What are the Distance Metrics for Local Features?

Zhendong Mao[1,2], Yongdong Zhang[2], Qi Tian[3]
[1]University of Chinese Academy of Sciences
[2]Institute of Computing Technology,Chinese Academy Of Sciences
[3]University of Texas at San Antonio
{maozhendong,zhyd}@ict.ac.cn, qitian@cs.utsa.edu

ABSTRACT

Previous research has found that the distance metric for similarity estimation is determined by the underlying data noise distribution. The well known Euclidean(L2) and Manhattan (L1) metrics are then justified when the additive noise are Gaussian and Exponential, respectively. However, finding a suitable distance metric for local features is still a challenge when the underlying noise distribution is unknown and could be neither Gaussian nor Exponential. To address this issue, we introduce a modeling framework for arbitrary noise distributions and propose a generalized distance metric for local features based on this framework. We prove that the proposed distance is equivalent to the L1 or the L2 distance when the noise is Gaussian or Exponential. Furthermore, we justify the Hamming metric when the noise meets the given conditions. In that case, the proposed distance is a linear mapping of the Hamming distance. The proposed metric has been extensively tested on a benchmark data set with five state-of-the-art local features: SIFT, SURF, BRIEF, ORB and BRISK. Experiments show that our framework better models the real noise distributions and that more robust results can be obtained by using the proposed distance metric.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering, Retrieval models, Search process, Selection process; I.4.7 [**Feature Measurement**]: Feature representation

Keywords

distance metric, local feature, noise distribution

1. INTRODUCTION

L2 and L1 metrics are frequently used in matching traditional local features such as SIFT[4], SURF[1], while Hamming metric is frequently used in matching binary local features such as BRIEF [2], ORB [6], and BRISK [3]. However, the connection between the three metrics and the prevalent local features has not been fully studied yet.

Sebe et al [7] is one of the first (if not first) efforts to relate distance metric to noise distribution from a maximum likelihood perspective. In [7], they justified that Gaussian, Exponential, and Cauchy distributions result in L2, L1, and Cauchy metrics, respectively. The common assumption is that the real noise distribution should fit either the Gaussian or the Exponential. But for local features (especially binary local features), Gaussian or Exponential assumption is often invalid. This arises the question: whether we could find a more accurate noise model for both traditional and binary local features? whether we could find a generalized distance metric which incorporates L1, L2 and Hamming metrics as the special cases?

To address this issue, we develop a noise modeling framework and a generalized distance metric for local features. When given a local feature, we first model its noise distribution using our framework, and then determine our distance metric. Our main contributions are:

- A framework for modeling arbitrary noise distributions is presented and a generalized distance metric for local features based on this framework is proposed.
- L1, L2 and Hamming distance are proved as special cases of the proposed distance. Specially, necessary conditions of using Hamming metric are specified and thus we give theoretical support for the use of Hamming metric in binary feature matching.
- Theoretical lower bound of distance threshold used in feature matching is justified as the product of feature dimension and noise entropy.

To validate our framework, we compute the real noise distributions of SIFT and SURF and apply different models to fit them. We also illustrate the efficiency of our distance metric by comparing it with traditional distance metrics on benchmark image set.

2. NOISE MODELING FRAMEWORK AND FEATURE DISTANCE METRIC

Maximum Likelihood theory allows us to relate a data distribution to a distance metric. Suppose there are two subsets of D features from the database A: $X \subset A, Y \subset A$ which according to the ground truth are matching: $X \equiv Y$. This can be presented in detail:

$$x_i = y_i + d_i, i = 1, ..., D \tag{1}$$

MM'13, October 21–25, 2013, Barcelona, Spain.
Copyright 2013 ACM 978-1-4503-2404-5/13/10 ...$15.00.
http://dx.doi.org/10.1145/2502081.2502134.

where d_i is the noise which represented by the difference between corresponding features x_i and y_i. In this context, the similarity probability between X and Y can be defined:

$$P(X,Y) = \prod_{i=1}^{D} p(d_i) \tag{2}$$

where function p is the probability density of the noise, and $p(d_i)$ describes the similarity between x_i and y_i. According to (2) we have to find the probability density function of the noise that maximizes the similarity probability. This is the maximum likelihood estimator for X, given Y.

Taking the negative logarithm of (2) we find that we have to minimize the expression:

$$\sum_{i=1}^{D} \rho(d_i), \quad \rho(d_i) = -\log(p(d_i)) \tag{3}$$

Expression (3) is the metric which maximizes the similarity probability [7].

According to (3), which measure to use can be determined if the underlying noise distribution is known or well estimated. Sebe [7] focus on modeling the noise by continuous probability distributions and then determined by the probability density function which distance measure to use. However, in many applications, finding a suitable continuous probability distribution for the feature noise is an even more difficult problem. Instead, we found that the real distribution could be approximated by a multinomial distribution sampled from it and the accuracy of the approximation could be adjusted by sampling density. Therefore, we propose a framework to model the noise by multinomial distribution and a generalized distance metric based on this framework.

2.1 Distance Metric for Local Features

Without loss of generality, we suppose a local feature has N elements (N dimensions) and each element is stored in m bits. The special case $m = \infty$ denotes that the elements of features are stored in an unlimited space and thus the noise distributions of features are continuous. The distribution of the noise d_i is a joint distribution of each dimension:

$$p(d_i) = p(d_{i0}, d_{i1}, ..., d_{iN}) \tag{4}$$

where d_{ij} is the noise of j-th dimension. Since d_{ij} is represented by the difference between two corresponding elements and each element is stored in m bits, we can define (we suppose the noise of each dimension is independent and is identically distributed, which is the most common case):

$$\Pr(d_{ij} = c) = P_c, \quad (c \in [1-2^m, 2^m-1], \quad j = 1, ..., N) \tag{5}$$

Obviously, d_{ij} is multinomial distributed. Then $p(d_i)$ representing the probability of the noise $d_i (d_i = [d_{i0}, d_{i1}, ..., d_{iN}])$ can be written as:

$$p(d_i) = \prod_{c=1-2^m}^{2^m-1} P_c^{k_c} \tag{6}$$

where

$$k_c = \sum_{j=1}^{N} \delta_c(d_{ij}), \quad \delta_c(d_{ij}) = \begin{cases} 1 & if \quad d_{ij} = c \\ 0 & otherwise \end{cases} \tag{7}$$

k_c represents the number of d_{ij} that equals to c. So we have:

$$\sum_{c=1-2^m}^{2^m-1} k_c = N \tag{8}$$

Our distance metric between two local features x and y is defined as the negative logarithm of (6):

$$T_m(x,y) = \sum_{c=1-2^m}^{2^m-1} -k_c \log(P_c) \tag{9}$$

When the noise is distributed as Gaussian or double Exponential ($m = \infty$ is underlying):

$$P_c \sim e^{-c^2} \quad or \quad P_c \sim e^{-|c|} \tag{10}$$

our distance is equivalent to L2 or L1 distance.

$$T_m(x,y) = \sum_{c=-\infty}^{+\infty} k_c \cdot c^2 \quad or \quad T_m(x,y) = \sum_{c=-\infty}^{+\infty} k_c \cdot |c| \tag{11}$$

Therefore, L1 and L2 distance are special cases of the proposed distance. It is also worth mentioning that arbitrary continuous distributions of noise could be discretized by sampling and then result in a distance metric by using (9). Besides, quantization could be applied to noise for reducing the computation cost.

2.2 Distance Metric for Binary Local Features

By introducing (8) into (9) and let $m = 1$, we get the distance metric for binary local features:

$$T(x,y) = \log(\frac{P_0}{P_{-1}})k_{-1} + \log(\frac{P_0}{P_1})k_1 - N \log P_0 \tag{12}$$

Since the noises of binary features are intrinsically multinomial distributed (all possible values of element noise are -1, 0, and 1), (12) is the theoretical distance metric for binary local features according to Maximum Likelihood theory. To investigate the relationship between Hamming distance and the proposed distance, We define two conditions C1 and C2 as follows:

$$C1 : P_{-1} < P_0, P_1 < P_0 \qquad C2 : P_{-1} = P_1$$

If these two conditions are met, we can rewrite the proposed distance as:

$$T(x,y) = h \cdot (k_{-1} + k_1) + t \tag{13}$$

where

$$h = \log(\frac{P_0}{P_{-1}}), h > 0, t = -N \log(P_0) \tag{14}$$

As we all know, Hamming distance can be defined as:

$$H = k_{-1} + k_1 \tag{15}$$

One can note that when condition $C1$ and $C2$ are satisfied, Hamming distance is a strictly positive linear mapping of the proposed distance. In this case, Hamming metric is the theoretical distance metric. In case that $C1$ is satisfied but $C2$ is not strictly satisfied, Hamming metric is still a suitable distance metric. If $C1$ is not met, Hamming distance is opposite to the similarity of features. On that condition, Hamming distance should not be used to match features .

2.3 Distance Threshold

In section 2.2, the distance between two corresponding local features x and y is given by (9). Let x_i and y_i represent the i-th elements of x and y, respectively. Let $E(c)$ denotes the event $x_i - y_i = c, i = 1, ...N$. Note that k_c is the number of times $E(c)$ occurs and P_c is the probability of $E(c)$'s occurrence. According to Borel's law of large numbers:

$$\frac{k_c}{N} \rightarrow P_c \quad as \quad N \rightarrow \infty \tag{16}$$

By introducing (16) into (9) :

$$T_m(x, y) \rightarrow u(m) \quad as \quad N \rightarrow \infty, \tag{17}$$

where

$$u(m) = -N \cdot \sum_c P_c \log(P_c) \tag{18}$$

(17) demonstrates that the proposed distance of two corresponding features converges in probability to $u(m)$, which is the product of feature dimension and noise entropy. For most local features, dimension N is very large (SIFT and SURF have 128 dimensions, binary features have even more), which means the proposed distances of corresponding features are very close to $u(m)$. If the distance threshold used in feature matching is larger than $u(m)$, most correct matches would be excluded. Therefore, $u(m)$ is the theoretical lower bound of distance threshold when using the proposed distance metric. Since we have demonstrated that L2, L1 and Hamming distance are special cases of the proposed distance, $u(m)$ is also the theoretical lower bound of distance threshold when using these distance metrics.

Based on analysis above, we further develop our distance metric as:

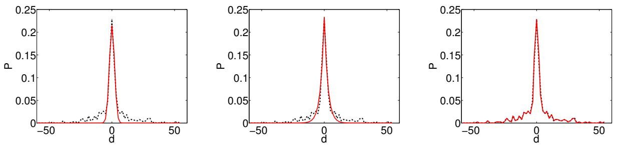
$$T'_m(x, y) = T_m(x, y) - u(m) \tag{19}$$

3. EXPERIMENT AND ANALYSIS

The proposed distance metric has been extensively tested on five well-known local features: SIFT, SURF, BRIEF, ORB and BRISK. For each feature, we inspect how proposed distance metric performs comparing to several traditional distance metrics such as L1, L2 and Hamming metrics. Besides, Gaussian, Exponential and our models are applied to fit the noise distributions of SIFT and SURF to evaluate our modeling framework.

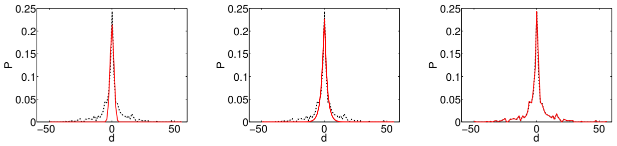
3.1 Experimental setup

Data set: The testing environment is the benchmark image set proposed by Mikolajczyk et al [5]. The image set consists of 8 categories. Each category contains a sequence of six images exhibiting an increasing amount of transformation. We find that the latter images in each category contain larger transformation, which is too hard for some features. Therefore, only the first three images are picked from each category. The first image and the second image in each category are selected as training set for computing real noise distribution, while the first image and the third image in each category are selected as test set for matching.



(a) Gaussian (b) Exponential (c) Multinomial

Figure 1: Real noise distribution of SIFT on training set modeled by three distributions. The dashed black line represents the real noise distribution and the solid red line represents the estimation. (R-square score representing goodness of fit is: (a) 0.94 ; (b) 0.96 ; (c) 0.99).



(a) Gaussian (b) Exponential (c) Multinomial

Figure 2: Real noise distribution of SURF on training set modeled by three distributions. The dashed black line represents the real noise distribution and the solid red line represents the estimation. (R-square score representing goodness of fit is: (a) 0.87 ; (b) 0.92 ; (c) 0.99).

Matching strategy: Following [5], two features A and B are matched if B is the nearest neighbor to A and if the distance ratio between the first and the second nearest neighbor is below a threshold: $\|A - B\|/\|A - C\| < t$, where B is the first and C is the second nearest neighbor to A. For SIFT and SURF , we set $t = 0.8$. For binary features, we set $t = 0.95$.

All detectors and descriptors are implemented using public OpenCV source code with default parameters.

3.2 Distance Metric for SIFT and SURF

We consider the following steps to evaluate the modeling framework and the proposed distance metric:

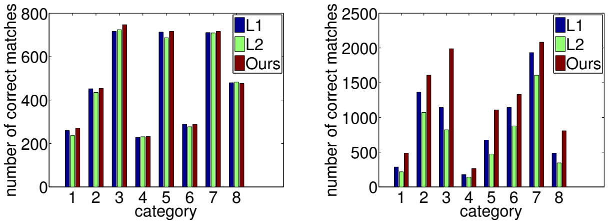
Step1: Compute SIFT and SURF features and find correct matches from the training set. Normalize all features into a fixed range(from -60 to 60).

Step2: Compute the noise distribution from the differences between corresponding elements of matched features.

Step3: Quantize the noises by round-off (-60 to 60, 121bins), hence get a multinomial distribution.

Step4: Fit the Exponential, the Gaussian, and the multinomial distributions to the real distribution. Compare each of the model distribution M to the real noise distribution R using R-square (Coefficient of determination):

$$R\text{-square} = \frac{\sum_i (R_i - M_i)^2}{\sum_i (R_i - \bar{R})^2} \tag{20}$$



(a) SIFT (b) SURF

Figure 3: Performance of distance metrics on local features.

Table 1: Real noise distribution of binary features on training set

Categories	$p(d=-1)$	$p(d=0)$	$p(d=1)$
BRIEF	6.12%	87.77%	6.11%
ORB	6.10%	87.44%	6.46%
BRISK	5.24%	89.53%	5.23%

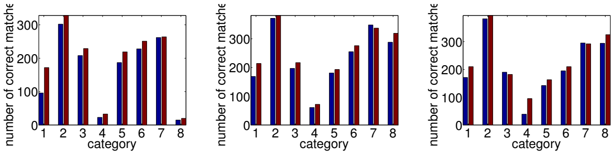
Step5: Apply the corresponding distance metric of each model distribution on test set (L1, L2 and our distance metric). Compare the matching performance.

In Figure 1 and 2 we have the matchings of the real noise distributions of SIFT and SURF considering the three distributions. The quantitative results of the estimation accuracy are also given by R-square score in Figure 1 and 2. One could notice that our framework is more effective to find a better model for the real noise distribution.

Figure 3 shows the matching results of SIFT and SURF when different distance metrics are used. The vertical scale of these two figures records the number of correct matches which represents matching performance. From the Figure 3 we can find that the proposed distance metric performs the best. For SIFT, the performances of the three distance metric have no big difference. The reason is that the R-square score of the three distributions modeling SIFT noise are very close. For SURF, the R-square score of the multinomial distribution is much higher than the other distributions. As a result, the proposed distance metric performs much better than L1 and L2 distance metrics. As a conclusion, the matching performance is in the same order of the R-square score, which means finding a better distance metric can be understood as establishing a better model for the real noise distribution. This conclusion is in accordance with the theory described in the Section 2.

3.3 Distance Metric for Binary Features

The real noise distributions of BRIEF, ORB, and BRISK are also computed based on the training set. These are given in table 1. We can see that for all three binary features, the condition C1 is satisfied ($P_{-1} < P_0, P_1 < P_0$) and the condition C2 is approximately satisfied ($P_{-1} \approx P_1$). According to the theory described in section 2.3, Hamming distance metric is then an approximate positive linear mapping of the theoretical distance on these features. Therefore, we give theoretical support for using Hamming distance metric on BRIEF, ORB, and BRISK. Furthermore, we also illustrate a framework to test whether Hamming metric should be used for matching a new binary feature.



(a) BRIEF (b) ORB (c) BRISK

Figure 4: Performance of distance metrics on binary local features. Blue bars represent performance of Hamming metric and Red bars represent performance of ours.

The comparison results of our metric and Hamming metric on test set are given in Figure 4. We can see that our metric outperforms Hamming metric on all three features. The reason is that our metric can better reflect the real noise model on this data set. In addition, the experimental results illustrate that Hamming metric is also an effective distance metric for these three binary features.

4. CONCLUSION

In this paper, we propose a generalized distance metric for local features according to Maximum Likelihood theory. The proposed metric is based on a framework which approximates the feature noise distribution by multinomial distribution. Prevalent traditional distances such as L1, L2 and Hamming distance are justified as the special cases of the generalized distance.

We investigate the noise distributions of state-of-the-art local features. Our modeling framework is identified to be very effective to model their noise distributions. We also compare the matching performances of our distance metric and prevalent traditional ones. In all the experiments, better results are obtained based on our distance metric.

5. ACKNOWLEDGMENTS

This work is supported by National Nature Science Foundation of China (61273247, 61271428), National Key Technology Research and Development Program of China (2012BAH-39B02). This work was supported in part to Dr. Qi Tian by ARO grant W911BF-12-1-0057, NSF IIS 1052851, Faculty Research Awards by Google, FXPAL, and NEC Laboratories of America, and 2012 UTSA START-R Research Award respectively. This work was supported in part by NSFC 61128007.

6. REFERENCES

[1] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *ECCV*, 2006.

[2] M. Calonder, V. Lepetit, and et al. Brief: Binary robust independent elementary features. In *ECCV*, 2010.

[3] S. Leutenegger, M. Chli, and R. Siegwart. Brisk: Binary robust invariant scalable keypoints. In *ICCV*, 2011.

[4] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[5] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 2005.

[6] E. Rublee, V. Rabaud, and et al. Orb: an efficient alternative to sift or surf. In *ICCV*, 2011.

[7] N. Sebe, M. Lew, and D. Huijsmans. Toward improved ranking metrics. *PAMI*, 2000.