatic metric adaptation and element selection in our boosted distance metric bridge the gap between the high-level similarity concept and low-level features. The experimental results have proven the proposed method is more effective and efficient than traditional distance metrics.

In the future we would like to incorporate our new metric into state-of-the-art classification techniques and evaluate the performance improvement.

**Acknowledgement:** This work was supported in part by the 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. The work of Nicu Sebe was done within the MUSCLE-NOE. We thank Mark Doderer for polishing the paper.

## 6. REFERENCES

[1] M. Zakai, "General distance criteria," *IEEE Trans. on Information Theory*, pp. 94-95, January 1964.

[2] N. Sebe, M. S. Lew, and D. P. Huijsmans, "Toward improved ranking metrics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1132-1143, Oct. 2000.

[3] J. Amores, N. Sebe and P. Radeva, "Boosting the distance estimation: application to the K-nearest neighbor classifier," *Pattern Recognition Letters*, Feb. 2006

[4] R. E. Schapire and Y. Singer, "Improved boosting using confidence-rated predictions," *Machine Learning* 37 (3) 297–336, 1999.

[5] J. R. Smith and S. F. Chang, "Transform features for texture classification and discrimination in large image database," *IEEE Intl. Conf. on Image Proc.*, 1994.

[6] S. Zhou, Y. Rui, and T. S. Huang, "Water-filling algorithm: a novel way for image feature extraction based on edge maps," *IEEE Intl. Conf. on Image Proc.*, 1999.

[7] L. Tang, *et al*., "Performance evaluation of a facial feature tracking algorithm," *Proc. NSF/ARPA Workshop: Performance vs. Methodology in Computer Vision*, 1994.

[8] K. Fukunaga and J. Mantock, "Nonparametric discriminant analysis," *IEEE Trans. PAMI*, Vol. 5, 1983.