Distance Functions: Theory, Algorithms and Applications

Thesis submitted for the degree of
"Doctor of Philosophy"

By
Ofir Pele

Submitted to the Senate of the Hebrew University
July, 2011

This work was carried out under the supervision of:
Prof. Michael Werman

Acknowledgments

It is difficult to overstate my gratitude to my Ph.D. supervisor, Prof. Michael Werman with his encouragement, support, and his great efforts through the research and most importantly his passion for research. Mike helped to make this thesis feasible and was a friend through the process.

I also thank my many friends in the corridor and lab with whom it was always fun to talk and discuss various issues.

In addition, I would like to thank many persons at the School of Engineering and Computer Science, both academic and administrative staff, for creating a stimulating environment to learn and grow. Special thanks to Prof. Shmuel Peleg for his priceless advises.

Lastly, and most importantly, I wish to thank my family for their encouragement, support, and love. My late father, whose dream was to see his children study. My mother, who did everything she could to allow her children to realize their full potential. My brothers, Oved and Eli for their encouragement, great support, and love. My wife, Liat, and my children, Itamar and Maayan who are the source of my happiness. Liat, you are the most perfect life partner I can imagine. I dedicate this thesis to them.

Abstract

Distance functions are at the core of numerous computer vision and machine learning tasks. For example, a multiple view geometry application (such as homography estimation), often includes a descriptor matching step. First, small image patches are selected. Then, descriptors (vectors) are computed to describe these patches. To decide if two descriptors (from two different images) are from the same real-world location the descriptors are compared using a distance function. The choice of the distance function determines both the accuracy and the speed of the method. Distance functions are also used for image retrieval and determine both the percent of similar images returned and the time it takes to return them. In machine learning, k-nearest neighbor classification performance is effected by the chosen distance function.

This thesis introduces several new distance functions and investigates their properties. Special emphasis is on practical applicability with theoretical insights. The thesis presents efficient algorithms to compute several distances. The distances are designed to be robust to common image noise such as occlusion, geometrical transforms, light changes and non-rigid deformations. On the other hand, they are designed to be as distinctive as possible. Our proposed methods have been successfully used both by computer vision researchers and by researchers in other fields. The success of the methods in other fields is probably because the noise characteristics in those fields are similar to image noise characteristics.

The first and second parts of the thesis are devoted to the Earth Mover's Distance (EMD) [1]. The EMD is a generalization of the transportation distance (also known as Monge-Kantorovich, Mallows, 1^{st} Wasserstein and match distance) for non-normalized histograms. We introduce a new generalization of the transportation distance for non-normalized histograms called, $\widehat{EMD}$. Unlike the classic EMD, $\widehat{EMD}$ is a metric even when the histograms are non-normalized. Additionally, in practice, $\widehat{EMD}$ is often better than EMD. We also show that $\widehat{EMD}$ is a generalization of the L_1 distance.

Histograms of oriented gradient descriptors are ubiquitous tools in numerous computer vision tasks. Unfortunately, the distances that are used to match such descriptors do not take into account the special structure of these histograms. One problem with these distances is that they only compare corresponding bins of histograms. Although there are cross-bin distances that take into account the amount of similarity between all bins, they are not adequate to the histograms of oriented gradients that are computed from images. Additionally, most cross-bin distances are too slow. We present a

linear time algorithm for the comparison of orientation histograms. This method outperforms state-of-the-art distances on a widely used, publicly available benchmark.

In the second part of the thesis we present a robust family of Earth Mover's Distances which is applicable to any type of histograms (not just orientation histograms). We developed an algorithm that computes these robust EMDs by an order of magnitude faster than the original approach. The resulting method has excellent image retrieval performance.

In the third part of the thesis, we present a new histogram distance family, the Quadratic-Chi (QC). QC members are Quadratic-Form distances with a cross-bin χ^2-like normalization. The cross-bin χ^2-like normalization reduces the effect of large bins having undue influence. Normalization was shown to be helpful in many cases, where the χ^2 histogram distance outperformed the L_2 norm. However, χ^2 is sensitive to quantization effects, such as caused by light changes, shape deformations etc. The Quadratic-Form part of QC members takes care of cross-bin relationships (*e.g.* red and orange), alleviating the quantization problem. We present two new cross-bin histogram distance properties: *Similarity-Matrix-Quantization-Invariance* and *Sparseness-Invariance* and show that QC distances have these properties. We also show experimentally that the properties improve performance. QC distances computation time complexity is linear in the number of non-zero entries in the bin-similarity matrix and can easily be parallelized. We present results for image retrieval and shape classification. We show that the new QC members outperforms state-of-the-art distances for these tasks, and have a short running time.

Finally, in the fourth part of the thesis, we describe a novel method for non-Mahalanobis metric learning denoted *Interpolated-Discretized Metric Learning*. This part of the work was motivated by the success of our non-Mahalanobis distance, the QC. A QC distance has parameters that can be learned. However, a serious limitation is that it can only model χ^2-like distances. In addition, it is applicable only to non-negative vectors. Finally, it is non-convex with respect to its parameters, so learning them is hard. To overcome these problems we suggest a new non-Mahalanobis distance family, the *Interpolated-Discretized* (ID) distance family, and show how to efficiently learn its parameters. An ID distance can approximate any *unknown* distance. Many metric learning methods can be used with our ID distances. We demonstrate one possible way of learning ID distances using the Large Margin Nearest Neighbor (LMNN) framework [2]. With ID distances, the LMNN optimization problem is a linear program. Finally, we show that our method often outperforms other state-of-the-art metric learning approaches.

Contents

Chapter 1

Introduction

Distance functions are at the core of numerous computer vision and machine learning tasks. For example, a multiple view geometry application (such as homography estimation), often includes a descriptor matching step. First, small image patches are selected. Then, descriptors (vectors) are computed to describe these patches. To decide if two descriptors (from two different images) are from the same real-world location the descriptors are compared using a distance function. On one hand, the distance function should be robust to common image noise or transformation (such as occlusion and light changes) so that the distance between two patches from the same real-world location will be small. On the other hand, the distance function should be as distinctive as possible, so that two patches from different real-world locations will be well separated. Additionally, we need efficient algorithms for the computation of the distance function as usually the number of distance computations is very large.

This thesis introduces several new distance functions and investigates their properties. Special emphasis is on theory that is built upon real-world assumptions. This introduction gives an overview of distance functions and their properties. Then, we describe the novel aspects of the thesis.

1.1 Distance Functions Overview

1.1.1 Bin-to-Bin Distance Functions

Bin-to-Bin distance functions compare corresponding bins of a vector to its exact corresponding bin in the second vector. That is, let $P, Q \in \mathbb{R}^N$ be two vectors, P_i is only compared to Q_i.

The most widely used distance family is the Minkowski-Form, also known

as L_p norms:

$$L_p(P,Q) = (\sum_i |P_i - Q_i|^p)^{\frac{1}{p}} \tag{1.1}$$

For $p \geq 1$, L_p is a metric (see section 1.1.2). The most widely used distances from this family are the L_1 distance also known as the Manhattan distance and the L_2 distance also known as the Euclidean distance.

The Kullback–Leibler (KL) divergence is a distance function for probability distributions (although it is used also for non-negative vectors which are not normalized). The KL divergence is defined as:

$$\mathrm{KL}(P,Q) = \sum_i P_i \log \frac{P_i}{Q_i} \tag{1.2}$$

The KL divergence measures the expected number of extra bits required to code samples from P when using a code based on Q. It is non-symmetric. A serious limitation of the distance is that when $P_i \neq 0$ and $Q_i = 0$, KL equals infinity.

To overcome KL problems, the Jensen-Shannon (JS) divergence was suggested by Lin [3]. It is defined as:

$$\begin{aligned} \mathrm{JS}(P,Q) &= \mathrm{KL}(P,M) + \mathrm{KL}(Q,M) \\ M &= \frac{P+Q}{2} \end{aligned} \tag{1.3}$$

The square root of a Jensen-Shannon divergence is a metric [4, 5] and it can be embeded isometrically as a subspace of a real Hilbert space [6]. Using the Taylor extension we get that JS is equal to [7]:

$$\mathrm{JS}(P,Q) = \sum_{n=1}^{\infty} \frac{1}{2n(2n-1)} \sum_i \frac{(P_i - Q_i)^{2n}}{(P_i + Q_i)^{2n-1}} \tag{1.4}$$

where $n = 1$ corresponds to Taylor degrees of one and two and $n = 2$ correspond to taylor degrees of three and four, etc. The first term is known in the computer vision community as χ^2 distance:

$$\chi^2(P,Q) = \frac{1}{2} \sum_i \frac{(P_i - Q_i)^2}{(P_i + Q_i)} \tag{1.5}$$

The practical results of χ^2 and JS are almost identical [8–10]. χ^2 is more efficiently computed than the JS. The square root of χ^2 is a metric [7].

The χ^2 histogram distance also emerged independently from the χ^2 test-statistic [11] where it is used to test the fit between a distribution and observed frequencies. χ^2 distance has recently received considerable attention in the computer vision literature. It has been used, for example, in state-of-the-art: contour detection and segmentation algorithms [12], descriptors matching [13], texture and object categories classification [14–17], near duplicate image identification[18] and shape classification [19, 20].

1.1.2 Metric Properties

There are four conditions for a distance function, $\mathcal{D}$, to be a *metric*:

1. $\mathcal{D}(P, Q) \geq 0$ (*non-negativity*).
2. $\mathcal{D}(P, Q) = 0$ if and only if $P = Q$ (*identity of indiscernibles*).
3. $\mathcal{D}(P, Q) = \mathcal{D}(Q, P)$ (*symmetry*).
4. $\mathcal{D}(P, Q) \leq \mathcal{D}(P, K) + \mathcal{D}(K, Q)$ (*subadditivity*).

Non-negativity is implied by the others: $2\mathcal{D}(P, Q) = \mathcal{D}(P, Q) + \mathcal{D}(Q, P) \geq \mathcal{D}(P, P) = 0$.

Two useful generalizations of metrics are the *semi-metric*, which drops the triangle inequality property and *pseudo-metric* which weakens the second condition to:

2. $\mathcal{D}(P, Q) = 0$ if $P = Q$ (but possibly $\mathcal{D}(P, Q) = 0$ for some $P \neq Q$).

1.1.3 Cross-Bin Distances

Bin-to-Bin distance functions such as L_2, L_1 and χ^2 compare only corresponding bin's of a vector to its exact corresponding bin in the second vector. The assumption when using these distances is that the histogram domains are aligned. However this assumption is violated in many cases due to quantization, shape deformation, light changes, etc. Bin-to-bin distances depend on the number of bins. If it is low, the distance is robust, but not discriminative, if it is high, the distance is discriminative, but not robust. Distances that take into account cross-bin relationships (cross-bin distances) can be both robust and discriminative.

The Quadratic-Form Distance

Let $A \in \mathbb{R}^{N \times N}$ the bin-similarity matrix. That is, a_{ij} encodes how much bin i is similar to bin j. The *Quadratic-Form* (QF) distance is defined as [21]:

$$\mathrm{QF}^A(P,Q) = \sqrt{(P-Q)^T A (P-Q)} \tag{1.6}$$

When the bin-similarity matrix A is the inverse of the covariance matrix, the QF distance is called the Mahalanobis distance [22]. If the bin-similarity matrix is positive-semidefinitive, the A matrix can be expressed as $A = LL^T$ for some real matrix L. Thus, the distance can be computed as the Euclidean norm between linearly transformed vectors:

$$\mathrm{QF}^A(P,Q) = ||LP - LQ||_2 \tag{1.7}$$

In this case the QF distance is a psuedo-metric (as it is possible that $QF^A(P,Q) = 0$ for some $P \neq Q$). If A is positive-definitive, the QF distance is a metric.

Recently there have been a lot of research effort on learning a positive-semidefinitive matrix, A, from labeled examples or given relative distances constraints [2,23–31]. Although much of the metric-learning work has focused on the QF distance[1], this measure has some inherent limitations. The most important one is that it is a function only of the difference between input vectors $\vec{x}^i - \vec{x}^j$. Thus, it will give the same distance whenever two vectors have the same difference vector. For example, a QF distance will not able to differentiate between $(\vec{x}^i = 42, \vec{x}^j = 40)$ and $(\vec{x}^i = 2, \vec{x}^j = 0)$. To overcome this limitation several metric learning methods have been proposed [2, 27, 29, 32, 33]. Chopra et al. [32] proposed to learn a convolutional neural net as a non-linear transformation before applying the ℓ_2 norm. Babenko et al. [33] suggested a boosting framework for learning non-Mahalanobis metrics. These methods are non-convex and thus they might suffer from local minimas and training is sensitive to parameters. Kernel methods were also proposed in order to learn a QF distance over non-linear transformations of the data [2,27,29]. Computing such a distance between two vectors scales linear in the number of training examples, which makes it impractical for large datasets.

The Transportation and Earth Mover's Distances

The second type of distance that takes into account cross-bin relationships is the transportation distance (also known as Monge-Kantorovich, Mallows,

[1]The QF distance is called Mahalanobis distance in machine learning community even when A is not the inverse of the covariance matrix.

1st Wasserstein and match distance) which is defined between probability distributions:

$$\begin{aligned} \text{transportation}^D(P,Q) = \min_{\{F_{ij}\}} \sum_{i,j} F_{ij} D_{ij} \quad s.t \\ \sum_j F_{ij} = P_i \\ \sum_i F_{ij} = Q_j \\ \sum_{i,j} F_{ij} = \sum_i P_i = \sum_j Q_j = 1 \\ F_{ij} \geq 0 \end{aligned} \tag{1.8}$$

where $\{F_{ij}\}$ denotes the flows. Each F_{ij} represents the amount transported from the ith supply to the jth demand. We call D_{ij} the *ground distance* between bin i and bin j. When D_{ij} is a metric the transportation distance is also a metric.

The transportation problem was formalized in 1781 by Gaspard Monge [34]. Major advances were made in the field by Leonid Kantorovich [35]. Consequently, the problem is sometimes known as the Monge–Kantorovich transportation problem. Wasserstein proposed to use the definition of the transportation problem as a distance function [36]. Mallows also studied this distance [37].

Early work using the transportation distances for histogram comparison in the computer vision community can be found in [38–41]. Shen and Wong [38] proposed to unfold two integer histograms, sort them and then compute the L_1 distance between the unfolded histograms. To compute the modulo matching distance between cyclic histograms they proposed taking the minimum from all cyclic permutations. This distance is equivalent to transportation distance between two normalized histograms. Werman et al. [39] showed that this distance is equal to the L_1 distance between the cumulative histograms. They also proved that matching two cyclic histograms by examining only cyclic permutations is in effect optimal. Werman et al. [40] proposed an $O(M \log M)$ algorithm for finding a minimal matching between two sets of M points on a circle. Peleg et al. [41] suggested using the EMD for grayscale images and using linear programming to compute it.

Rubner et al. [1] suggested a new generalization of the transportation distance to non-normalized histograms, which he called the Earth Mover's Distance (EMD):

$$
\begin{aligned}
\mathrm{EMD}^D(P,Q) = \min_{\{F_{ij}\}} \frac{\sum_{i,j} F_{ij} D_{ij}}{\min(\sum_i P_i, \sum_j Q_j)} \quad s.t \\
\sum_j F_{ij} \leq P_i \\
\sum_i F_{ij} \leq Q_j \\
\sum_{i,j} F_{ij} = \min(\sum_i P_i, \sum_j Q_j) \\
F_{ij} \geq 0
\end{aligned}
\tag{1.9}
$$

The Earth Mover's Distance is the least amount of work needed in order to transform the smaller histogram into *part* of the bigger histogram (best partial matching), normalized with minimum between the sums of the histograms. It is a metric only if the histograms are normalized (then it equals the transportation distance) and D_{ij} is a metric. Rubner et al. suggested using EMD for color and texture images. They computed the EMD using a specific linear programming algorithm - the transportation simplex. The algorithm worst case time complexity is exponential. Practical run time was shown to be super-cubic ($\Omega(N^3) \cap O(N^4)$). Interior-point algorithms or Orlin's algorithm [42] both have a time complexity of $O(N^3 \log N)$ and can also be used.

A major research effort has been done on computing EMD-L_1, that is, EMD with L_1 as the ground distance. Ling and Okada [43] showed that if the points lie on a Manhattan network (*e.g.* an image), the number of variables in the LP problem can be reduced from $O(N^2)$ to $O(N)$. To execute the EMD-L_1 computation, they employed a tree-based algorithm, Tree-EMD. Tree-EMD exploits the fact that a basic feasible solution of the simplex algorithm-based solver forms a spanning tree when the EMD-L_1 is modeled as a network flow optimization problem. The worst case time complexity is exponential. Empirically, they showed that this algorithm has an average time complexity of $O(N^2)$. Gudmundsson et al. [44] also put forward this simplification of the LP problem. They suggested an $O(N \log^{d-1} N)$ algorithm that creates a Manhattan network for a set of N points in $\mathbb{R}^d$. The Manhattan network has $O(N \log^{d-1} N)$ vertices and edges. Thus, using Orlin's algorithm [42] the EMD-L_1 can be computed with a time complexity of $O(N^2 \log^{2d-1} N)$. Indyk and Thaper [45] proposed approximating EMD-L_1 by embedding it into the L_1 norm. Embedding time complexity is $O(Nd \log \Delta)$, where N is the feature set size, d is the feature space dimension and Δ is the diameter of the union of the two feature sets. Grauman and

Darrell [46] substituted L_1 with histogram intersection in order to approximate partial matching. Shirdhonkar and Jacobs [47] presented a linear-time algorithm for approximating EMD-L_1 for low dimensional histograms using the sum of absolute values of the weighted wavelet coefficients of the difference histogram. Khot and Naor [48] showed that any embedding of the EMD over the d-dimensional Hamming cube into L_1 must incur a distortion of $\Omega(d)$, thus losing practically all distance information. Andoni et al. [49] showed that for sets with cardinalities upper bounded by a parameter s, the distortion reduces to $O(\log s \log d)$. A practical reduction in accuracy due to the approximation was reported by [47, 50].

A major limitation of EMD-L_1 is its sensitivity to outliers. That is, L_1, as a ground distance between bins, assign different outliers different distances. Thus, in real-world applications the accuracy of EMD-L_1 (and same for EMD with L_2 ground distance) is often very low (often even lower than simple L_1 or χ^2 distances) [10, 51–54].

1.2 Novel Aspects

The main novelties of this thesis and several successful applications are:

I. We present $\widehat{EMD}$, a new generalization of the transportation distance for non-normalized histograms. Unlike the classic generalization (EMD), $\widehat{EMD}$ is a metric even when the histograms are non-normalized. In practice, $\widehat{EMD}$ is often better than EMD.

II. We present a new linear time algorithm for the computation of $\widehat{EMD}$ or EMD between two normalized cyclic histograms (*e.g.* histograms of oriented gradients) with L_1 as the ground distance. However, we show that using this distance for image descriptors matching, the accuracy is low. We explain the theoretical reason which is sensitivity to outliers.

III. We present a new linear time algorithm for the computation of a $\widehat{EMD}$ or EMD between two not necessarily normalized cyclic histograms with a robust ground distance. We show that this distance has excellent image descriptors matching performance on a widely used and publicly available benchmark. Gupta et al. [54] also reported excellent image descriptors matching results using our distance.

IV. We present a new efficient algorithm for the computation of robust $\widehat{EMD}$ or EMD between general non-negative vectors. The algorithm is faster by an order of magnitude than the classic algorithm. We show

excellent results for image retrieval. This algorithm was successfully used by other researchers both from the computer vision community and from other communities. For example, Boltz et al. used it for superpixels matching. In a comparison study between various image retargeting algorithms, Rubinstein et al. [55] reported that our distance had the best correlation with human judgements of image perceptual similarity (between an image and its retargeted version). Macindoe and Richards [56] used our method for social graph comparisons.

V. We present a new family of histogram distances, called *Quadratic-Chi* (QC). Members of this family are robust to quantization effects and have a χ^2-like normalization. Excellent image retrieval and shape classification results are demonstrated. Xu et al. [57] reported state-of-the-art wood recognition results using our QC distances.

VI. We present two new cross-bin distances properties *Similarity-Matrix-Quantization-Invariance* and *Sparseness-Invariance* and a proof that $\widehat{EMD}$ and QC distances have these properties. We show, experimentally, that these properties are practically important.

VII. We present a new family of distances called *Interpolated-Discretized* (ID) distances. These distances can approximate any kind of bin-to-bin distance and that their parameters are easy to learn. We show that our method often outperforms state-of-the-art metric learning approaches.

I., II. and III. are presented in chapter 2, IV. is presented in chapter 3, V. and VI. are presented in chapter 4 and VII. is presented in chapter 5.

Chapter 2

A Linear Time Histogram Metric for Improved SIFT Matching

A Linear Time Histogram Metric for Improved SIFT Matching

r ele and Michael erman

School of Computer Science and Engineering
The Hebrew University of Jerusalem
{ofirpele,werman}@cs.huji.ac.il

Abstract. We present a new metric between histograms such as SIFT descriptors and a linear time algorithm for its computation. It is common practice to use the L_2 metric for comparing SIFT descriptors. This practice assumes that SIFT bins are aligned, an assumption which is often not correct due to quantization, distortion, occlusion etc.

In this paper we present a new Earth Mover's Distance (EMD) variant. We show that it is a metric (unlike the original EMD [1] which is a metric only for normalized histograms). Moreover, it is a natural extension of the L_1 metric. Second, we propose a linear time algorithm for the computation of the EMD variant, with a robust ground distance for oriented gradients. Finally, extensive experimental results on the Mikolajczyk and Schmid dataset [2] show that our method outperforms state of the art distances.

1 Introduction

istograms of oriented gradient descri tors [3 4 5 6] are ubi uitous tools in nu merous com uter vision tas s. ne of the most successful is the cale nvariant Feature Transform (FT) [3]. n a recent erformance evaluation [2] the FT descri tor was shown to out erform other local descri tors. The FT descri tor has roven to be successful in a lications such as ob ect recognition [3 8 9] ob ect class detection [10 11 12] image retrieval [13 14 15 16] robot locali ation [1] building anoramas [18] and image classi cation [19].

t is common ractice to use the L_2 metric for com aring FT descri tors. This ractice assumes that the histogram domains are aligned. owever this assum tion is violated through uanti ation sha e deformation detector local i ation errors etc. lthough the FT algorithm has ste s that reduce the e ect of uanti ation this is still a liability as can be seen by the fact that increasing the number of orientation bins negatively a ects erformance [3].

The arth Mover s istance (M) [1] is a cross bin distance that addresses this alignment roblem. M is de ned as the minimal cost that must be aid to transform one histogram into the other where there is a ground distance between the basic features that are aggregated into the histogram. There are two main roblems with the M . First it is not a metric between non normali ed histograms. econd for a general ground distance it has a high run time.

D. Forsyth, P. Torr, and A. Zisserman (Eds.): ECCV 2008, Part III, LNCS 5304, pp. 495–508, 2008.

n this a er we resent an arth Mover s istance (M) variant. e show that it is a metric if it is used with a metric ground distance. econd we resent a linear time algorithm for the com utation of the M variant with a robust ground distance for oriented gradients. Finally we resent ex erimental results for FT matching on the Mi ola c y and chmid dataset [2] showing that our method out erforms state of the art distances such as L_2 M L_1 [20] di usion distance [21] and M $_{\text{MOD}}$ [22 23].

This a er is organi ed as follows. ection 2 is an overview of revious wor . ection 3 introduces the new M variant. ection 4 introduces the new FT metric and the linear time algorithm for its com utation. ection 5 describes the ex erimental setu and ection 6 resents the results. Finally conclusions are drawn in ection .

2 Previous Work

arly wor using cross bin distances for histogram com arison can be found in [22 24 25 26]. hen and ong [24] ro osed to unfold two integer histograms sort them and then com ute the L_1 distance between the unfolded histograms. To com ute the modulo matching distance between cyclic histograms they ro osed ta ing the minimum from all cyclic ermutations. This distance is e uivalent to M between two normali ed histograms. erman et al. [25] showed that this distance is e ual to the L_1 distance between the cumulative histograms. They also roved that matching two cyclic histograms by examining only cyclic er mutations is in e ect o timal. ha and rihari [2] rediscovered these algorithms and described a character writer identi cation a lication. erman et al. [22] ro osed an $O(M \log M)$ algorithm for nding a minimal matching between two sets of M oints on a circle. The algorithm can be ada ted to com ute the M between two N bin normali ed histograms with time com lexity $O(N)$ (endix in [23]).

eleg et al. [26] suggested using the M for grayscale images and using linear rogramming to com ute it. ubner et al. [1] suggested using M for color and texture images. They com uted the M using a s eci c linear ro gramming algorithm the trans ortation sim lex. The algorithm worst case time com lexity is ex onential. ractical run time was shown to be su er cubic ($(N^3) \cap O(N^4)$). nterior oint algorithms with time com lexity $O(N^3 \log N)$ can also be used. ll of these algorithms have high com utational cost.

ndy and Tha er [28] ro osed a roximating the M by embedding it into an uclidean s ace. mbedding time com lexity is $O(Nd \log$) where N is the feature set si e d is the feature s ace dimension and is the diameter of the union of the two feature sets.

ecently ing and ada ro osed general cross bin distances for histogram descri tors. The rst is M L_1[20] *i.e.* M with L_1 as the ground distance. To execute the M L_1 com utation they ro ose a tree based algorithm Tree M . Tree M ex loits the fact that a basic feasible solution of the sim lex algorithm based solver forms a s anning tree when the M L_1 is modeled as a networ ow o timi ation roblem. The worst case time com lexity is

ex onential. m irically they show that this new algorithm has an average time com lexity $O(N^2)$. ing and ada also ro osed the di usion distance [21]. They de ned the di erence between two histograms to be a tem erature eld. The di usion distance was derived as the sum of dissimilarities over scales. The algorithm run time is linear.

For a com rehensive review of M and its a lications in com uter vision we refer the reader to ing and ada s a er [20].

3 The New EMD ariant - $\widehat{EMD}$

This section introduces $\widehat{EMD}$ a new arth Mover s istance variant. e show that it is a metric (unli e the original M [1] which is a metric only for nor mali ed histograms). Moreover it is a natural extension of the L_1 metric.

The arth Mover s istance (M) [1] is de ned as the minimal cost that must be aid to transform one histogram into the other where there is a ground distance between the basic features that are aggregated into the histogram.

iven two histograms P, Q the M as de ned by ubner et al. [1] is

$$EMD(P,Q) = \min_{\{f_{ij}\}} \frac{\sum_{i,j} f_{ij} d_{ij}}{\sum_{i,j} f_{ij}} \quad s.t \tag{1}$$

$$\sum_j f_{ij} \le P_i \ , \ \sum_i f_{ij} \le Q_j \ , \ \sum_{i,j} f_{ij} = \min(\sum_i P_i, \sum_j Q_j) \ , \ f_{ij} \ge 0 \tag{2}$$

where $\{f_{ij}\}$ denotes the ows. ach f_{ij} re resents the amount trans orted from the ith su ly to the jth demand. e call d_{ij} the *ground distance* between bin i and bin j in the histograms.

e ro ose $\widehat{EMD}$

$$\widehat{EMD}_\alpha(P,Q) = (\min_{\{f_{ij}\}} \sum_{i,j} f_{ij} d_{ij}) + |\sum_i P_i - \sum_j Q_j| \times \alpha \max_{i,j}\{d_{ij}\} \ \ s.t \quad . 2 \tag{3}$$

ote that for two robability histograms (*i.e.* total mass e ual to one) EMD and $\widehat{EMD}$ are e uivalent. owever if the masses are not e ual $\widehat{EMD}$ adds one su lier or demander such that the masses on both sides becomes e ual. The ground distance between this su lier or demander to all other demanders or su liers res ectively is set to be α times the maximum ground distance. n addition the $\widehat{EMD}$ is not normali ed by the total ow.

ote that $\widehat{EMD}$ with $\alpha = 0.5$ and with the ronc er δ ground distance multi lied by two ($d_{ij} = 0$ if $i = j$ 2 otherwise) is e ual to the L_1 metric.

f $\alpha \ge 0.5$ and the ground distance is a metric $\widehat{EMD}$ is a metric (unli e the original EMD [1] which is a metric only for normali ed histograms). roof is given in endix [23]. eing a metric can lead to more e cient data structures and search algorithms.

e now give two exam les of when the usage of $\widehat{EMD}$ is more a ro riate (both of which are the case for the FT descri tors). The rst is when the total mass of the histograms is im ortant. For exam le let $P = (1, 0)$ $Q = (0, 1)$ $P' = (9, 0)$ $Q' = (0, 9)$. sing L_1 as a ground distance and $\alpha = 1$ $EMD(P, Q) = 1 = EMD(P', Q')$ while $\widehat{EMD}(P, Q) = 1 < 9 = \widehat{EMD}(P', Q')$. The second is when the di erence in total mass between histograms is a distinctive cue. For exam le let $P = (1, 0)$ $Q = (1,)$. sing L_1 as a ground distance and $\alpha = 1$ $EMD(P, Q) = 0$ while $\widehat{EMD}(P, Q) =$.

4 The SIFT$_{\text{DIST}}$ Metric

This section introduces FT$_{\text{DIST}}$ a new metric between FT descri tors. t is common ractice to use the L_2 metric for com aring FT descri tors. This ractice assumes that the FT histograms are aligned so that a bin in one histogram is only com ared to the corres onding bin in the other histogram. This is often not the case due to uanti ation distortion occlusion etc. ur distance has three instead of two matching costs ero cost for exact corres onding bins one cost for neighboring bins and two cost for farther bins and for the extra mass. Thus the metric is robust to small errors and outliers.

The section rst de nes FT$_{\text{DIST}}$ and then resents a linear time algorithm for its com utation.

4.1 SIFT$_{\text{DIST}}$ Definition

This section rst describes the FT descri tor. econd it ro oses Thresholded Modulo arth Mover s istance (M $_{\text{TMOD}}$) an M variant for oriented gra dient histograms. Finally it de nes the FT$_{\text{DIST}}$.

The FT descri tor [3] is a $M \times M \times N$ histogram. ach of the $M \times M$ s atial cells contains an N bin histogram of oriented gradients. ee Fig. 1 for a visuali ation of FT descri tors.

et $A = \{0, \ldots, N-1\}$ be N oints e ually s aced on a circle. The modulo L_1 distance for two oints $i, j \in A$ is

$$D_{MOD}(i, j) = \min(|i - j|, N - |i - j|) \tag{4}$$

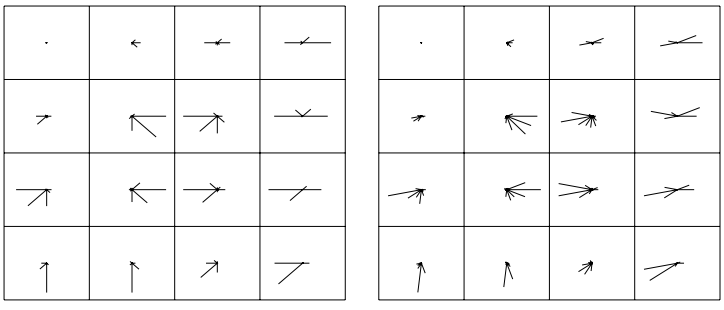


Fig. 1. (a) $4 \times 4 \times 8$ SIFT descriptor. (b) $4 \times 4 \times 16$ SIFT descriptor.

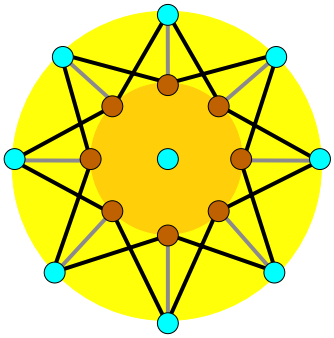


Fig. 2. The flow network of $\mathrm{EMD}_{\mathrm{TMOD}}$ for $N = 8$. The brown vertexes on the inner circle are the bins of the "supply" histogram, P. The cyan vertexes on the outer circle are the bins of the "demand" histogram, Q. We assume without loss of generality that $\sum_i P_i \geq \sum_j Q_j$; thus we add one infinite sink in the middle. The short gray edges are *zero-cost* edges. The long black edges are *one-cost* edges. *Two-cost* edges that turn the graph into a full bi-partite graph between sinks and sources are not colored for visibility.

The thresholded modulo L_1 distance is de ned as

$$D_{TDMO}(i,j) = \min\left(D_{MOD}(i,j), 2\right) \tag{5}$$

M $_{\mathrm{TMOD}}$ is de ned as $\widehat{EMD}$ (. 3) with ground distance $D_{TDMO}(i,j)$ and $\alpha = 1$. D_{TDMO} is a metric. t follows from endix [23] that M $_{\mathrm{TMOD}}$ is also a metric. sing M $_{\mathrm{TMOD}}$ the trans ortation cost to the two nearby bins is 1 while for farther bins and for the extra mass it is 2 (see Fig. 2). For exam le for $N = 16$ we assume that all di erences larger than 22.5° are caused by outliers and should be assigned the same trans ortation cost.

The $\mathrm{FT}_{\mathrm{DIST}}$ between two FT descri tors is de ned as the sum over the M $_{\mathrm{TMOD}}$ between all the $M \times M$ oriented gradient histograms. $\mathrm{FT}_{\mathrm{DIST}}$ is also an M where edges between s atial bins have an in nite cost.

$\widehat{EMD}$ (. 3) has two advantages over ubner s M (. 1) for com aring FT descri tors. First the di erence in total gradient magnitude between FT s atial cells is an im ortant distinctive cue. sing ubner s de nition this cue is ignored. econd $\widehat{EMD}$ is a metric even for non normali ed histograms.

4.2 A Linear Time $\mathrm{SIFT}_{\mathrm{DIST}}$ Algorithm

s $\mathrm{FT}_{\mathrm{DIST}}$ is a sum of M $_{\mathrm{TMOD}}$ solutions we resent a linear time algorithm for M $_{\mathrm{TMOD}}$.

i e all other arth Mover s istances M $_{\mathrm{TMOD}}$ can be solved by a max ow min cost algorithm. ach bin i in the rst histogram is connected to bin i in the second histogram with a *zero-cost* edge two nearby bins with *one-cost* edges and to all other bins with *two-cost* edges. ee Fig. 2 for an illustration of this ow networ .

The algorithm starts by saturating all *zero-cost* edges. s the ground distance (. 5) obeys the triangle ine uality this ste does not change the minimum cost

solution [22]. ote that after the rst ste nishes from each of the su lier demander airs that were connected with a *zero-cost edge* either the su lier is em ty or the demander is full.

To minimi e the cost after the rst ste the algorithm needs to maximi e the ow through the *one-cost* edges *i.e.* the roblem becomes a max ow roblem on a gra h with N vertexes and at most N edges where the maximum ath length is 1 as ow goes only from su ly to demand (see Fig. 2). This ste starts by chec ing whether all vertexes are degree 2 if yes we remove an arbitrary edge. econd we traverse the su liers vertexes cloc wise twice. For each degree one vertex we saturate its edge. f we did not remove an edge at the beginning of this ste there will be no augmenting aths. f an edge was removed the algorithm ows through all augmenting aths. ll these aths can be found by returning the edge and ex anding from it.

The algorithm nishes by owing through all *two-cost* edges which is e uiva lent to multi lying the maximum of the total remaining su ly and demand by two and adding it to the distance.

5 Experimental Setup

e evaluate $\mathrm{FT}_{\mathrm{DIST}}$ using a test rotocol similar to that of Mi ola c y and chmid [2]. The dataset was downloaded from [29]. The test data contain eight folders each with six images with di erent geometric and hotometric transfor mations and for di erent scene ty es. Fig. 3 shows the rst and the third image from each folder. ix image transformations are evaluated view oint change scale and rotation image blur light change and com ression. The images are either of lanar scenes or the camera osition was xed during ac uisition. The images are therefore always related by a homogra hy. The ground truth homogra hies are su lied with the dataset. For further details about the dataset we refer the reader to [2].

The evaluation criterion is based on the number of correct and false matches obtained for an image air. The match de nition de ends on the matching strat egy. The matching strategy we use is a symmetric version of owe s ratio match ing [3]. et $a \in A$ and $b \in$ be two descri tors and $n(a, A)$ and $n(b, \ )$ be their s atial neighbors that is all descri tors in A and res ectively such that the ratio of the intersection and union of their regions with a and b res ectively is larger than 0.5. a and b are matched if all the following conditions hold $a = \arg\min_{a' \in A} D(a', b)$ $a_2 = \arg\min_{a' \in A \setminus n(a,A)} D(a', b)$ $b = \arg\min_{b' \in B} D(a, b')$ $b_2 = \arg\min_{b' \in B \setminus n(b,B)} D(a, b')$ and $\min \ \frac{D(a_2,b)}{D(a,b)}, \frac{D(a,b_2)}{D(a,b)} \ \geq \ .$ value is varied to obtain the curves. ote that for $= 1$ the matching strategy is the symmetric nearest neighbor strategy. The techni ue of not using overly close regions as a second best match was used by Forssen and owe [30].

The match correctness is determined by the *overlap error* [2]. t measures how well regions A and corres ond under a nown homogra hy and is de ned by the ratio of the intersection and union of the regions $_S = 1 - \frac{A \bigcap H^T B H}{A \bigcup H^T B H}$.

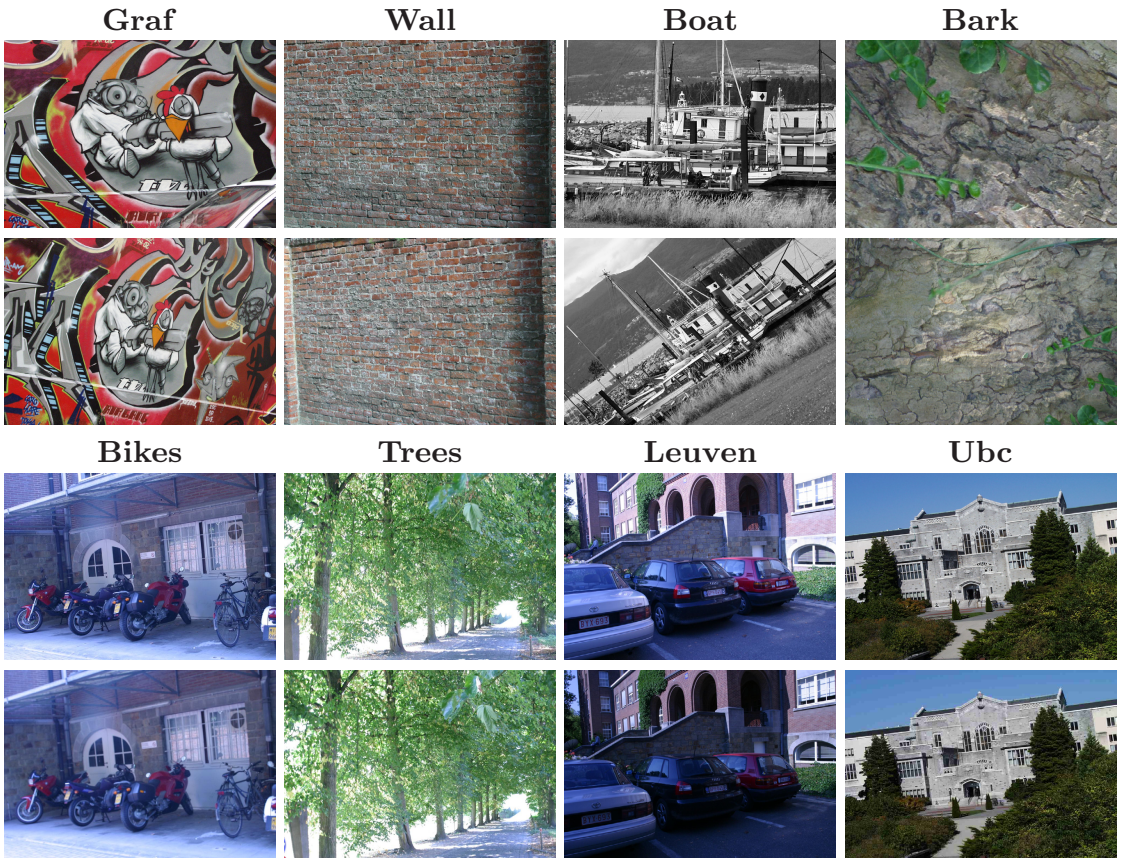


Fig. 3. Examples of test images (left): **Graf** (viewpoint change, structured scene), **Wall** (viewpoint change, textured scene), **Boat** (scale change + image rotation, structured scene), **Bark** (scale change + image rotation, textured scene), **Bikes** (image blur, structured scene), **Trees** (image blur, textured scene), **Leuven** (light change, structured scene), and **Ubc** (JPEG compression, structured scene)

s in [2] a match is assumed to be correct if $_S < 0.5$. The corres ondence number (ossible correct matches) is the maximum matching si e in the correct match bi artite gra h. The results are resented with *recall* versus 1 *precision* $recall = \frac{\#\text{correct matches}}{\#\text{correspondences}}$ $1 - precision = \frac{\#\text{false matches}}{\#\text{all matches}}$.

n all ex eriments we used edaldi s FT detector and descri tor im lemen tation [31]. ll arameters were set to the defaults exce t the oriented gradient bins number where we also tested our method with a FT descri tor having 16 oriented gradient bins (see Fig. 1 (b) in age 498).

6 Results

n this section we resent and discuss the ex erimental results. The erformance of FT_{DIST} is com ared to that of L_2 M L_1 [20] di usion distance [21] and M $_{\text{MOD}}$[1] [22 23] for view oint change scale and rotation image blur light

[1] Note that as the fast algorithm for EMD_{MOD} assumes normalized histograms, we normalized each histogram for its computation.

change and com ression. The matching was done between FT descri tors with eight and sixteen orientation bins (see Fig. 1 in age 498). Finally we resent run time results.

Figs. 4 5 6 are 1 recision vs. recall gra hs. ue to s ace constraints we resent gra hs for the matching of the rst to the third and fth images from each folder in the Mi ola c y and chmid dataset [29]. esults for the rest of the data are similar and are in [32].

n all of the ex eriments $\mathrm{FT}_{\mathrm{DIST}}$ com uted between FT descri tors with sixteen oriented gradient bins (FT 16) out erforms all other methods.

$\mathrm{FT}_{\mathrm{DIST}}$ com uted between the original FT descri tors with eight oriented gradient bins (FT 8) is usually the second. lso using $\mathrm{FT}_{\mathrm{DIST}}$ consistently roduces results with greater recision and recall for the symmetric nearest neighbor matching (the rightmost oint of each curve).

ncreasing the number of orientation bins from eight to sixteen decreases the erformance of L_2 and increases the erformance of $\mathrm{FT}_{\mathrm{DIST}}$. This can be ex lained by the $\mathrm{FT}_{\mathrm{DIST}}$ robustness to uanti ation errors.

Fig. 4 shows matching results for view oint change on structured (**Graf**) and textured (**Wall**) scenes . $\mathrm{FT}_{\mathrm{DIST}}$ has the highest ran ing. s in [2] erformance is better on the textured scene. erformance decreases with view oint change (com are **Graf-3** to **Graf-5** and **Wall-3** to **Wall-5**) while the distance ran ing remains the same. For large view oint change erformance is oor (see **Graf-5**) and the resulting gra hs are not smooth. This can be ex lained by the fact that the FT detector and descri tor are not a ne invariant.

Fig. 5 shows matching results for similarity transformation on structured (**Boat**) and textured (**Bark**) scenes. $\mathrm{FT}_{\mathrm{DIST}}$ out erforms all other distances. s in [2] erformance is better on the textured scene.

Fig. 6 shows matching results for image blur on structured (**Bikes**) and tex tured (**Trees**) scenes. $\mathrm{FT}_{\mathrm{DIST}}$ has the highest ran ing. The FT descri tor is a ected by image blur. similar observation was made by Mi ola c y and chmid [2].

Fig. **Leuven** shows matching results for light change. $\mathrm{FT}_{\mathrm{DIST}}$ obtains the best matching score. The erformance decreases with lac of light (com are **Leuven-3** to **Leuven-5**) although not drastically.

Fig. **Ubc** shows matching results for com ression. $\mathrm{FT}_{\mathrm{DIST}}$ out er forms all other distances. erformance decreases with com ression level (com are **Ubc-3** to **Ubc-5**).

Table 1 resent run time results. ll runs were conducted on a ual ore M teron 2.6 rocessor. The table contains the run time in seconds of

Table 1. Run time results in seconds of 10 distance computations

	$\mathrm{SIFT}_{\mathrm{DIST}}$	$(L_2)^2$	$\mathrm{EMD}_{\mathrm{MOD}}$ [22,23]	Diffusion [21]	EMD-L_1 [20]
SIFT-8	1.5	0.35	18	28	192
SIFT-16	2.2	1.3	29	56	637

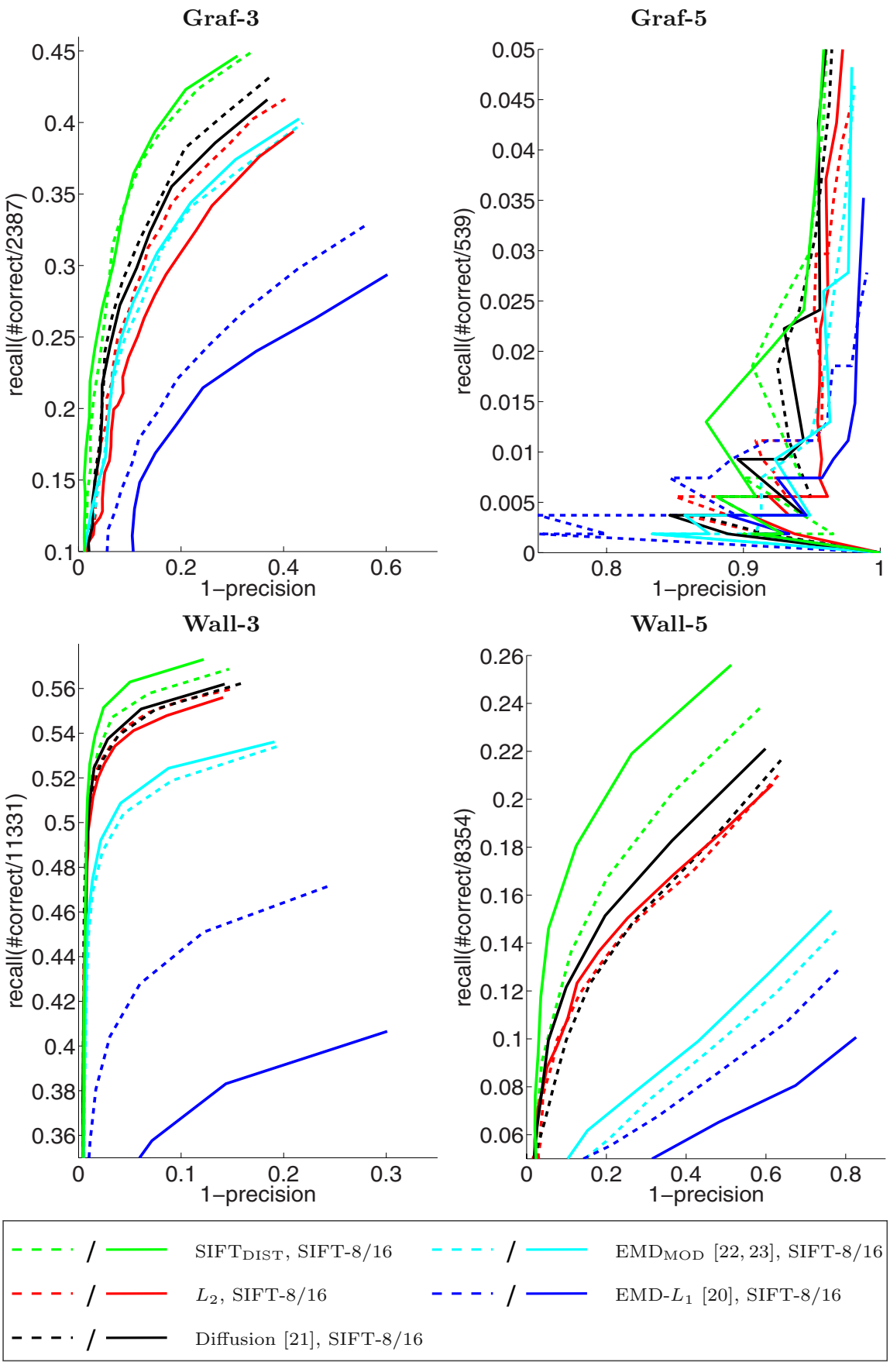


Fig. 4. Results on the Mikolajczyk and Schmid dataset [2]. Should be viewed in color.

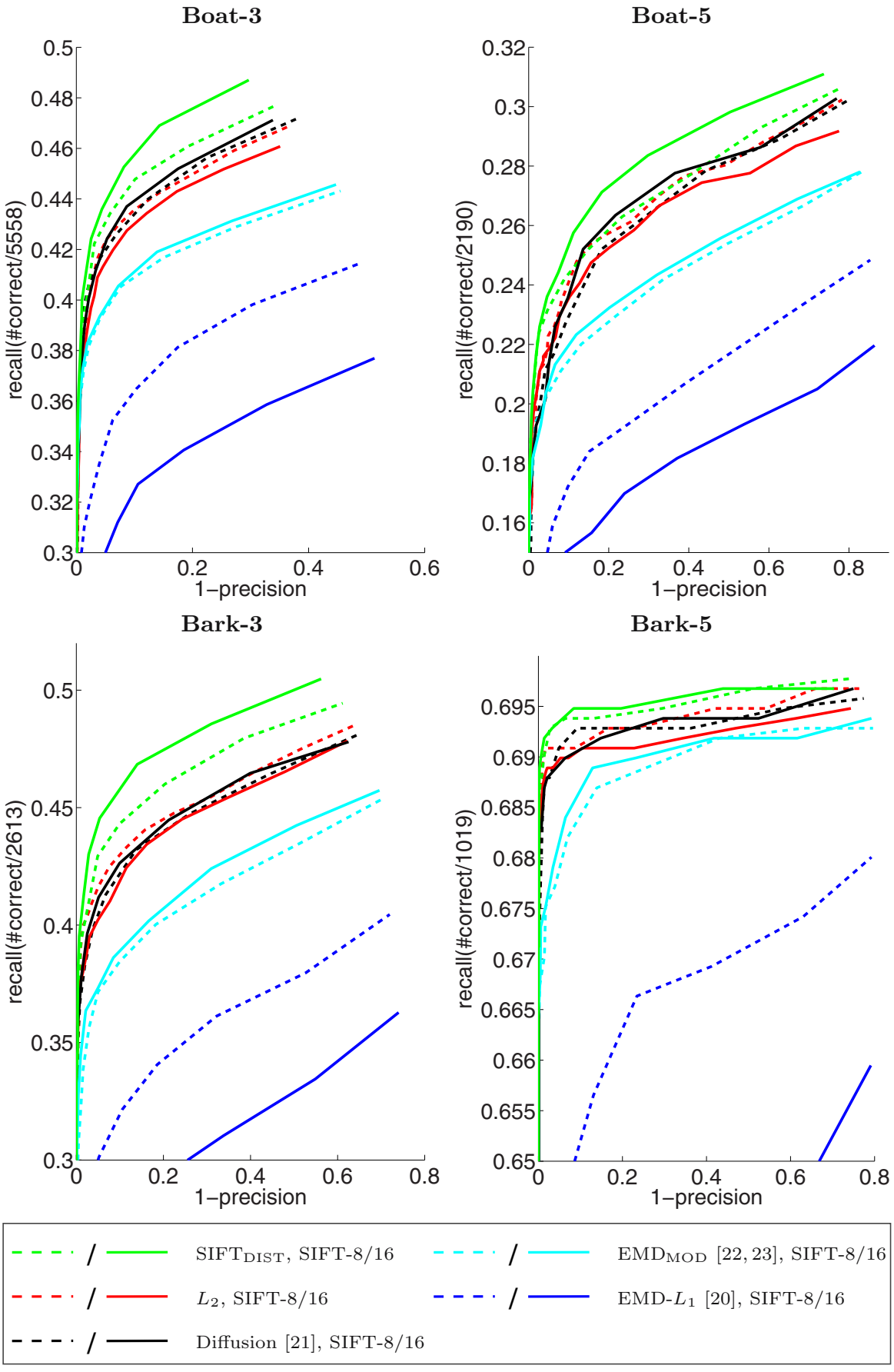


Fig. 5. Results on the Mikolajczyk and Schmid dataset [2]. Should be viewed in color.

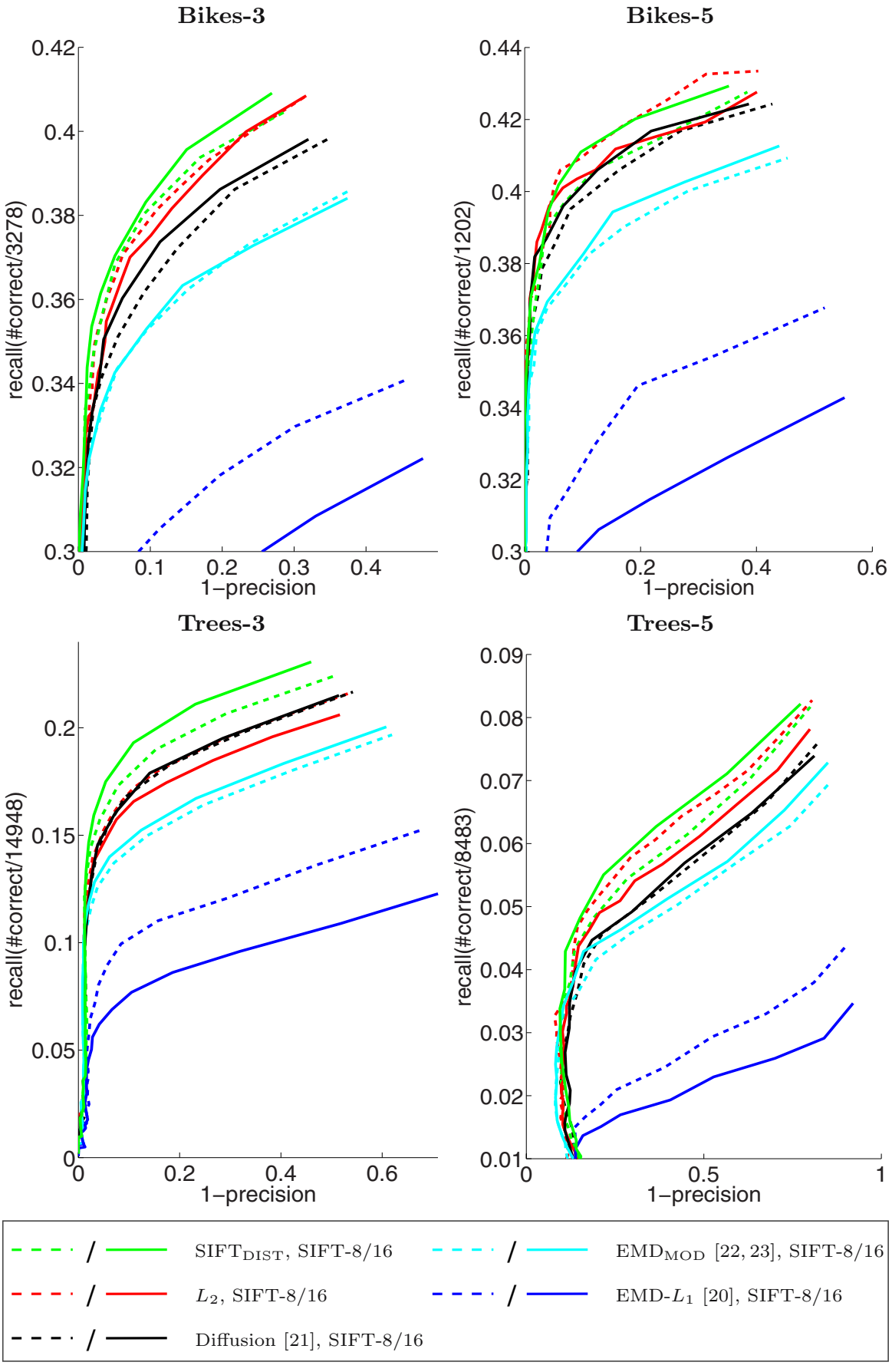


Fig. 6. Results on the Mikolajczyk and Schmid dataset [2]. Should be viewed in color.

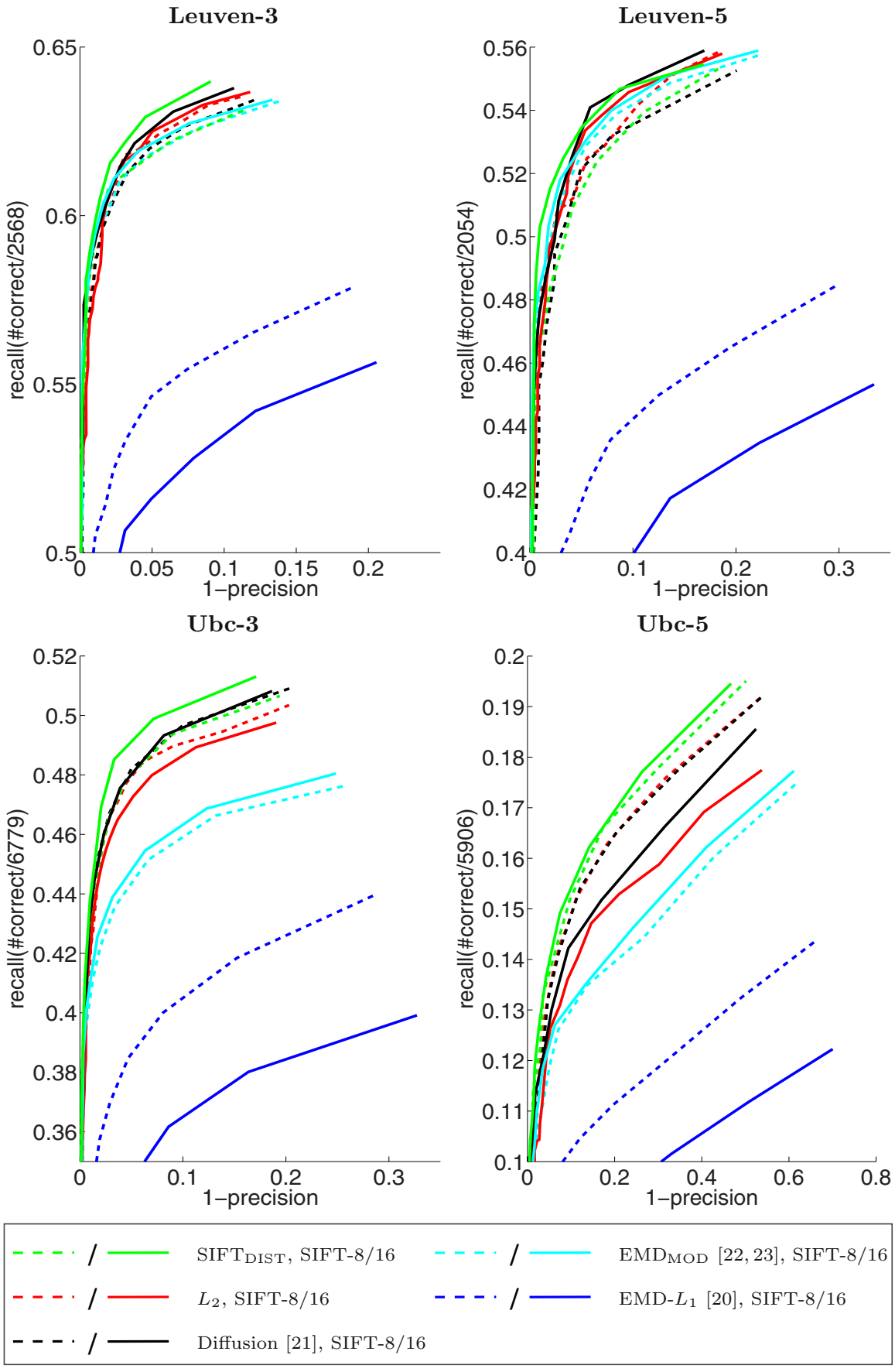


Fig. 7. Results on the Mikolajczyk and Schmid dataset [2]. Should be viewed in color.

each distance com utation between two sets of 1000 FT descri tors with eight and sixteen oriented gradient bins. ote that we measured the run time of $(L_2)^2$ and not L_2 as com uting the root does not change the order of elements and is time consuming. FT_{DIST} is the fastest cross bin distance.

7 Conclusions

e resented a new cross bin metric between histograms and a linear time algo rithm for its com utation. xtensive ex erimental results for FT matching on the Mi ola c y and chmid dataset [2] showed that our method out erforms state of the art distances.

The s eed can be further im roved using techni ues such as ayesian se uen tial hy othesis testing [33] sub linear indexing [34] and a roximate nearest neighbor [35 36]. The new cross bin histogram metric may also be useful for other histograms either cyclic (*e.g.* hue in color images) or non cyclic (*e.g.* in tensity in grayscale images). The ro ect home age including code (++ and Matlab wra ers) is at `http://www.cs.huji.ac.il/~ofirpele/SiftDist`.

References

1. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. International Journal of Computer Vision 40(2), 99–121 (2000)
2. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Trans. Pattern Analysis and Machine Intelligence 27(10), 1615–1630 (2005)
3. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60(2), 91–110 (2004)
4. Bay, H., Tuytelaars, T., Gool, L.J.V.: Surf: Speeded up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)
5. Dalai, N., Triggs, B., Rhone-Alps, I., Montbonnot, F.: Histograms of oriented gradients for human detection. In: CVPR, vol. 1 (2005)
6. Heikkila, M., Pietikainen, M., Schmid, C.: Description of Interest Regions with Center-Symmetric Local Binary Patterns. In: ICVGIP, pp. 58–69 (2006)
7. Ferrari, V., Tuytelaars, T., Van Gool, L.: Simultaneous object recognition and segmentation by image exploration. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3021, pp. 40–54. Springer, Heidelberg (2004)
8. Sudderth, E., Torralba, A., Freeman, W., Willsky, A.: Learning hierarchical models of scenes, objects, and parts. In: ICCV, vol. 2, pp. 1331–1338 (2005)
9. Arth, C., Leistner, C., Bischof, H.: Robust Local Features and their Application in Self-Calibration and Object Recognition on Embedded Systems. In: CVPR (2007)
10. Mikolajczyk, K., Leibe, B., Schiele, B.: Multiple object class detection with a generative model. In: CVPR (2006)
11. Dorko, G., Schmid, C., Gravir-Cnrs, I., Montbonnot, F.: Selection of scale-invariant parts for object class recognition. In: ICCV, pp. 634–639 (2003)
12. Opelt, A., Fussenegger, M., Pinz, A., Auer, P.: Weak hypotheses and boosting for generic object detection and recognition. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3022. Springer, Heidelberg (2004)

13. Sivic, J., Zisserman, A.: Video Google: a text retrieval approach to object matching in videos. In: ICCV, pp. 1470–1477 (2003)
14. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: CVPR (2007)
15. Snavely, N., Seitz, S., Szeliski, R.: Photo tourism: exploring photo collections in 3D. ACM Transactions on Graphics (TOG) 25(3), 835–846 (2006)
16. Sivic, J., Everingham, M., Zisserman, A.: Person Spotting: Video Shot Retrieval for Face Sets. In: Leow, W.-K., Lew, M., Chua, T.-S., Ma, W.-Y., Chaisorn, L., Bakker, E.M. (eds.) CIVR 2005. LNCS, vol. 3568, pp. 226–236. Springer, Heidelberg (2005)
17. Se, S., Lowe, D., Little, J.: Local and global localization for mobile robots using visuallandmarks. In: IROS, vol. 1 (2001)
18. Brown, M., Lowe, D.: Recognising panoramas. In: ICCV, p. 3 (2003)
19. Nowak, E., Jurie, F., Triggs, B.: Sampling strategies for bag-of-features image classification. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 490–503. Springer, Heidelberg (2006)
20. Ling, H., Okada, K.: An Efficient Earth Mover's Distance Algorithm for Robust Histogram Comparison. IEEE Trans. Pattern Analysis and Machine Intelligence 29(5), 840–853 (2007)
21. Ling, H., Okada, K.: Diffusion distance for histogram comparison. In: CVPR, vol. 1, pp. 246–253 (2006)
22. Werman, M., Peleg, S., Melter, R., Kong, T.: Bipartite graph matching for points on a line or a circle. Journal of Algorithms 7(2), 277–284 (1986)
23. http://www.cs.huji.ac.il/~ofirpele/publications/ECCV2008.pdf
24. Shen, H., Wong, A.: Generalized texture representation and metric. Computer vision, graphics, and image processing 23(2), 187–206 (1983)
25. Werman, M., Peleg, S., Rosenfeld, A.: A distance metric for multidimensional histograms. Computer Vision, Graphics, and Image Processing 32(3) (1985)
26. Peleg, S., Werman, M., Rom, H.: A unified approach to the change of resolution: Space and gray-level. IEEE Trans. Pattern Analysis and Machine Intelligence 11(7), 739–742 (1989)
27. Cha, S., Srihari, S.: On measuring the distance between histograms. Pattern Recognition 35(6), 1355–1370 (2002)
28. Indyk, P., Thaper, N.: Fast image retrieval via embeddings. In: 3rd International Workshop on Statistical and Computational Theories of Vision (October 2003)
29. http://www.robots.ox.ac.uk/~vgg/research/affine/index.html
30. Forssén, P., Lowe, D.: Shape Descriptors for Maximally Stable Extremal Regions. In: ICCV, pp. 1–8 (2007)
31. http://vision.ucla.edu/~vedaldi/code/sift/sift.html
32. http://www.cs.huji.ac.il/~ofirpele/publications/ECCV2008addRes.pdf
33. Pele, O., Werman, M.: Robust real time pattern matching using bayesian sequential hypothesis testing. IEEE Trans. Pattern Analysis and Machine Intelligence 30(8), 1427–1443 (2008)
34. Obdrzalek, S., Matas, J.: Sub-linear indexing for large scale object recognition. In: BMVC, vol. 1, pp. 1–10 (2005)
35. Arya, S., Mount, D., Netanyahu, N., Silverman, R., Wu, A.: An optimal algorithm for approximate nearest neighbor searching fixed dimensions. Journal of the ACM (JACM) 45(6), 891–923 (1998)
36. Beis, J., Lowe, D.: Shape indexing using approximate nearest-neighbour search in high-dimensional spaces. In: CVPR, pp. 1000–1006 (1997)

Chapter 3

Fast and Robust Earth Mover's Distances

Fast and Robust Earth Mover's Distances

Ofir Pele
The Hebrew University of Jerusalem
ofirpele@cs.huji.ac.il

Michael Werman
The Hebrew University of Jerusalem
werman@cs.huji.ac.il

Abstract

We present a new algorithm for a robust family of Earth Mover's Distances - EMDs with thresholded ground distances. The algorithm transforms the flow-network of the EMD so that the number of edges is reduced by an order of magnitude. As a result, we compute the EMD by an order of magnitude faster than the original algorithm, which makes it possible to compute the EMD on large histograms and databases. In addition, we show that EMDs with thresholded ground distances have many desirable properties. First, they correspond to the way humans perceive distances. Second, they are robust to outlier noise and quantization effects. Third, they are metrics. Finally, experimental results on image retrieval show that thresholding the ground distance of the EMD improves both accuracy and speed.

1. Introduction

Histograms are ubiquitous tools in numerous computer vision tasks. It is common practice to use distances such as L_2 or χ^2 for comparing histograms. This practice assumes that the histogram domains are aligned. However this assumption is violated through quantization, shape deformation, light changes, etc.

The Earth Mover's Distance (EMD) [29] is a cross-bin distance that addresses this alignment problem. EMD is defined as the minimal cost that must be paid to transform one histogram[1] into the other, where there is a "ground distance" between the basic features that are aggregated into the histogram. The EMD as defined by Rubner is a metric only for normalized histograms. However, recently Pele and Werman [26] suggested $\widehat{EMD}$ and showed that it is a metric for all histograms.

A major issue that arises when using EMD is which ground distance to use for the basic features. This, of course, depends on the histograms, the task and practical considerations. In many cases we would like the distance to correspond to the way humans perceive distances (image retrieval, for example). In other cases we would like the distance to fit the distribution of the noise (keypoint matching, for example). Practical considerations include speed of computation and the metric property that enables fast algorithms for nearest neighbor searches [41, 7], fast clustering [10] and large margin classifiers [15, 36].

We propose using thresholded ground distances. *i.e.* distances that saturate to a constant value. These distances have many desirable properties. First, saturated distances correspond to the way humans perceive distances [34]. Second, many natural noise distributions have a heavy tail; *i.e.* outlier noise. Thresholded distances assign different outliers the same large distance. Finally, we present an algorithm that computes EMD with a thresholded ground distance faster by an order of magnitude than the original algorithm. The algorithm transforms the flow-network of the EMD so that the number of edges is reduced by an order of

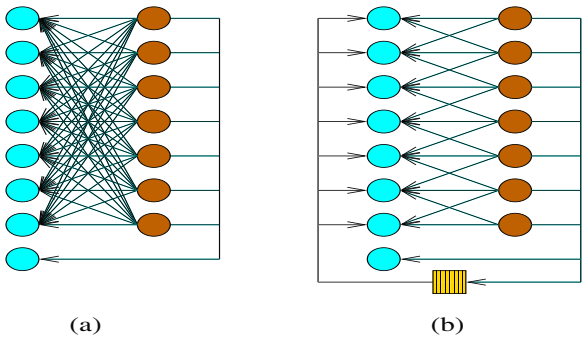


Figure 1. An example of the transformation on a flow network of an EMD or $\widehat{EMD}$ with a ground distance of: $d(a,b) = min(2, |a-b|)$. (a) is the original flow network with N^2+N edges. Note that $N(N-3)$ of these edges have cost 2. The bottom cyan vertex on the left is the sink that handles the difference between the total mass of the two histograms (ingoing edges cost is 0 for EMD and $\alpha \max_{ij} d_{ij}$ for $\widehat{EMD}$). (b) is the transformed flow network. The striped yellow square is the new transhipment vertex. Ingoing edge cost is the threshold (*e.g.* 2) and outgoing edge cost is 0.

[1] Rubner's noted that EMD can be used with sparse histograms which he coined *signatures*. Our algorithm is applicable to both histograms and signatures.

2009 IEEE 12th International Conference on Computer Vision (ICCV)
978-1-4244-4419-9/09/$25.00 ©2009 IEEE

magnitude (see Fig. 1).

The Earth Mover's Distance has been used successfully in many applications such as image retrieval [29, 23], edge and corner detection [30], keypoint matching [26, 8, 20], near duplicate image identification [40], classification of texture and object categories [42, 19], NMF [31] and contour matching [12]. Many of these works used saturated distances, usually the negative exponent function. The major contribution of this paper is a fast algorithm for the computation of the EMD with thresholded ground distance. We argue that thresholded distances have all the benefits of the negative exponent function that is typically used as a saturated distance; its big advantage is its much shorter computation time.

This paper is organized as follows. Section 2 is an overview of previous work. Section 3 describes the Earth Mover's Distance. Section 4 discusses thresholded distances and proves that they are metrics. Section 5 describes the fast algorithm. Section 6 presents the results. Finally, conclusions are drawn in Section 7.

2. Previous Work

This section first describes EMD algorithms. Second, it describes the use of saturated ground distances in the EMD framework.

2.1. EMD Algorithms

Early work using cross-bin distances for histogram comparison can be found in [33, 39, 38, 28]. Shen and Wong [33] suggested unfolding two integer histograms, sorting them and then computing the L_1 distance between the unfolded histograms. To compute the modulo matching distance between cyclic histograms they took the minimum from all cyclic permutations. This distance is equivalent to the EMD between two normalized histograms. Werman *et al*. [39] showed that this distance is equal to the L_1 distance between the cumulative histograms. They also proved that matching two cyclic histograms by only examining cyclic permutations is optimal. Werman *et al*. [38] proposed an $O(M \log M)$ algorithm for finding a minimal matching between two sets of M points on a circle. The algorithm was adapted by Pele and Werman [26] to compute the EMD between two N-bin, normalized histograms with time complexity $O(N)$. Peleg *et al*. [28] suggested using the EMD for grayscale images and using linear programming to compute it. Rubner *et al*. [29] suggested using the EMD for color and texture images and generalized the definition of the EMD to non-normalized histograms. They computed the EMD using a specific linear programming algorithm - the transportation simplex. The algorithm's worst case time complexity is exponential. Practical run time was shown to be super-cubic. Interior-point algorithms or Orlin's algorithm [25] both have a time complexity of $O(N^3 \log N)$ and can also be used.

Ling and Okada proposed EMD-L_1[20]; *i.e*. EMD with L_1 as the ground distance. They showed that if the points lie on a Manhattan network (*e.g*. an image), the number of variables in the LP problem can be reduced from $O(N^2)$ to $O(N)$. To execute the EMD-L_1 computation, they employed a tree-based algorithm, Tree-EMD. Tree-EMD exploits the fact that a basic feasible solution of the simplex algorithm-based solver forms a spanning tree when the EMD-L_1 is modeled as a network flow optimization problem. The worst case time complexity is exponential. Empirically, they showed that this algorithm has an average time complexity of $O(N^2)$. Gudmundsson *et al*. [14] also put forward this simplification of the LP problem. They suggested an $O(N \log^{d-1} N)$ algorithm that creates a Manhattan network for a set of N points in $\mathbb{R}^d$. The Manhattan network has $O(N \log^{d-1} N)$ vertices and edges. Thus, using Orlin's algorithm [25] the EMD-L_1 can be computed with a time complexity of $O(N^2 \log^{2d-1} N)$. Indyk and Thaper [17] proposed approximating EMD-L_1 by embedding it into the L_1 norm. Embedding time complexity is $O(Nd \log \Delta)$, where N is the feature set size, d is the feature space dimension and Δ is the diameter of the union of the two feature sets. Grauman and Darrell [13] substituted L_1 with histogram intersection in order to approximate partial matching. Shirdhonkar and Jacobs [35] presented a linear-time algorithm for approximating EMD-L_1 for low dimensional histograms using the sum of absolute values of the weighted wavelet coefficients of the difference histogram. Lv *et al*. [23] proposed embedding an EMD with thresholded ground distance into the L_1 norm. Khot and Naor [18] showed that any embedding of the EMD over the d-dimensional Hamming cube into L_1 must incur a distortion of $\Omega(d)$, thus losing practically all distance information. Andoni *et al*. [5] showed that for sets with cardinalities upper bounded by a parameter s, the distortion reduces to $O(\log s \log d)$. A practical reduction in accuracy due to the approximation was reported by [12, 23, 35]. In order to increase precision, Grauman and Darrell [12] and Lv *et al*. [23] used the approximation as a filter that returns a set of similar objects, and then used the exact EMD computation to rerank these objects. Khanh Do Ba *et al*. [6] presented optimal algorithms for estimating EMD-L_1 or EMD with a tree-metric as the ground distance.

Pele and Werman [26] proposed $\widehat{EMD}$ - a new definition of the EMD for non-normalized histograms. They showed that unlike Rubner's definition, the $\widehat{EMD}$ is also a metric for non-normalized histograms. In addition, they proposed a linear-time algorithm that computes the $\widehat{EMD}$ with a ground distance of 0 for corresponding bins, 1 for adjacent bins and 2 for farther bins and for the extra mass.

2.2. Saturated Ground Distances with the EMD

Rubner *et al*. [29] and Ruzon and Tomasi [30] used a negative exponent function to saturate their ground distance for the tasks of image retrieval and edge detection, respectively. The negative exponent function practically saturates large distances to a fixed threshold (see Fig. 2). The negative exponent function is used for saturating a metric, because it does not break the triangle inequality. We show that this is true as well for a thresholding function. Note that saturating with a negative exponent might have the drawback of changing the behavior of small distances (see Fig. 2).

Lv *et al*. [23] conducted image retrieval experiments from a database of 10000 images. They showed that thresholding the ground distance improves precision.

Pele and Werman [26] compared several distances for the task of SIFT matching. They proposed an $\widehat{EMD}$ variant. The ground distance of this $\widehat{EMD}$ is 0 for corresponding bins, 1 for adjacent bins and 2 for farther bins and for the extra mass; *i.e*. a thresholded distance. They showed that this distance improves SIFT matching, while $\widehat{EMD}$ with non-thresholded distances negatively affects performance.

3. The Earth Mover's Distance

The Earth Mover's Distance (EMD) [29] is defined as the minimal cost that must be paid to transform one histogram into the other, where there is a "ground distance" between the basic features that are aggregated into the histogram.

Given two histograms P, Q the EMD as defined by Rubner *et al*. [29] is:

$$\text{EMD}(P,Q) = (\min_{\{f_{ij}\}} \sum_{i,j} f_{ij} d_{ij}) / (\sum_{i,j} f_{ij}) \quad s.t \quad f_{ij} \geq 0$$

$$\sum_j f_{ij} \leq P_i \quad \sum_i f_{ij} \leq Q_j \quad \sum_{i,j} f_{ij} = \min(\sum_i P_i, \sum_j Q_j)$$

where $\{f_{ij}\}$ denotes the flows. Each f_{ij} represents the amount transported from the ith supply to the jth demand. We call d_{ij} the *ground distance* between bin i and bin j in the histograms. Pele and Werman [26] suggested $\widehat{EMD}$:

$$\widehat{EMD}_\alpha(P,Q) = (\min_{\{f_{ij}\}} \sum_{i,j} f_{ij} d_{ij}) + |\sum_i P_i - \sum_j Q_j| \alpha \max_{i,j} d_{ij}$$

$s.t$ EMD constraints

Pele and Werman proved that $\widehat{EMD}$ is a metric for any two histograms if the ground distance is a metric and $\alpha \geq 0.5$ [26]. The metric property enables fast algorithms for nearest neighbor searches [41, 7]), fast clustering [10] and large margin classifiers [15, 36]. Because of these advantages we will use the Pele and Werman definition in the remainder of this paper (with $\alpha = 1$).

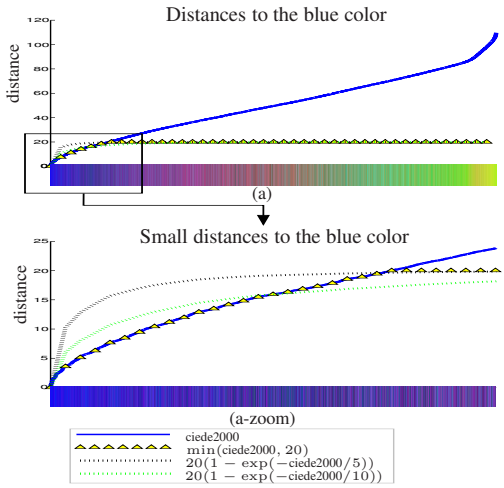


Figure 2. This figure should be viewed in color, preferably on a computer screen. The x-axes are colors, sorted by their distance to the blue color. The distances are the ciede2000 distance [22] and three monotonic saturating transformations applied to it: two negative exponent functions and a thresholding function; *i.e*. a minimum function. There are several observations we can derive from these graphs. First, although ciede2000 is considered as state of the art, it is still far from perfect, especially in the medium to large distance range. Second, color distances should be saturated. For example, although red and yellow are both simply different from blue, the ciede2000 distance between blue and red is 56, while the ciede2000 distance between blue and yellow is 102. This was already noted by Rubner *et al*. [29] and Ruzon and Tomasi [30] who suggested using a negative exponent function because it is a metric, if the distance that is raised to the power is a metric. In this paper we show that the thresholding function is also a metric if the thresholded distance is a metric. Finally, for most of the range, the negative exponent and thresholded functions are very similar. They mostly differ for small distances, where the negative exponent changes the original distance, while the thresholding function does not. Since the ciede2000 was designed and perceptually tested on this range [22], changing the distances on this range can negatively affect performance as was noted by Rubner *et al*. for the Euclidean distance on L*a*b* space [29].

4. Thresholded Distances

Thresholded distances are distances that saturate to a threshold; *i.e*. let $d(a,b)$ be a distance measure between two features - a, b. The thresholded distance with a threshold of $t > 0$ is defined as: $d_t(a,b) = \min(d(a,b), t)$.

We now prove that if d is a metric then d_t is also a metric. Non-negativity and symmetry hold trivially, so we only need to prove that the triangle inequality holds.

$d_t(a,b) + d_t(b,c) \geq d_t(a,c)$ if d is a metric.

We consider three cases:

1. $\Big(d_t(a,b) < t\Big) \wedge \Big(d_t(b,c) < t\Big) \wedge \Big(d_t(a,c) < t\Big) \Rightarrow$
$\Big(d_t(a,b) = d(a,b)\Big) \wedge \Big(d_t(b,c) = d(b,c)\Big) \wedge$
$\Big(d_t(a,c) = d(a,c)\Big) \Rightarrow d_t(a,b) + d_t(b,c) \geq d_t(a,c)$
2. $\Big(d_t(a,b) = t\Big) \vee \Big(d_t(b,c) = t\Big) \Rightarrow$
$d_t(a,b) + d_t(b,c) \geq t \geq d_t(a,c)$
3. $d_t(a,c) = t \Rightarrow$

Assume for contradiction that:

$$d_t(a,b) + d_t(b,c) < d_t(a,c) = t \Rightarrow$$
$$\Big(d_t(a,b) < t\Big) \wedge \Big(d_t(b,c) < t\Big) \Rightarrow$$
$$\Big(d_t(a,b) = d(a,b)\Big) \wedge \Big(d_t(b,c) = d(b,c)\Big) \Rightarrow$$
$$d(a,b) + d(b,c) < t = d_t(a,c) \leq d(a,c) \Rightarrow$$
$$d(a,b) + d(b,c) < d(a,c)$$

The last statement contradicts the triangle inequality of the metric d. The union of cases 2 and 3 is complement of case 1. Thus, cases 1-3 constitute the entire event space. Therefore we proved that d_t is a metric if d is a metric. It is noteworthy that d_t can be a metric even if d is not a metric.

5. Fast Computation of the EMD with a Thresholded Ground Distance

This section describes an algorithm that computes $\widehat{EMD}$ or EMD with a thresholded ground distance an order of magnitude faster than the original algorithm [2].

$\widehat{EMD}$ can be solved by a min-cost-flow algorithm. Our algorithm makes a simple transformation of the flow network that reduces the number of edges. If N is the number of bins in the histogram, the flow network of $\widehat{EMD}$ has exactly $N^2 + N$ edges (see (a) in Fig. 1). N^2 edges connect all sources to all sinks. The extra N edges connect all sources to the sink that handles the difference between the total mass of the two histograms (we assume without loss of generality that the source histogram total mass is greater or equal to the sink histogram total mass).

The transformation (see Fig. 1) first removes all edges with cost t. Second, it adds a new transhipment vertex. Finally we connect all sources to this vertex with edges of cost t and connect the vertex to all sinks with edges of cost 0.

Let K be the average number of edges going out of each bin that have a cost different than the threshold t. The new flow network has $NK + N$ edges from the original network, N edges connecting all sources to the transhipment vertex and N edges connecting the transhipment vertex to all sinks. Thus the total number of edges is $N(K + 3)$. If K is a constant the number of edges is $O(N)$ as opposed to the original $\Theta(N^2)$. Note that the new flow network is no longer a transportation problem, but a transhipment problem [2]. However, both are special cases of the min-cost-flow problem. Thus any algorithm that solves min-cost-flow can be used for both problems.

Let $K = O(1)$; the min-cost-flow optimization problem can be solved with a worst case time complexity of:

$$O(\min((N^2 \log\log U \log(NC)), \quad (N^2 \log U \sqrt{\log C}),$$
$$(N^2 \log U \log N), \quad (N^2 \log^2 N)))$$

The algorithms are taken from: Ahuja *et al*. [1], Edmonds and Karp [9], used with Ahuja *et al*.'s shortest path algorithm [3], Edmonds and Karp [9], used with Fredman and Tarjan's shortest path algorithm [11] and Orlin [25]. Algorithms with a C term assume integral cost coefficients that are bounded by C. Algorithms with a U term assume integral supply and demands that are bounded by U.

We now prove that the original and the transformed flow networks have the same minimum-cost solution. Let $\mathcal{O}$ be the original flow network and let $\mathcal{T}$ be the transformed flow network. We first show how to create a feasible flow in $\mathcal{T}$, given a feasible flow in $\mathcal{O}$. Both flows will have the same cost. This will prove that the min-cost-flow solution for $\mathcal{T}$ is smaller or equal to the min-cost-flow solution for $\mathcal{O}$.

Given a feasible flow in $\mathcal{O}$, all flows on edges that were not removed are copied to the new flow for $\mathcal{T}$. For flows on edges with the cost of the threshold, we transfer the flow through the transhipment vertex. This gives us a feasible flow in $\mathcal{T}$ with the same cost.

We now show how to create a feasible flow in $\mathcal{O}$, given a feasible flow in $\mathcal{T}$. The flow in $\mathcal{O}$ will have a cost smaller or equal to the cost of the flow in $\mathcal{T}$. This will prove that the min-cost-flow solution for $\mathcal{T}$ is greater or equal to the min-cost-flow solution for $\mathcal{O}$. Together with the previous proof, this shows that the two flow networks have the same min-cost solution.

Given a feasible flow in $\mathcal{T}$, all flows on edges not connected to the transhipment vertex are copied to $\mathcal{O}$. Second, each unit of mass that flows from vertex i to the transhipment vertex and then from the transhipment vertex to vertex j is transferred directly from vertex i to vertex j. We note that this is possible since $\mathcal{O}$ is fully bi-partite. We also note that the cost of the new flow will be smaller or equal to the cost of the flow in $\mathcal{T}$ as all edges in $\mathcal{O}$ have a cost smaller or equal to the threshold. This completes the proof.

It is noteworthy that the algorithmic technique presented in this paper can be applied not only with thresholded ground distance, but in any case where a group of vertexes can be connected to another group with the same cost. For example, we can add a transhipment vertex for all vertexes with a specific color (*e.g*. blue) such that the cost of the transportation between them will be smaller than the cost of the transportation to other colors.

[2]Note that the optimization problems in EMD and $\widehat{EMD}$ are exactly the same. Thus, any algorithm that computes EMD can compute $\widehat{EMD}$ and vice versa with the same time complexity.

5.1. Implementation notes

Flow-netowrk set-up time. For a fixed histogram configuration (*e.g*. SIFT) the flow-network can be pre-computed once. For sparse histograms (*signatures*), the flow-network set-up time complexity is $O(MN)$; where N is the number of non-zero bins and M is the average number of neighbors that need to be checked if the distance is lower than the threshold. M is at most N but it can be lower. For example, let t be the threshold. If we are comparing two images and the ground distance is a linear combination of the spatial distance and the color distance, then the distance computation to vertexes with L_1 distance bigger or equal to t can be skipped. That is, the set-up time in this case is $O(\min(t^2N, N^2))$.

Pre-flowing Monge sequences. A Monge sequence contains edges in the flow-network that can be pre-flowed (in the order of the sequence) without changing the min-cost solution [24, 16]. For example, if the ground-distance is a metric, zero-cost edges are Monge sequence [38]. Alon *et al*. [4] introduced an efficient algorithm which determines the longest Monge sequence.

Pre-flowing to/from isolated nodes. If a source is connected only to the new transhipment vertex, we can pre-flow all its mass to the transhipment vertex and eliminate it. If a sink is connected only to the transhipment vertex, we can add its deficit to the transhipment vertex and eliminate it.

6. Results

In this section we present results for image retrieval. However, note that we do not claim that this is the optimal way to achieve image retrieval. Image retrieval is used here as an example of an application where the EMD has already been used, to show that thresholded distances yield good results. The major contribution is the faster algorithm. We show that by using our algorithm the running time decreases by an order of magnitude.

We use a database that contains 773 landscape images from the COREL database, that were also used in Wang *et al*. [37]. The dataset contains 10 classes[3]: People in Africa, Beaches, Outdoor Buildings, Buses, Dinosaurs, Elephants, Flowers, Horses, Mountains and Food. The number of images in each class ranges from 50 to 100.

From each class we selected 5 images as query images (numbers $1, 10, \ldots, 40$). Then we searched for the 50 nearest neighbors for each query image. We computed the distance of each image to the query image and its reflection and took the minimum. We present results for three types of image representations: histograms of orientations, L*a*b* color space and finally linear combinations of the two. We conclude this section with running times.

[3]The original database contains some visually ambiguous classes such as Africa that also contains images of beaches in Africa. We manually removed these ambiguous images.

6.1. SIFT

Our first image representation is orientation histograms. The first representation - SIFT is a $6 \times 8 \times 8$ SIFT descriptor [21] computed globally on the whole image. The second representation - CSIFT is a SIFT-like descriptor. This descriptor tackles two problems related to the SIFT descriptor for color image retrieval. First, it takes into account color edges by computing the SIFT descriptor on the compass edge image [30]. Note that on an edge image there should be no distinction between opposite directions (0 and 180 for example). Thus, opposite directions are considered equal. The second drawback of the SIFT descriptor for color image retrieval is its normalization. The normalization is problematic as we lose the distinctive cue of the amount of edge points in the image. In the CSIFT computation we skip the normalization step. The final CSIFT descriptor has $6 \times 8 \times 8$ bins. We used the following distances for these descriptors: L_1, L_2, χ^2, EMD-L_1 [20], SIFT$_{\text{DIST}}$ [26] and $\widehat{EMD}$. Let M be the number of orientation bins. The ground distances between bins (x_i, y_i, o_i) and (x_j, y_j, o_j) we use are:

$$d_R = ||(x_i, y_i) - (x_j, y_j)||_2 + \min(|o_i - o_j|, M - |o_i - o_j|)$$
$$d_T = \min(d_R, T)$$

The results are given in Fig. 4(a). Due to lack of space we present for each distance measure, the descriptor with which it performed best. Results of all pairs of descriptors and distance measures can be found at: `www.cs.huji.ac.il/%7eofirpele/FastEMD/`. The $\widehat{EMD}$ with a thresholded ground distance performs much better than $\widehat{EMD}$ with a non-thresholded ground distance. In fact, while $\widehat{EMD}$ with a non-thresholded ground distance negatively affects performance, $\widehat{EMD}$ with a thresholded ground distance improves performance. L_1 is equivalent to $\widehat{EMD}$ with the Kroncker δ ground distance [26]. SIFT$_{\text{DIST}}$ is the sum of $\widehat{EMD}$ over all the spatial cells (each spatial cell contains one orientation histogram). The ground distance for the orientation histograms is: $\min(|o_i - o_j|, M - |o_i - o_j|, 2)$. It was shown in [26] and here that this addition of a small invariance to the orientations shifts improves performance. Our distance also adds a small invariance to spatial shifts and thus improves performance even more. However, using a non-thresholded distance adds too much invariance at the expense of distinctiveness, thus reducing performance.

6.2. L*a*b* Color Space

The state of the art color distance is Δ_{00} - ciede2000 on L*a*b* color space[22, 32] (see also Fig. 2). Thus, our second type of image representation is simply a resized color image in the L*a*b* space. We resized each image to 32×48 and converted them to L*a*b* space. Let I_1, I_2 be

Figure 3. Example of image retrieval using the best distance - $\widehat{EMD}$ with thresholded ground distances (top row) and the second best distance - L_1 like distance (bottom row). The nearest neighbor images are ordered from left to right by their distance from the query image. Note that by allowing small deformations in the $\widehat{EMD}$ we obtain results that are visually similar to the query image. Allowing larger deformations; *i.e.* using a non-thresholded distance negatively affects results.

the two L*a*b* images. We used the following distances:

$$L_1\Delta_{00} = \sum_{x,y}(\Delta_{00}(I_1(x,y), I_2(x,y)))$$

$$L_1\Delta_{00}^T = \sum_{x,y}(\min(\Delta_{00}(I_1(x,y), I_2(x,y)), T))$$

$$L_2\Delta_{00} = \sum_{x,y}(\Delta_{00}(I_1(x,y), I_2(x,y)))^2$$

$$L_2\Delta_{00}^T = \sum_{x,y}(\min(\Delta_{00}(I_1(x,y), I_2(x,y)), T))^2$$

We also used $\widehat{EMD}$, where the ground distance between two pixels $(x_i, y_i, L_i, a_i, b_i)$, $(x_j, y_j, L_j, a_j, b_j)$ is:

$$dc_T = \min(||(x_i, y_i) - (x_j, y_j)||_2 + \Delta_{00}((L_i, a_i, b_i), (L_j, a_j, b_j)), T)$$

The results for $\widehat{EMD}$ with a non-thresholded ground distance are not reported here since the experiments have not finished running (more than ten days). Results are presented in Fig. 4(b). As shown, $\widehat{EMD}$ with a thresholded ground distance outperforms all other distances.

6.3. Color and SIFT Combined

In the experiments described in this section, we used linear combinations of the orientation histograms and color. We combined the two best distances for each of the methods (note that each distance was normalized so that its average is 1):

$$dC = \alpha\left(\widehat{EMD}\ \ d_{T=2}, \text{CSIFT}\right) + (1-\alpha)\left(\widehat{EMD}\ \ dc_{T=20}\right)$$

We also used a combination of the two L_1-like distances:

$$dC2 = \alpha\left(L_1, SIFT\right) + (1-\alpha)\left(L_1\Delta_{00}^{T=20}\right)$$

We used three different α values: 0.1, 0.3, 0.5. The results appear in Fig. 4(c)-(f). The combination of two $\widehat{EMD}$ with thresholded distances performs best, especially for hard classes such as People in Africa and Food. The image results examples for one "Food" query image are given in Fig. 3. More image results can be found at:
`www.cs.huji.ac.il/%7eofirpele/FastEMD/`

6.4. Running Time Results

The algorithm we used for the computation of the $\widehat{EMD}$ is successive shortest path [2]. This algorithm has a worst time complexity of $O(N^2U \log N)$. All runs were conducted on a Pentium 2.8GHz. A comparison of the practical running time of our algorithm and other methods are given in tables 1,2 and in Fig. 5. It is noteworthy that EMD-L_1 accuracy is much lower than our method and even lower than the simple L_1 norm (see Fig. 4(a)). Indyk and Thaper [17] and Shirdhonkar and Jacobs [35] approximate EMD-L_1, so their accuracy is even lower. $\text{SIFT}_{\text{DIST}}$ gives good accuracy (second best, see Fig. 4(a)) and is faster than our method. Our method has better accuracy. More importantly, $\text{SIFT}_{\text{DIST}}$ is limited to a thresholded L1 norm with a threshold of 2 for 1-dimensional histograms. Our method is much more general. For example, $\text{SIFT}_{\text{DIST}}$ cannot be applied to colors. Lv *et al.* [23] report that their approximation runs 5 times faster than Rubner's. Our method runs 75-700 times faster and returns the exact distance.

Our	Rubner's [29]	EMD-L_1[20]	$\text{SIFT}_{\text{DIST}}$[26]
0.04s	3s	0.04s	0.00007s

Table 1. 384-dimensional SIFT-like descriptors matching time

Our	Rubner's [29]	EMD-L_1[20]	$\text{SIFT}_{\text{DIST}}$[26]
6s	4400s	N.A.	N.A.

Table 2. L*a*b* images matching time. Note that EMD-L_1 can be applied only to regular grids and $\text{SIFT}_{\text{DIST}}$ can be applied only to 1-dimensional histograms.

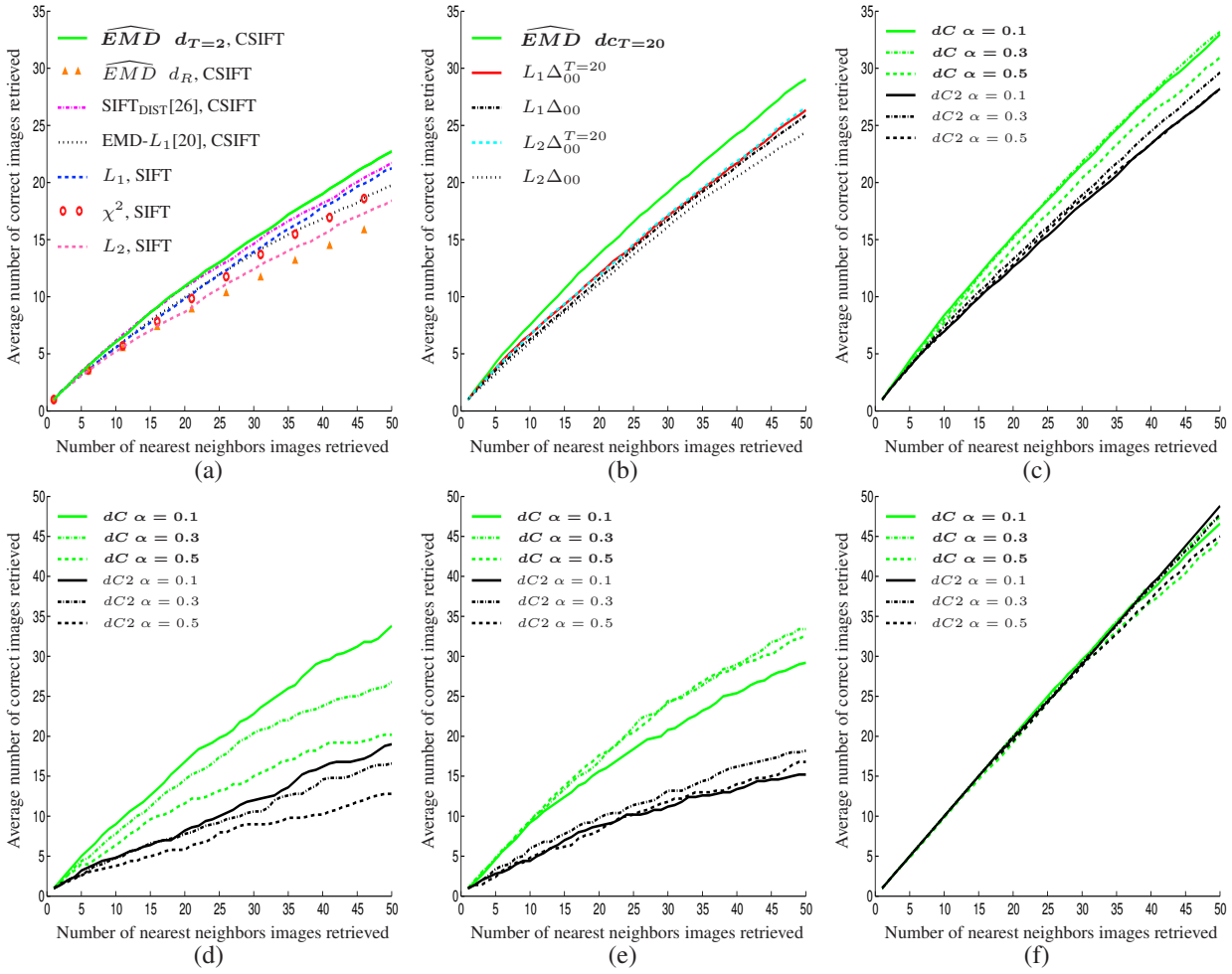


Figure 4. Results for image retrieval. Our method is in bold font. (a) Orientation histogram results. Due to lack of space we present for each distance measure, the descriptor with which it performed best. Results of all pairs of descriptors and distance measures can be found at:
`www.cs.huji.ac.il/%7eofirpele/FastEMD/`. (b) L*a*b* color space results. (c)-(f) Linear combinations of color and orientation results, where (c) is an average over all classes and (d)-(f) are for specific classes: (d) People in Africa (e) Food (f) Flowers. There are two key observations. First, $\widehat{EMD}$ with thresholded ground distances performs best. Second, some of the classes are easy, *e.g.* (f) Flowers. For these classes $\widehat{EMD}$ does not improve performance. However, for harder classes, such as (d) People in Africa and (e) Food, $\widehat{EMD}$ significantly improves results. Image result examples for one "Food" query image are given in Fig. 3.

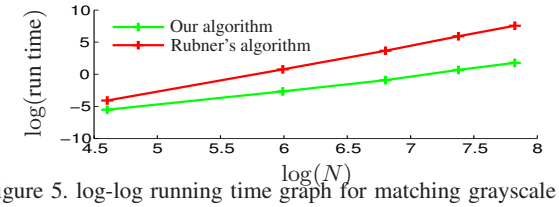


Figure 5. log-log running time graph for matching grayscale images with N pixels.. Linear fit for our algorithm: $\mathbf{2.3} \times \log(N) - 16$. Linear fit for Rubner's algorithm: $\mathbf{3.6} \times \log(N) - 21$.

7. Conclusions

We presented a new family of Earth Mover's Distances. Members of this family correspond to the way humans perceive distances, and are robust to outlier noise and quantization effects. We proved that they are metrics. We also proposed a fast algorithm. The algorithm runs an order of magnitude faster than the original algorithm, which makes it possible to compute the EMD on large histograms and databases. Experimental results show that EMD has the best performance when it is used with thresholded distances. This has also been shown by Rubner *et al.* [29] and Ruzon and Tomasi [30] for saturated distances, which are essentially thresholded. This has also been demonstrated for thresholded distances by Lv *et al.* [23] and Pele and Werman [26]. Our results strengthen these findings. Most im-

portantly, our paper shows that using a thresholded distance not only improves accuracy, but reduces the run time, using our algorithm. The speed can be further improved using techniques such as Bayesian sequential hypothesis testing [27]. The project homepage, including code (C++ and Matlab wrappers) is at:
`www.cs.huji.ac.il/%7eofirpele/FastEMD/`.

References

[1] R. Ahuja, A. Goldberg, J. Orlin, and R. Tarjan. Finding minimum-cost flows by double scaling. *Mathematical Programming*, 1992.

[2] R. Ahuja, T. Magnanti, and J. Orlin. *Network flows: theory, algorithms, and applications*. 1993.

[3] R. Ahuja, K. Mehlhorn, J. Orlin, and R. Tarjan. Faster algorithms for the shortest path problem. *JACM*, 1990.

[4] N. Alon, S. Cosares, D. S. Hochbaum, and R. Shamir. An algorithm for the detection and construction of monge sequences. *LAA*, 1989.

[5] A. Andoni, P. Indyk, and R. Krauthgamer. Earth mover distance over high-dimensional spaces. In *SODA*, 2008.

[6] K. D. Ba, H. L. Nguyen, H. N. Nguyen, and R. Rubinfeld. Sublinear time algorithms for earth mover's distance. *CoRR*, 2009.

[7] P. Ciaccia, M. Patella, and P. Zezula. M-tree: An Efficient Access Method for Similarity Search in Metric Spaces. In *ICVLDB*, 1997.

[8] R. Collins and W. Ge. CSDD Features: Center-Surround Distribution Distance for Feature Extraction and Matching. In *ECCV*, 2008.

[9] J. Edmonds and R. Karp. Theoretical Improvements in Algorithmic Efficiency for Network Flow Problems. *JACM*, 1972.

[10] C. Elkan. Using the Triangle Inequality to Accelerate k-Means. In *ICML*, 2003.

[11] M. Fredman and R. Tarjan. Fibonacci heaps and their uses in improved network optimization algorithms. *JACM*, 1987.

[12] K. Grauman and T. Darrell. Fast contour matching using approximate earth mover's distance. In *CVPR*, 2004.

[13] K. Grauman and T. Darrell. The pyramid match kernel: Efficient learning with sets of features. *JMLR*, 2007.

[14] J. Gudmundsson, O. Klein, C. Knauer, and M. Smid. Small Manhattan Networks and Algorithmic Applications for the Earth Movers Distance. In *EWCG*, 2007.

[15] M. Hein, O. Bousquet, and B. Schölkopf. Maximal margin classification for metric spaces. *JCSS*, 2005.

[16] A. Hoffman. On simple linear programming problems. In *SPM*, 1963.

[17] P. Indyk and N. Thaper. Fast image retrieval via embeddings. In *IWSCTV*, 2003.

[18] S. Khot and A. Naor. Nonembeddability theorems via Fourier analysis. *Mathematische Annalen*, 2006.

[19] S. Lazebnik, C. Schmid, and J. Ponce. A sparse texture representation using local affine regions. *PAMI*, 2005.

[20] H. Ling and K. Okada. An Efficient Earth Mover's Distance Algorithm for Robust Histogram Comparison. *PAMI*, 2007.

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

[22] M. Luo, G. Cui, and B. Rigg. The Development of the CIE 2000 Colour-Difference Formula: CIEDE2000. *CRA*, 2001.

[23] Q. Lv, M. Charikar, and K. Li. Image similarity search with compact data structures. In *ICIKM*, 2004.

[24] G. Monge. Déblai et remblai. *Mémoires de l'Académie des Sciences*, 1781.

[25] J. Orlin. A faster strongly polynomial minimum cost flow algorithm. In *STOC*, 1988.

[26] O. Pele and M. Werman. A linear time histogram metric for improved sift matching. In *ECCV*, 2008.

[27] O. Pele and M. Werman. Robust real time pattern matching using bayesian sequential hypothesis testing. *PAMI*, 2008.

[28] S. Peleg, M. Werman, and H. Rom. A unified approach to the change of resolution: Space and gray-level. *PAMI*, 1989.

[29] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *IJCV*, 2000.

[30] M. Ruzon and C. Tomasi. Edge, Junction, and Corner Detection Using Color Distributions. *PAMI*, 2001.

[31] R. Sandler and M. Lindenbaum. Nonnegative Matrix Factorization with Earth Movers Distance Metric. In *CVPR*, 2009.

[32] G. Sharma, W. Wu, and E. Dalal. The CIEDE2000 color-difference formula: implementation notes, supplementary test data, and mathematical observations. *CRA*, 2005.

[33] H. Shen and A. Wong. Generalized texture representation and metric. *CVGIP*, 1983.

[34] R. Shepard. Toward a universal law of generalization for psychological science. *Science*, 1987.

[35] S. Shirdhonkar and D. Jacobs. Approximate earth movers distance in linear time. In *CVPR*, 2008.

[36] U. von Luxburg and O. Bousquet. Distance–Based Classification with Lipschitz Functions. *JMLR*, 2004.

[37] J. Wang, J. Li, and G. Wiederhold. SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture LIbraries. *PAMI*, 2001.

[38] M. Werman, S. Peleg, R. Melter, and T. Kong. Bipartite graph matching for points on a line or a circle. *JoA*, 1986.

[39] M. Werman, S. Peleg, and A. Rosenfeld. A distance metric for multidimensional histograms. *CVGIP*, 1985.

[40] D. Xu, T. Cham, S. Yan, and S. Chang. Near Duplicate Image Identification with Spatially Aligned Pyramid Matching. In *CVPR*, 2008.

[41] P. Yianilos. Data structures and algorithms for nearest neighbor search in general metric spaces. In *SODA*, 1993.

[42] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *IJCV*, 2007.

Chapter 4

The Quadratic-Chi Histogram Distance Family

The Quadratic-Chi Histogram Distance Family

Ofir Pele and Michael Werman

School of Computer Science
The Hebrew University of Jerusalem
{ofirpele,werman}@cs.huji.ac.il

Abstract. We present a new histogram distance family, the Quadratic-Chi (QC). QC members are Quadratic-Form distances with a cross-bin χ^2-like normalization. The cross-bin χ^2-like normalization reduces the effect of large bins having undo influence. Normalization was shown to be helpful in many cases, where the χ^2 histogram distance outperformed the L_2 norm. However, χ^2 is sensitive to quantization effects, such as caused by light changes, shape deformations etc. The Quadratic-Form part of QC members takes care of cross-bin relationships (*e.g.* red and orange), alleviating the quantization problem. We present two new cross-bin histogram distance properties: *Similarity-Matrix-Quantization-Invariance* and *Sparseness-Invariance* and show that QC distances have these properties. We also show that experimentally they boost performance. QC distances computation time complexity is linear in the number of non-zero entries in the bin-similarity matrix and histograms and it can easily be parallelized. We present results for image retrieval using the Scale Invariant Feature Transform (SIFT) and color image descriptors. In addition, we present results for shape classification using Shape Context (SC) and Inner Distance Shape Context (IDSC). We show that the new QC members outperform state of the art distances for these tasks, while having a short running time. The experimental results show that both the cross-bin property and the normalization are important.

1 Introduction

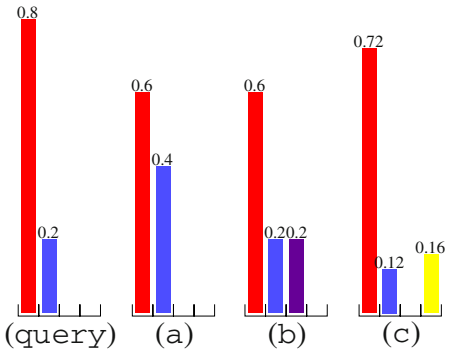
It is common practice to use bin-to-bin distances such as the L_1 and L_2 norms for comparing histograms. This practice assumes that the histogram domains are aligned. However this assumption is violated in many cases due to quantization, shape deformation, light changes, etc. Bin-to-bin distances depend on the number of bins. If it is low, the distance is robust, but not discriminative, if it is high, the distance is discriminative, but not robust. Distances that take into account cross-bin relationships (cross-bin distances) can be both robust and discriminative.

There are two kinds of cross-bin distances. The first is the Quadratic-Form distance [1]. Let P and Q be two histograms and A the bin-similarity matrix. The Quadratic-Form distance is defined as:

$$\text{QF}\ \ (P,Q) = \sqrt{(P-Q)^T A (P-Q)} \tag{1}$$

When the bin-similarity matrix A is the inverse of the covariance matrix, the Quadratic-Form distance is called the Mahalanobis distance. If the bin-similarity matrix is positive-definitive, then the Quadratic-Form distance is a metric. In this case the

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part II, LNCS 6312, pp. 749–762, 2010.



distance	(a)-(c) ordered by their distance (in small fonts) to the (query)					
QCN(our)	(a)	0.35	(b)	0.62	(c)	0.86
QCS(our)	(a)	0.31	(b)	0.41	(c)	0.43
QF	(c)	0.20	(a)	0.28	(b)	0.28
EMD	(c)	3.20	(a)	4.00	(b)	4.00
L_1	(c)	0.32	(a)	0.40	(b)	0.40
χ^2	(a)	0.05	(c)	0.09	(b)	0.11

Fig. 1. This figure should be viewed in color, preferably on a computer screen. A toy example showing the behavior of distances that reduce the effect of large bins and the behavior of distances that take cross-bin relationships into account. We show four color histograms, each histogram has four colors: red, blue, purple, and yellow. The Quadratic-Form (QF), the Earth Mover Distance (EMD) and the L_1 norm do not reduce the effect of large bins. Thus, they rank (query) to be more similar to (c) than to (a). χ^2 considers (a) to be more similar, but as it does not take cross-bin relationships into account it fails with (b). Our proposed members of the Quadratic-Chi histogram distance family, QCN and QCS consider (a) to be most similar, (b) the second and (c) the least similar as they take into account cross-bin relationships and reduce the effect of large bins, using an appropriate normalization.

Quadratic-Form distance is the L_2 norm between linear transformations of P and Q. If the bin-similarity matrix is positive-semidefinite, then the Quadratic-Form distance is a semi-metric.

The second type of distance that takes into account cross-bin relationships is the Earth Mover's Distance (EMD). EMD was defined by Rubner et al. [2] as the minimal cost that must be paid to transform one histogram (P) into the other (Q):

$$\begin{aligned} \text{EMD} \quad (P,Q) &= (\min_{\{\ _{ij}\}} \sum_{i,j} F_{ij} D_{ij}) / (\sum_{i,j} F_{ij}) \quad s.t \quad F_{ij} \geq 0 \\ \sum_j F_{ij} \leq P_i \quad \sum_i F_{ij} \leq Q_j \quad \sum_{i,j} F_{ij} &= \min(\sum_i P_i, \sum_j Q_j) \end{aligned} \tag{2}$$

where $\{\ _{ij}\}$ denotes the flows. Each $_{ij}$ represents the amount transported from the ith supply to the jth demand. We call D_{ij} the *ground distance* between bin i and bin j. If D_{ij} is a metric, the EMD as defined by Rubner is a metric only for normalized histograms. Recently Pele and Werman [3] suggested EMD:

$$\begin{aligned} \text{EMD}_\alpha (P,Q) &= (\min_{\{\ _{ij}\}} \sum_{i,j} F_{ij} D_{ij}) + |\sum_i P_i - \sum_j Q_j| \alpha \max_{i,j} D_{ij} \\ s.t \ \ &\text{EMD constraints} \end{aligned} \tag{3}$$

If D_{ij} is a metric and $\alpha \geq \frac{1}{2}$, EMD is a metric for all histograms [3]. For normalized histograms EMD and EMD are equal (*e.g.* Fig. 1).

In many natural histograms the difference between large bins is less important than the difference between small bins and should be reduced. See for example Fig. 1. The Chi-Squared (χ^2) is a histogram distance that takes this into account. It is defined as:

$$\chi^2(P,Q) = \frac{1}{2}\sum_i \frac{(P_i - Q_i)^2}{(P_i + Q_i)} \tag{4}$$

The χ^2 histogram distance comes from the χ^2 test-statistic [4] where it is used to test the fit between a distribution and observed frequencies. In this paper the histograms are not necessarily normalized, and thus not probabilities vectors. χ^2 was successfully used for texture and object categories classification [5,6,7], near duplicate image identification[8], local descriptors matching [9], shape classification [10,11] and boundary detection [12]. The χ^2, like other bin-to-bin distances such as the L_1 and the L_2 norms, is sensitive to quantization effects.

2 Our Contribution

In this paper we present a new cross-bin histogram distance family: Quadratic-Chi (QC). Like the Quadratic-Form, its members take cross-bin relationships into account. Like the χ^2, its members reduce the effect of differences caused by bins with large values. We discuss QC members' properties, including a formalization of a two new cross-bin histogram distance properties: *Similarity-Matrix-Quantization-Invariance* and *Sparseness-Invariance*. We show that all QC members and the EMD have these properties. We also show importance experimentally.

For full histograms QC distances computation time is linear in the number of non-zero entries in the bin-similarity matrix. In this case, QC distances can be implemented with 5 lines of Matlab code (see Algorithm 1). For two sparse histograms (for example bag-of-words histograms) with a total of non-zeros entries and an average of K non-zeros entries in each row of the similarity matrix, a QC distance computation time complexity is $O(\ K)$. See code (C++ and Matlab wrappers) at:
`http://www.cs.huji.ac.il/~ofirpele/QC/`. Finally, QC distances' parallelization is trivial.

We present results for image retrieval on the Corel dataset using the SIFT descriptor [13] and small color images. We also present results for shape classification using

Algorithm 1. Quadratic-Chi Matlab Code for Full Histograms

```
function dist= QC(P,Q,A,m)

Z= (P+Q)*A;
% 1 can be any number as Z_i==0 iff D_i=0
Z(Z==0)= 1;
Z= Z.^m;
D= (P-Q)./Z;
% max is redundant if A is positive-semidefinite
dist= sqrt( max(D*A*D',0) );
```

Shape Context (SC) [10] and Inner Distance Shape Context (IDSC) [11]. QC members performance is excellent. They outperform state of the art distances including χ^2, QF, L_1, L_2, EMD[14], $\text{SIFT}_{\text{DIST}}$[3], EMD-$L_1$[15], Diffusion[16], Bhattacharyya [17], Kullback-Leibler[18] and Jensen-Shannon[19] while having a short running time. We have found that the normalization is very important. Surprisingly, excellent performance was achieved using a new bin-to-bin distance from the QC family, that has a large normalization factor. Its cross-bin version yielded an additional improvement, outperforming all other distances for SIFT, SC and IDSC.

3 The Quadratic-Chi Histogram Distance Family

3.1 The Quadratic-Chi Histogram Distance Definition

Let P and Q be two non-negative bounded histograms. That is, $P, Q \in [0, U]^N$. Let A be a non-negative symmetric bounded bin-similarity matrix such that each diagonal element is bigger or equal to every other element in its row (this demand is weaker than being a strongly dominant matrix). That is, $A \in [0, U]^N \times [0, U]^N$ and $\forall i, j \; A_{ii} \geq A_{ij}$. Let $0 \leq \quad < 1$ be the normalization factor. A Quadratic-Chi (QC) histogram distance is defined as:

$$\text{QC}_m^A(P,Q) = \sum_{ij} \frac{P_i - Q_i}{\sum_c (P_c + Q_c) A_{ci}{}^m} \frac{P_j - Q_j}{\sum_c (P_c + Q_c) A_{cj}{}^m} A_{ij} \quad (5)$$

where we define $\frac{0}{0} = 0$. If A is positive-semidefinite, the argument inside the square root (the sum) is non-negative. If A is not positive-semidefinite we can get non-real (complex) distances. This is true also for the Quadratic-Form (Eq. 1). We prefer not to restrict ourselves to positive-semidefinite matrices. On the other hand, we don't want non-real distances. So, we define a complex distance as zero. In practice, this was never needed, even with non-positive-semidefinite matrices. This is due to the fact that the eigenvectors of the similarity matrices corresponding to negative eigenvalues were very far from smooth, while the difference vector for natural histograms P and Q is usually very smooth, see Fig. 2.

Each addend's denominator inside the square root is zero if and only if the addend's numerator is zero. A $\text{QC}_m^A(P,Q)$ distance is continuous. In particular, if the addend's denominator tends to zero, the whole addend tends to zero. Proofs are in [20].

The Quadratic-Chi distance family generalizes both the Quadratic-Form (QF) and a monotonic transformation of χ^2. That is, $\text{QC}_0^A(P,Q) = \text{QF}^A(P,Q)$ and if I is the identity matrix, $\text{QC}_{0.5}^I(P,Q) = \sqrt{2\chi^2(P,Q)}$.

3.2 Metric Properties

There are three conditions for a distance function, $\mathcal{D}$, to be a semi-metric. The first is *non-negativity* (*i.e.* $\mathcal{D}(P,Q) \geq 0$), the second is *symmetry* (*i.e.* $\mathcal{D}(P,Q) = \mathcal{D}(Q,P)$) and the third is *subadditivity* (*i.e.* $\mathcal{D}(P,Q) \leq \mathcal{D}(P,K) + \mathcal{D}(K,Q)$). $\mathcal{D}$ is a metric if it is

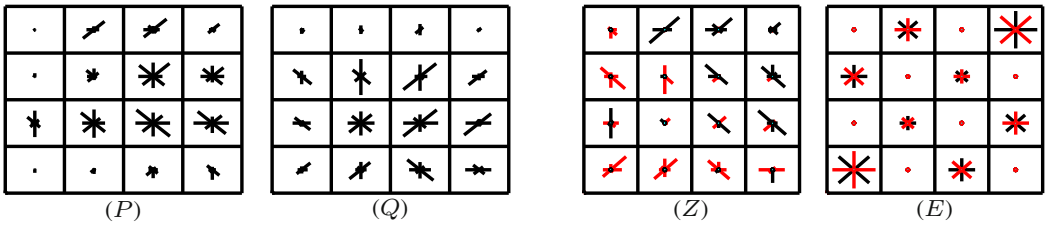


Fig. 2. This figure illustrates why it is not likely to get negative values in the square root argument of a QC distance for natural histograms and a typical similarity matrix. P and Q are two SIFT histograms. Z is the normalized difference vector. That is: $Z_i = \frac{}{\sum_c \quad c \quad c \quad ci \quad m} - \frac{}{\sum_c \quad c \quad c \quad ci \quad m}$. Negative values are represented with red, positive values are represented with black. E is one of the eigenvectors of the similarity matrix that we used in the experiments which correspond to a negative eigenvalue. Z is very smooth while E is very non-smooth. This is typical of eigenvectors with negative values with typical parameters.

a semi-metric and it also has the property of *identity of indiscernibles* (*i.e.* $\mathcal{D}(P,Q) = 0$ if and only if $P = Q$).

A QC_m^A distance without the square root, is non-negative if the bin-similarity matrix, A, is positive-semidefinite. If A is positive-definitive, then it also has the property of *identity of indiscernibles*. This follows directly from the fact that the argument inside the square root in a QC histogram distance is a quadratic-form between two vectors. A QC histogram distance is symmetric if the bin-similarity matrix, A, is symmetric.

We now discuss subadditivity (*i.e.* $\mathcal{D}(P,Q) \leq \mathcal{D}(P,K) + \mathcal{D}(K,Q)$) for several distances. The χ^2 histogram distance is not subadditive. For example let $i = 0$, $k = 1$, $j = 2$ we get $\chi^2(i,j) = 1 > \chi^2(i,k) + \chi^2(k,j) = \frac{2}{3}$. However, $\sqrt{\chi^2}$ is subadditive for one and two dimensional non-negative histograms (verified by analysis). Experimentally it appears that $\sqrt{\chi^2}$ is subadditive for an N-dimensional non-negative histograms. Experimentally, QC members with the identity matrix seems to be subadditive for non-negative histograms. However, QC members with some positive-definitive bin-similarity matrices are not subadditive. The question when the QC histogram distances are subadditive is currently unresolved. An additional discussion about triangle inequality can be found in Jacobs et al. [21].

4 Cross-Bin Histogram Distance Properties

4.1 The Similarity-Matrix-Quantization-Invariance Property

The *Similarity-Matrix-Quantization-Invariance* property ensures that if two bins in the histograms have been erroneously quantized, this will not affect the distance. Mathematically we define this as:

Definition 1. *Let $\mathcal{D}$ be a cross-bin histogram distance between two histograms P and Q and let A be the bin-similarity/distance matrix. We assume P, Q and A are non-negative and that A is symmetric. Let $A_{k,:}$ be the kth row of A. Let* $= [\ _1, \ldots, \ _N]$

be a non-negative vector and $0 \leq \alpha \leq 1$. *We define* $^{\alpha,k,b} = [\ldots, \alpha\ _k, \ldots,\ _b + (1-\alpha)\ _k, \ldots]$. *That is,* $^{\alpha,k,b}$ *is a transformation of* $$ *where* $(1-\alpha)\ _k$ *mass has moved from bin* k *to bin* b. *We define* $\mathcal{D}$ *to be* Similarity-Matrix-Quantization-Invariant *if:*

$$A\ _{,} = A\ _{,} \Rightarrow \forall\ 0 \leq \alpha \leq 1,\ 0 \leq \beta \leq 1\ \ \mathcal{D}\ \ (P,Q) = \mathcal{D}\ \ (P^{\alpha,\ ,}, Q\ ^{,\ ,}) \tag{6}$$

We prove that EMD, EMD and all the Quadratic-Chi histogram distances are *Similarity-Matrix-Quantization-Invariant* in the appendix [20].

4.2 The Sparseness-Invariance Property

The *Sparseness-Invariance* property ensures that distances between sparse histograms will be equal to distances between full histograms. Mathematically we define this as:

Definition 2. *Let* $\mathcal{D}$ *be a cross-bin histogram distance between two histograms* $P \in\ ^N$ *and* $Q \in\ ^N$ *and let* A *be the* $N \times N$ *bin similarity/distance matrix. Let* A' *be any* $(N+1) \times (N+1)$ *matrix whose upper-left sub-matrix equals* A. *We define* $\mathcal{D}$ *to be* Sparseness-Invariant *if:*

$$\mathcal{D}\ \ ([P_1, \ldots, P\ ], [Q_1, \ldots, Q\ ]) = \mathcal{D}\ ^{'}([P_1, \ldots, P\ , \mathbf{0}], [Q_1, \ldots, Q\ , \mathbf{0}]) \tag{7}$$

QC members, EMD and the EMD are *Sparseness-Invariant* directly from their definitions. A stronger property called *Extension-Invariance* was proposed by D'Agostino and Dardanoni for bin-to-bin distances [22]. This property requires that, if both histograms are extended by concatenating each of them with the same vector (not necessarily zeros), the distance is left unaltered. Cross-bin distances assumes dependence between histogram bins, thus this requirement is too strong for them.

4.3 Cross-Bin Histogram Distance Properties Discussion

A *Sparseness-Invariant* cross-bin histogram distance does not depend on the specific representation of the histograms (full or sparse). A *Similarity-Matrix-Quantization-Invariant* cross-bin histogram distance encompass its cross-bin relationships only in the bin-similarity matrix. Intuitively such properties are desirable. In the appendix [20], we compare experimentally distances which resembles QC distances, but are either not *Similarity-Matrix-Quantization-Invariant* or not *Sparseness-Invariant*. The comparison shows that these properties considerably boost performance (especially for sparse color histograms).

Rubner et al. [2,23] claim that one of the key advantages of the Earth Mover's Distance is that each compared object may be represented by an individual (possibly with a different number of bins) binning that is adapted to its specific distribution. The Quadratic-Form is regarded as not having this property (see for example, Table 1 in [23]). Since all the Quadratic-Chi histogram distances (including the Quadratic-Form) are both *Similarity-Matrix-Quantization-Invariant* and *Sparseness-Invariant* there is no obstacle to using them with individual binning; *i.e.* to use them to compare histograms that were adapted to each object individually.

Similarity-Matrix-Quantization-Invariant and *Sparness-Invariant* can contradict. For example, any distance applied to the transformed vectors $P'_i = \sum_c (P_c) A_{ci}$ and $Q'_i = \sum_c (Q_c) A_{ci}$ is *Similarity-Matrix-Quantization-Invariant*. However the χ^2 distance between P' and Q' is not *Sparseness-Invariant* (with respect to P and Q).

5 Implementation Notes

5.1 The Similarity Matrix and the Normalization Factor

It is desirable to have a transformation from a distance matrix into a similarity matrix, as many spaces are equipped with a useful distance (*e.g.* color space [24]). Hafner et al. [1] proposed this transformation:

$$A_{ij} = 1 - \frac{D_{ij}}{\max_{ij}(D_{ij})} \tag{8}$$

Another possibility for choosing a similarity matrix is by using cross validation. However, we think that like for the Quadratic-Form, learning the similarity matrix (and for QC also the normalization factor) will be the best way to adjust them. This is left for future work. Currently we suggest to use thresholded ground distances as was used in [2,25,3,14] and choosing the normalization factor by cross validation.

5.2 Efficient Online Bin-Similarity Matrix Computation

For a fixed histogram configuration (*e.g.* SIFT, SC and IDSC) the bin-similarity matrix can be pre-computed once. Then, each distance computation is linear in the number of non-zero entries in the bin-similarity matrix.

There are cases where the bin-similarity matrix can not be pre-computed. For example, in our color experiments (Section 6.1), we used $N \times M$ color images as sparse histograms. That is, the query histogram was: $[1, \ldots, 1, 0, \ldots, 0]$ and each image being compared to the query was represented by the histogram: $[0, \ldots, 0, 1, \ldots, 1]$. Note that the full histogram dimension is $M \times N \times 256^3$, computing an $(M \times N \times 256^3)^2$ similarity matrix offline is not feasible. We can compute the similarity online for each pair of sparse histograms in $O((NM)^2)$ time. We now discuss how to do it more efficiently.

If we are comparing two images (as in Section 6.1) we can use a similarity matrix that gives far-away pixels zero similarity (see Eq. 10). Then, we can simply compare each pixel in one image to its corresponding $T \times T$ spatial neighbors in the second image. This reduces running time to $O(NMT^2)$. Using this technique, it is important to use a sparse representation for the bin-similarity matrix.

6 Results

We present results using the newly defined distances and state of the art distances, for image retrieval using SIFT-like descriptors and color image descriptors. In addition, we present results for shape classification using Inner Distance Shape Context (IDSC). More results for shape classification using SC, can be found in the appendix [20].

6.1 Image Retrieval Results

In this section we present results for image retrieval using the same benchmark as Pele and Werman [14]. We employed a database that contained 773 landscape images from the COREL database that were also used in Wang et al. [26]. The dataset has 10 classes[1]:

[1] The original database contains some visually ambiguous classes such as Africa that also contains images of beaches in Africa. We used the filtered image dataset that was downloaded from: `http://www.cs.huji.ac.il/~ofirpele/FastEMD/`

People in Africa, Beaches, Outdoor Buildings, Buses, Dinosaurs, Elephants, Flowers, Horses, Mountains and Food. The number of images in each class ranges from 50 to 100. From each class we selected 5 images as query images (images $1, 10, \ldots, 40$). Then we searched for the 50 nearest neighbors for each query image. We computed the distance of each image to the query image and its reflection and took the minimum. We present results for two types of image representations: SIFT-like descriptors and small L*a*b* images.

SIFT-like Descriptors. The first representation - SIFT is a $6 \times 8 \times 8$ SIFT descriptor [13] computed globally on the whole image. The second representation - CSIFT is a SIFT-like descriptor on a color-edge image. See [14] for more details.

We experimented with two new types of QC distances. The first is $\text{QC}^A_{0.5}$, which is a cross-bin generalization of $\sqrt{2\chi^2}$, which we call Quadratic-Chi-Squared (QCS). The second is $\text{QC}^A_{0.9}$, which has a larger normalization factor, which we call Quadratic-Chi-Normalized (QCN). We do not use QC^A_m with $\quad \geq 1$ due to discontinuity problems, see appendix [20] (practically, QC^A_1 had slightly poorer results compared to $\text{QC}^A_{0.9}$). We also experimented with the Quadratic-Form (QF) distance which is QC^A_0. For all of these distances we used the bin-similarity matrix in Eq. 8. Let $M = 8$ be the number of orientation bins, as in Pele and Werman [14], the ground distance between bins (x_i, y_i, o_i) and (x_j, y_j, o_j) is:

$$d_T(i,j) = \min \quad ||(x_i, y_i) - (x_j, y_j)||_2 + \min(|o_i - o_j|, M - |o_i - o_j|) \quad , \quad T \qquad (9)$$

We also used the identity matrix as a similarity matrix for all the above distances. We also compared to L_2 and χ^2. $\text{QF}^I = L_2$, and nearest neighbors of χ^2 and QCS^I are the same.

We also compared to four EMD variants. The first was EMD^D_1 with $D = d_T$ (Eq. 9) as in Pele and Werman [3]. The second was the L_1 norm which is equal to $\text{EMD}^D_{0.5}$ with D equals to the Kronecker delta multiplied by two. The third is $\text{SIFT}_{\text{DIST}}$[3] which is the sum of EMD over all the spatial cells (each spatial cell contains one orientation histogram). The ground distance for the orientation histograms is: $\min(|o_i - o_j|, M - |o_i - o_j|, 2)$ (M is the number of orientation bins). The fourth was the EMD-L_1[15] which is EMD with L_1 as the ground distance. We also tried non-thresholded ground distances (which produce non-sparse similarity matrices). However, the results were poor. This is in line with Pele and Werman's findings that cross-bin distances should be used with thresholded ground distances [14]. Finally, we compared to the Diffusion distance proposed by Ling and Okada [16] and to three probabilistic based distances: Bhattacharyya [17], Kullback-Leibler (KL) [18] and Jensen-Shannon (JS) [19] (we added Matlab's epsilon to all histogram bins when computing KL and JS throughout the paper, as they are not well defined if there is a zero bin, without doing so accuracy was very low).

For each distance measure, we present the descriptor (SIFT/CSIFT) with which it performed best. The results for all the pairs of descriptors and distance measures can be found in the appendix[20]. The results are presented in Fig. 3(a) and show that $\text{QCN}^{1-\frac{d_{T=2}}{2}}$ (QCN with the similarity matrix: $A_{ij} = 1 - \frac{d_{T=2}(i,j)}{2}$) outperformed all other methods. $\text{EMD}^{d_{T=2}}_1$ ranked second. The computation of $\text{QCN}^{1-\frac{d_{T=2}}{2}}$ was 266

times faster than $\text{EMD}_1^{d_{T=2}}$, see Table 2 in page 760. QCN^I ranked third, which shows the importance of the normalization factor.

All cross-bin distances that use thresholded ground distances outperformed their bin-by-bin versions. The figure also shows that χ^2 and QF improve upon L_2. QCN and QCS which are mathematically sound combinations of χ^2 and QF outperformed both.

L*a*b* Images. Our second type of image representation is a small L*a*b* image. We resized each image to 32×48 and converted them to L*a*b* space. The state of the art color distance is $_{00}$ - CIEDE2000 on L*a*b* color space[24,27]. As it is meaningful only for small distances we threshold it (as in [2,25,14]).

Again, we experimented with QCS, QCN and QF distances using the bin-similarity matrix in Eq. 8. The ground distance between two pixels $(x_i, y_i, L_i, a_i, b_i)$, $(x_j, y_j, L_j, a_j, b_j)$:

$$\begin{aligned} \text{s}(i,j) &= ||(x_i, y_i) - (x_j, y_j)||_2 \\ \text{dc}_{T_1,T_2}(i,j) &= \begin{cases} \min\left((\text{s}(i,j) + \Delta_{00}((L_i, a_i, b_i), (L_j, a_j, b_j))), T_1\right) & \text{if } \text{s}(i,j) \le T_2 \\ T_1 & \text{otherwise} \end{cases} \end{aligned} \quad (10)$$

This distance is similar to the one used by [14], except that distances with spatial difference larger than the threshold T_2 are set to the maximum threshold T_1. This was done to accelerate the online computation of the bin-similarity matrix. The accuracy using this distance is the same as using the distance from Pele and Werman [14]. See appendix [20]. We also used EMD with dc_{T_1,T_2} (Eq. 10) as a ground distance. Let I_1, I_2 be the two L*a*b* images. We also used the following distances:

$$L_1\Delta_{00} = \sum_{x,y}(\Delta_{00}(I_1(x,y), I_2(x,y))) \qquad L_1\Delta_{00}^T = \sum_{x,y}(\min(\Delta_{00}(I_1(x,y), I_2(x,y)), T))$$

$$L_2\Delta_{00} = \sum_{x,y}(\Delta_{00}(I_1(x,y), I_2(x,y)))^2 \qquad L_2\Delta_{00}^T = \sum_{x,y}(\min(\Delta_{00}(I_1(x,y), I_2(x,y)), T))^2$$

QCN^I, χ^2, L_2, L_1, $\text{SIFT}_{\text{DIST}}$[3], EMD-$L_1$[15], the Diffusion[16], Bhattacharyya [17], KL [18] and JS [19] distances cannot be applied to L*a*b* images as they are either bin-to-bin distances or applicable only to Manhattan networks.

We present results in Fig. 3. As shown, $\text{QCS}^{1-\frac{\text{dc}_{T_1=20,T_2=5}}{20}}$ and $\text{EMD}_1^{\text{dc}_{T_1=20,T_2=5}}$ [14] distances ranked first. $\text{QCS}^{1-\frac{\text{dc}_{T_1=20,T_2=5}}{20}}$ ran 300 times faster (see Table 2). However, since the computation of the bin-similarity matrix cannot be offline here, the real gain is a factor of 17. The $\text{QF}^{1-\frac{\text{dc}_{T_1=20,T_2=5}}{20}}$ distance ranked last, which shows the importance of the normalization factor of the QC histogram members.

Although a QC distance alleviates quantization problems, EMD does it better, instead of matching everything to everything it finds the optimal matching. EMD however, does not reduce the effect of large bins. We conjecture that a variant of EMD which will reduce the effect of large bins will have an excellent performance.

6.2 Shape Classification Results

In this section we present results for shape classification using the same framework as Ling et al. [11,15,28]. We test for shape classification with the Inner Distance Shape

Table 1. Shape classification results. $\text{QCN}^{1-\frac{\text{dsc}_{T=2}}{2}}$ outperformed all other distances.

	Top 1	Top 2	Top 3	Top 4	AUC%
$\text{QCN}^{1-\frac{\text{dsc}_{T=2}}{2}}$	39	38	38	34	0.950
QCN^{I}	40	37	36	33	0.940
$\text{QCS}^{1-\frac{\text{dsc}_{T=2}}{2}}$	39	35	38	28	0.912
QCS^{I}	40	34	37	27	0.907
χ^2	40	36	36	21	0.902
$QF^{1-\frac{\text{dsc}_{T=2}}{2}}$	40	34	39	19	0.897
L_2	39	35	35	18	0.873

	Top 1	Top 2	Top 3	Top 4	AUC%
$\widehat{\text{EMD}}_1^{\text{dsc}_{T=2}}$	39	36	35	27	0.902
L_1	39	35	35	25	0.890
$\text{SIFT}_{\text{DIST}}$[3]	38	37	27	22	0.848
EMD-L_1[15]	39	35	38	30	0.917
Diffusion[16]	39	35	34	23	0.880
Bhattacharyya[17]	40	37	32	23	0.895
KL[18]	40	38	36	29	0.938
JS[19]	40	35	37	21	0.900

Context (IDSC) [11]. The original Shape Context (SC) descriptor was proposed by Belongie et al. [10]. Belongie et al. [10] and Ling and Jacobs [11] used the χ^2 distance for comparing shape context histograms. Ling and Okada [15] showed that replacing χ^2 with EMD-L_1 improves results. We show that QC members yields the best results.

We tested on the articulated shape data set [11,28], that contains 40 images from 8 different objects. Each object has 5 images articulated to different degrees. The dataset is very challenging because of the similarity between different objects. The original SC had a very poor performance on this dataset, see appendix [20].

Again, we experimented with QCS, QCN and QF distances with the bin-similarity matrix in Eq. 8. The ground distance between two bins $(\ _i, o_i)$, $(\ _i, o_i)$ was (M is the number of orientation bins):

$$\text{dsc}_T(i,j) = \min\left((|d_i - d_j| + \min(|o_i - o_j|, M - |o_i - o_j|), T\right) \tag{11}$$

We also used the identity matrix as a similarity matrix, and thus we also compare to L_2. χ^2 and QCS^I distances are not equivalent here as the distance is not used for nearest neighbors. We refer the reader to Belongie et al. paper to see its usage [10]. Practically, QCS^I slightly outperformed χ^2 in this task, see Table 1.

We also compared to four EMD variants: EMD_1^D with $D = \text{dsc}_T$ (Eq. 11), the L_1 norm, $\text{SIFT}_{\text{DIST}}$[3] and EMD-$L_1$[15]. Finally, we compared to the Diffusion distance proposed by Ling and Okada [16] and to three probabilistic based distances: Bhattacharyya [17], Kullback-Leibler (KL) [18] and Jensen-Shannon (JS) [19].

To evaluate results, for each image, the four most similar matches are chosen from other images in the dataset. The retrieval result is summarized as the number of 1st, 2nd, 3rd and 4th most similar matches that come from the correct object. Table 1 shows the retrieval results. The $\text{QCN}^{1-\frac{\text{dsc}_{T=2}}{2}}$ outperformed all the other methods. QCN^I performance is again excellent, which shows the importance of the normalization factor.

Again all cross-bin distances outperformed their bin-by-bin versions. Again, χ^2 and QF improved upon L_2. QCN and QCS which are mathematically sound combinations of χ^2 and QF outperformed both.

6.3 Running Time Results

All runs were conducted on a Pentium 2.8GHz. A comparison of the practical running time of all distances is given in Table 2. Clearly QCN and QCS distances are fast to

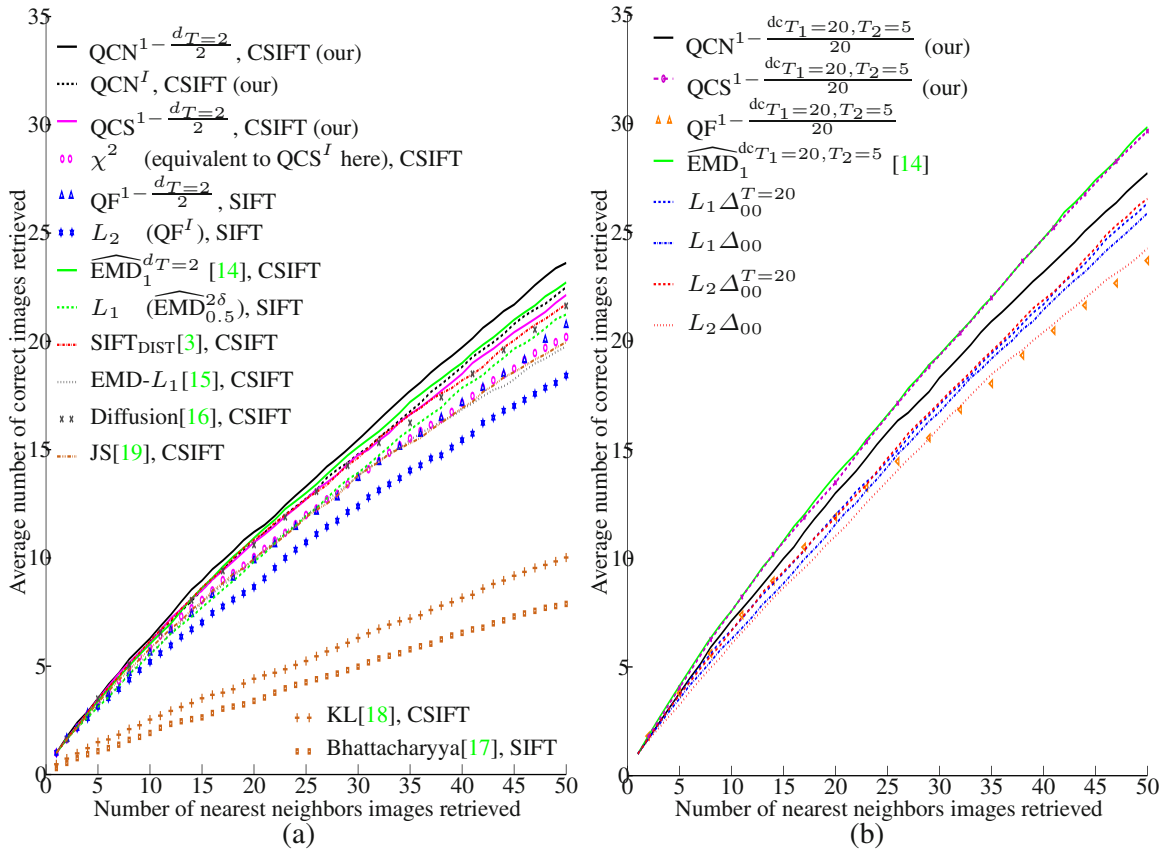


Fig. 3. Results for image retrieval.

(a) **SIFT-like descriptors.** For each distance measure, we present the descriptor (SIFT/CSIFT) with which it performed best. The results for all the pairs of descriptors and distance measures can be found in the appendix[20]. There are several key observations. First, the QC members performance is excellent. $\mathrm{QCN}^{1-\frac{d_{T=2}}{2}}$ (QCN with the similarity matrix: $A_{ij} = 1 - \frac{T=2\ i,j}{2}$) outperformed all other distances. ${\mathrm{EMD}_1}^{T=2}$ ranked second, but its computation was 266 times slower than $\mathrm{QCN}^{1-\frac{d_{T=2}}{2}}$ computation (see Table 2). Second, all cross-bin versions of the distances (with d_T or a transformation of it) performed better than their bin-by-bin versions (with the identity matrix or the Kronecker delta function). Third, QCN ranked third, although its a bin-to-bin distance. This shows the importance of the normalization factor. Finally, χ^2 and QF improve upon L_2. However, χ^2 does not take cross-bin relationships into account and QF does not reduce the effect of large bins. QCS and QCN histogram distances, which are mathematically sound combinations of χ^2 and QF have the two properties and outperformed both.

(b) **L*a*b* images results.** QCN , χ^2, L_2, L_1, $\mathrm{SIFT}_{\mathrm{DIST}}$[3], EMD-$L_1$[15], Diffusion[16], Bhattacharyya[17], KL[18] and JS[19] distances are not applicable here. $\mathrm{QCS}^{1-\frac{\mathrm{dc}_{T_1=20,T_2=5}}{20}}$ and $\mathrm{EMD}_1^{\mathrm{dc}_{T_1=20,T_2=5}}$ [14] ranked first. $\mathrm{QCS}^{1-\frac{\mathrm{dc}_{T_1=20,T_2=5}}{20}}$ computation is 300 times faster than $\mathrm{EMD}_1^{\mathrm{dc}_{T_1=20,T_2=5}}$ without taking the bin-similarity matrix computation into account and 17 times faster when it is taken into account (see Table 2). $\mathrm{QF}^{1-\frac{\mathrm{dc}_{T_1=20,T_2=5}}{20}}$ ranked last, which shows the importance of the normalization factor in QC members.

Table 2. (SIFT) 384-dimensional SIFT-like descriptors matching time (in *milliseconds*). The distances from left to right are the same as the distances in Fig. 3 (a) from up to down.
(IDSC) 60-dimensional IDSC histograms matching time (in *microseconds*). The distances from left to right are the same as the distances in Table 1 from up to down.
(L*a*b*) 32×48 L*a*b* images matching time (in *milliseconds*). The distances from left to right are the same as the distances in Fig. 3 (b) from up to down. In parentheses is the time it takes to compute the distance and the bin-similarity matrix as it cannot be computed offline.

Descriptor	QCN^{A_2}	QCN^{I}	QCS^{A_2}	QCS^{I}	χ^2	QF^{A_2}	L_2	$\widehat{\text{EMD}}^{D_2}$ [14]	L_1	$\text{SIFT}_{\text{DIST}}$[3]
(SIFT)	**0.15**	**0.1**	**0.07**	**0.014**	**0.013**	**0.05**	**0.011**	**40**	**0.011**	**0.07**
(IDSC)	**6.41**	**2.99**	**2.32**	**0.35**	**0.34**	**1.25**	**0.14**	**133.75**	**0.32**	**0.31**

Descriptor	EMD-L_1[15]	Diffusion[16]	JS[19]	KL[18]	Bhattacharyya[17]
(SIFT)	**40**	**0.27**	**0.088**	**0.048**	**0.015**
(IDSC)	**20.57**	**3.15**	**1.40**	**8.53**	**17.17**

Descriptor	$\text{QCN}^{A_{20}}$	$\text{QCS}^{A_{20}}$	$\text{QF}^{A_{20}}$	$\widehat{\text{EMD}}^{D_{20}}$ [14]	$L_1\Delta_{00}^{T=20}$	$L_1\Delta_{00}$	$L_2\Delta_{00}^{T=20}$	$L_2\Delta_{00}$
(L*a*b*)	**20** (370)	**19** (369)	**11** (361)	**6000** (6350)	**3.2**	**3.2**	**3.2**	**3.2**

compute. This is consistent with their linear time complexity. The only non-linear time distances are EMD [14] and EMD-L_1[15] which are also practically much slower than the other methods. Our method can be easily parallelized, taking advantage of multi-core computers or the GPU.

7 Conclusions

We presented a new cross-bin distance family - the Quadratic-Chi (QC). QC distances have many desirable properties. Like the Quadratic-Form histogram distance they take into account cross-bin relationships. Like χ^2 they reduce the effect of large bins. We formalized two new cross-bin properties, *Similarity-Matrix-Quantization-Invariance* and *Sparseness -Invariance*. QC members were shown to have both. Finally, QC distance computation time is linear in the number of non-zero entries in the bin-similarity matrix. Experimentally, QC outperformed state of the art distances, while having a very short run-time.

There are several open questions that we still need to explore. The first is for which QC distances does the the triangle inequality holds for. The second is whether we can change the Earth Mover's Distance so that it will also reduce the effect of large bins. Concave-cost network flow [29] seems to be the right direction for future work although it presents two major obstacles. First, the concave-cost network flow optimization is NP-hard [29]. However, there are available approximations [29,30]. Second, simply using concave-cost flow networks will result in a distance which is not *Similarity-Matrix-Quantization-Invariant*. We would also like to explore whether metric learning methods such as [31,32,33,34,35,36,37,38] can be generalized for the Quadratic-Chi histogram distance. Assent et al. [39] have suggested methods that accelerate database retrieval

that uses Quadratic-Form distances. Generalizing these methods for the Quadratic-Chi distances is of interest. Finally, other computer vision applications such as tracking can use the QC distances. The project homepage, including code (C++ and Matlab wrappers) is at: `http://www.cs.huji.ac.il/~ofirpele/QC/`.

References

1. Hafner, J., Sawhney, H., Equitz, W., Flickner, M., Niblack, W.: Efficient color histogram indexing for quadratic form distance functions. PAMI (1995)
2. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. IJCV (2000)
3. Pele, O., Werman, M.: A linear time histogram metric for improved sift matching. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 495–508. Springer, Heidelberg (2008)
4. Snedecor, G., Cochran, W.: Statistical Methods, Ames, Iowa, 6th edn. (1967)
5. Cula, O., Dana, K.: 3D texture recognition using bidirectional feature histograms. IJCV (2004)
6. Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: A comprehensive study. IJCV (2007)
7. Varma, M., Zisserman, A.: A statistical approach to material classification using image patch exemplars. PAMI (2009)
8. Xu, D., Cham, T., Yan, S., Duan, L., Chang, S.: Near Duplicate Identification with Spatially Aligned Pyramid Matching. In: CSVT (accepted)
9. Forssén, P., Lowe, D.: Shape Descriptors for Maximally Stable Extremal Regions. In: ICCV (2007)
10. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. PAMI (2002)
11. Ling, H., Jacobs, D.: Shape classification using the inner-distance. PAMI (2007)
12. Martin, D., Fowlkes, C., Malik, J.: Learning to detect natural image boundaries using local brightness, color, and texture cues. PAMI (2004)
13. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV (2004)
14. Pele, O., Werman, M.: Fast and robust earth mover's distances. In: ICCV (2009)
15. Ling, H., Okada, K.: An Efficient Earth Mover's Distance Algorithm for Robust Histogram Comparison. PAMI (2007)
16. Ling, H., Okada, K.: Diffusion distance for histogram comparison. In: CVPR (2006)
17. Bhattacharyya, A.: On a measure of divergence between two statistical populations defined by their probability distributions. BCMS (1943)
18. Kullback, S., Leibler, R.: On information and sufficiency. AMS (1951)
19. Lin, J.: Divergence measures based on the Shannon entropy. IT (1991)
20. Pele, O., Werman, M.: The quadratic-chi histogram distance family - appendices (2010), `http://www.cs.huji.ac.il/ ofirpele/publications/ECCV2010app.pdf`
21. Jacobs, D., Weinshall, D., Gdalyahu, Y.: Classification with nonmetric distances: Image retrieval and class representation. PAMI (2000)
22. D'Agostino, M., Dardanoni, V.: What's so special about Euclidean distance? SCW (2009)
23. Rubner, Y., Puzicha, J., Tomasi, C., Buhmann, J.: Empirical evaluation of dissimilarity measures for color and texture. CVIU (2001)
24. Luo, M., Cui, G., Rigg, B.: The Development of the CIE 2000 Colour-Difference Formula: CIEDE2000. CRA (2001)

25. Ruzon, M., Tomasi, C.: Edge, Junction, and Corner Detection Using Color Distributions. PAMI (2001)
26. Wang, J., Li, J., Wiederhold, G.: SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture LIbraries. PAMI (2001)
27. Sharma, G., Wu, W., Dalal, E.: The CIEDE2000 color-difference formula: implementation notes, supplementary test data, and mathematical observations. CRA (2005)
28. Ling, H.: Articulated shape benchmark and idsc code (2010), `http://www.ist.temple.edu/ hbling/code/ inner-dist-articu-distrbution.zip`
29. Guisewite, G., Pardalos, P.: Minimum concave-cost network flow problems: Applications, complexity, and algorithms. AOR (1990)
30. Amiri, A., Pirkul, H.: New formulation and relaxation to solve a concave-cost network flow problem. JORS (1997)
31. Xing, E.P., Ng, A.Y., Jordan, M.I., Russell, S.: Distance metric learning with application to clustering with side-information. In: NIPS (2003)
32. Bar-Hillel, A., Hertz, T., Shental, N., Weinshall, D.: Learning distance functions using equivalence relations. In: ICML (2003)
33. Goldberger, J., Roweis, S., Hinton, G., Salakhutdinov, R.: Neighbourhood components analysis. In: NIPS (2005)
34. Globerson, A., Roweis, S.: Metric learning by collapsing classes. In: NIPS (2006)
35. Yang, L., Jin, R.: Distance metric learning: A comprehensive survey. MSU (2006)
36. Davis, J., Kulis, B., Jain, P., Sra, S., Dhillon, I.: Information-theoretic metric learning. In: ICML (2007)
37. Yu, J., Amores, J., Sebe, N., Radeva, P., Tian, Q.: Distance learning for similarity estimation. PAMI (2008)
38. Weinberger, K., Saul, L.: Distance metric learning for large margin nearest neighbor classification. JMLR (2009)
39. Assent, I., Wichterich, M., Seidl, T.: Adaptable Distance Functions for Similarity-based Multimedia Retrieval. DSN (2006)

Chapter 5

Interpolated-Discretized Metric Learning using Linear Programming

This chapter presents results which have not been published yet

Interpolated-Discretized Metric Learning using Linear Programming

Abstract

Non-Mahalanobis distances such as χ^2 and robust estimators increase accuracy in many machine learning and computer vision applications. However, the distances are usually hand picked and cannot be learned efficiently from data. In this paper, we present a new non-Mahalanobis distance family, the *Interpolated-Discretized* (ID) distance family, and show how to efficiently learn its parameters. Once an ID distance is learned, computation of the distance is linear in the dimension and does not depend on the number of training examples (unlike kernel methods in which computing a distance scales linearly in the number of training examples). Many metric learning methods can be used with our ID distances, We demonstrate one possible way of learning ID distances using the Large Margin Nearest Neighbor (LMNN) framework [1]. With ID distances, the LMNN optimization problem is a linear program. Finally, we show that our method often outperforms other state-of-the-art metric learning approaches.

1 Introduction

Distance learning is of fundamental importance in machine learning. Learned distances can be used to improve algorithms which rely on distance computations such as nearest neighbor classification [1, 2], energy-based models [1, 3], supervised kernel machines (such as GPs or SVMs) and even unsupervised clustering algorithms [4].

Most of the work on metric learning [1, 2, 4, 5, 6, 7, 8, 9, 10, 11] has concentrated on learning a Mahalanobis distance, which is equivalent to learning a linear transformation of the data and then using the standard Euclidean distance. Let $\vec{x}^i$, $\vec{x}^j$ be two vectors, the squared Mahalanobis distance between them is parameterized by a positive-semidefinite (PSD) matrix, M:

$$\text{Mahalanobis}(\vec{x}^i, \vec{x}^j) = (\vec{x}^i - \vec{x}^j)^T M (\vec{x}^i - \vec{x}^j) \tag{1}$$

The PSD constraint on M is required for obtaining non-negative distances. Also, it implies that the M matrix can be expressed as $M = LL^T$ for some real matrix L. Thus, the distance can be computed as the squared Euclidean norm between linearly transformed vectors:

$$\text{Mahalanobis}(\vec{x}^i, \vec{x}^j) = ||L\vec{x}^i - L\vec{x}^j||_2^2 \tag{2}$$

Note that, as defined above, the square root of a Mahalanobis measure is a pseudo-metric, as the distance between two different vectors might be zero. Computing the distance between never-seen examples using Mahalanobis distances scales quadratically in the dimension. The *Interpolated Discretized* (ID) distance we introduce in the current work scales linearly in the dimension with just a small constant memory overhead.

Although much of the metric-learning work has focused on the Mahalanobis distance, this measure has some inherent limitations. The most important one is that it is a function only of the difference between input vectors $\vec{x}^i - \vec{x}^j$. Thus, it will give the same distance whenever two vectors have the same difference vector. For example, a Mahalanobis distance will not able to differentiate between

$(\vec{x}^i = 42, \vec{x}^j = 40)$ and $(\vec{x}^i = 2, \vec{x}^j = 0)$. To overcome this limitation several metric learning methods have been proposed [1, 3, 7, 9, 12]. We discuss these methods in section 4.

Another approach to overcoming the above limitation is to apply a function directly to $\vec{x}^i, \vec{x}^j$ such that the distance is not invariant to $\vec{x}^i - \vec{x}^j$. One particularly successful instance of this approach is the χ^2 distance, which has recently received considerable attention in the computer vision literature. It has been used in state-of-the-art: contour detection and segmentation algorithms [13], descriptors matching [14, 15], shape classification [15, 16] and image retrieval [15] to name a few. Another related approach is robust estimators which were successfully used optical flow computation [17]. Although successful, χ^2 and robust estimators are *hard-coded* and cannot be naturally learned or improved using data.

Another related shortcoming of the Mahalanobis distance is that it is invariant to where $\vec{x}^i$ and $\vec{x}^j$ are in space, as long as their difference vector is the same. This is undesirable in many cases, as we would often like to normalize the distances by the range of the input vectors, e.g., as in the χ^2 metric. In other words, the Mahalanobis distance is limited in terms of the bin to bin (or feature to feature) relationships it can capture.

In this work, we propose the *Interpolated-Discretized* (ID) distance. Our measure can overcome some of the key inherent limitations of the Mahalnobis distance, and is also more flexible than other approaches such as χ^2. Specifically, an ID distance can model different distances over different parts of the space.

Our approach contains two key elements. The first is a discretization of each feature into C values such that the interaction between any two features is represented by an arbitrary $C \times C$ matrix. This allows us to model arbitrary relations between the two features. However, such a coarsely quantized representation neglects relevant information in the original value, and is thus undesirable. To overcome this, we employ a careful interpolation approach, which does take into account the original feature values.

The resulting ID measure is continuous in feature space but highly expressive due to its different treatment of different feature regimes. Before going into details, we provide an illustrative example in Fig. 1. The figure shows how the χ^2 measure can be well approximated via ID with only $C = 10$ clusters per feature. Note that in general, the goal is not to approximate a known distance such as χ^2, but to learn to approximate the best *unknown* distance. A nice property of ID, that can also be seen in Fig. 1 is that for $C = 2$ it exactly equals the l_1 norm. This again highlights the fact that ID is continuous, but becomes more and more detailed as C is increased.

In the current work, we do not represent cross-bin dependencies. It is possible to extend our method such that ID distances will also be able to model relationships between features, so it will be able to learn both Mahalanobis type of distances and bin-to-bin distances and variants in between. This can be done by discretizing pairs of values. However, this will increase the computational complexity considerably. We decided to concentrate on bin-to-bin ID distance since it is interesting to check

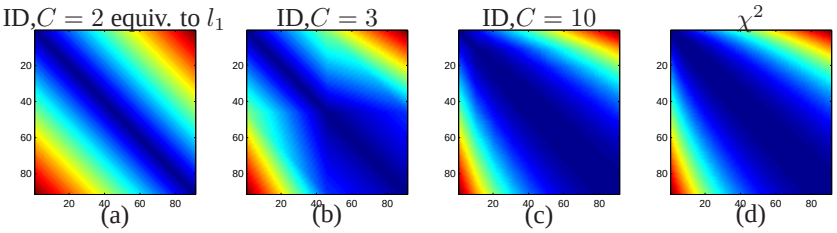


Figure 1: An illustration of a 1-dimensional χ^2 distance approximation using ID distances with different numbers of discretization points. Panel (d) shows the original χ^2 distance. In (a) we see that using just two points of discretization (i.e., $C = 2$ corresponding to the range $[0, 90]$), the ID distance in this case is equal to the l_1 norm. In (b), using three points of discretization $[0, 45, 90]$ (i.e., $C = 3$), the ID distance starts to resemble χ^2, although the difference is noticeable. In (c), using $C = 10$ corresponding to the range $[0, 10, \ldots, 90]$, the ID distance is almost identical to χ^2. We emphasize that the goal is not to approximate a known distance such as χ^2, but to learn to approximate the best *unknown* distance.

what can be achieved with those, especially since other works already showed that the diagonal Mahalanobis Metric Learning achieves competitive results [2, 18] and they are very fast to compute.

A useful property of an ID distance is that it is a linear combination of its parameters. Thus, many of the proposed metric learning optimization problems are especially suited for use with an ID distance. We demonstrate learning ID distance parameters using the Large Margin Nearest Neighbor (LMNN) framework [1]. LMNN with an ID distance is a linear program.

We evaluate our method on a variety of standard datasets and show that it often outperforms state-of-the-art metric learning approaches. We thus believe that ID distances are very useful distances which can be learned effectively from data.

2 The Interpolated-Discretized Distance Family

We begin by describing the ID distance between two scalars x^i, x^j. The vector case will easily follow. An ID distance between two scalar features is parameterized via a vector and a matrix. The first parameter is a sorted vector of C cluster centers of possible values of the feature: $[v_1, \ldots, v_C]$. The second parameter is a matrix $M \in \mathbb{R}^{C \times C}$, where $M_{a,b}$ corresponds to the distance between v_a and v_b. See Fig. 2.

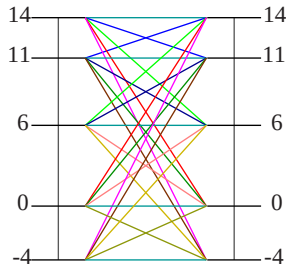


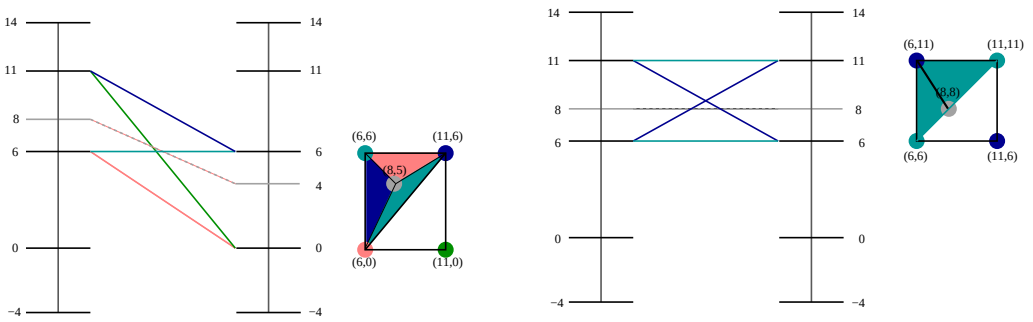
Figure 2: An Illustration of 1-dimensional ID distance parameters. Feature values are clustered into $C = 5$ cluster centers: $[-4, 0, 6, 11, 14]$. The full 5×5 matrix of distances between pairs of cluster centers values is illustrated as edges.

However, we would not like the distance to have the same value when x^i, x^j belong to a given pair of clusters. Rather, we would like the measure to smoothly vary with the feature values. One natural way to achieve this is via interpolation. Given two scalars, the two closest cluster centers are found for each. Next, four coefficients are computed. Each coefficient corresponds to one pair of cluster centers and is a normalized triangle area. See Fig. 3 for description and illustration of how the coefficients are computed. Finally, the ID distance is the sum of the coefficients multiplied by their corresponding entries from the matrix of the distances between pairs of cluster centers values:

$$\begin{aligned} \mathrm{ID}(x^i, x^j) &= \sum_{t=1}^{4} \alpha_{a(t),b(t)} M_{a(t),b(t)} \\ a(1) = a(2) &= \operatorname*{argmax}_c \{v_c \leq x^i\} \\ a(3) = a(4) &= \operatorname*{argmin}_c \{v_c \geq x^i\} \\ b(1) = b(3) &= \operatorname*{argmax}_c \{v_c \leq x^j\} \\ b(2) = b(4) &= \operatorname*{argmin}_c \{v_c \geq x^j\} \end{aligned} \tag{3}$$

In the above the coefficients $\alpha_{a(t),b(t)}$ correspond to the interpolation coefficient. See Fig. 3.

We now discuss properties of the ID distance. There are three conditions for a distance function, D, to be a semimetric. The first is *non-negativity* (*i.e.* $D(\vec{x}^i, \vec{x}^j) \geq 0$). The second is *identity* (*i.e.* $D(\vec{x}^i, \vec{x}^j) = 0$ iff $\vec{x}^i = \vec{x}^j$). The third is *symmetry* (*i.e.* $D(\vec{x}^i, \vec{x}^j) = D(\vec{x}^j, \vec{x}^i)$). A natural question is what properties does the M matrix need to satisfy in order for the resulting ID measure to be a



(a1) (a2) (b1) (b2)

Figure 3: An illustration of how the coefficients are computed for $x^i = 8, x^j = 4$ where the cluster centers are: $[-4, 0, 6, 11, 14]$ (as in Fig. 2). In (a1), the closest cluster centers are $(6, 11)$ for $x^i = 8$ and $(0, 6)$ for $x^j = 4$. Thus, the four discrete distances are: $M_{3,2}$, $M_{3,3}$, $M_{4,2}$ and $M_{4,3}$ which correspond to the four cluster center pairs: $(6, 0)$, $(6, 6)$, $(11, 0)$ and $(11, 6)$. To compute the α coefficients (the interpolation step), we check in (a2) whether the point $(x^i = 8, x^j = 4)$ is above or below the line connecting $(6, 0)$ with $(11, 6)$. Then, the normalized triangle areas are the coefficients of the corresponding discrete distances. Note that since the point was above the line, the coefficient of $M_{4,2}$ is zero.
In (b1), (b2) we see a degenerate case where the interpolation is between $M_{3,3}$ and $M_{4,4}$. If we enforced the identity property on M, these will be zero and the resulting distance is also zero.
This interpolation has several advantages. First, the resulting distance will have the identity property if we enforce the distance between two identical cluster centers to be zero. Second, using this interpolation results in a distance which is always continuous and differentiable except at the discretization points. This means, for example, that an ID distance is not very sensitive to the number of clusters chosen. Finally, an ID distance is equivalent to a weighted l_1 norm if the number of cluster centers is two and the distance between identical cluster centers is set to zero.

semimetric. It's straightforward to see that the conditions on M are simply that it satisfies $M_{i,j} > 0$ for all $i \neq j$, as well as $M_{i,i} = 0$ for all i and $M_{i,j} = M_{j,i}$ for all i, j. All these can be easily enforced via linear constraints on the elements of M.

A semimetric is also a metric if it has the property of *subadditivity* (*i.e.* $D(\vec{x}^i, \vec{x}^j) \leq D(\vec{x}^i, \vec{x}^k) + D(\vec{x}^k, \vec{x}^j)$). We can transform a semimetric into a non-continuous metric by linearly mapping all strictly bigger than zero distances to the range $[1, 2]$. Thus it is very easy to transform an ID distance into a metric. An additional discussion about triangle inequality can be found in Jacobs et al. [19].

To generalize to N features we define a separate matrix, M^n for each feature.

3 Interpolated-Discretized Large Margin Nearest Neighbor Metric Learning

In this section we show how to learn an ID distance using the Large Margin Nearest Neighbor Metric Learning framework [1]. We start by describing the LMNN formulation.

Let $\{(\vec{x}^i, y^i)\}_{i=1}^T$ denote a set of T labeled examples with inputs $\vec{x}^i \in \mathbb{R}^N$ and discrete (but not necessarily binary) class label y^i. We use the binary matrix $y_{ij} \in \{0, 1\}$ to indicate whether or not the labels y^i and y^j match.

In addition to the class label y_i, for each input $\vec{x}^i$ we also specify k "target" neighbors, that is, k other inputs with the same label y^i that we wish to have minimal distance to $\vec{x}^i$. In the absence of any knowledge, the target neighbors can be the k closest nearest neighbors using the Euclidean distance. This is also what we use in this paper. We denote that $\vec{x}^j$ is a target neighbor of $\vec{x}^i$ as $j \rightsquigarrow i$.

The LMNN framework has a loss function with two terms. The first, tries to minimize the distance between all $\vec{x}^i$ and their corresponding target neighbors. The second, tries to push vectors with different labels away, such that the distance to them minus the distance to the target neighbor (the margin) will be bigger or equal to 1. Since this might be unfeasible, slack variables are added and

the second term is the sum of slack variables. The two terms are weighted with a parameter μ. Using ID distances the optimization of LMNN is:

$$\begin{aligned}
\min_{M,\xi} \quad & (1-\mu)\sum_{i,j\rightsquigarrow i}\mathrm{ID}(\vec{x}^i,\vec{x}^j)+\mu\sum_{i,j\rightsquigarrow i,l}(1-y_{il})\xi_{ijl} \quad \text{subject to:} \\
& \forall\ i,j\rightsquigarrow i,l \quad \mathrm{ID}(\vec{x}^i,\vec{x}^l)-\mathrm{ID}(\vec{x}^i,\vec{x}^j)\geq 1-\xi_{ijl} \\
& \forall\ i,j\rightsquigarrow i,l \quad \xi_{ijl}\geq 0 \\
& \mathrm{ID}(\vec{x}^i,\vec{x}^j)=\sum_{n=1}^{N}\sum_{t=1}^{4}\alpha^n_{a(t),b(t)}M^n_{a(t),b(t)} \\
& \forall\ n,a,b \quad M^n_{a,b}\geq 0
\end{aligned} \tag{4}$$

This linear program allows to learn any bin-to-bin distance function. In order to learn a semi-metric, the following constraints should be added:

$$\begin{aligned}
\forall\ n,a \quad & M^n_{a,a}=0 \\
\forall\ n,a\neq b \quad & M^n_{a,b}\geq\varepsilon \\
\forall\ n,a,b \quad & M^n_{a,b}=M^n_{b,a}
\end{aligned} \tag{5}$$

where ε is some strictly positive small scalar. We use $\varepsilon=10^{-5}$. We can also use an upper bound on M^n entries as a regularizer. We used one as an upper bound.

As we discussed before, we can transform a semimetric into a non-continuous metric by linearly mapping all strictly bigger than zero distances to the range $[1,2]$. Thus it is very easy to transform an ID distance into a metric.

Finally, we can add monotone constraints as a regularizer:

$$\forall\ n,\ a<b<c, \quad M^n_{a,b}\leq M^n_{a,c} \quad M^n_{b,c}\leq M^n_{a,c} \tag{6}$$

4 Related Work

Our method builds in a novel direction on the success of previous metric learning approaches. As Weinberger and Saul [1] conjectured, more adaptive transformations of the input space can lead to improved performance. Our method allows to enlarge the number of the learned parameters, while the computation of the distance between two never-seen examples is only linear in the dimension.

Chopra et al. [3] proposed to learn a convolutional neural net as a non-linear transformation before applying the ℓ_2 norm. They showed excellent results on image data. Babenko et al. [12] suggested a boosting framework for learning non-Mahalanobis metrics. They also presented excellent results on image data. These methods are non-convex and thus they might suffer from local minimas and training is sensitive to parameters.

Kernel methods were also proposed in order to learn a Mahalanobis distance over non-linear transformations of the data [1, 7, 9]. Computing such a distance between two vectors scales linear in the number of training examples, which makes it impractical for large datasets. Computing our ID distances does not depend on the number of training examples.

A family of non-Mahalanobis distances recently proposed is the *Qudaratic-Chi* (QC) [15]. The QC family generalizes both the Mahalanobis distance and the χ^2 distance. A QC distance have parameters that can be learned. However, a serious limitation is that it can only model χ^2-like distances. In addition, it is applicable only to non-negative vectors. Finally, it is non-convex with respect to its parameters, so learning them is hard.

Rosales and Fung [5] also propose learning metrics via linear programming. However, while we learn a non-Mahalanobis distance, their method learns a subfamily of Mahalanobis distance. That is, their method is restricted to learning a Mahalanobis distance which is parameterized with a diagonal dominant matrix.

5 Results

We evaluated our approach on ten UCI data sets of varying size and difficulty. Table 1 summarizes the data sets properties. Except for Spectf all experimental results are averaged over several runs of randomly generated stratified 70/30 splits of the data. The Spectf data set has a predefined training/test split. All experiments are averaged over 100 runs, except for the Steel database where it is averaged over 20 runs and for the Cardio dataset where it is averaged over 5 runs.

	Iris	Wine	Ion	Sonar	Transf	Spectf	Breast	Glass	Steel	Cardio
examples	150	178	351	208	748	267	106	214	1941	2126
classes	3	3	2	2	2	2	6	6	7	10
dimension	4	13	34	60	9	44	9	9	27	21

Table 1: Properties of data sets

We applied kNN classifiers with different distances. We compared using our learned ID distances to using Mahalanobis distance learned with ITML [9] and with the original LMNN [1].

All methods were used with target neighbors $k = 3$. ITML was used with the default parameters in the distribution of the code. That is, the slack variable parameter γ which inversely correspond to the regularization magnitude was tuned using cross-validation over the values $\{0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000\}$. The LMNN and ID-LMNN μ parameter was set as 0.5 (the default in LMNN code). Our ID-LMNN approach was tested with two different numbers of clusters C. First, $C = 2$ which is equivalent to learning weighted l_1 norm using the LMNN framework. Second, $C = 10$ which allows to capture the non-linearities in the data.

We used Mosek [20] as the LP solver, except for Steel and Cardio where we used a subgradient solver as the problems do not fit in memory. We are working on accelerating the subgradient method using active set method as in LMNN and using stochastic subgradient methods. We believe that it will allow us to work on much larger datasets.

The results are presented in Fig. 4. There are several observations. First, for almost all databases, metric learning is better than naive Euclidean distance. Second, learned ID distance with $C = 2$, which is equivalent to a learned weighted l_1 distance, has an excellent performance. Learned diagonal Mahalanobis distances (equivalent to weighted l_2) is known to have competitive results [2, 18]. In our experiments, learned weighted l_1 outperforms the full Mahalanobis distances in many cases.

Finally, learning our proposed ID distances with $C = 10$ is a strong competitor on all datasets and outperformed other methods on most of the datasets, especially the larger datasets and those with a large number of classes.

6 Conclusions

We presented a new distance family - the *Interpolated-Discretized* (ID) distances. ID distances have many desirable properties. First, they can approximate any bin-to-bin distance. Second, as they are linear combination of parameters, they can be learned efficiently. Third, using linear constraints we can enforce them to be a metric without needing more complex semi-definite requirements on the parameters as in Mahalanobis distances. Finally, we have shown that experimentally learned ID distances have excellent performance and they often outperform state-of-the-art learned distances.

We note that on data with a large quantization problems, such as image pixels values, our method is likely not to perform well. However, modern computer vision methods do not use image pixels values as a feature, but use state-of-the-art descriptors such as SIFT [21] which do not suffer too much from quantization problems.

Future work will include the generalization of ID distance to model "cross-bin" relationships. This can be done by discretizing pairs of values. However, this will be much more time demanding and this issue need to be resolved.

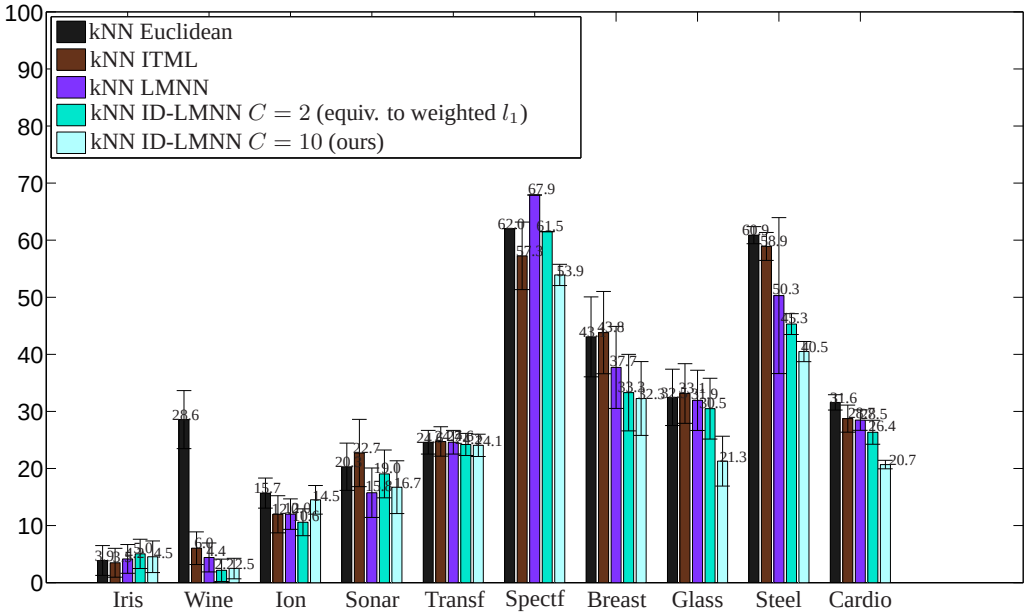


Figure 4: Test errors with standard deviations for kNN classification using several distances. In our experiments, learned weighted l_1 outperforms the full Mahalanobis distances in many cases. Our proposed ID distances with $C = 10$ is a strong competitor on all datasets and outperformed other methods on most of the datasets, especially the larger datasets and those with a large number of classes.

References

[1] K.Q. Weinberger and L.K. Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 2009.

[2] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. Neighbourhood components analysis. *NIPS*, 2005.

[3] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, 2005.

[4] E.P. Xing, A.Y. Ng, M.I. Jordan, and S. Russell. Distance metric learning with application to clustering with side-information. *NIPS*, 2003.

[5] R. Rosales and G. Fung. Learning sparse metrics via linear programming. In *ICKDM*, 2006.

[6] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning distance functions using equivalence relations. In *ICML*, 2003.

[7] A. Globerson and S. Roweis. Metric learning by collapsing classes. *NIPS*, 2006.

[8] L. Yang and R. Jin. Distance metric learning: A comprehensive survey. *MSU*, 2006.

[9] J.V. Davis, B. Kulis, P. Jain, S. Sra, and I.S. Dhillon. Information-theoretic metric learning. In *ICML*, 2007.

[10] D. Ramanan and S. Baker. Local distance functions: A taxonomy, new algorithms, and an evaluation. *PAMI*, 2010.

[11] B. McFee and G.R.G. Lanckriet. Learning multi-modal similarity. *JMLR*, 2011.

[12] B. Babenko, S. Branson, and S. Belongie. Similarity metrics for categorization: from monolithic to category specific. In *ICCV*, 2009.

[13] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *PAMI*, 2010.

[14] P.E. Forssén and D.G. Lowe. Shape Descriptors for Maximally Stable Extremal Regions. In *ICCV*, 2007.

[15] O. Pele and M. Werman. The quadratic-chi histogram distance family. In *ECCV*, 2010.

[16] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *PAMI*, 2002.

[17] A. Bab-Hadiashar and D. Suter. Optic flow calculation using robust statistics. In *CVPR*, 1997.

[18] M.P. Kumar, PHS Torr, and A. Zisserman. An invariant large margin nearest neighbour classifier. In *ICCV*, 2007.

[19] D.W. Jacobs, D. Weinshall, and Y. Gdalyahu. Classification with nonmetric distances: Image retrieval and class representation. *PAMI*, 2000.

[20] The mosek optimization software. `http://www.mosek.com/`.

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

Chapter 6

Conclusions

We presented several new distance functions and efficient algorithms for their computation. The distance functions were designed to be robust to common image noise and deformation. On the other hand, the distance functions were designed to be as distinctive as possible. The distances which balance between robustness and distinctiveness improved on the state-of-the-art of several computer vision applications such as image retrieval and shape classification. Improved results using our distance functions were also reported by other researchers, both from the computer vision community and outside of it (*e.g.* [54–57]).

Our first conjecture was that real-world noise has two main components. First, limited deformation noise. Second, not limited outlier noise. For example, deformation noise can change blue to some other shade of blue; On the other hand, outlier noise (*e.g.* occlusion) can change blue to any other color. We proposed using saturated cross-bin distances, that is, distances where different outliers receive the same value.

We also presented $\widehat{EMD}$, a new generalization of the transportation distance for non-normalized histograms. Unlike the classic generalization (EMD), $\widehat{EMD}$ is a metric even when the histograms are non-normalized. Additionally, $\widehat{EMD}$ is often better than EMD in practice. We have shown that these distances outperformed state-of-the-art distance functions in many computer vision applications.

A second observation was that image histograms have a unique statistical behavior since large values usually indicates large variance. χ^2 was proposed as a distance function that overcomes this problem [8, 9]. χ^2 is the second order Taylor expansion of the Jensen Shannon (JS) divergence [7] and thus χ^2 and JS results are almost identical [8–10]. We suggested a new family of histogram distances, called *Quadratic-Chi* (QC). Members of this family are robust to quantization effects and have a χ^2-like normaliza-

tion. We also proposed two new cross-bin distances properties *Similarity-Matrix-Quantization-Invariance* and *Sparseness-Invariance* and a proof that $\widehat{EMD}$ and QC distances have these properties. We demonstrated excellent experimental results, also showing the practical importance of the two new properties.

Finally, we described a novel method for non-Mahalanobis metric learning, *Interpolated-Discretized Metric Learning*. This was motivated by the success of our non-Mahalanobis distance, the QC. A QC distance has parameters that can be learned. However, a serious limitation is that it can only model χ^2-like distances. In addition, it is applicable only to non-negative vectors. Finally, it is non-convex with respect to its parameters, so learning them is hard. To overcome these problems we suggested a new non-Mahalanobis distance family, the *Interpolated-Discretized* (ID) distance family, and showed how to efficiently learn its parameters. An ID distance can approximate any *unknown* distance. Many metric learning methods can be used with our ID distances. We demonstrated one possible way of learning ID distances using the Large Margin Nearest Neighbor (LMNN) framework [2]. With ID distances, the LMNN optimization problem is a linear program. Finally, we showed that our method often outperforms other state-of-the-art metric learning approaches.

Future work

The work described in this thesis can be extended in several directions:

- Several algorithmic improvements are planned for the algorithm presented in chapter 3.
- The algorithms speed can be further improved using techniques such as Bayesian sequential hypothesis testing [58], metric trees [59–61] and approximate nearest neighbor [62, 63].
- An interesting question is whether we can change the Earth Mover's Distance so that it will also reduce the effect of large values. Concave-cost network flow [64] seems to be the right direction for future work although it presents two major obstacles. First, the concave-cost network flow optimization is NP hard [64]. However, there are available approximations [64, 65]. Second, simply using concave-cost flow networks will result in a distance which is not *Similarity-Matrix-Quantization-Invariant*.

- Generalization of the *Interpolated-Discretized* distance to model cross-bin relationships. This can be done by discretizing pairs of values. However, this will be much more time demanding and this issue need to be resolved.

Bibliography

[1] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval.," *IJCV*, 2000.

[2] K. Weinberger and L. Saul, "Distance metric learning for large margin nearest neighbor classification," *JMLR*, 2009.

[3] J. Lin, "Divergence measures based on the shannon entropy," *TIT*, 1991.

[4] F. Ôsterreicher and I. Vajda, "A new class of metric divergences on probability spaces and its statistical applications," *AISM*, 2003.

[5] D. M. Endres and S. J.E., "A new metric for probability distributions," *IT*, 2003.

[6] F. Topsøe, "Jenson-shannon divergence and norm-based measures of discrimination and variation," *Preprint*, 2003.

[7] F. Topsoe, "Some inequalities for information divergence and related measures of discrimination," *IT*, 2000.

[8] J. Puzicha, T. Hofmann, and J. Buhmann, "Non-parametric similarity measures for unsupervised texture segmentation and image retrieval," in *CVPR*, 1997.

[9] Y. Rubner, J. Puzicha, C. Tomasi, and J. Buhmann, "Empirical evaluation of dissimilarity measures for color and texture," *CVIU*, 2001.

[10] O. Pele and M. Werman, "The quadratic-chi histogram distance family," in *ECCV*, 2010.

[11] G. Snedecor and W. Cochran, "Statistical Methods, ed 6. Ames, Iowa," 1967.

[12] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *PAMI*, 2010.

[13] P. Forssén and D. Lowe, "Shape Descriptors for Maximally Stable Extremal Regions," in *ICCV*, 2007.

[14] O. Cula and K. Dana, "3D texture recognition using bidirectional feature histograms," *IJCV*, 2004.

[15] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *IJCV*, 2007.

[16] M. Varma and A. Zisserman, "A statistical approach to material classification using image patch exemplars," *PAMI*, 2009.

[17] A. Vedaldi and A. Zisserman, "Efficient additive kernels via explicit feature maps," in *CVPR*, 2010.

[18] D. Xu, T. Cham, S. Yan, L. Duan, and S. Chang, "Near Duplicate Identification with Spatially Aligned Pyramid Matching," *CSVT*, 2010.

[19] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *PAMI*, 2002.

[20] H. Ling and D. Jacobs, "Shape classification using the inner-distance," *PAMI*, 2007.

[21] J. Hafner, H. Sawhney, W. Equitz, M. Flickner, and W. Niblack, "Efficient color histogram indexing for quadratic form distance functions," *PAMI*, 1995.

[22] P. Mahalanobis, "On the generalized distance in statistics," in *PNIS*, 1936.

[23] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," *NIPS*, 2005.

[24] E. Xing, A. Ng, M. Jordan, and S. Russell, "Distance metric learning with application to clustering with side-information," *NIPS*, 2003.

[25] R. Rosales and G. Fung, "Learning sparse metrics via linear programming," in *ICKDM*, 2006.

[26] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, "Learning distance functions using equivalence relations," in *ICML*, 2003.

[27] A. Globerson and S. Roweis, "Metric learning by collapsing classes," *NIPS*, 2006.

[28] L. Yang and R. Jin, "Distance metric learning: A comprehensive survey," *MSU*, 2006.

[29] J. Davis, B. Kulis, P. Jain, S. Sra, and I. Dhillon, "Information-theoretic metric learning," in *ICML*, 2007.

[30] D. Ramanan and S. Baker, "Local distance functions: A taxonomy, new algorithms, and an evaluation," *PAMI*, 2010.

[31] B. McFee and G. Lanckriet, "Learning multi-modal similarity," *JMLR*, 2011.

[32] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *CVPR*, 2005.

[33] B. Babenko, S. Branson, and S. Belongie, "Similarity metrics for categorization: from monolithic to category specific," in *ICCV*, 2009.

[34] G. Monge, "Déblai et remblai," *Mémoires de l'Académie des Sciences*, 1781.

[35] L. Kantorovitch, "On the translocation of masses," *MS*, 1958.

[36] L. Wasserstein, "Markov processes over denumerable products of spaces, describing large systems of automata," *PIT*, 1969.

[37] C. Mallows, "A note on asymptotic joint normality," *AMS*, 1972.

[38] H. Shen and A. Wong, "Generalized texture representation and metric," *CVGIP*, 1983.

[39] M. Werman, S. Peleg, and A. Rosenfeld, "A distance metric for multidimensional histograms.," *CVGIP*, 1985.

[40] M. Werman, S. Peleg, R. Melter, and T. Kong, "Bipartite graph matching for points on a line or a circle," *JOA*, 1986.

[41] S. Peleg, M. Werman, and H. Rom, "A unified approach to the change of resolution: Space and gray-level.," *PAMI*, 1989.

[42] J. Orlin, "A faster strongly polynomial minimum cost flow algorithm," in *STOC*, 1988.

[43] H. Ling and K. Okada, "An Efficient Earth Mover's Distance Algorithm for Robust Histogram Comparison," *PAMI*, 2007.

[44] J. Gudmundsson, O. Klein, C. Knauer, and M. Smid, "Small Manhattan Networks and Algorithmic Applications for the Earth Mover's Distance," in *EWCG*, 2007.

[45] P. Indyk and N. Thaper, "Fast image retrieval via embeddings," in *IWSCTV*, 2003.

[46] K. Grauman and T. Darrell, "The pyramid match kernel: Efficient learning with sets of features," *JMLR*, 2007.

[47] S. Shirdhonkar and D. Jacobs, "Approximate earth mover's distance in linear time," in *CVPR*, 2008.

[48] S. Khot and A. Naor, "Nonembeddability theorems via Fourier analysis," *Mathematische Annalen*, 2006.

[49] A. Andoni, P. Indyk, and R. Krauthgamer, "Earth mover distance over high-dimensional spaces," in *SODA*, 2008.

[50] K. Grauman and T. Darrell, "Fast contour matching using approximate earth mover's distance," in *CVPR*, 2004.

[51] T. Goh, R. West, and K. Okada, "Robust detection of semantically equivalent visually dissimilar objects," in *SLAM*, 2008.

[52] O. Pele and M. Werman, "A linear time histogram metric for improved sift matching," in *ECCV*, 2008.

[53] O. Pele and M. Werman, "Fast and robust earth mover's distances," in *ICCV*, 2009.

[54] R. Gupta, H. Patil, and A. Mittal, "Robust order-based methods for feature description," in *CVPR*, 2010.

[55] M. Rubinstein, D. Gutierrez, O. Sorkine, and A. Shamir, "A comparative study of image retargeting," in *SIGGRAPH Asia*, 2010.

[56] O. Macindoe and W. Richards, "Graph comparison using fine structure analysis," in *ICSC*, 2010.

[57] T. Xu, H. Wang, and L. Sun, "Wood recognition based on histogram matching algorithms," *Applied Mechanics and Materials*, vol. 58, pp. 1744–1748, 2011.

[58] O. Pele and M. Werman, "Robust real time pattern matching using bayesian sequential hypothesis testing," *PAMI*, 2008.

[59] P. Yianilos, "Data structures and algorithms for nearest neighbor search in general metric spaces," in *SODA*, 1993.

[60] P. Ciaccia, M. Patella, and P. Zezula, "M-tree: An Efficient Access Method for Similarity Search in Metric Spaces," in *ICVLDB*, 1997.

[61] A. Beygelzimer, S. Kakade, and J. Langford, "Cover trees for nearest neighbor," in *ICML*, 2006.

[62] S. Arya, D. Mount, N. Netanyahu, R. Silverman, and A. Wu, "An optimal algorithm for approximate nearest neighbor searching fixed dimensions," *JACM*, 1998.

[63] M. Muja and D. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," in *VISSAPP*, 2009.

[64] G. Guisewite and P. Pardalos, "Minimum concave-cost network flow problems: Applications, complexity, and algorithms," *AOR*, 1990.

[65] A. Amiri and H. Pirkul, "New formulation and relaxation to solve a concave-cost network flow problem," *JORS*, 1997.

פונקציות מרחק: תיאוריה, אלגורתמים ואפליקציות

חיבור לשם קבלת תואר
"דוקטור לפילוסופיה"

מאת
אופיר פלא

הוגש לסינט של האוניברסיטה העברית
יולי 2011

עבודה זו נעשתה בהדרכתו של
פרופ' מיכאל ורמן

תקציר

אלגורתמים רבים בתחום הראייה הממוחשבת ובתחום הלמידה החישובית כול-לים בתוכם חישוב מרחקים. לדוגמה, שערוך תנועת מצלמה מתמונה אחת לאחרת, כולל שלב של התאמת מתארים (מתארים הם וקטורים של מספרים המתאר-ים איזור מסויים בתמונה). השלב הראשון הוא בחירת איזורים קטנים בשתי התמונות. השלב השני הוא חישוב מתאר עבור כל אחד מהאיזורים הללו. על מנת להחליט אם שני מתארים (משתי תמונות שונות) הם למעשה תמונה של אותו האיזור התלת מימדי בעולם, המתארים מושווים בעזרת פונקציה המחשבת מרחק בינהם. בחירת פונקצית המרחק משפיעה על דיוק ומהירות השיטה. שימוש אחר לפונקציות מרחק הוא חיפוש תמונות דומות. המרחק יקבע את איכות התמונות הדומות שיוחזרו למשתמש וכמה זמן ייקח להחזיר אותן. בתחום הל-מידה חישובית, איכות תיוג על פי k השכנים הקרובים ביותר נקבעת על ידי פונקצית המרחק המשמשת לקביעת השכנים.

תיזה זו מציגה מספר פונקציות מרחק חדשות וחוקרת את תכונותיהן. דגש מיוחד מושם על ישימות מעשית המבוססת על תובנות תיאורטיות. התיזה מציגה אלגורתמים יעילים לחישוב פונקציות המרחק. פונקציות המרחק המוצעות אדישות לרעש בתמונות, מצד שני, הן בעלות ערכים גבוהים עבור שתי מתארים שונים (לדוגמא שתי מתארים שלא מתארים את אותו האיזור בעולם). חוקרים אחרים השתמשו בשיטות המוצעות בתיזה זו על מנת לפתור בעיות מראייה ממוחשבת ומתחומים אחרים וקיבלו תוצאות טובות יותר משיטות אחרות. הצלחת השי-טות בתחומים אחרים נובעת כנראה מכך שתכונות הרעש בתחומים הללו דומות לתכונות הרעש של תמונות.

החלק הראשון והחלק השני של תיזה זו מוקדשים ל- Earth Mover's Distance (EMD) [1]. מרחק זה הוא הכללה של מרחק ה- transportation להיסטוגרמות לא מנורמלות. אנו מציגים הכללה חדשה למרחק ה- transportation להיסטוגרמות לא מנורמלות, $\widehat{EMD}$. שלא כמו ה-EMD הקלאסי, $\widehat{EMD}$ הוא מטריקה אפילו עבור היסטוגרמות לא מנורמלות. בנוסף לכך, שימוש ב- $\widehat{EMD}$ נותן בדרך כלל תוצאות ניסוייות טובות יותר משימוש ב-EMD. אנחנו גם מראים כי $\widehat{EMD}$ הינו הכללה טבעית של מרחק ה-L_1.

מתארים שהם היסטוגרמות של זויות משמשות למספר רב מאד של משימות ראייה ממוחשבת. רב המרחקים שמשמשים להשוואת שתי מתארים כאלו לא לוקחים בחשבון את המבנה המעגלי. לעומת זאת חישוב המרחקים שכן לוקחים את המבנה המעגלי בחשבון הוא ארוך מדי. אנו מציגים אלגוריתם לינארי המשווה בין היסטוגרמות של זויות ולוקח בחשבון את המבנה המעגלי. שיטה זו משיגה תוצאות טובות יותר מהשיטות הקודמות על בוחן ביצועים שהשתמשו בו רבים וזמין לכולם.

בחלק השני של התיזה אנו מציגים משפחה של Earth Mover's Distances אדישה לרעש וניתן להשתמש בה לכל סוג של היסטוגרמות (לא רק היסטוגרמות של זויות). פיתחנו אלגוריתם המחשב את המרחקים הללו בסדר גודל מהר יותר

מאשר חישוב מרחקים דומים. שיטה זו משיגה תוצאות מצויינות בחיפוש תמונות דומות.

בחלק השלישי של התיזה אנו מציגים משפחה חדשה של מרחקים להיסטוגרמות – Quadratic-Chi (QC). מרחקי QC הם מרחקי תבנית ריבועית עם נרמול דומה לנרמול של מרחק ה- χ^2. נרמול זה מוריד את השפעת התאים עם הערכים הגבוהים ונורמליזציה זו הוכחה כמועילה במקרים רבים. עם זאת, χ^2 רגיש לקוונטיזציה (שנגרמת לדוגמא על ידי שינויי תאורה ועיוותי צורה). הפרמטרים של התבנית הריבועית ממדלים את הקשר בין התאים השונים של ההיסטוגרמה (לדוגמא אדום וכתום) וכך בעיית הקוונטיזציה משוככת. אנו מציגים שתי תכונות חדשות למרחקים הממדלים קשר בין תאי ההיסטוגרמה: *Similarity-Matrix-Quantization-Invariance* ו- *Sparseness-Invariance* ומראים כי לכל מרחקי ה-QC יש את התכונות הללו. אנו מראים גם כי תכונות אילו חשובות מבחינה מעשית. חישוב מרחק QC נעשה בזמן לינארי. כמו כן קל למקבל את חישוב המרחק. מרחקי QC משיגים תוצאות טובות יותר ממה שהיו פונקציות המרחק הטובות ביותר עד כה וזמן חישובם קצר.

בפרק הרביעי של התיזה אנו מתארים שיטה חדשנית ללמידת מרחקים שקראנו לה - *"Interpolated-Discritized Metric Leraning"*. הצלחת מרחקי ה-QC נתנה מוטיבציה לחלק זה של העבודה. אמנם למרחקי QC ישנם פרמטרים שניתן לכוונן מאוד אותם, אך מגבלה חמורה היא שהמרחקים הללו מתאימים רק לוקטורים לא שליליים. בנוסף, QC לא קמור ביחס לפרמטרים שלו ולכן קשה ללמוד אותם. על מנת להתגבר על בעיות אלו אנו מציעים משפחה חדשה של מרחקים לה קראנו "Interpolated-Discritized" - או בקיצור ID. אנו מראים כיצד ניתן ללמוד את הפרמטרים של מרחק כזה בצורה יעילה. מרחק ID יכול לקרב כל מרחק. קל לשנות את רוב שיטות למידת המרחקים הקיימות כיום כך שילמדו מרחקי ID. אנו מדגימים דרך אחת אפשרית ללימוד מרחקי ID על ידי שימוש בשיטה Large Margin Nearest Neighbor [2]. עם מרחקי ID בעיית האופטימזציה של שיטה זו היא בעיית תכנון לינארי. אנו מראים כי במקרים רבים שיטת למידת המרחקים החדשה שלנו משיגה תוצאות טובות יותר משיטות למידת מרחקים שהיו השיטות הטובות ביותר עד כה.

Distance Functions: Theory, Algorithms and Applications - Appendices

Thesis submitted for the degree of
"Doctor of Philosophy"

By
Ofir Pele

Submitted to the Senate of the Hebrew University
July, 2011

This work was carried out under the supervision of:
Prof. Michael Werman

Contents

Chapter 1

A Linear Time Histogram Metric for Improved SIFT Matching - Appendices

Appendix A. Linear Time $\mathrm{EMD}_{\mathrm{MOD}}$ Algorithm

This appendix presents a linear time algorithm for the computation of the Modulo Earth Mover's Distance between normalized cyclic histograms. The algorithm is an adaption of Werman et al. algorithm for bipartite graph matching for points on a circle [22].

Let $A = \{0, \ldots, N-1\}$ be N points, equally spaced, on a circle. Let Q and P be two normalized histograms of such points. The Modulo Earth Mover's Distance ($\mathrm{EMD}_{\mathrm{MOD}}$) between Q and P is defined as:

$$\mathrm{EMD}_{\mathrm{MOD}}(Q,P) = \min_{F=\{f_{ij}\}} \sum_{i,j} f_{ij} D_{MOD}(i,j) \quad s.t \tag{6}$$

$$\sum_j f_{ij} \leq P_i \ , \ \sum_i f_{ij} \leq Q_j \ , \ \sum_{i,j} f_{ij} = \sum_i P_i = \sum_j Q_j \ , \ f_{ij} \geq 0 \tag{7}$$

where $D_{MOD}(i,j)$ is the modulo L_1 distance:

$$D_{MOD}(i,j) = \min(|i-j|, N-|i-j|) \tag{8}$$

The linear time $\mathrm{EMD}_{\mathrm{MOD}}$ pseudo code is given in Alg. 1. Lines 2-8 compute the F function from [22] for masses instead of points; *i.e.* $F[i]$ is the total mass in bins smaller or equal to i in Q, minus the total mass in bins smaller or equal to i in P. As was noted by Cabrelli and Molter [37] the cutting point of the circle for the match can be found with a weighted median algorithm when the points on the circles are sorted. In histograms the points are sorted and equally spaced; thus it is enough to find the median. Lines 10-16 in Alg. 1 cut the circles into two lines. Finally, lines 17-31 match groups of points instead of single points.

Algorithm 1 $\mathrm{EMD}_{\mathrm{MOD}}(Q, P, N)$

```
1: D ⇐ 0
2: cQ ⇐ Q[0]
3: cP ⇐ P[0]
4: F[0] ⇐ Q[0] − P[0]
5: for i = 1 to N − 1 do
6:    cQ ⇐ cQ + Q[i]
7:    cP ⇐ cP + P[i]
8:    F[i] ⇐ cQ − cP
9: i ⇐ ((index of the median of F) + 1)  mod N
10: for t = 0 to N − 1 do
11:    I[t] ⇐ i
12:    i ⇐ (i + 1)  mod N
13: tQ ⇐ 0
14: tP ⇐ 0
15: iQ ⇐ I[tQ]
16: iP ⇐ I[tP]
17: while true do
18:    while Q[iQ]=0 do
19:       tQ ⇐ tQ + 1
20:       if tQ = N then
21:          return D
22:       iQ ⇐ I[tQ]
23:    while P[iP]=0 do
24:       tP ⇐ tP + 1
25:       if tP = N then
26:          return D
27:       iP ⇐ I[tP]
28:    f ⇐ min(Q[iQ], P[iP])
29:    Q[iQ] ⇐ Q[iQ] − f
30:    P[iP] ⇐ P[iP] − f
31:    D ⇐ f × min(|iQ − iP|, N − |iQ − iP|)
```

Appendix B. $\widehat{EMD}$ Metric Proof

In this appendix we prove that if $\alpha \geq 0.5$ and the ground distance is a metric, $\widehat{EMD}$ (see Eq. 3 in page 3) is a metric. Non-negativity and symmetry hold trivially in all cases, so we only need to prove that the triangle inequality holds.

Theorem 1. *Let A, B, C be three vectors in $(\mathbb{R}^+)^N$ and let $\widehat{EMD}$ be with a metric ground distance and $\alpha \geq 0.5$, then:*

$$\widehat{EMD}_\alpha(A,B) + \widehat{EMD}_\alpha(B,C) \geq \widehat{EMD}_\alpha(A,C) \tag{9}$$

Proof. Given an $N \times N$ distance matrix d_{ij}, let $\tilde{d}_{ij}$ be an $(N+1) \times (N+1)$ distance matrix such that:

$$\tilde{d}_{ij} = \begin{cases} d_{ij} & 1 \leq i,j \leq N \\ \alpha \max_{i,j}\{d_{ij}\} & i = N+1, j \neq N+1 \\ \alpha \max_{i,j}\{d_{ij}\} & i \neq N+1, j = N+1 \\ 0 & i = N+1, j = N+1 \end{cases} \tag{10}$$

Let $S = \max(\sum_{i=1}^N A_i, \sum_{i=1}^N B_i, \sum_{i=1}^N C_i)$. We define three vectors in $\mathbb{R}^{N+1}$:

$$\tilde{A} = [A, S - \sum_{i=1}^N A_i] \tag{11}$$

$$\tilde{B} = [B, S - \sum_{i=1}^N B_i] \tag{12}$$

$$\tilde{C} = [C, S - \sum_{i=1}^N C_i] \tag{13}$$

The sum of each of the three vectors equals S. In addition:

$$\widehat{EMD}_\alpha(A,B) = EMD(\tilde{A},\tilde{B}) \times S \tag{14}$$

$$\widehat{EMD}_\alpha(B,C) = EMD(\tilde{B},\tilde{C}) \times S \tag{15}$$

$$\widehat{EMD}_\alpha(A,C) = EMD(\tilde{A},\tilde{C}) \times S \tag{16}$$

where $\widehat{EMD}$ is with the ground distance d_{ij} and EMD is with the ground distance $\tilde{d}_{ij}$. Rubner et al. [1] proved that the triangle inequality holds for EMD when the ground distance is a metric and the histograms are normalized. $\tilde{A}, \tilde{B}, \tilde{C}$ are normalized histograms. If the ground distance d_{ij} is a metric and $\alpha \geq 0.5$, the ground distance $\tilde{d}_{ij}$ is a metric (an enumeration of all cases proves this). Thus we get:

$$EMD(\tilde{A},\tilde{B})+EMD(\tilde{B},\tilde{C}) \geq EMD(\tilde{A},\tilde{C}) \tag{17}$$

$$\Downarrow (S \geq 0) \tag{18}$$

$$(EMD(\tilde{A},\tilde{B}) \times S)+(EMD(\tilde{B},\tilde{C}) \times S) \geq (EMD(\tilde{A},\tilde{C}) \times S) \tag{19}$$

$$\Downarrow (\text{Eqs. 14, 15, 16}) \tag{20}$$

$$\widehat{EMD}_\alpha(A,B)+\widehat{EMD}_\alpha(B,C) \geq \widehat{EMD}_\alpha(A,C) \tag{21}$$

Appendix C. A Linear Time Histogram Metric for Improved SIFT Matching Additional Results

Ofir Pele[1] and Michael Werman[1]

School of Computer Science and Engineering
The Hebrew University of Jerusalem
{ofirpele,werman}@cs.huji.ac.il

This document contains all the 1-precision vs. recall graph results for the experiments in the paper: "A Linear Time Histogram Metric for Improved SIFT Matching" [1].

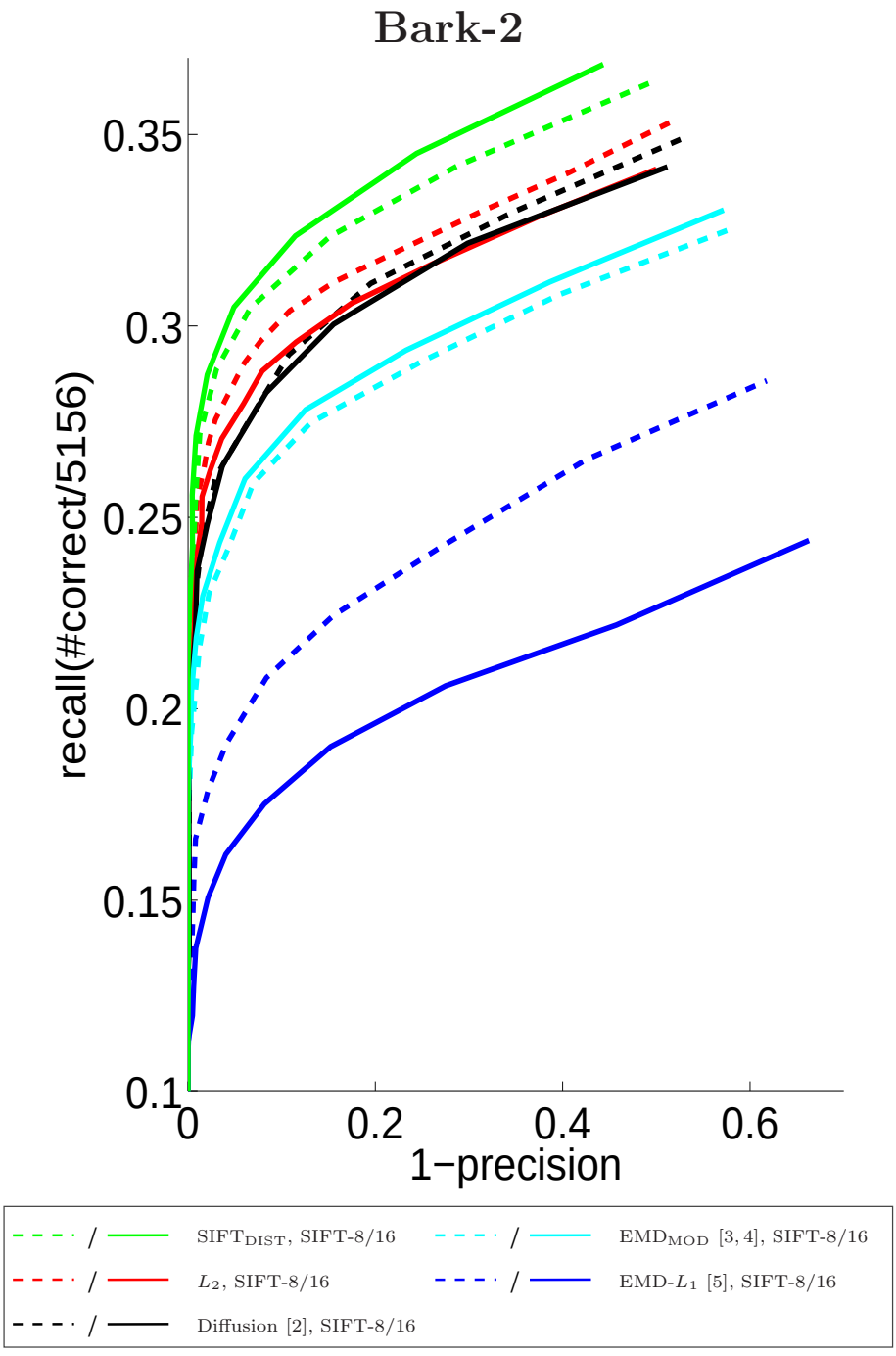


Fig. 1. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

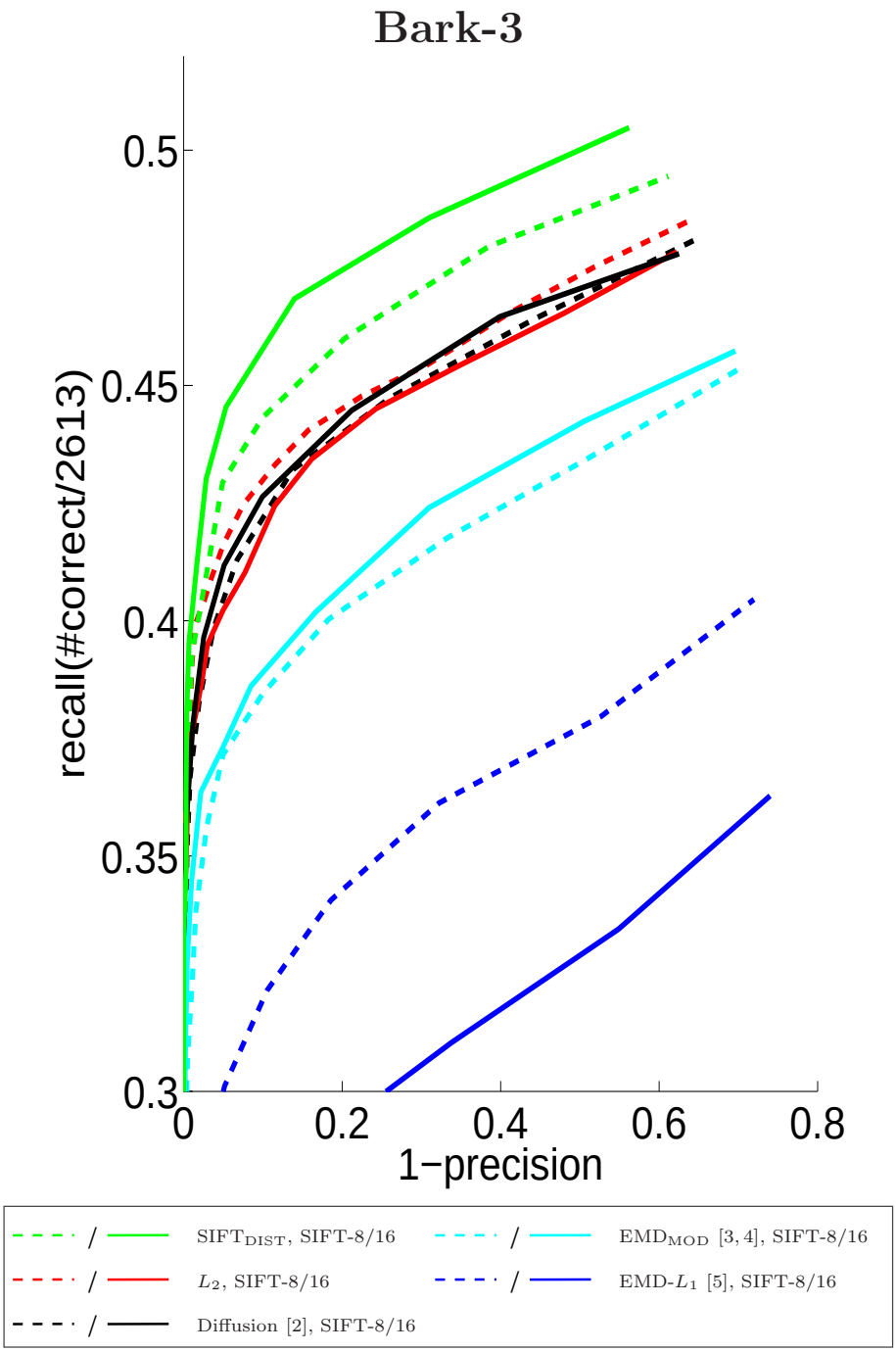


Fig. 2. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

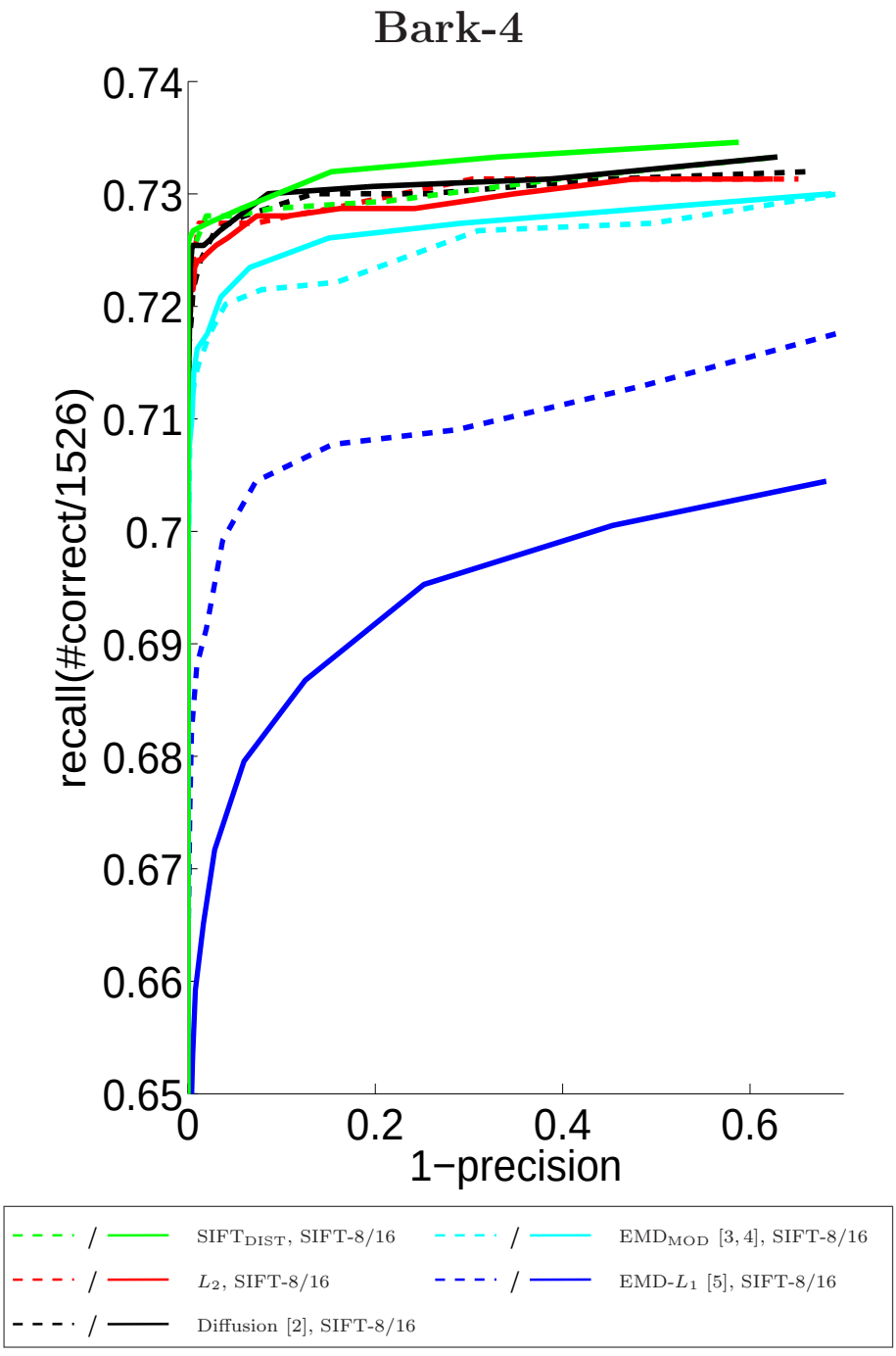


Fig. 3. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

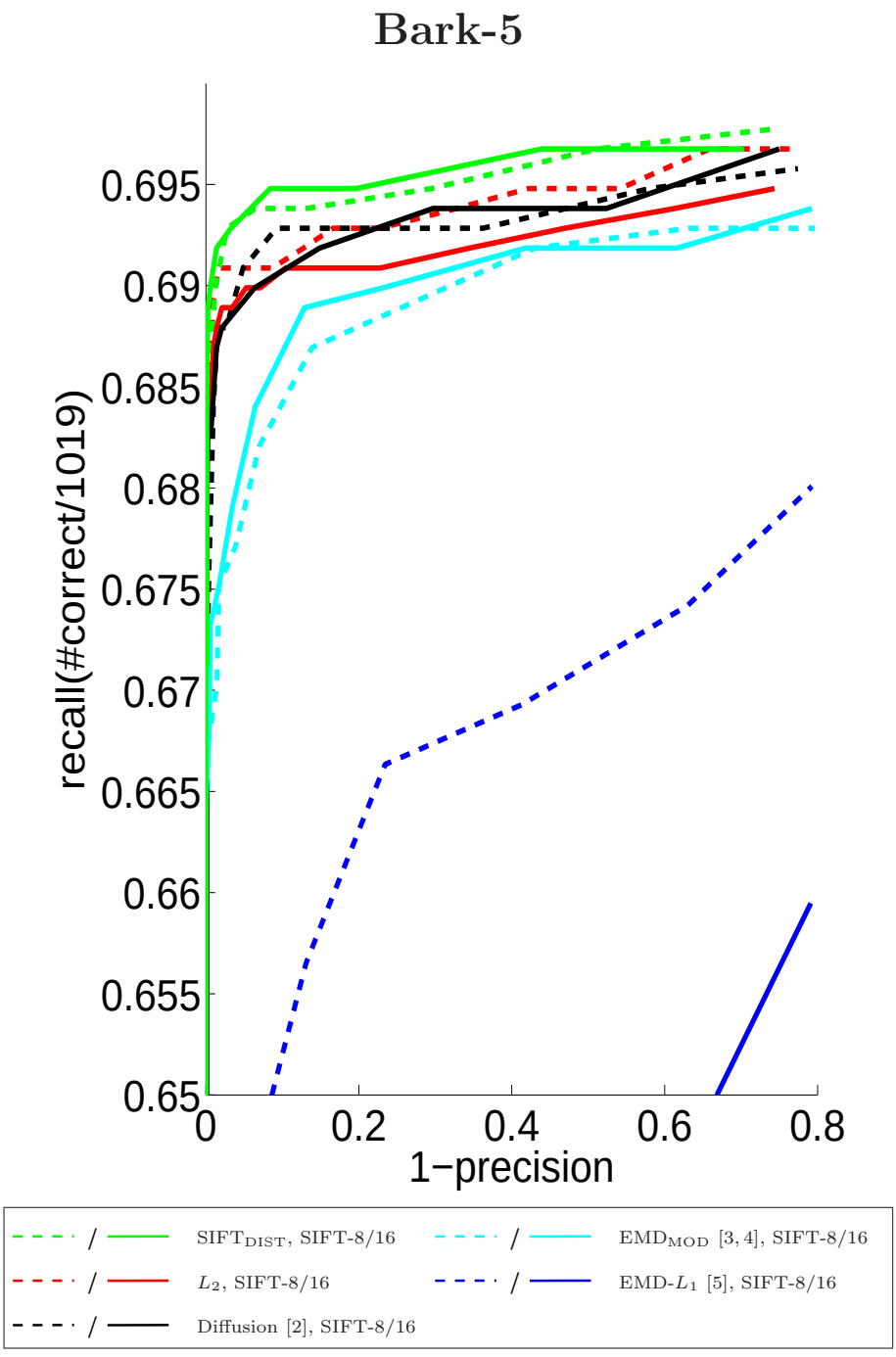


Fig. 4. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

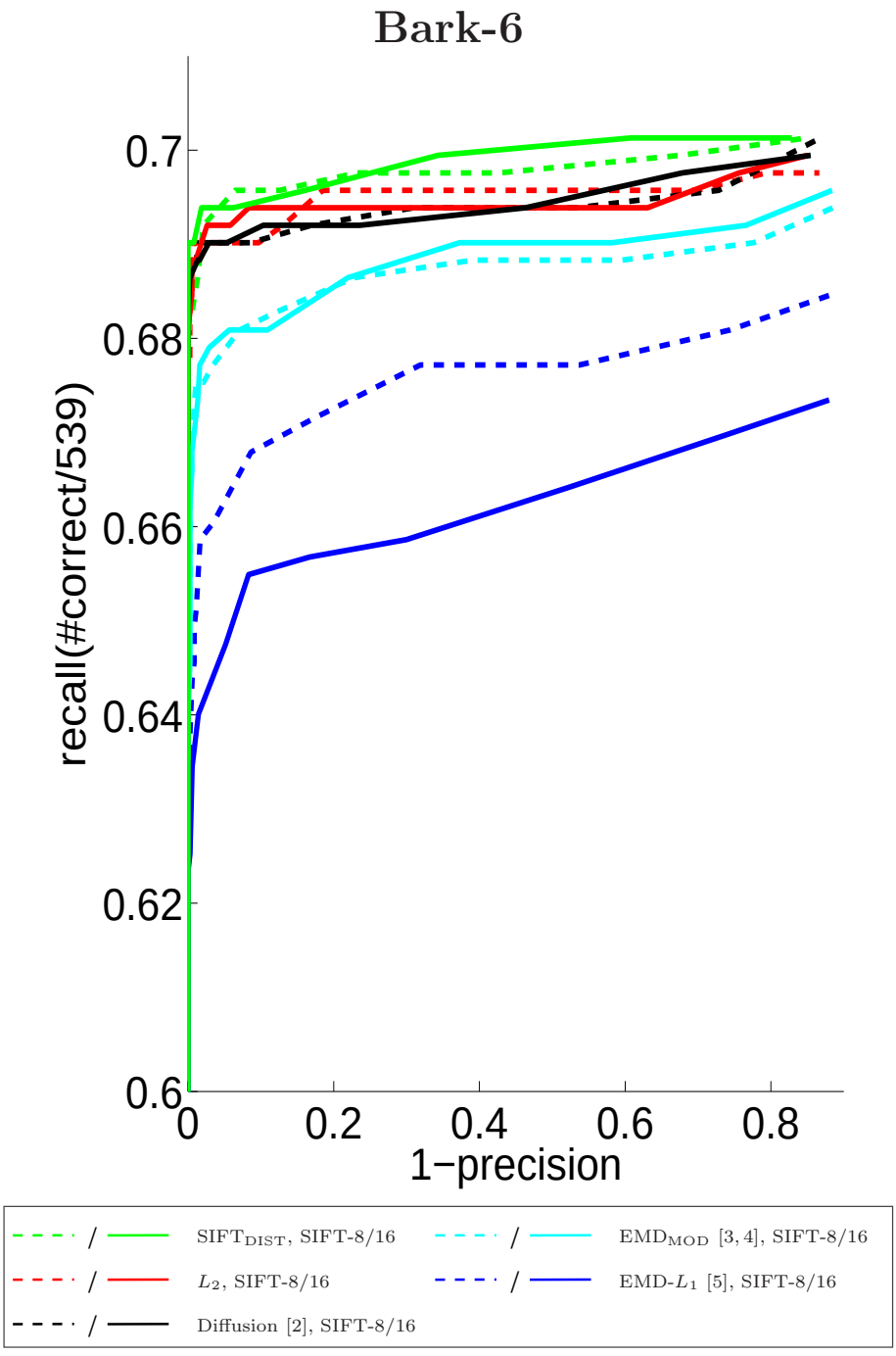


Fig. 5. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

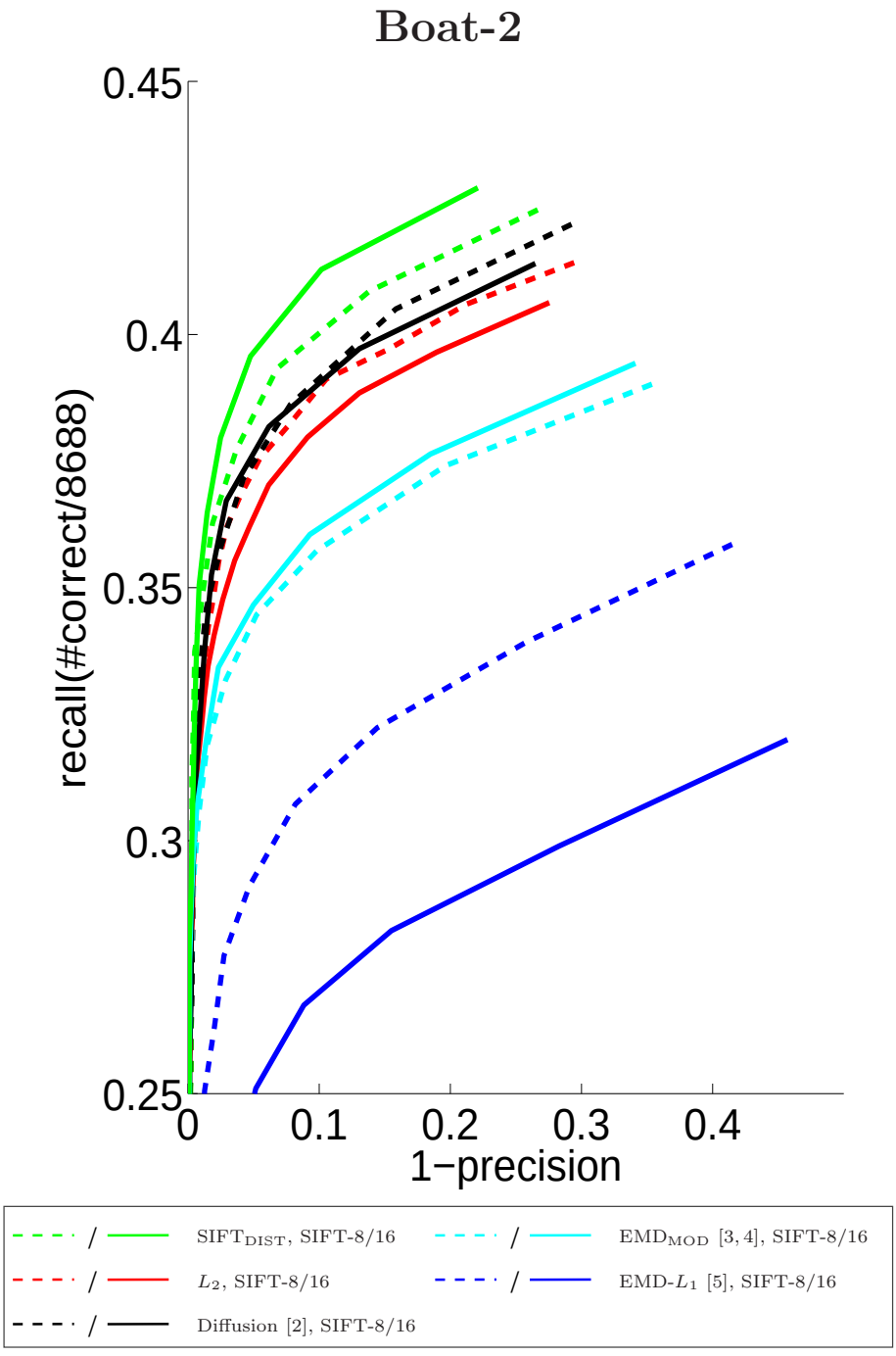


Fig. 6. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

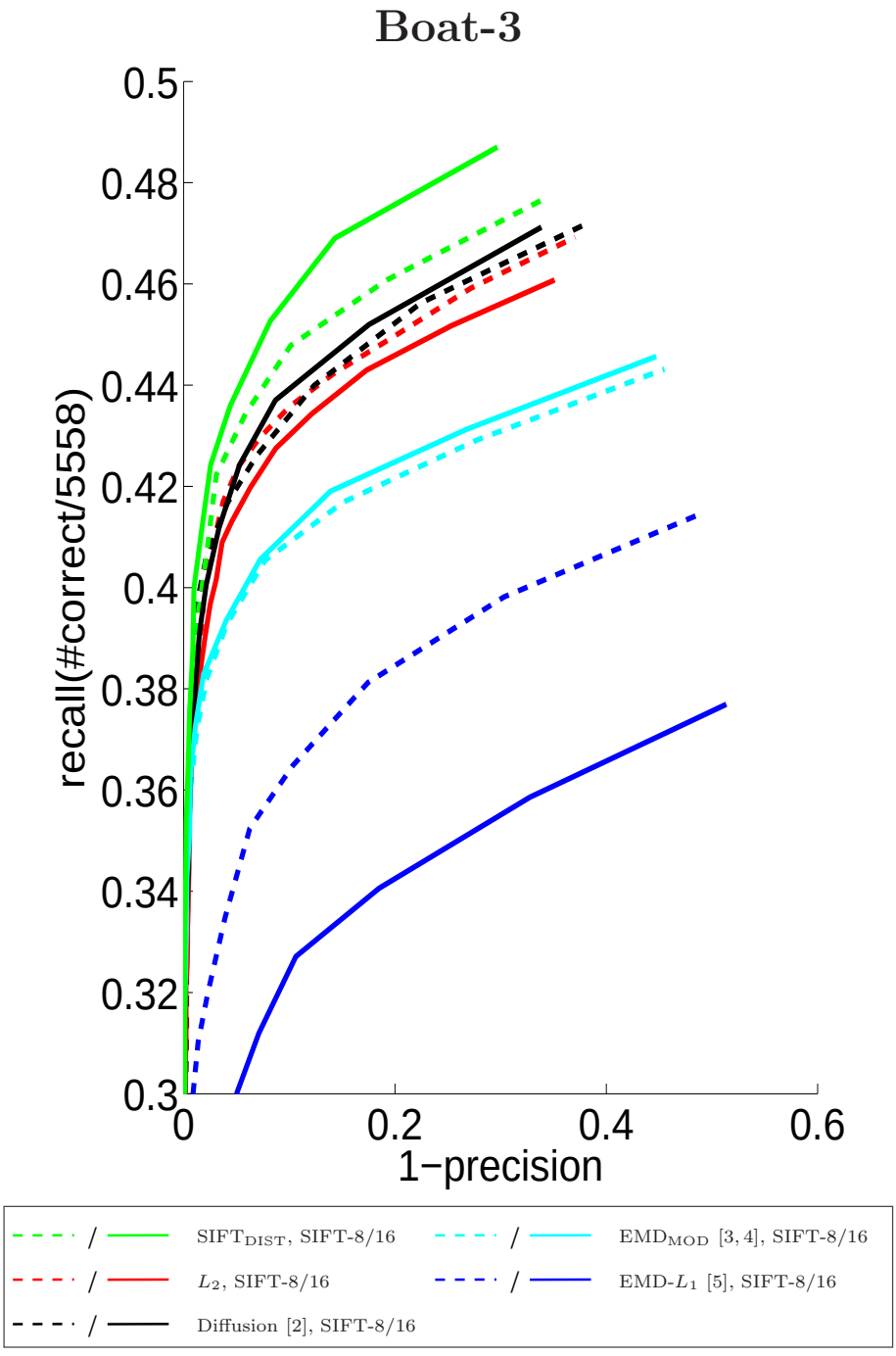


Fig. 7. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

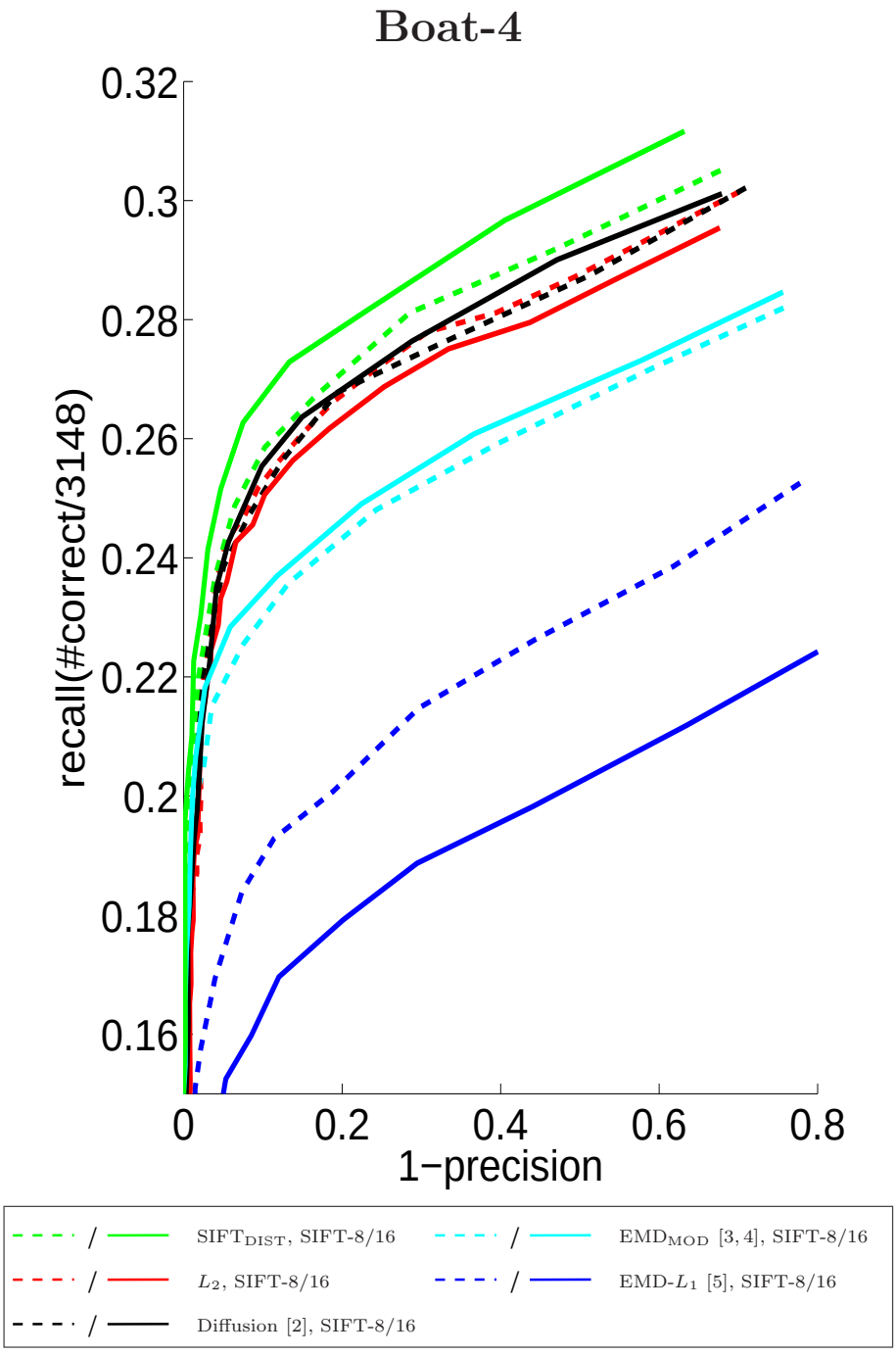


Fig. 8. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

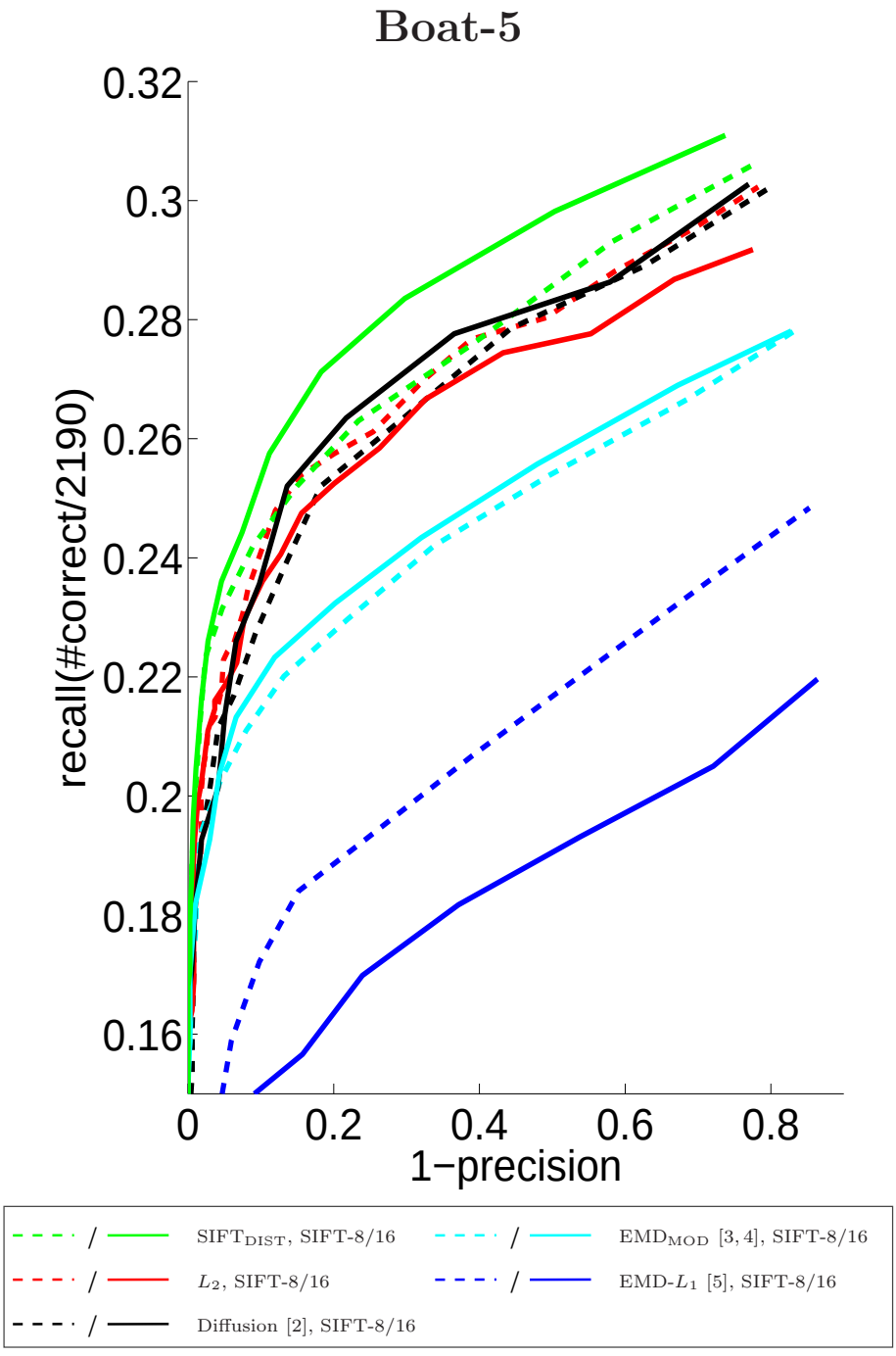


Fig. 9. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

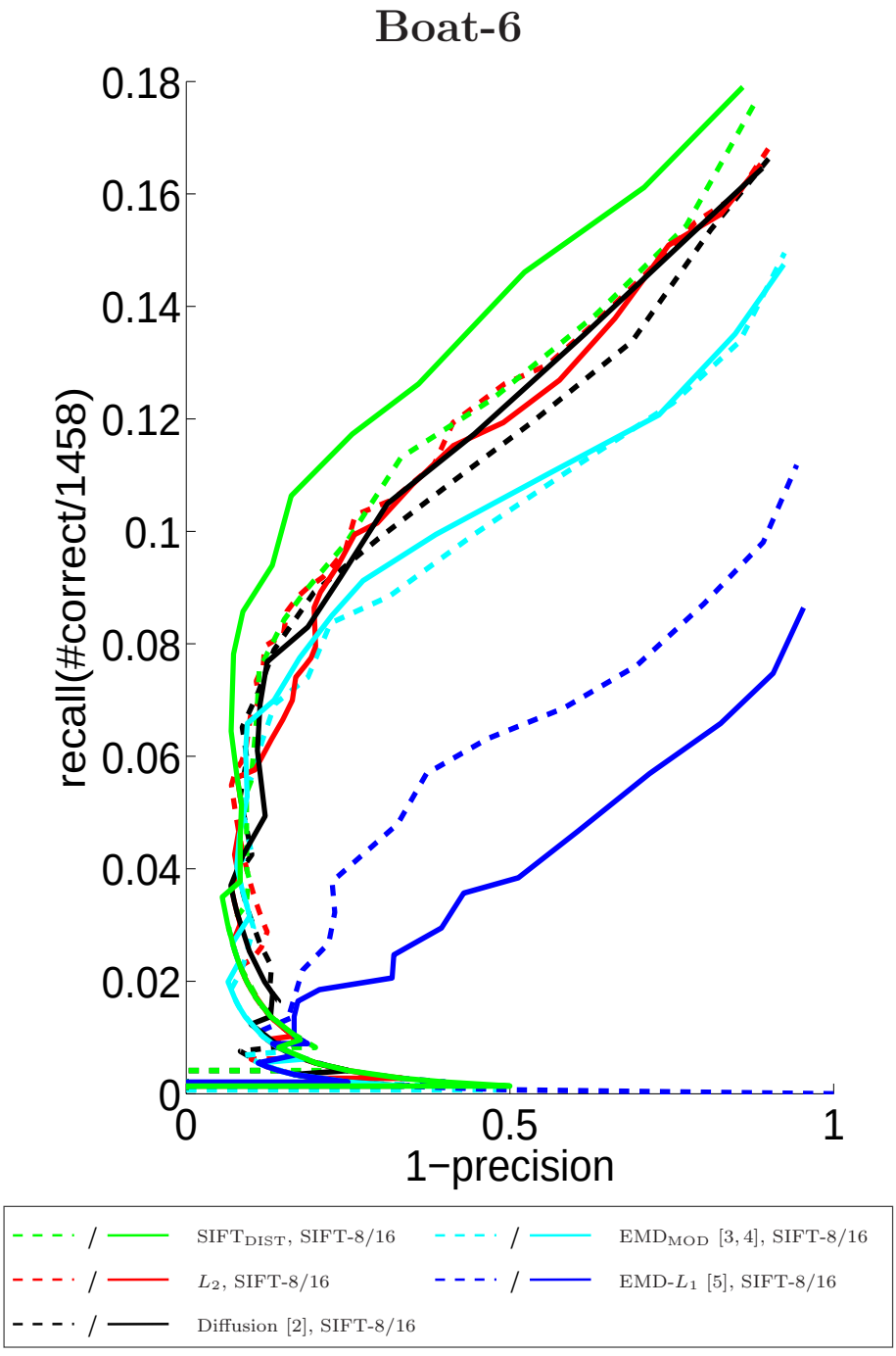


Fig. 10. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

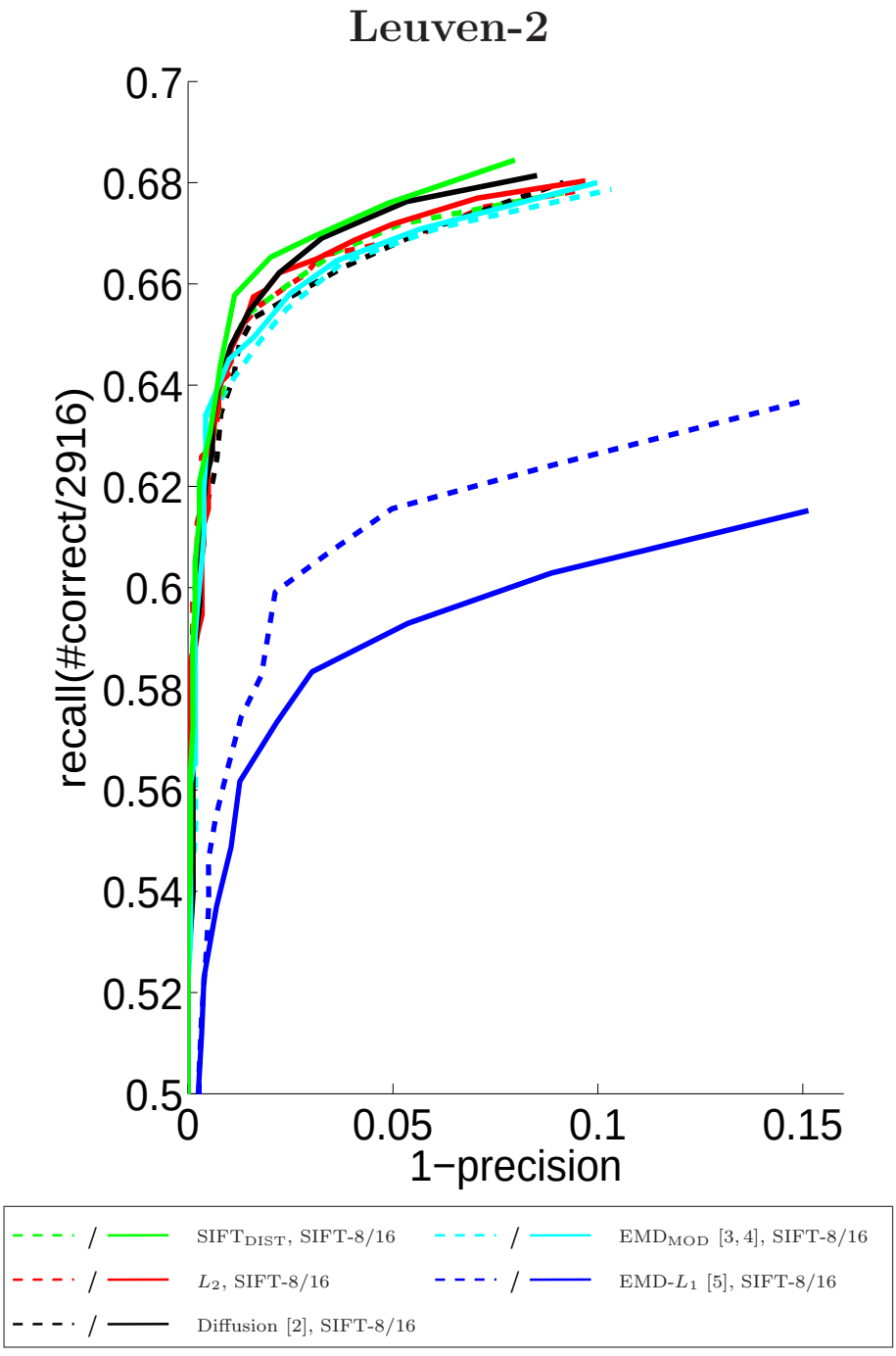


Fig. 11. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

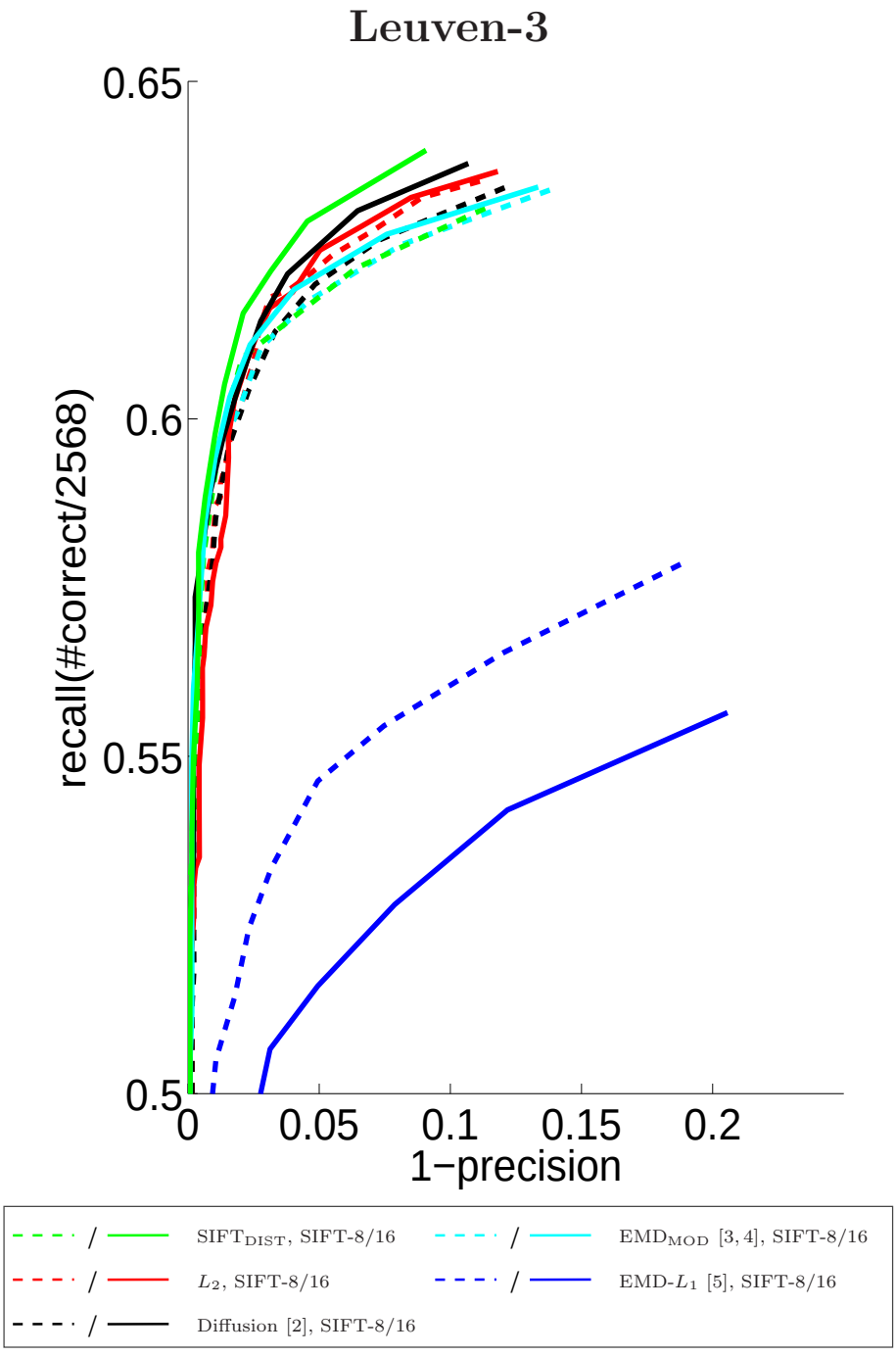


Fig. 12. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

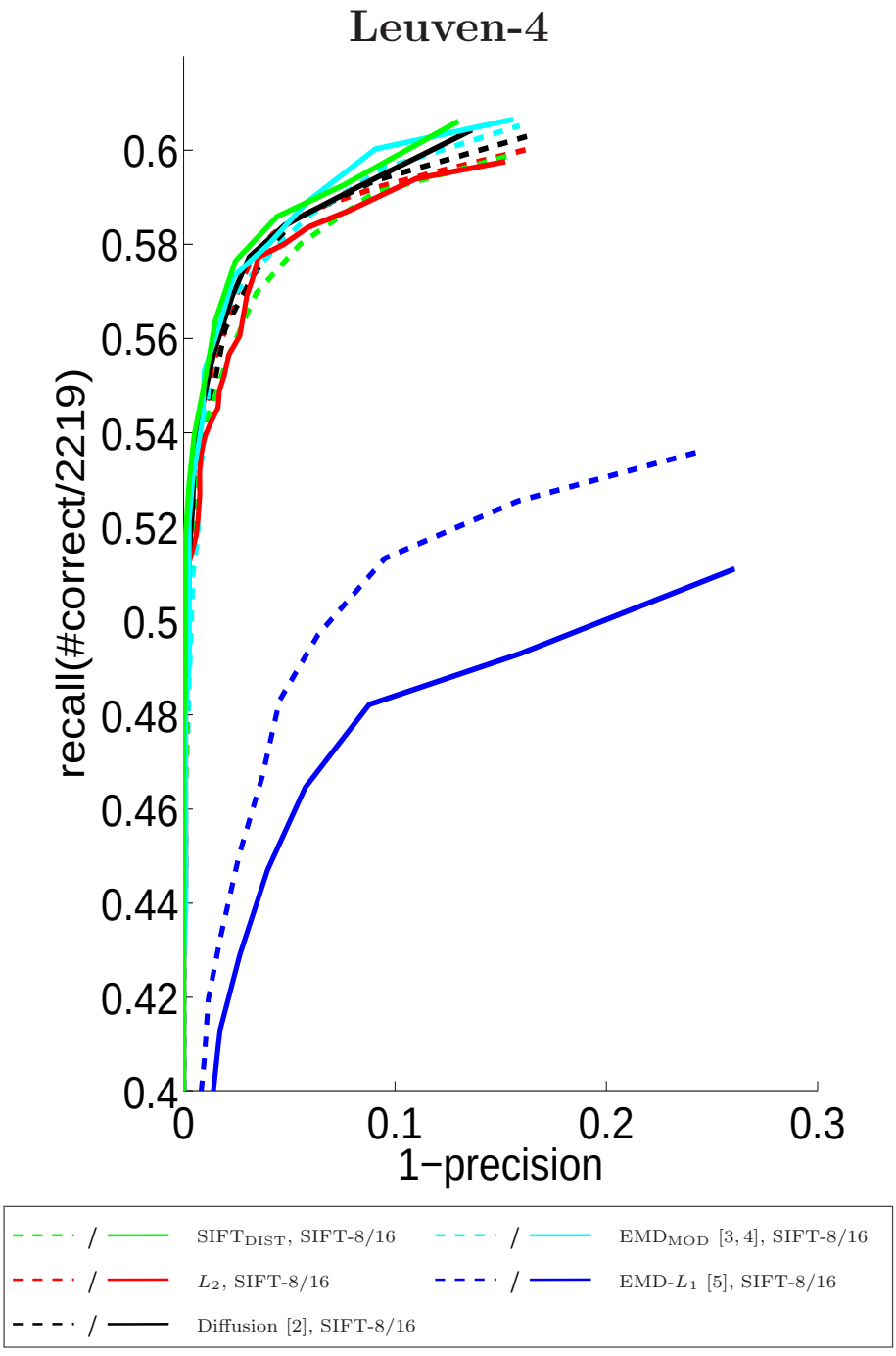


Fig. 13. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

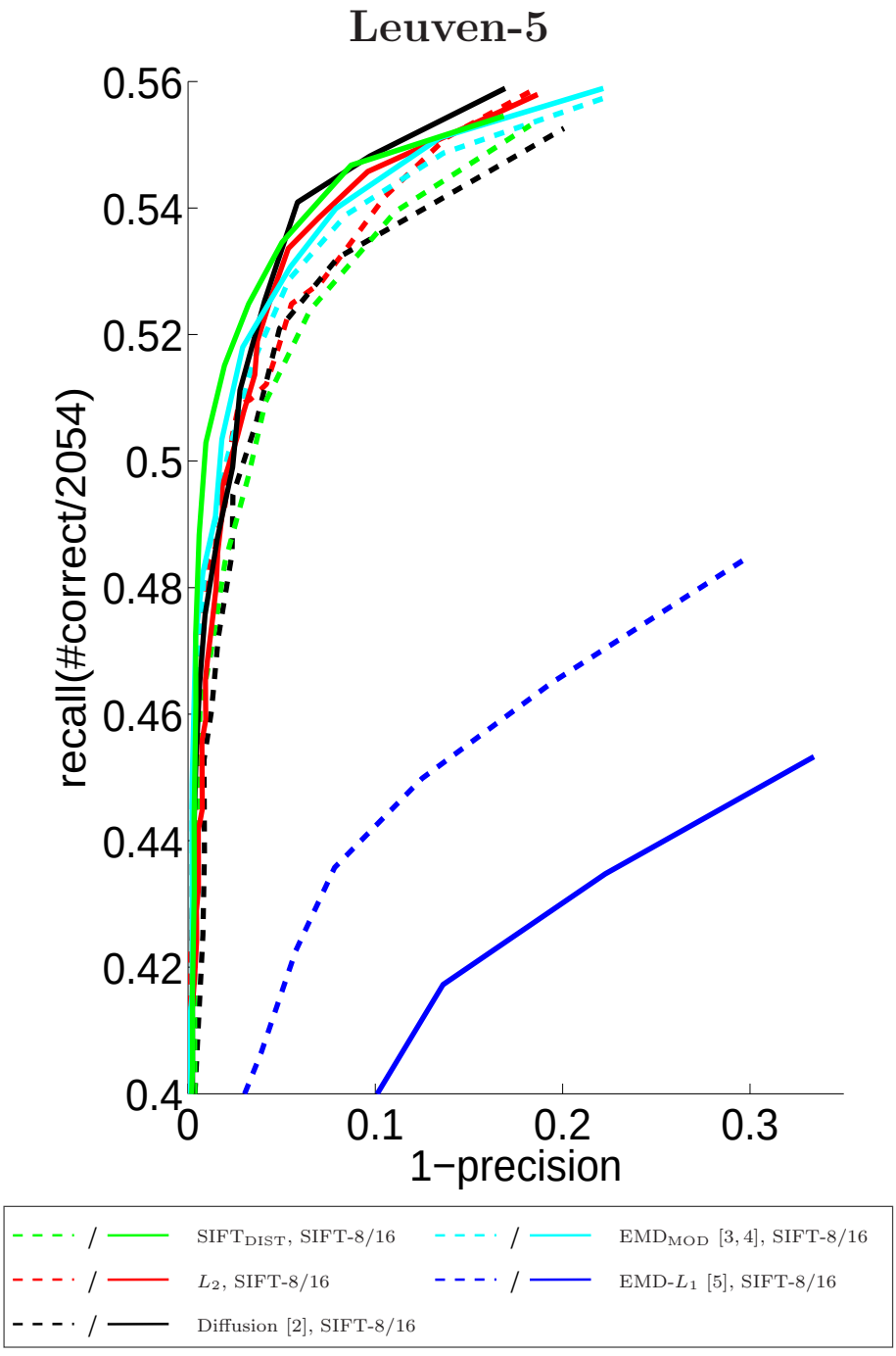


Fig. 14. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

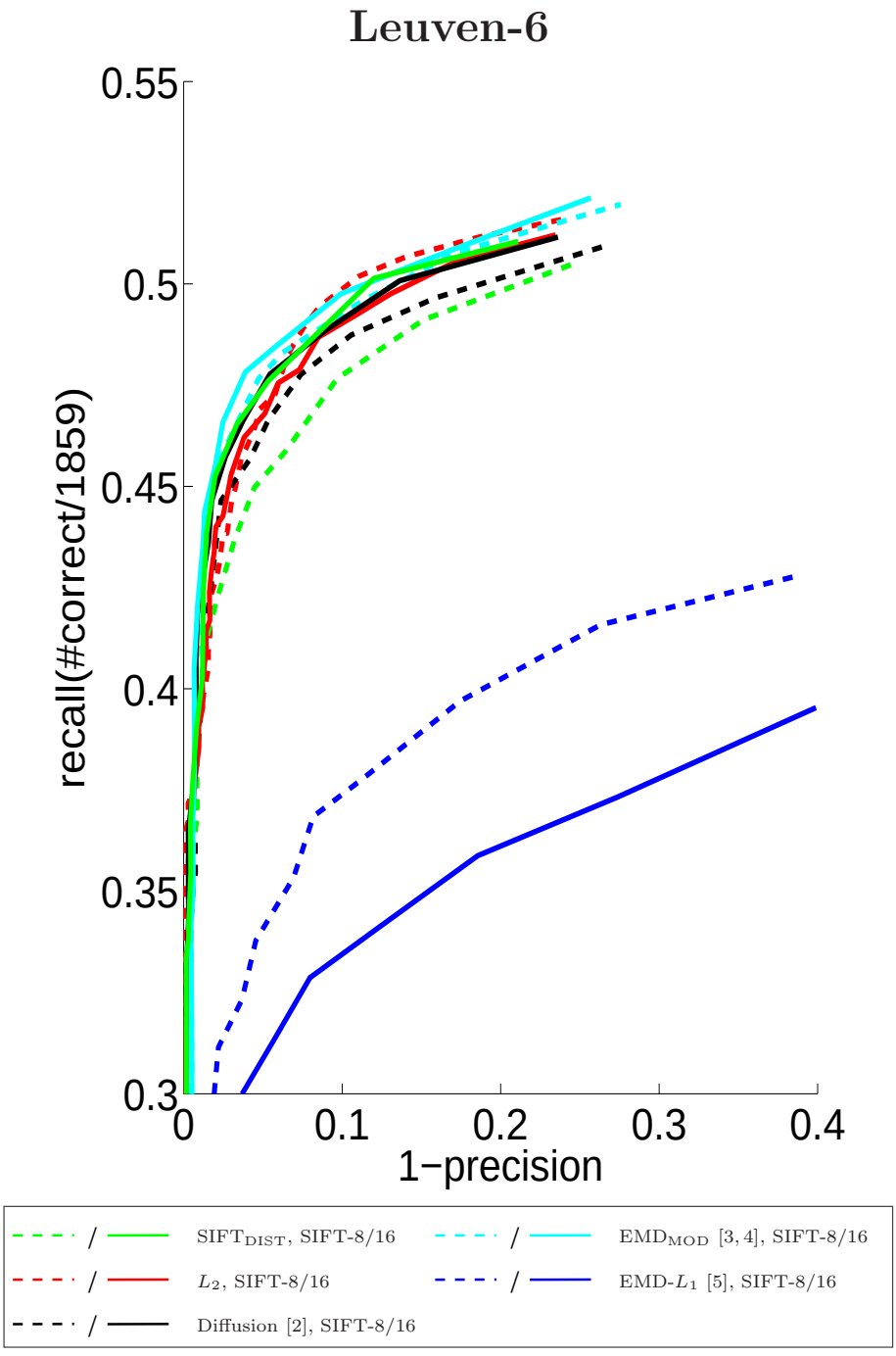


Fig. 15. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

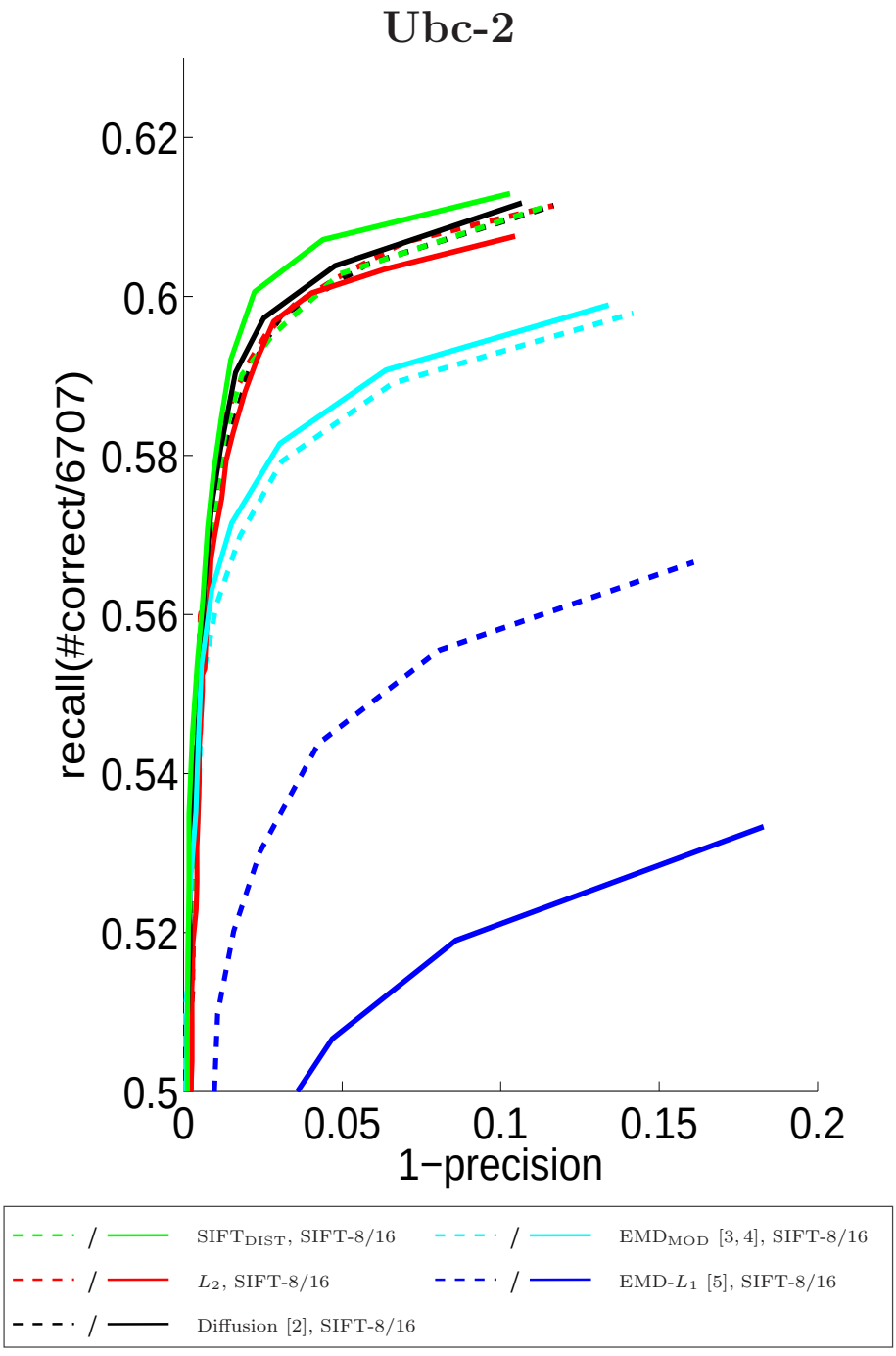


Fig. 16. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

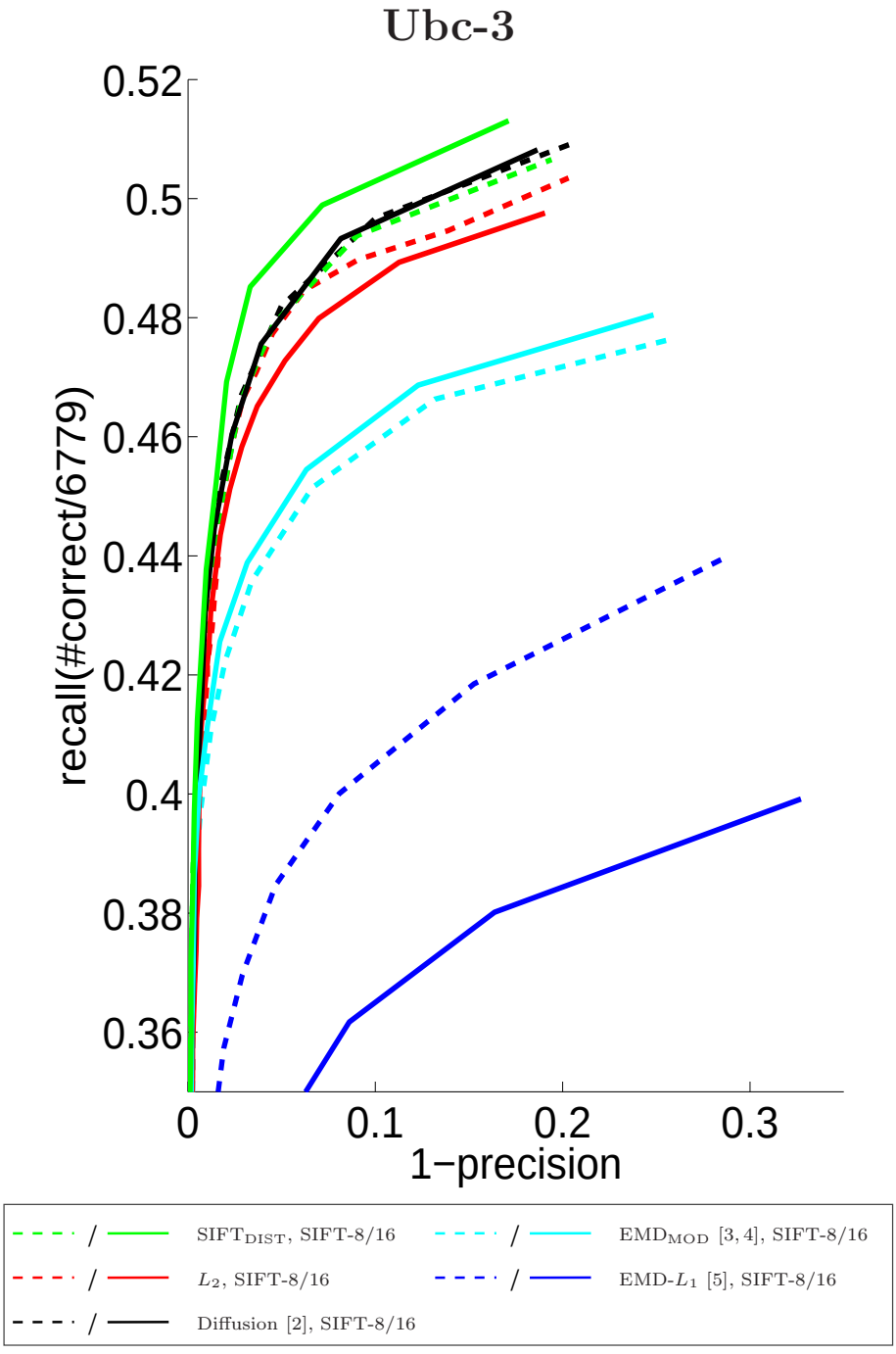


Fig. 17. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

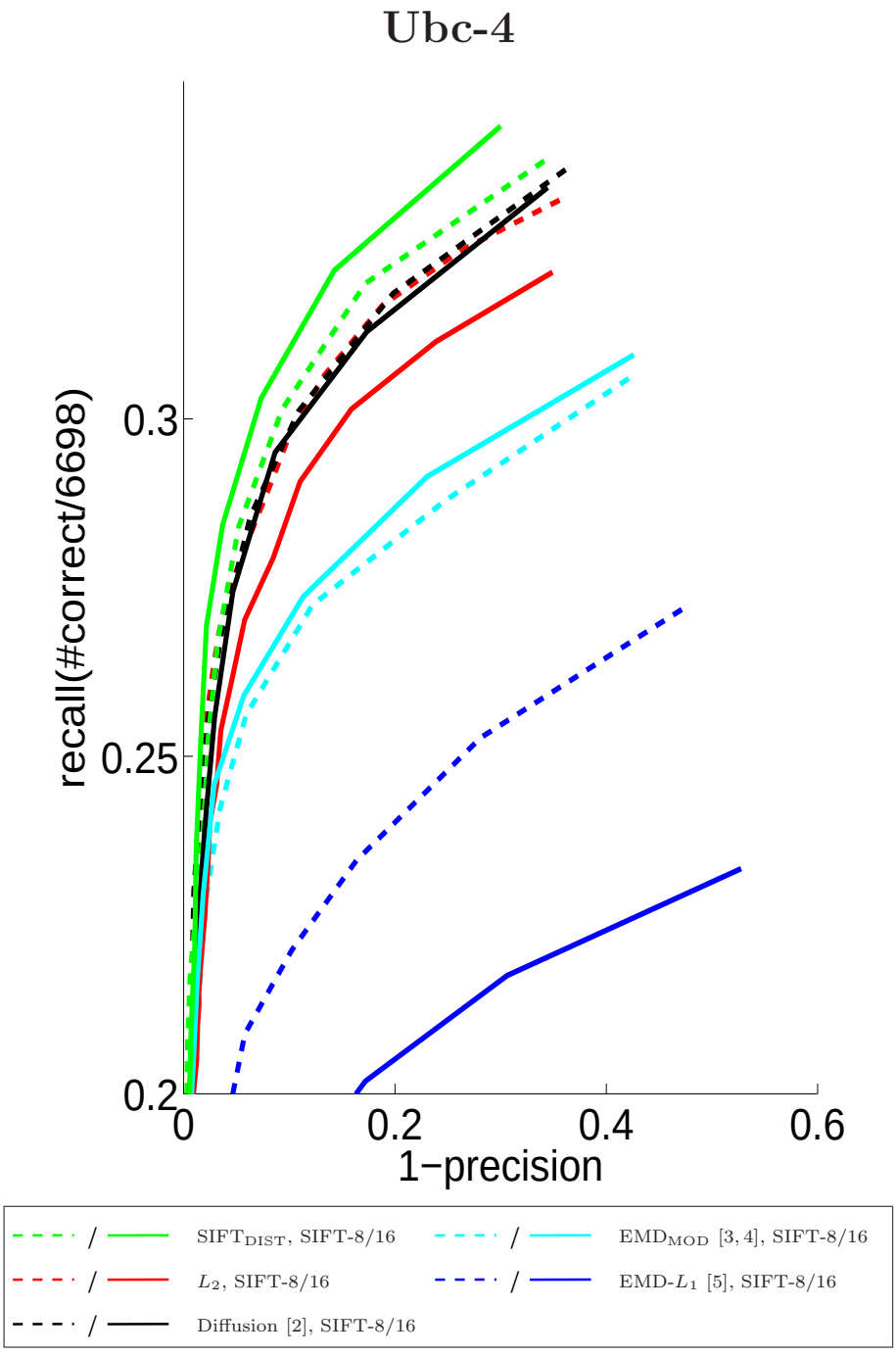


Fig. 18. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

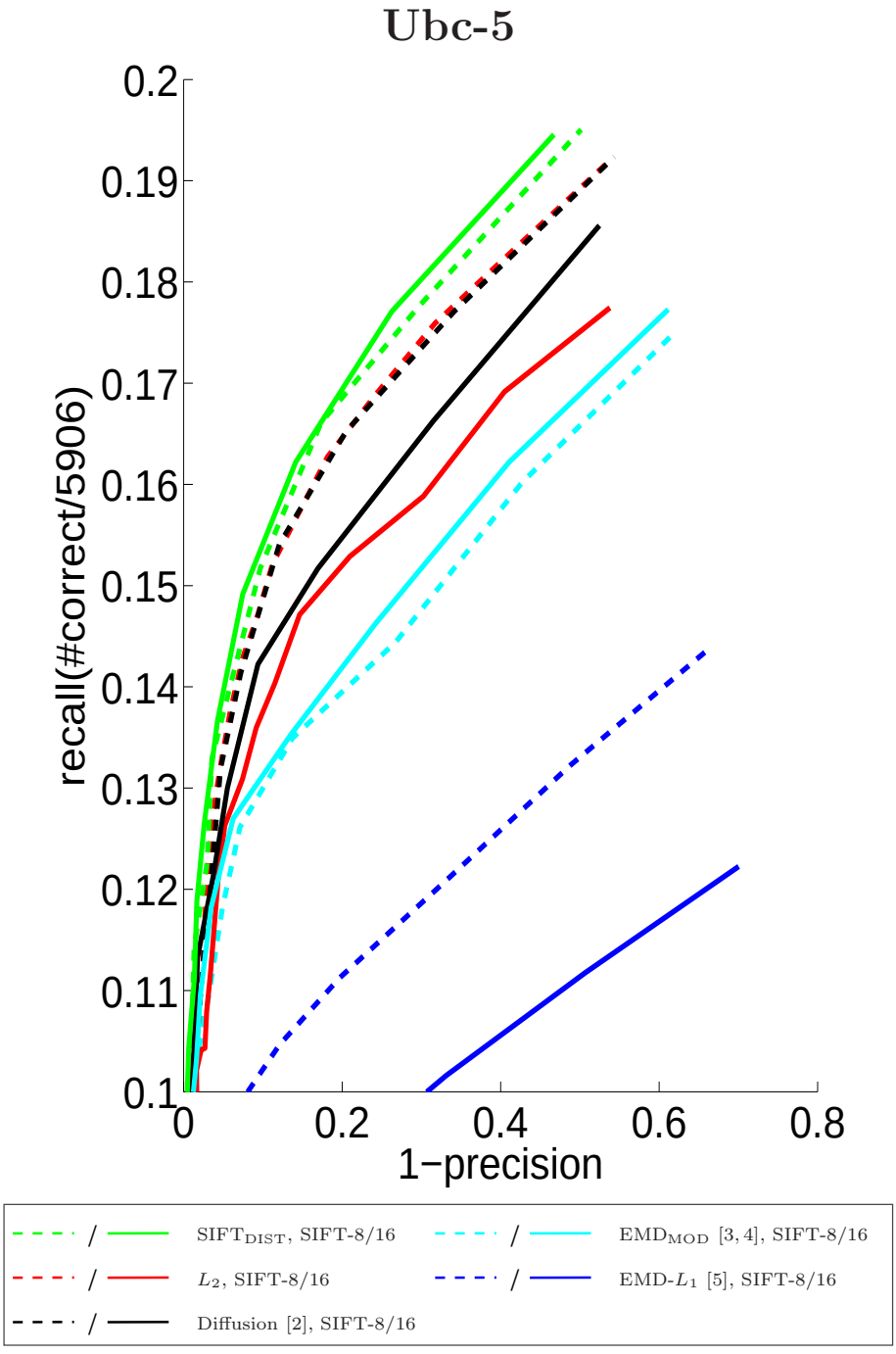


Fig. 19. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

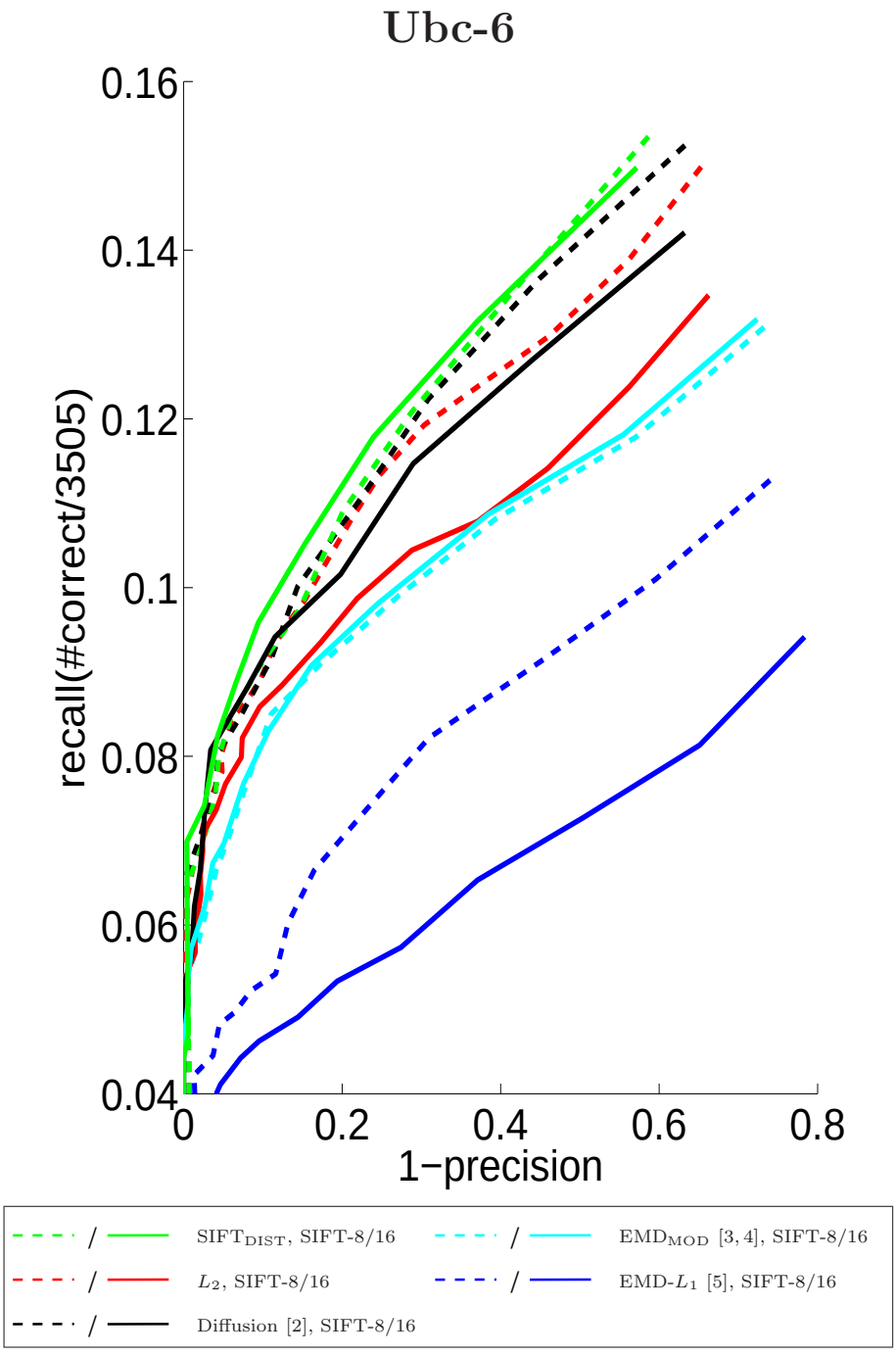


Fig. 20. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

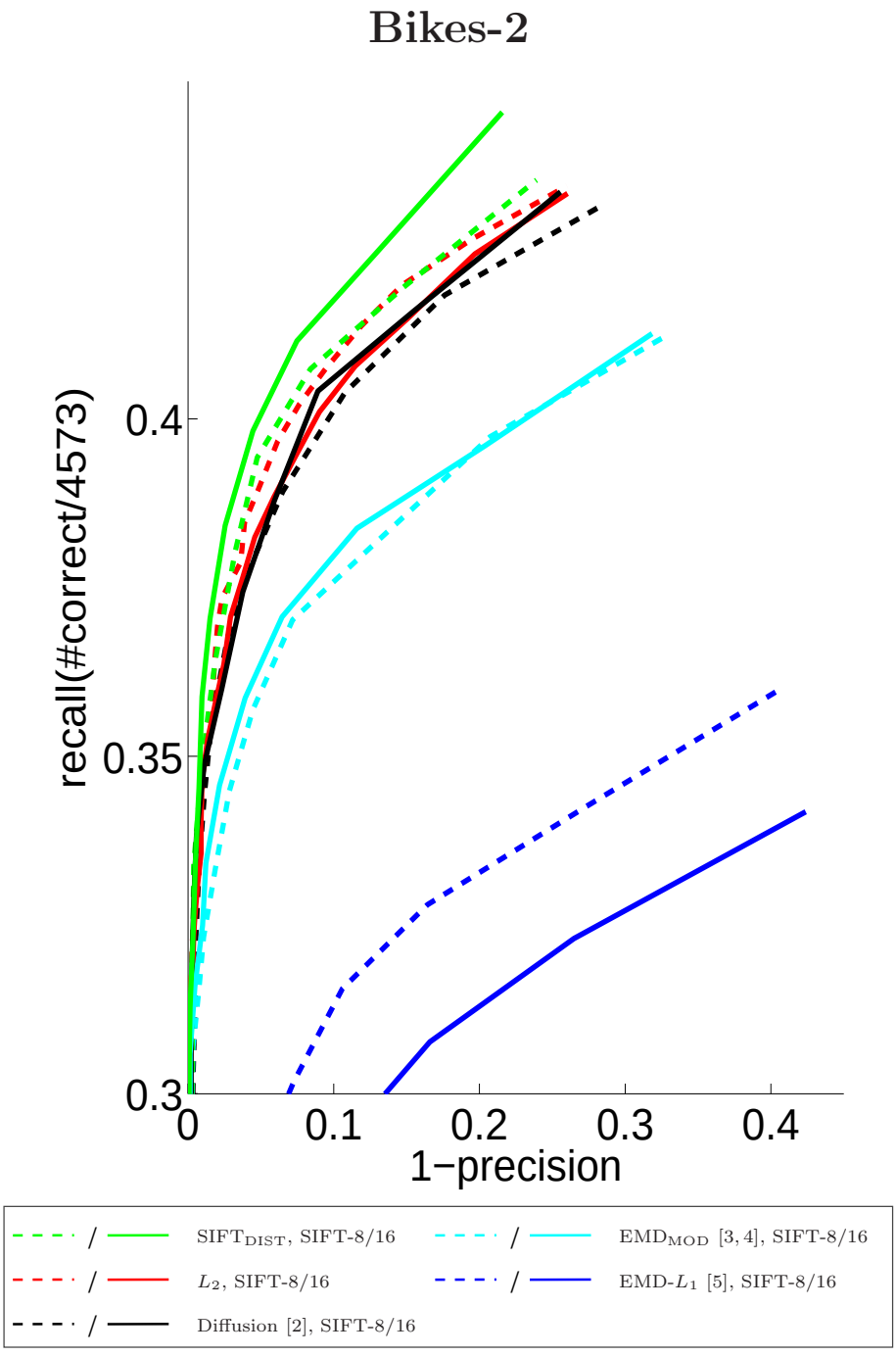


Fig. 21. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

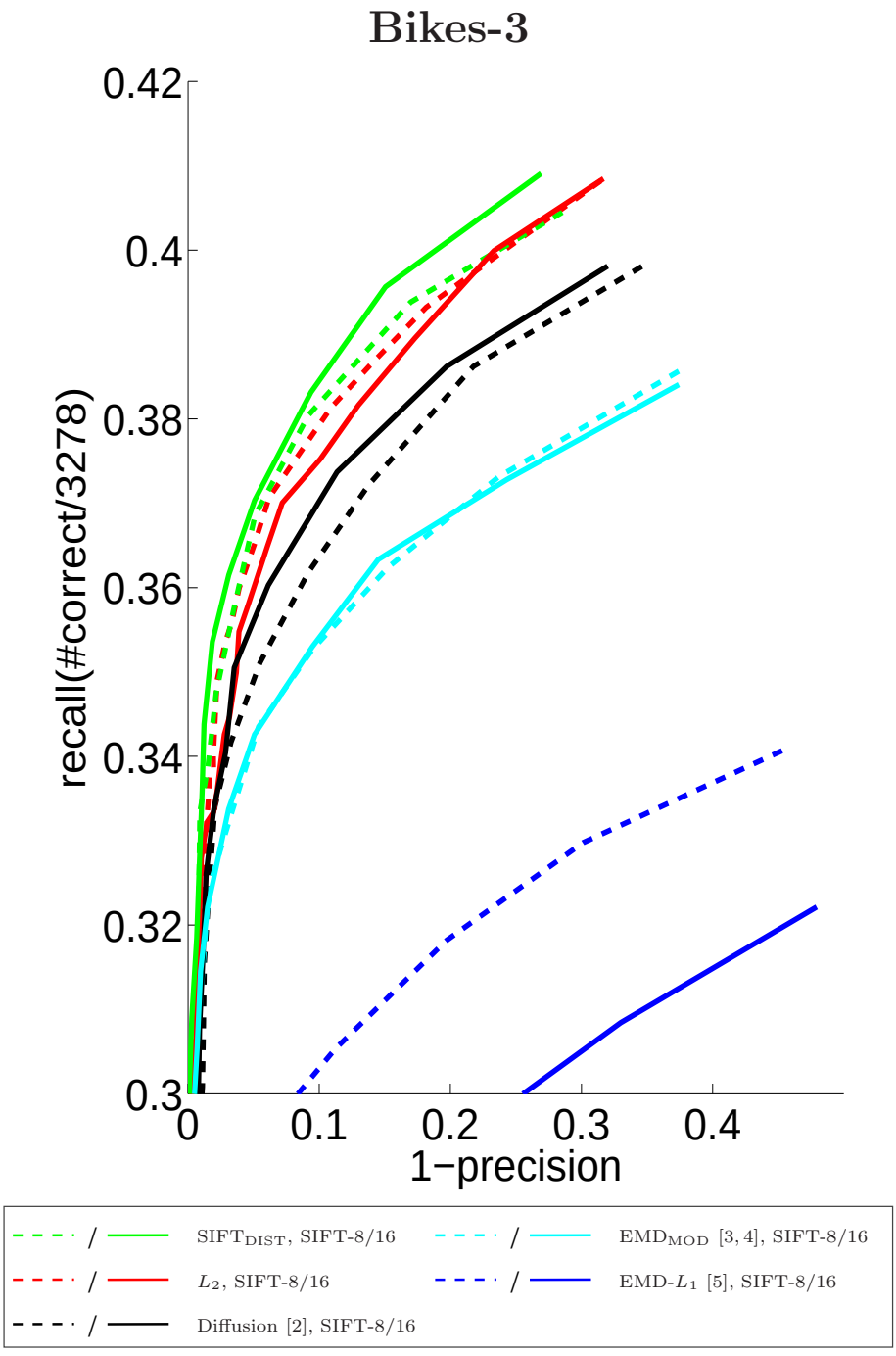


Fig. 22. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

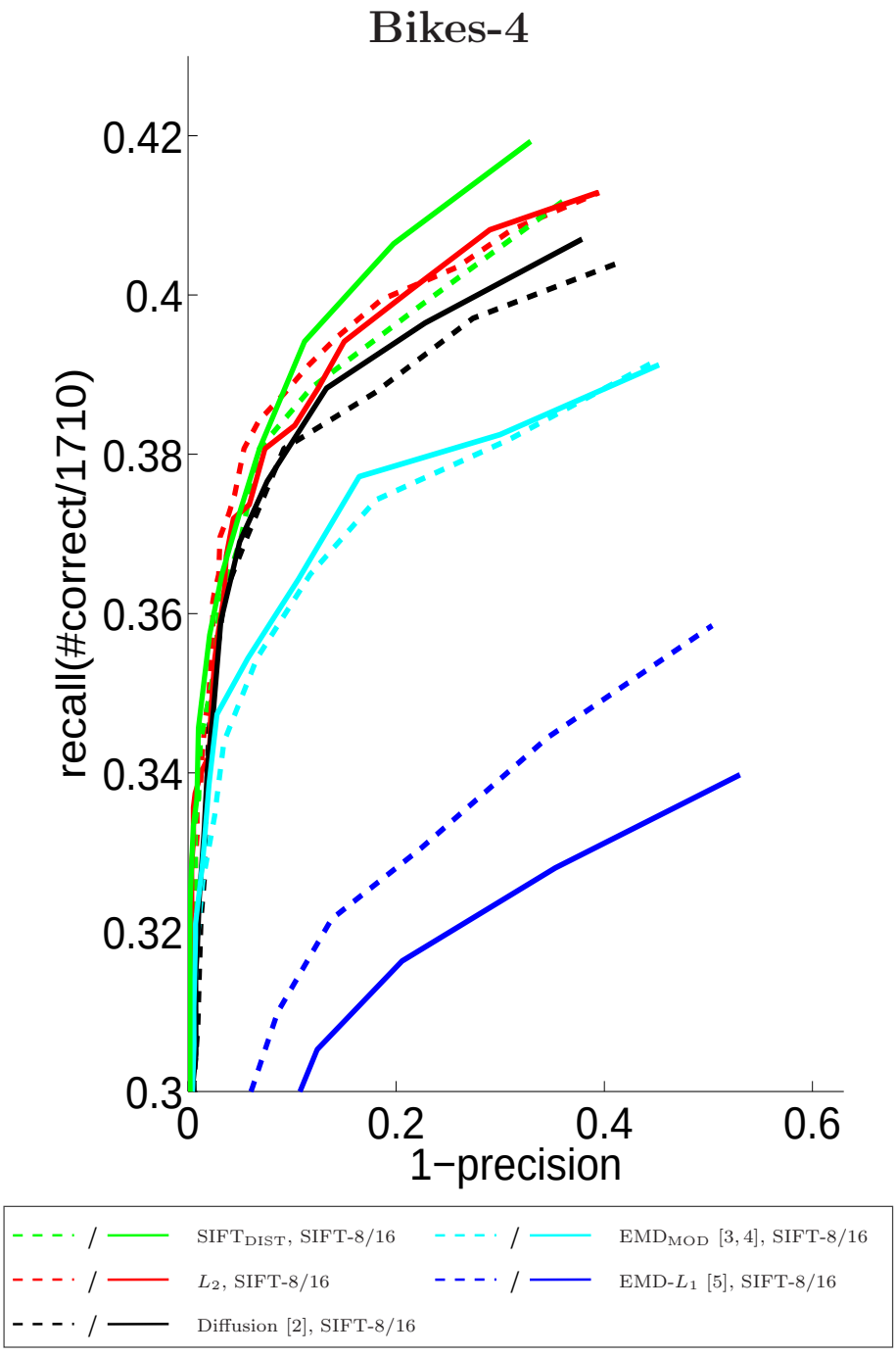


Fig. 23. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

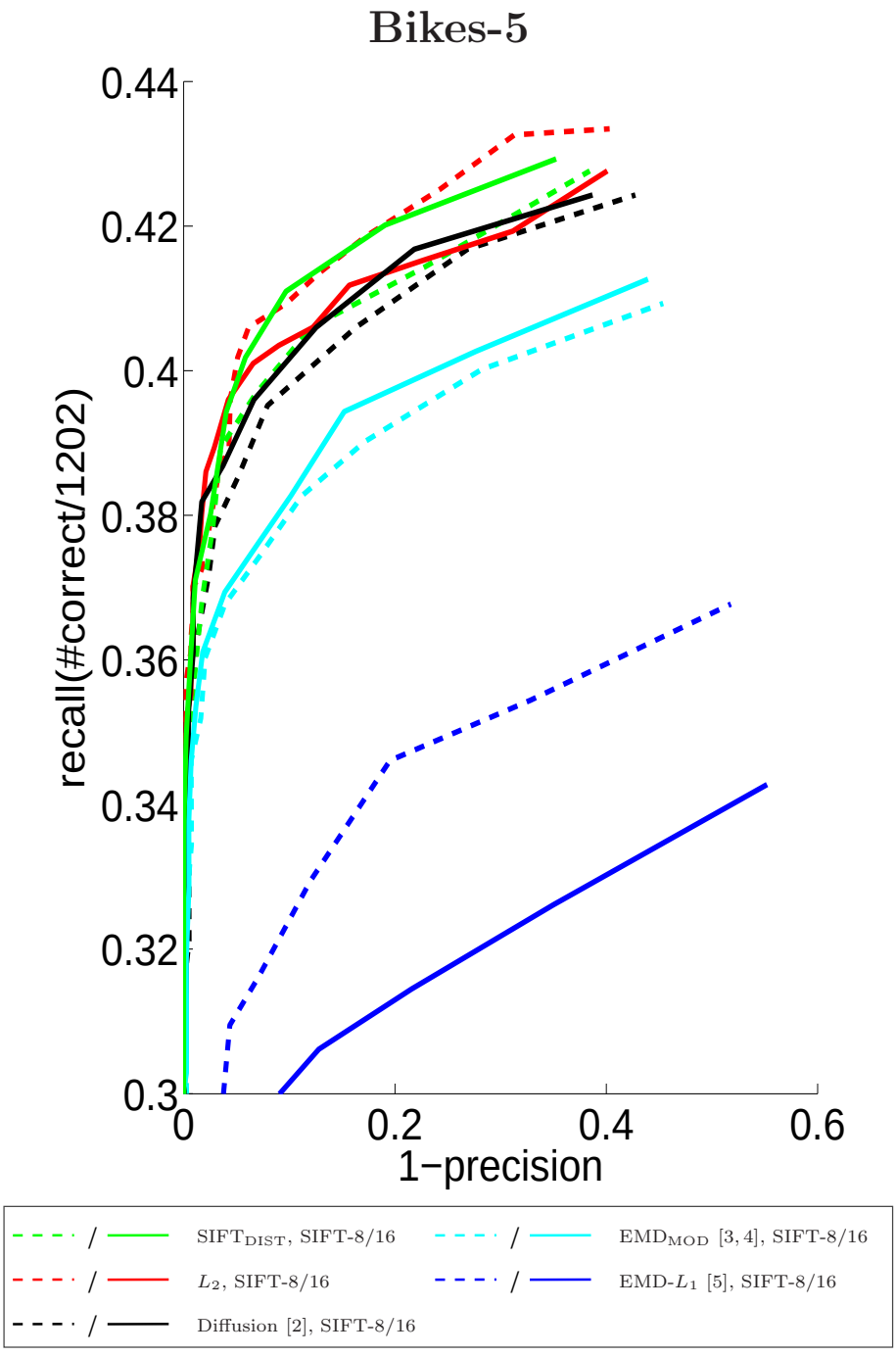


Fig. 24. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

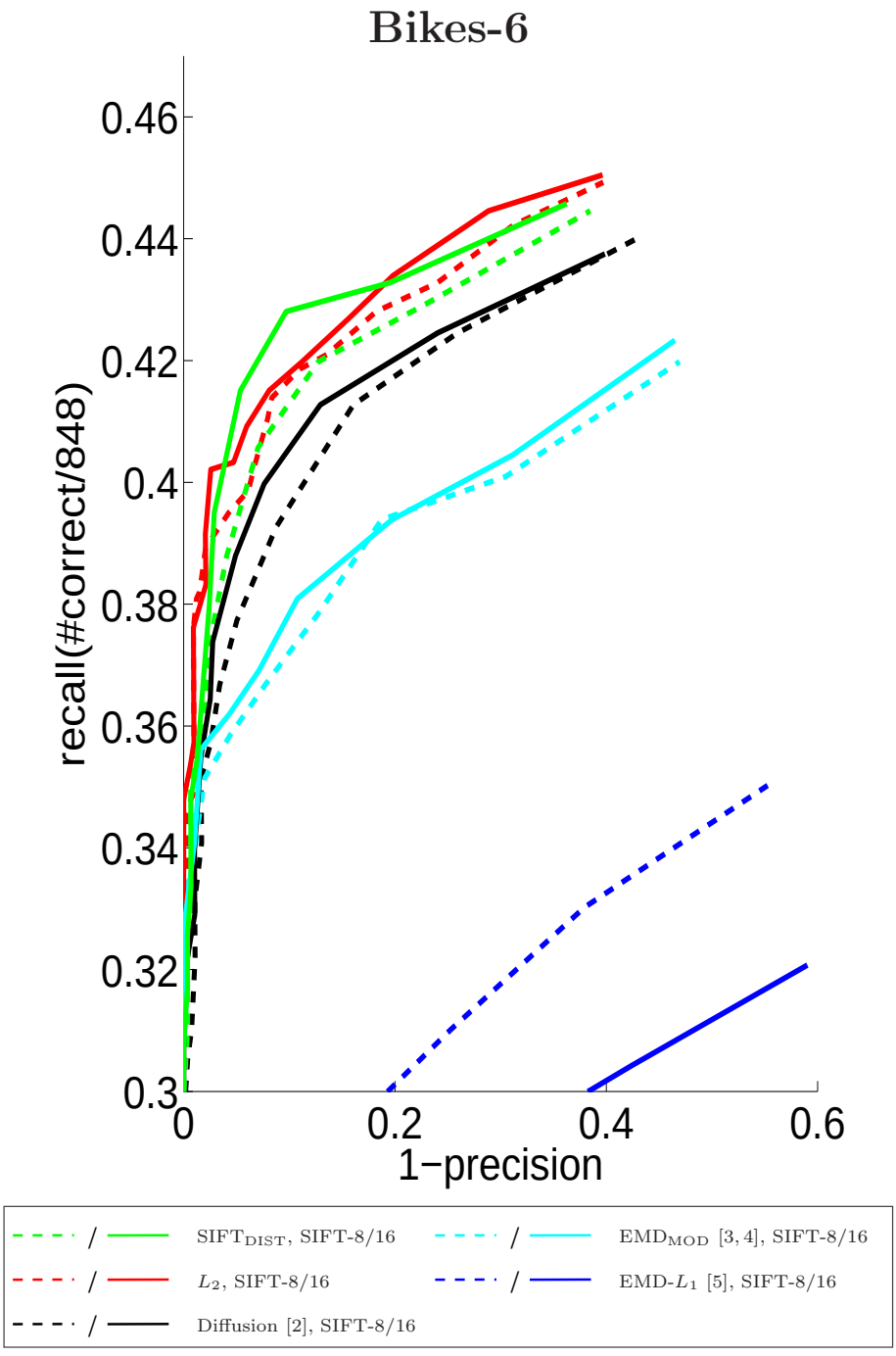


Fig. 25. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

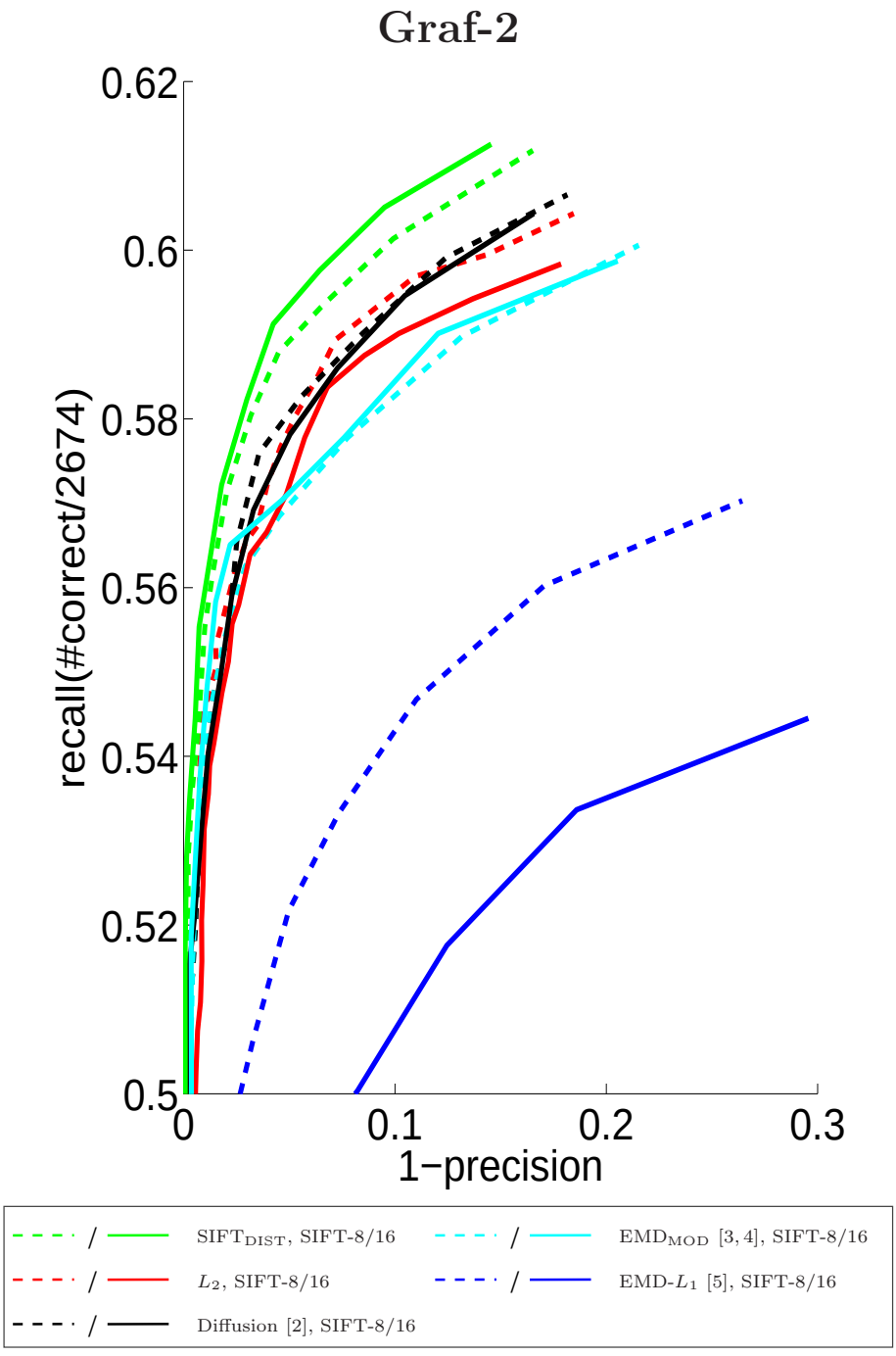


Fig. 26. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

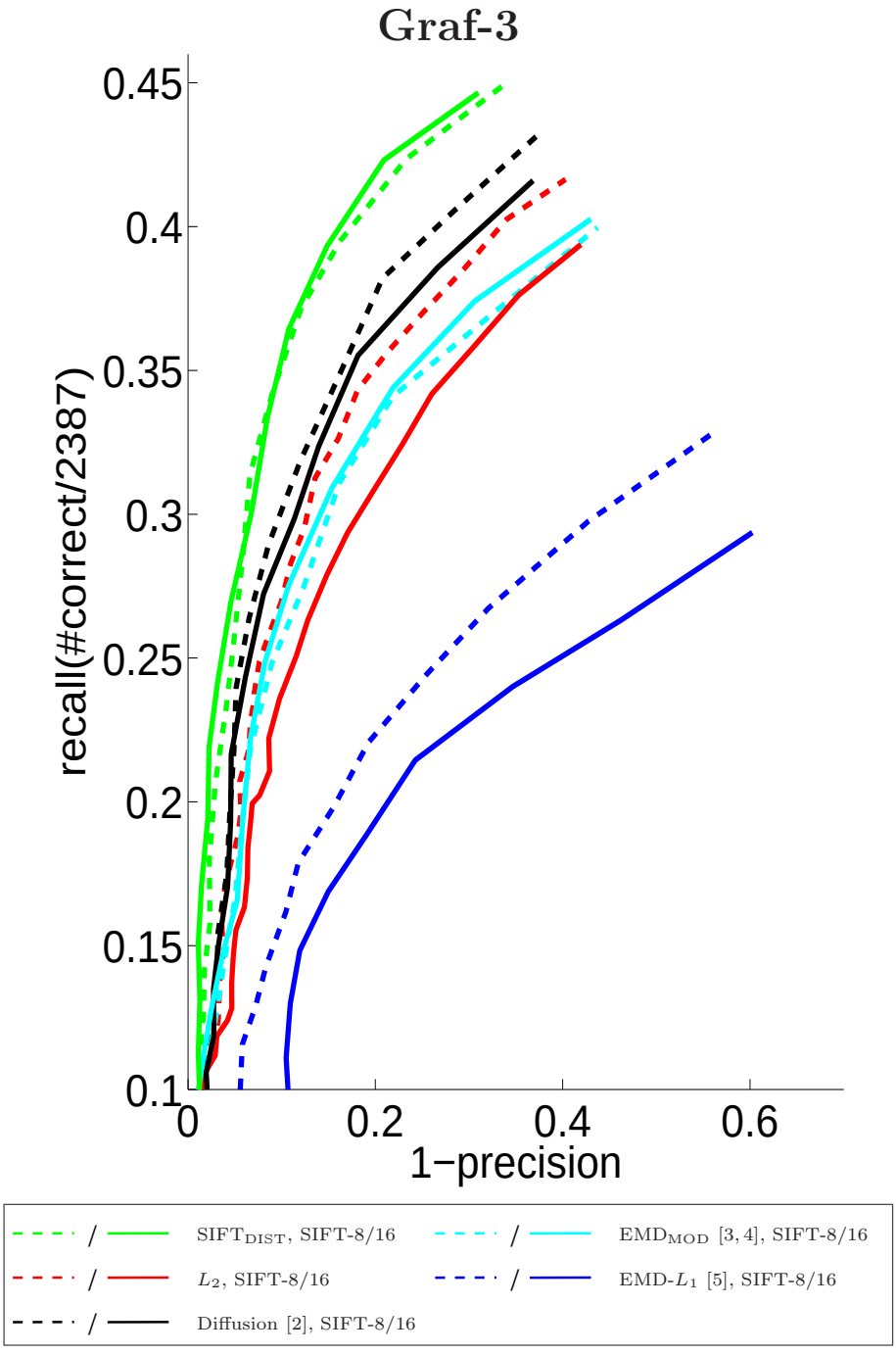


Fig. 27. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

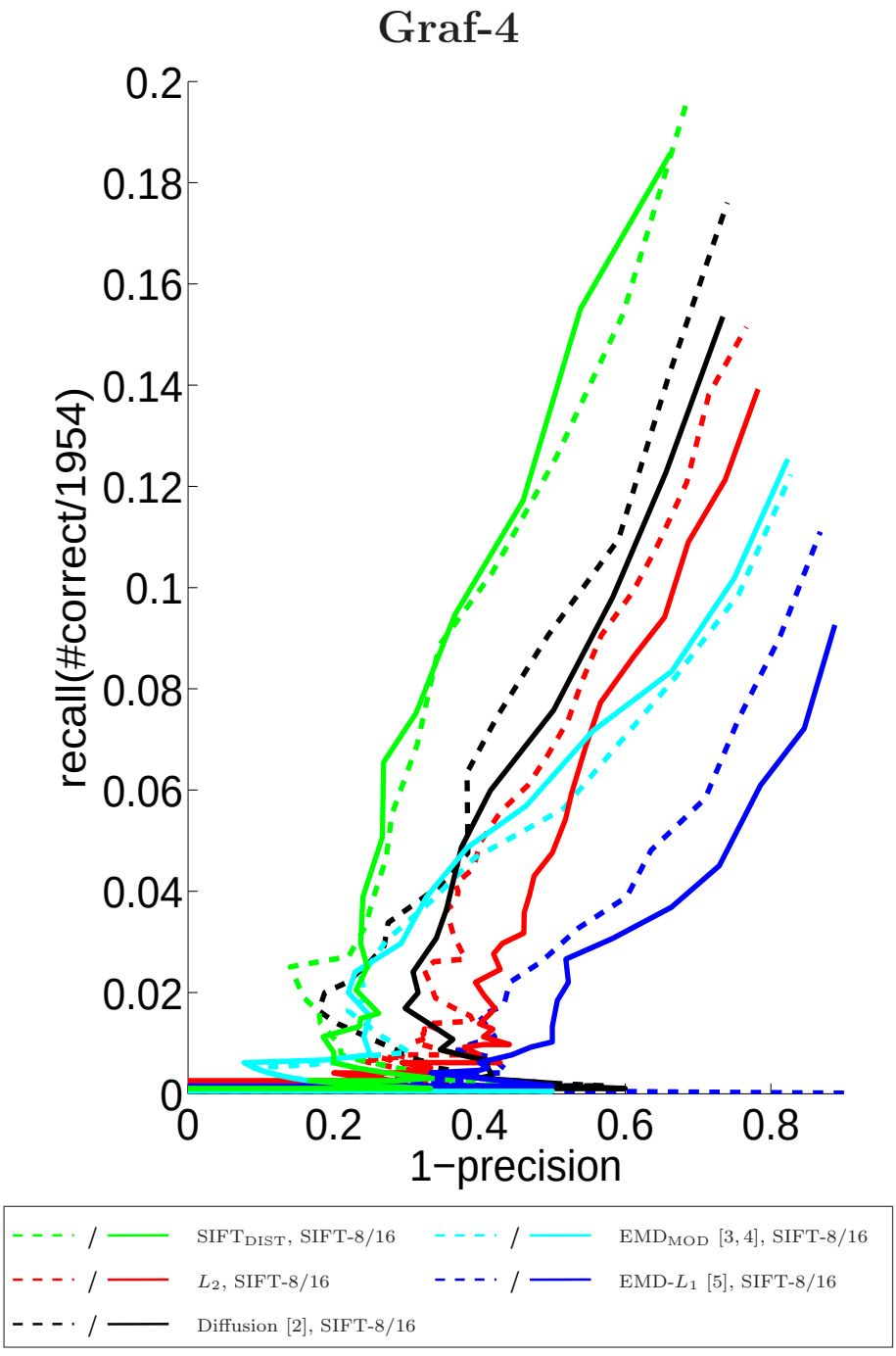


Fig. 28. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

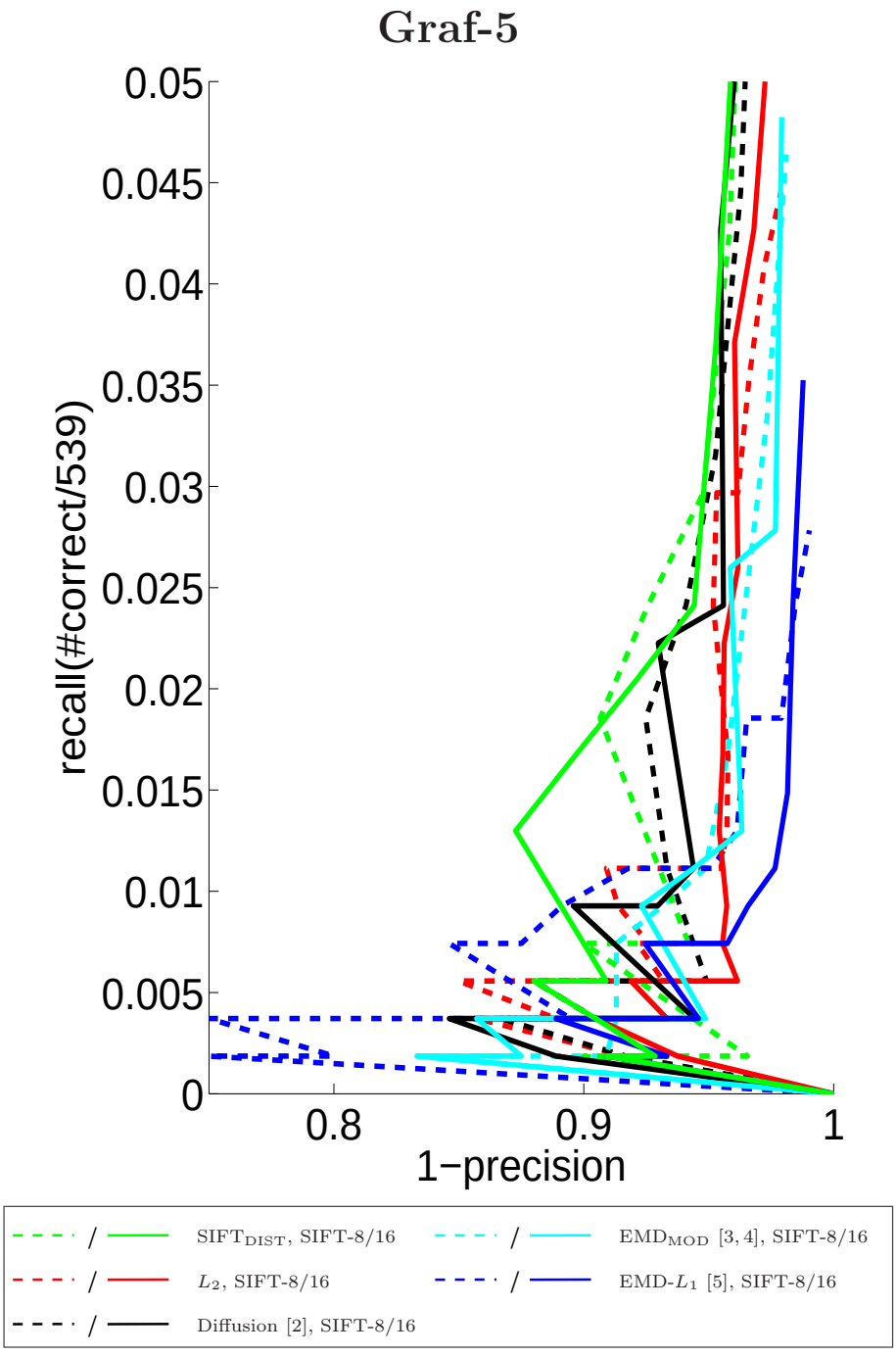


Fig. 29. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

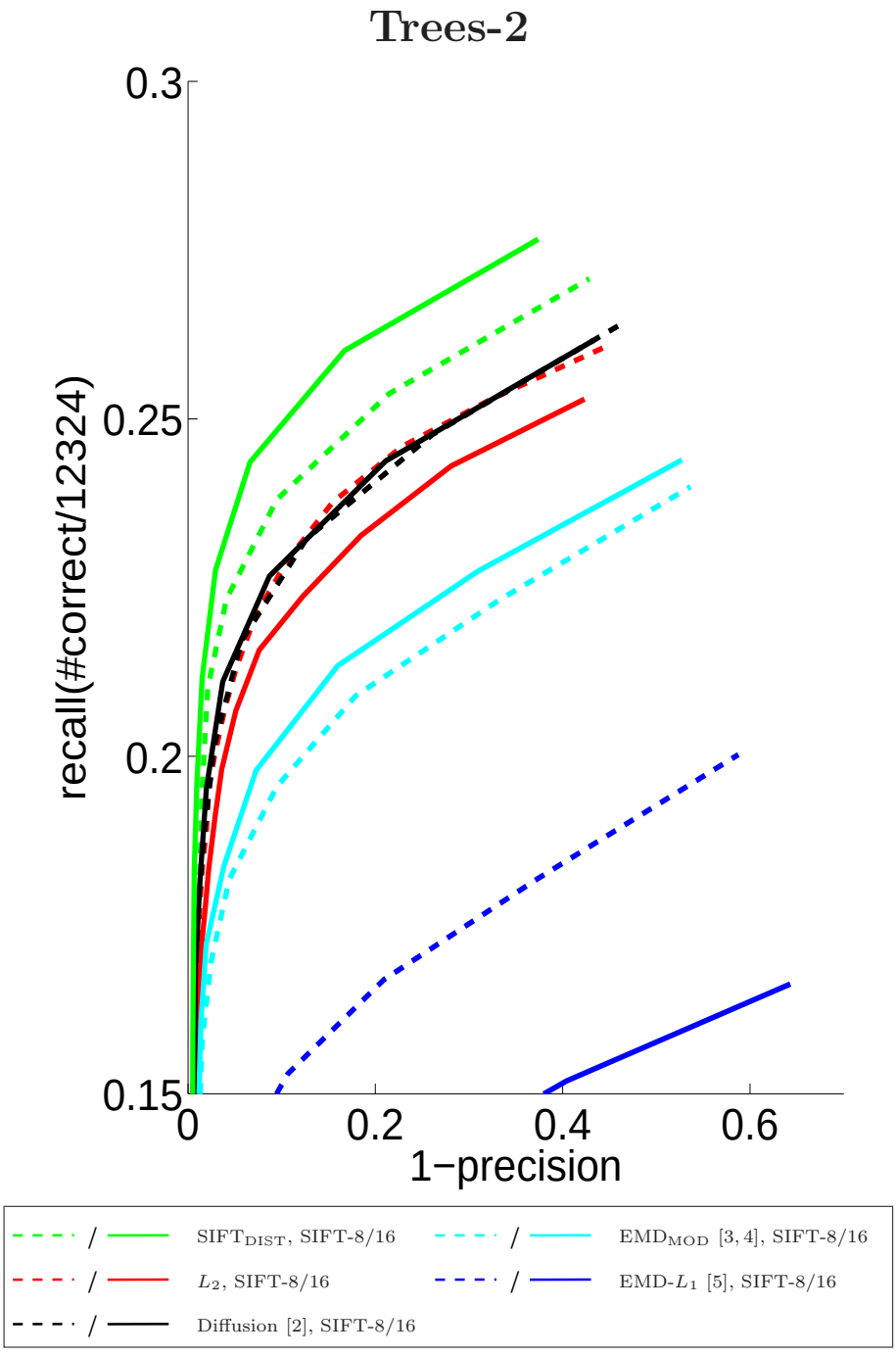


Fig. 30. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

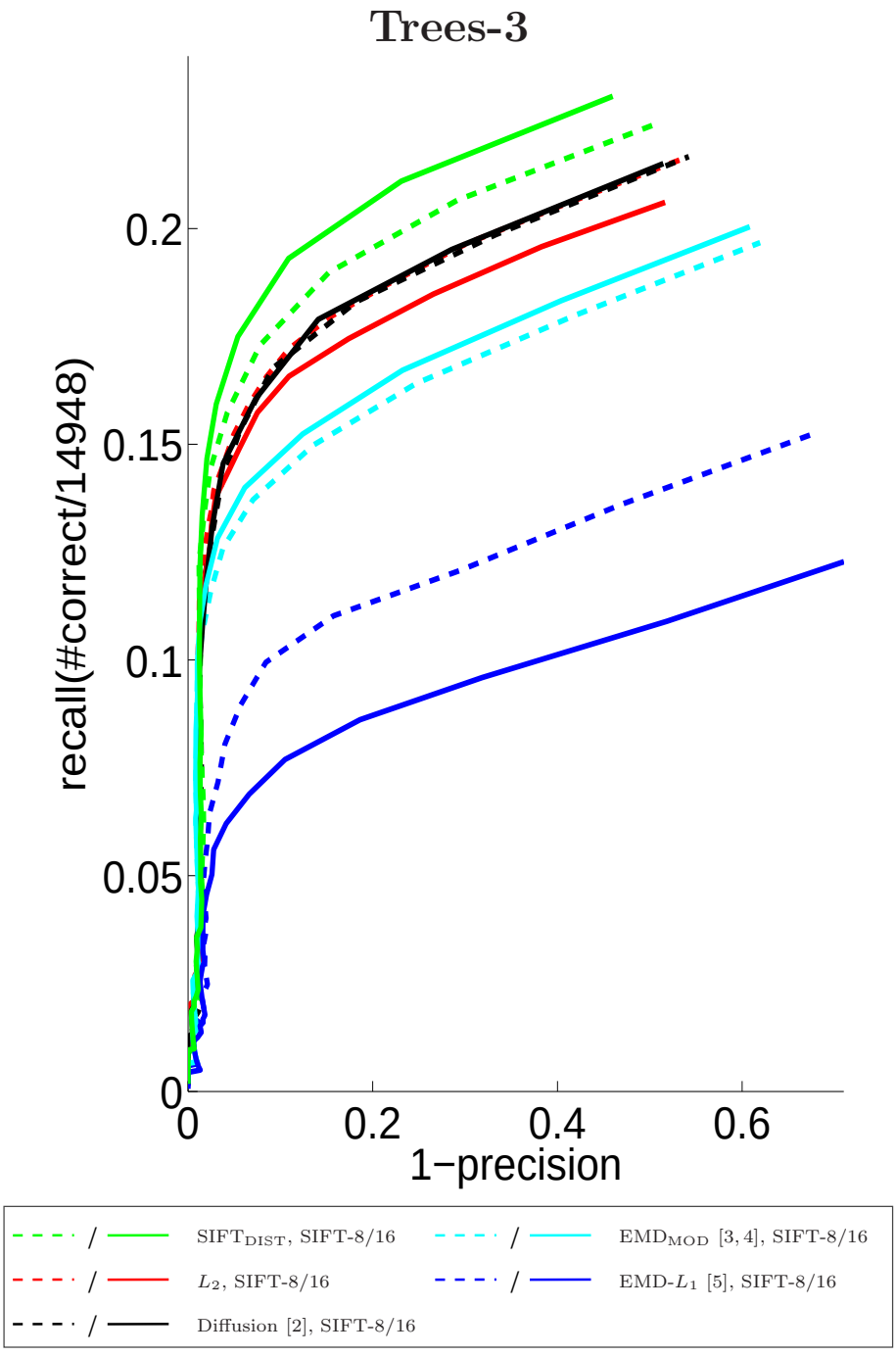


Fig. 31. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

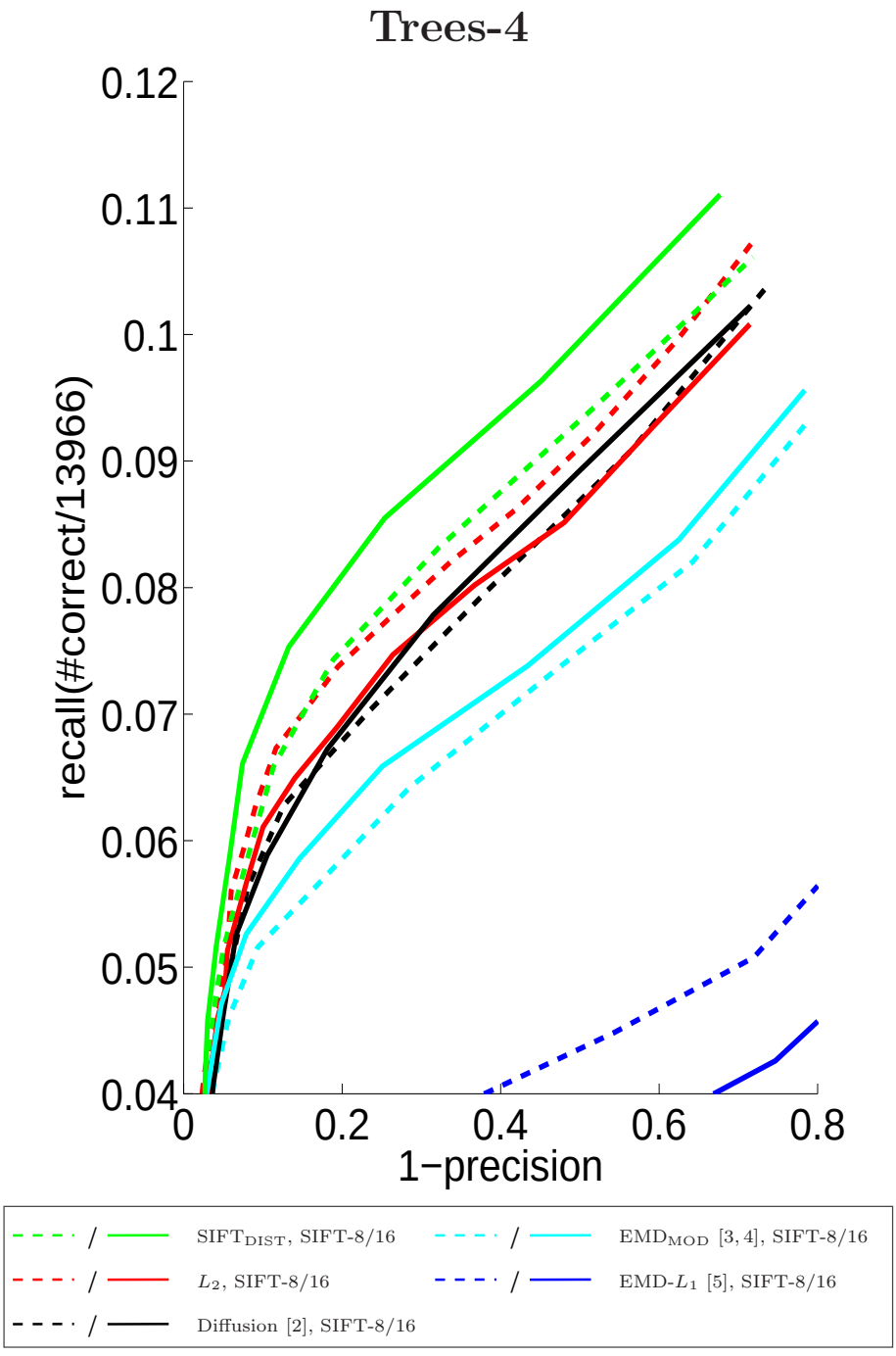


Fig. 32. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

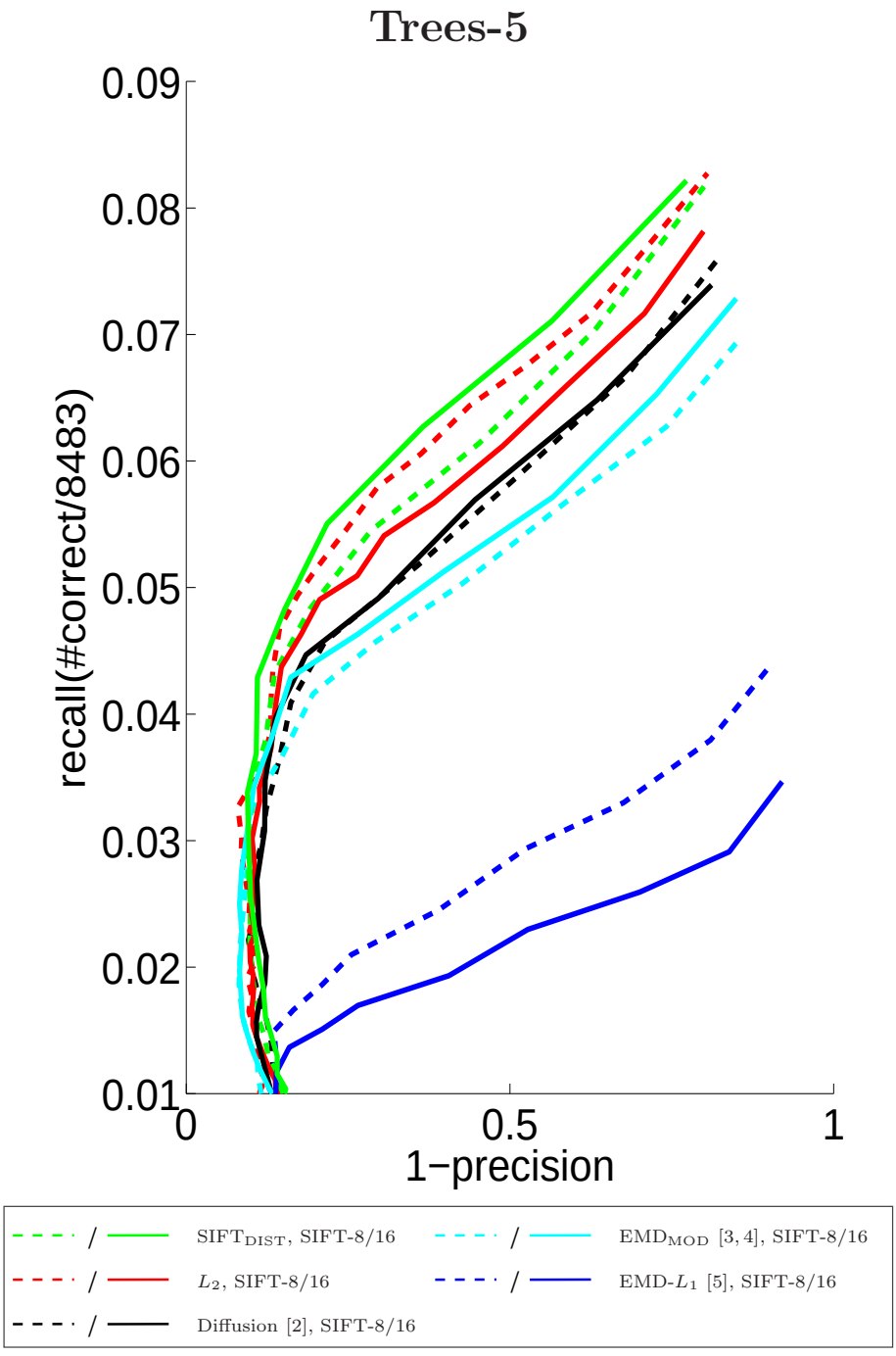


Fig. 33. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

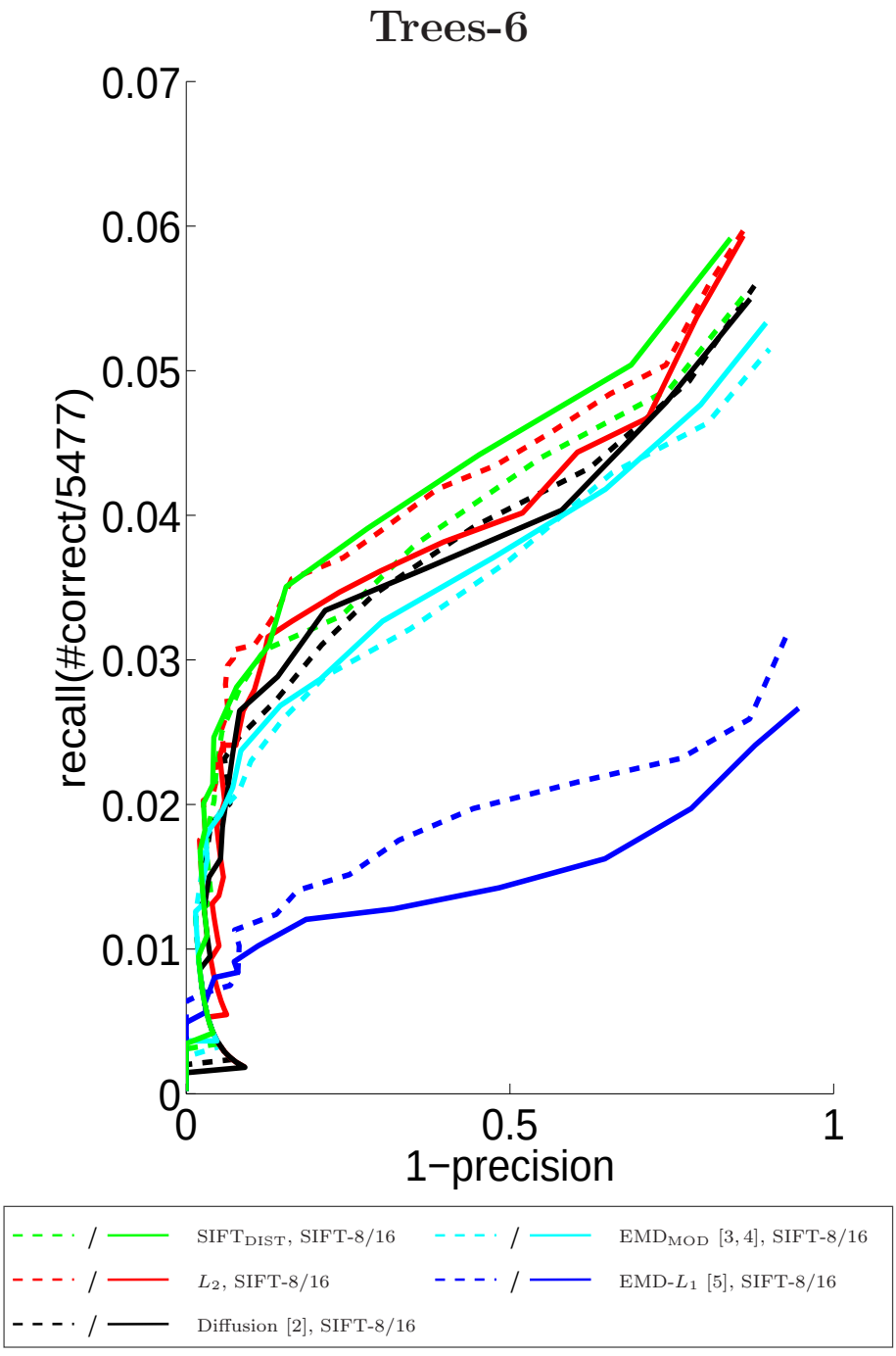


Fig. 34. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

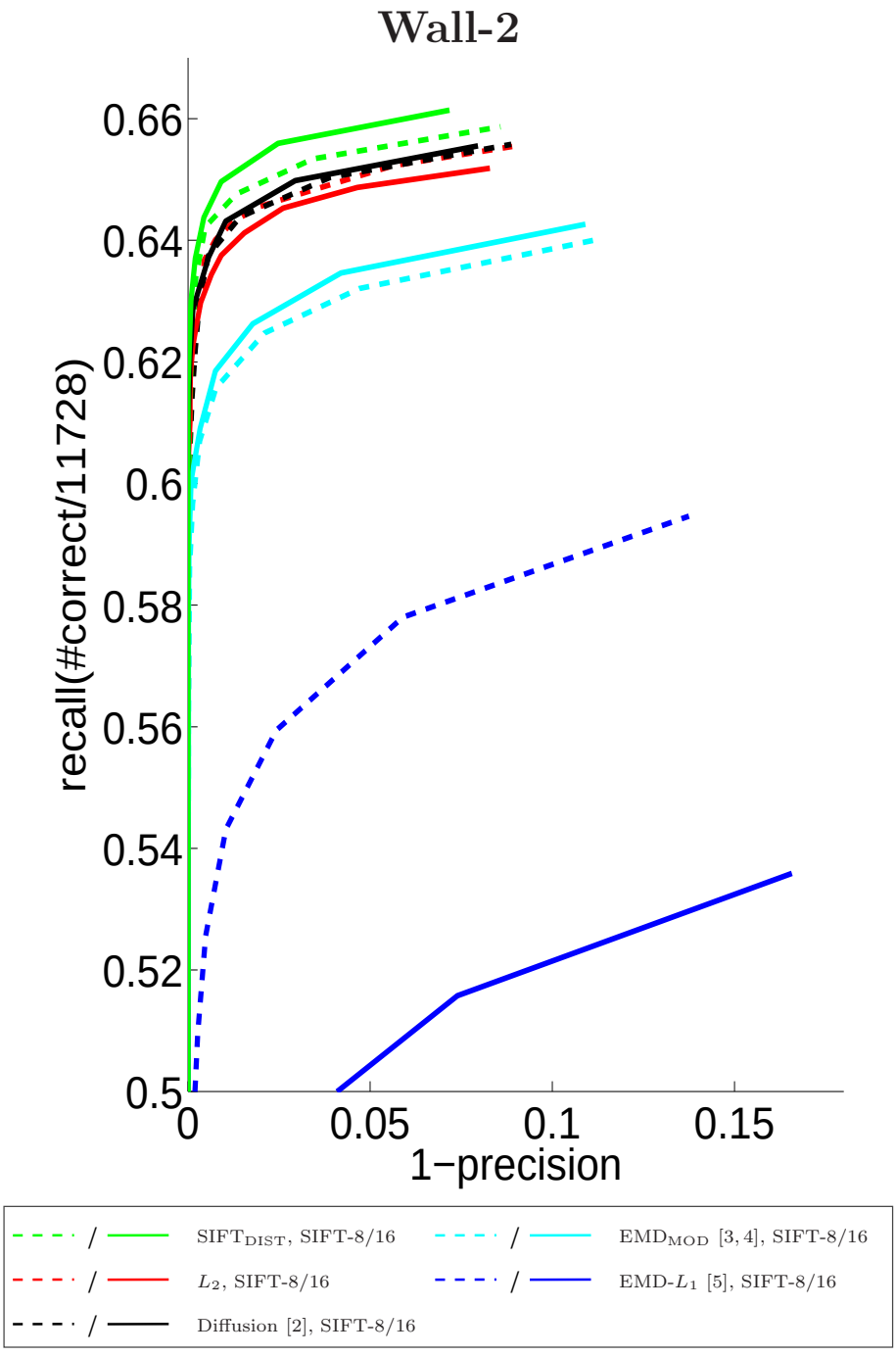


Fig. 35. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

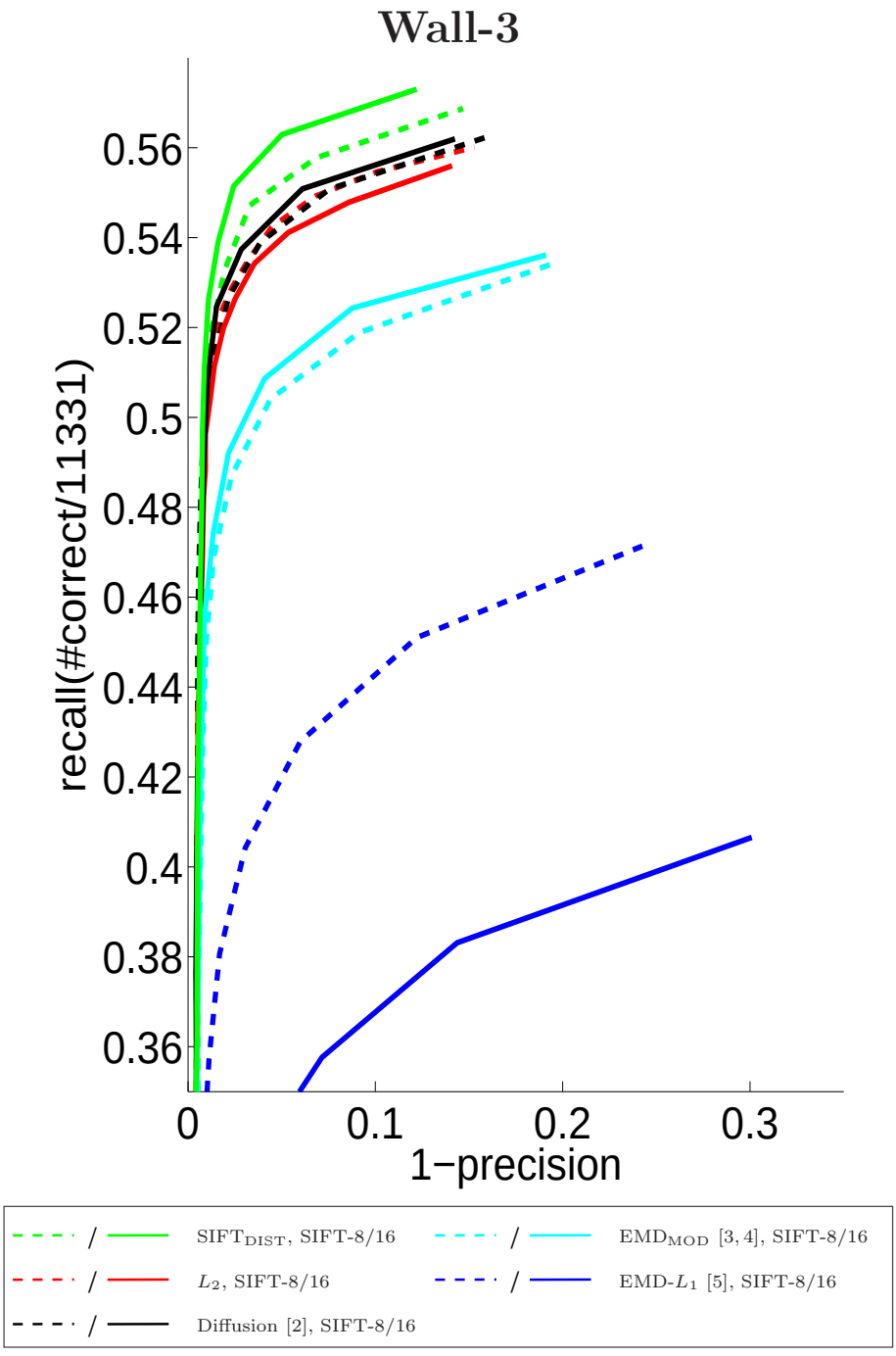


Fig. 36. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

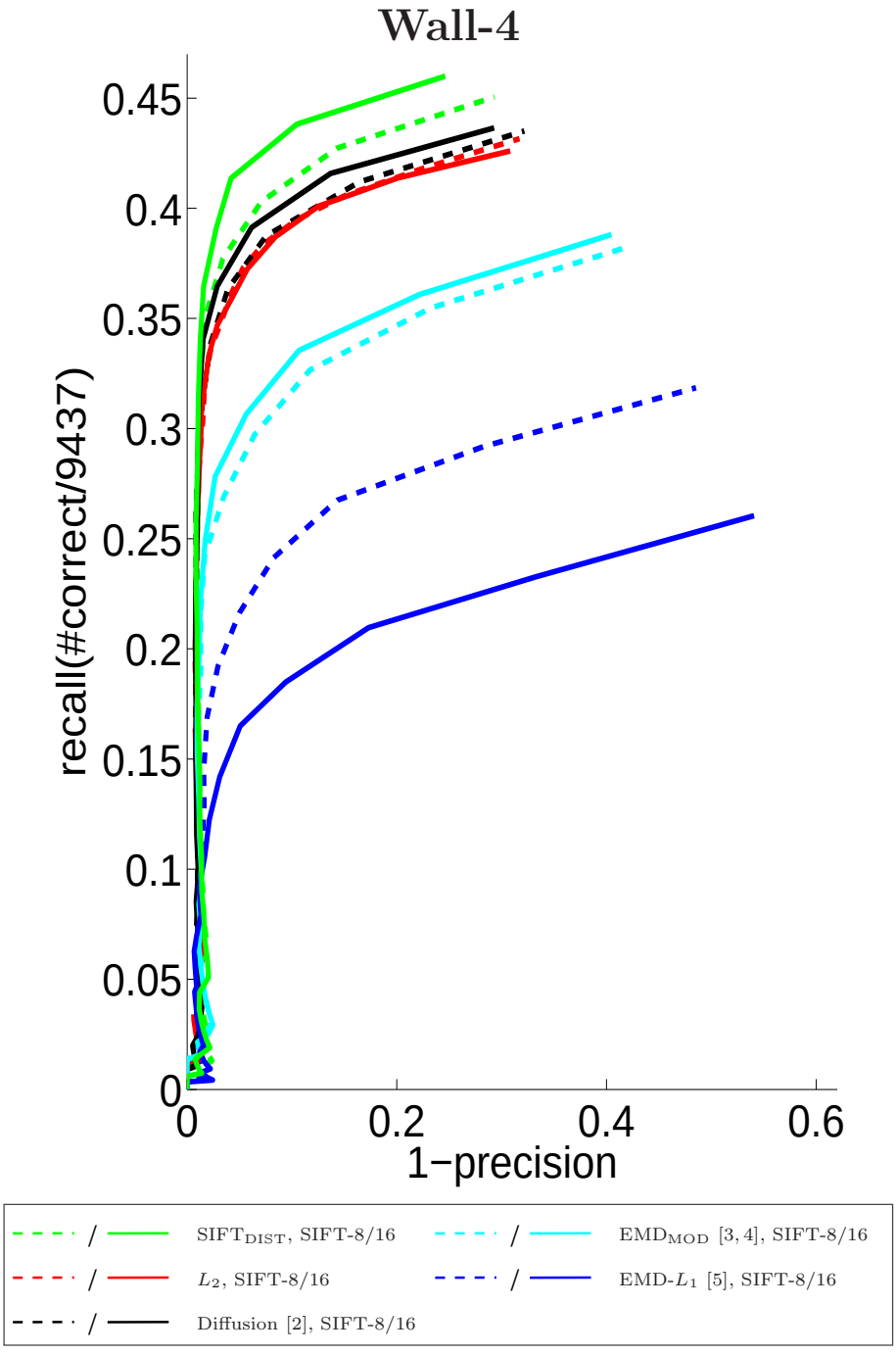


Fig. 37. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

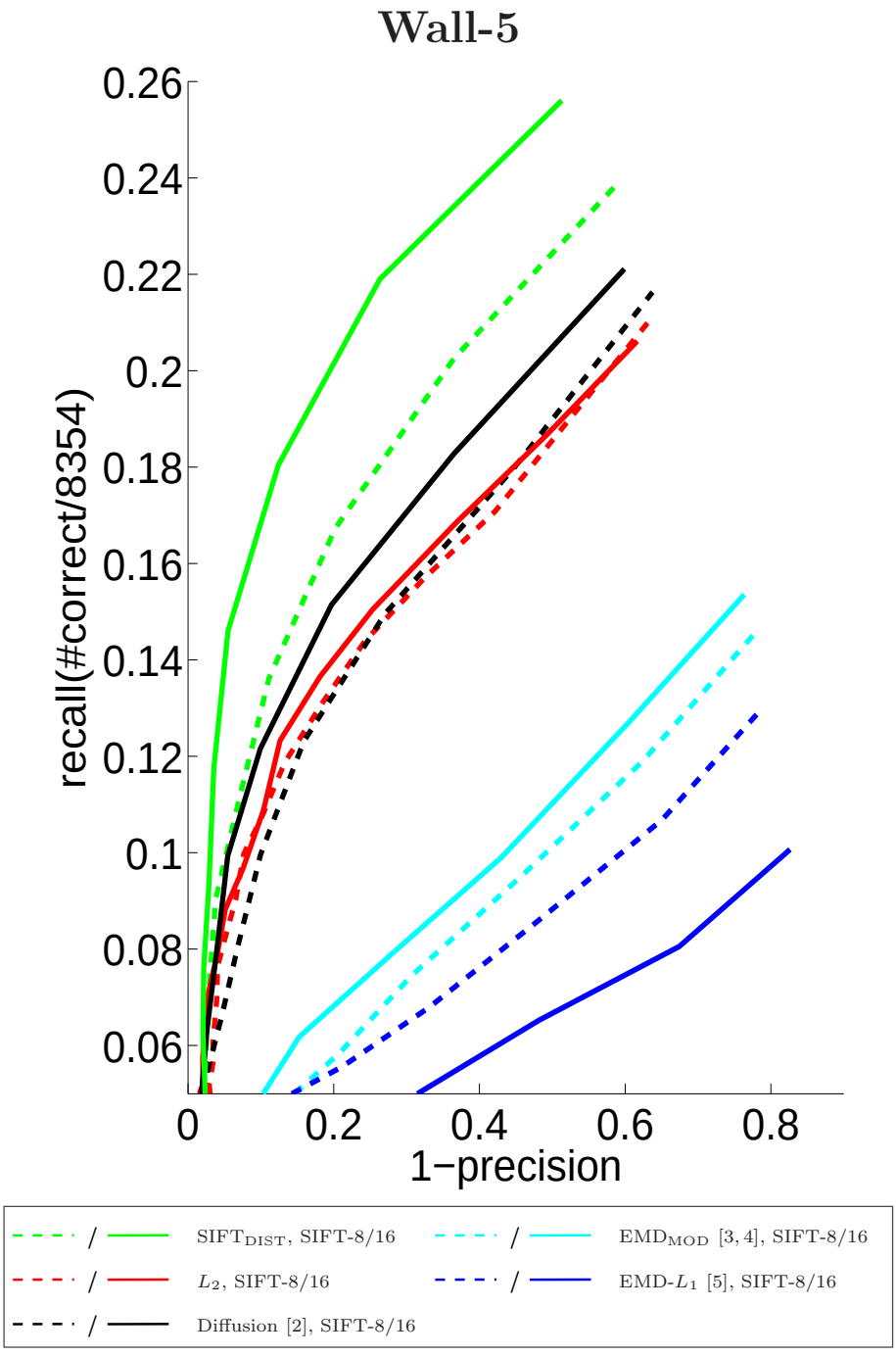


Fig. 38. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

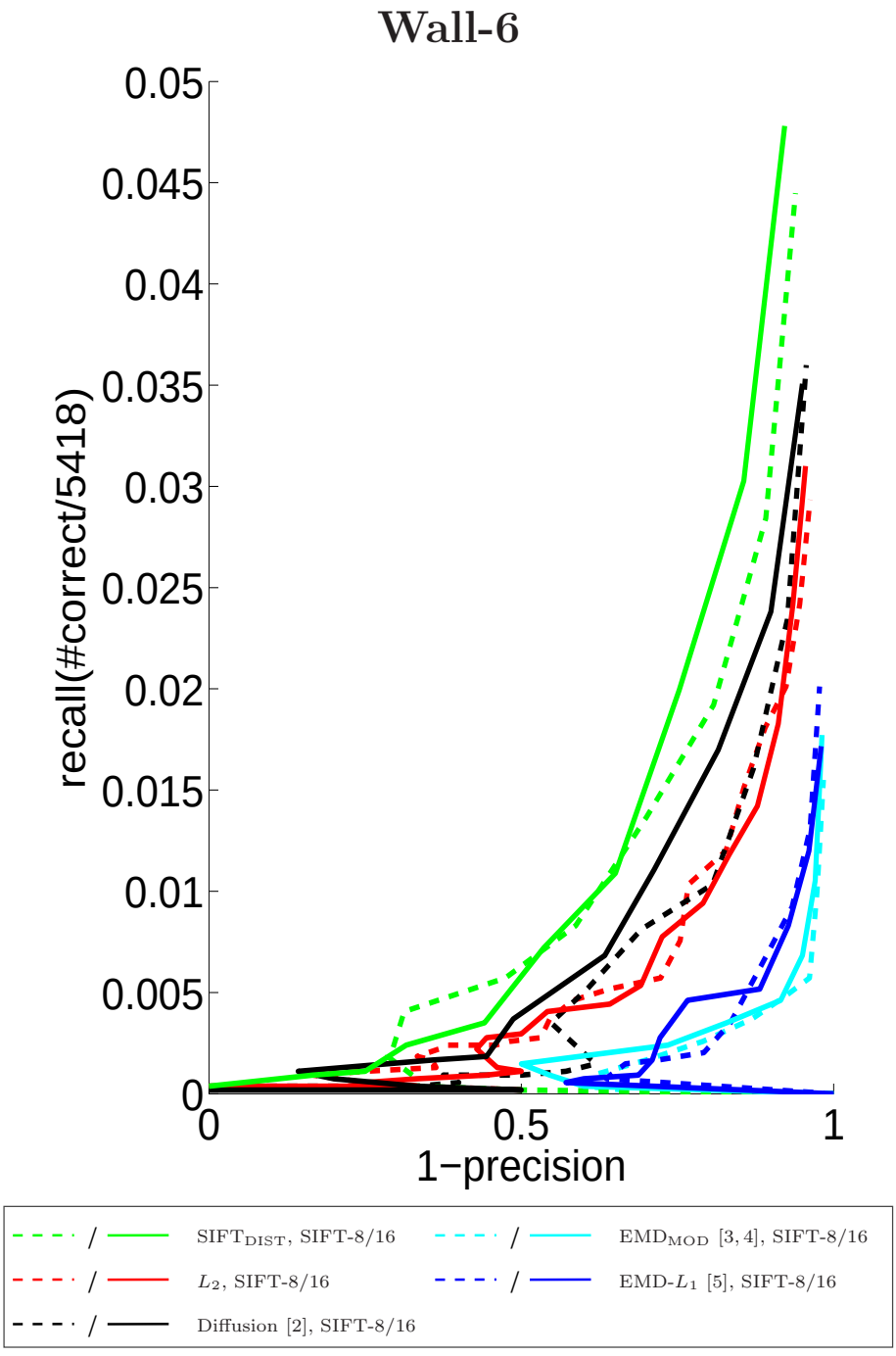


Fig. 39. Results on the Mikolajczyk and Schmid dataset [6]. Should be viewed in color.

References

1. Pele, O., Werman, M.: A linear time histogram metric for improved sift matching. In: ECCV. (2008)
2. Ling, H., Okada, K.: Diffusion distance for histogram comparison. In: CVPR. Volume 1. (2006) 246–253
3. Werman, M., Peleg, S., Melter, R., Kong, T.: Bipartite graph matching for points on a line or a circle. Journal of Algorithms **7**(2) (1986) 277–284
4. `http://www.cs.huji.ac.il/~ofirpele/publications/ECCV2008.pdf`
5. Ling, H., Okada, K.: An Efficient Earth Mover's Distance Algorithm for Robust Histogram Comparison. IEEE Trans. Pattern Analysis and Machine Intelligence **29**(5) (2007) 840–853
6. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Trans. Pattern Analysis and Machine Intelligence **27**(10) (2005) 1615–1630

Chapter 2

Fast and Robust Earth Mover's Distances - Appendices

Fast and Robust Earth Mover's Distances Additional Results - Image Retrieval Examples

Ofir Pele
The Hebrew University of Jerusalem
ofirpele@cs.huji.ac.il

Michael Werman
The Hebrew University of Jerusalem
werman@cs.huji.ac.il

This document contains additional image retrieval results for the paper - "Fast and Robust Earth Mover's Distances" [1]. The results are for image retrieval as described in the paper. Each figure contain a query image on the left and nearest neighbors images which are ordered from left to right by their distance from the query image. Top row is for the best distance - $\widehat{EMD}$ with thresholded ground distances. Bottom row is for the second best distance - L_1 like distance. By allowing small deformations in the $\widehat{EMD}$ we obtain results that are visually similar to the query image. It is noteworthy that in some cases the returned images are not from the same class, but are still visually similar. For example, Fig. 2.

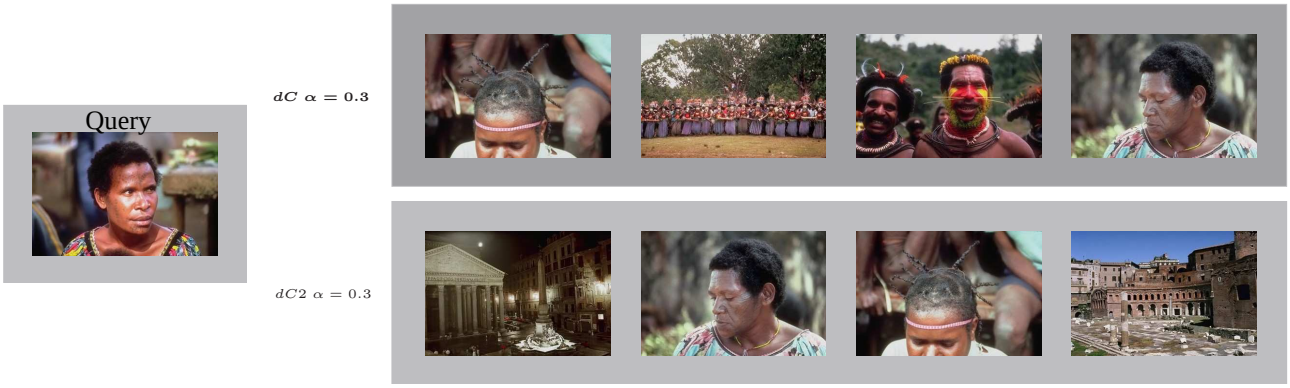


Figure 1.

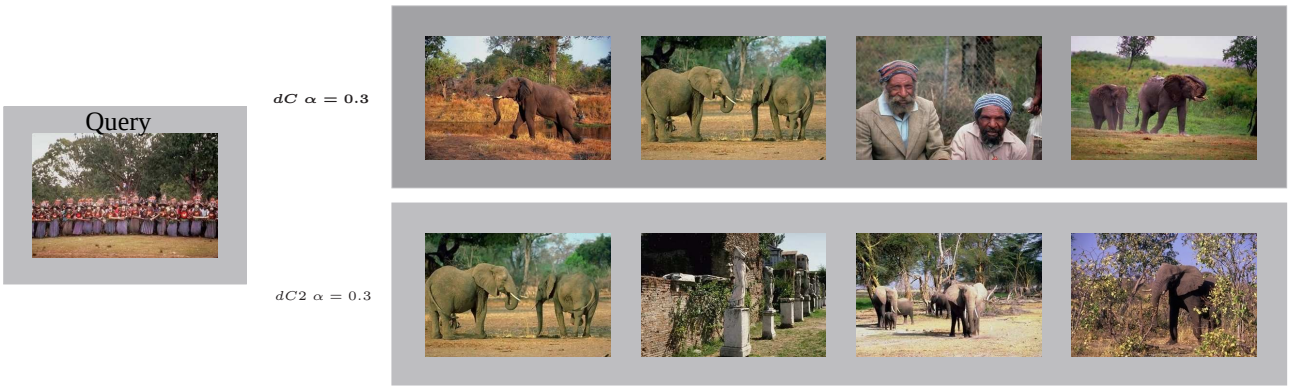


Figure 2.

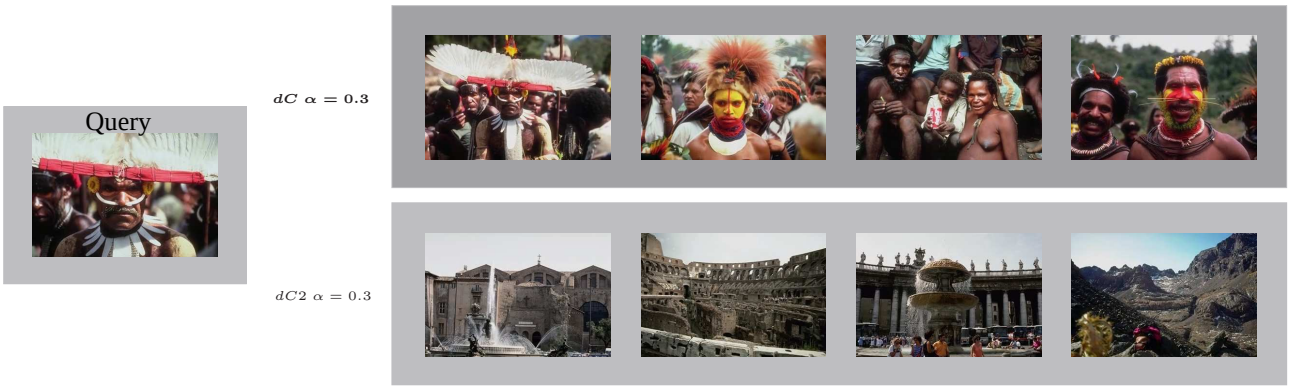


Figure 3.

Figure 4.

Figure 5.

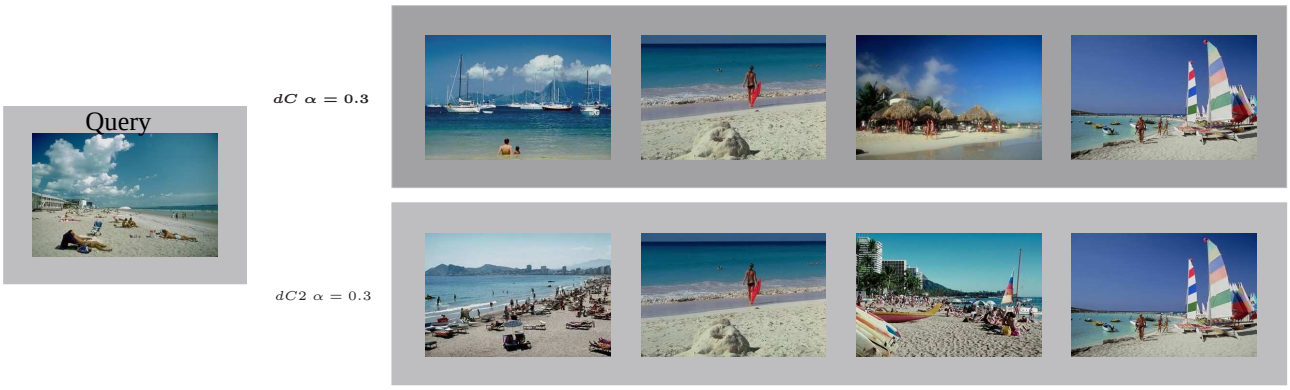


Figure 6.

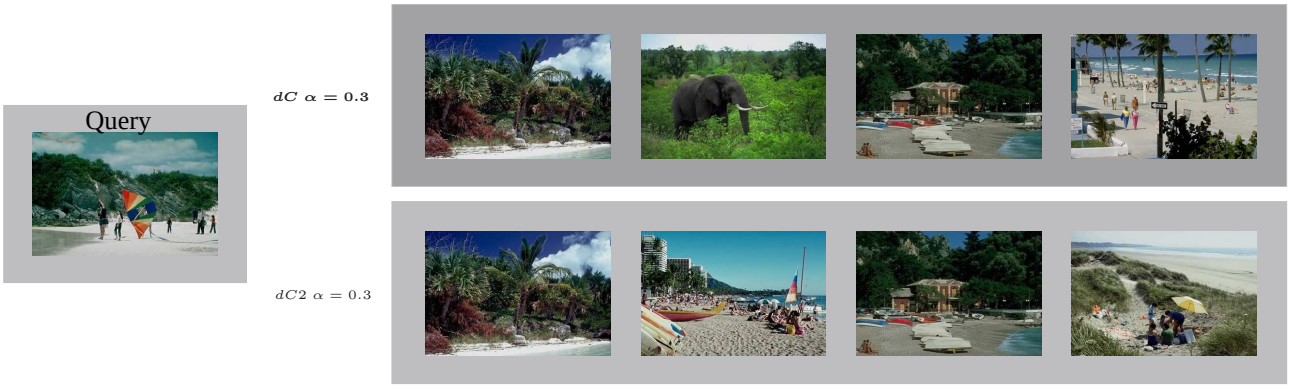


Figure 7.

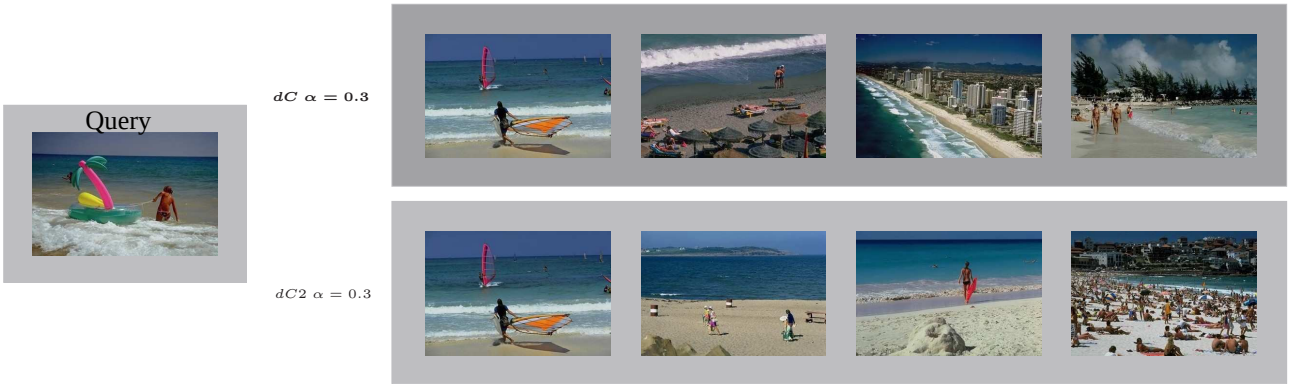


Figure 8.

Figure 9.

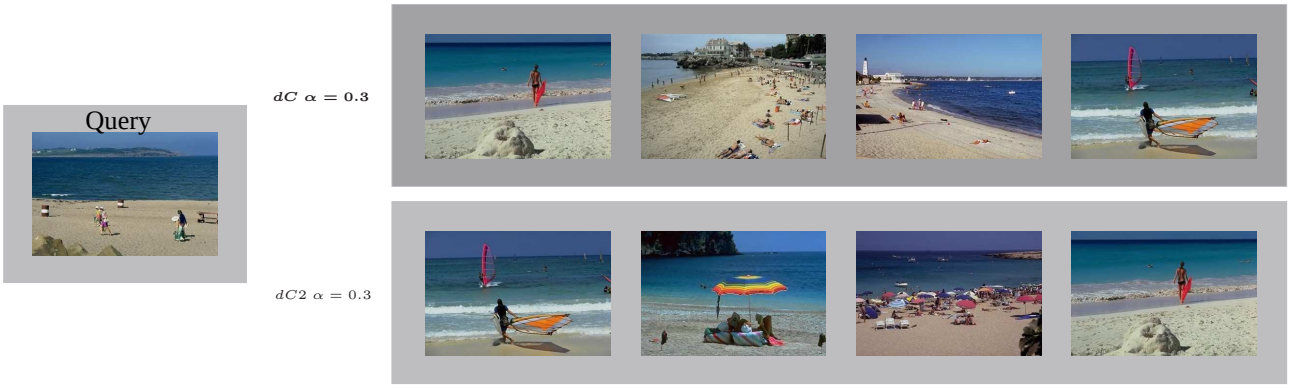


Figure 10.

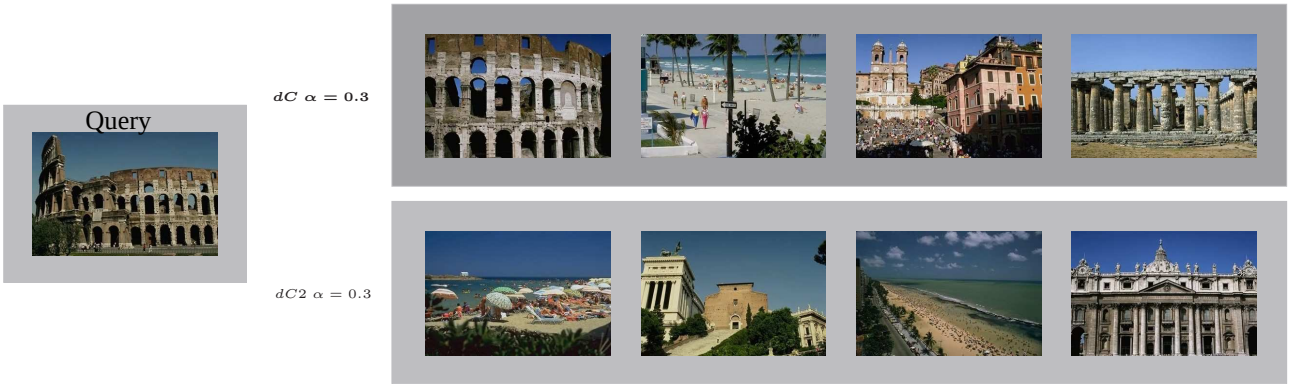


Figure 11.

Figure 12.

Figure 13.

Figure 14.

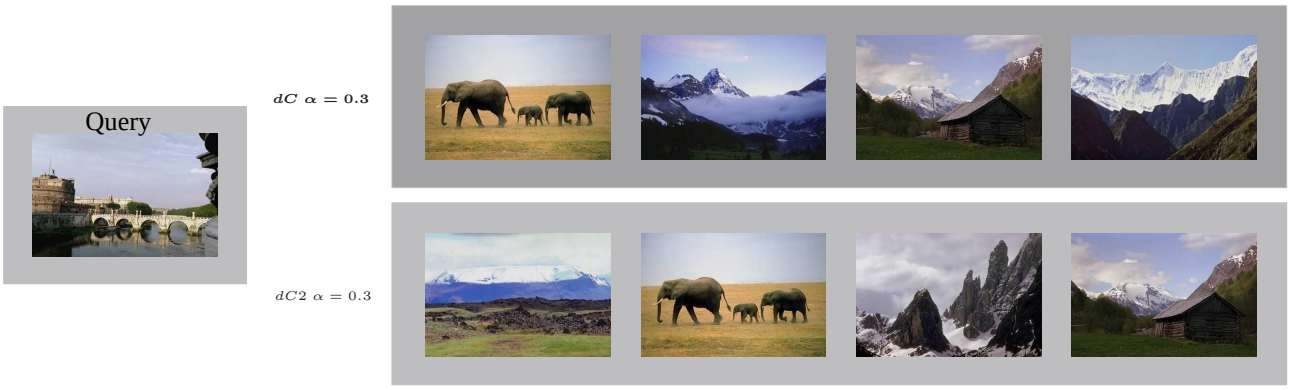


Figure 15.

Figure 16.

Figure 17.

Figure 18.

Figure 19.

Figure 20.

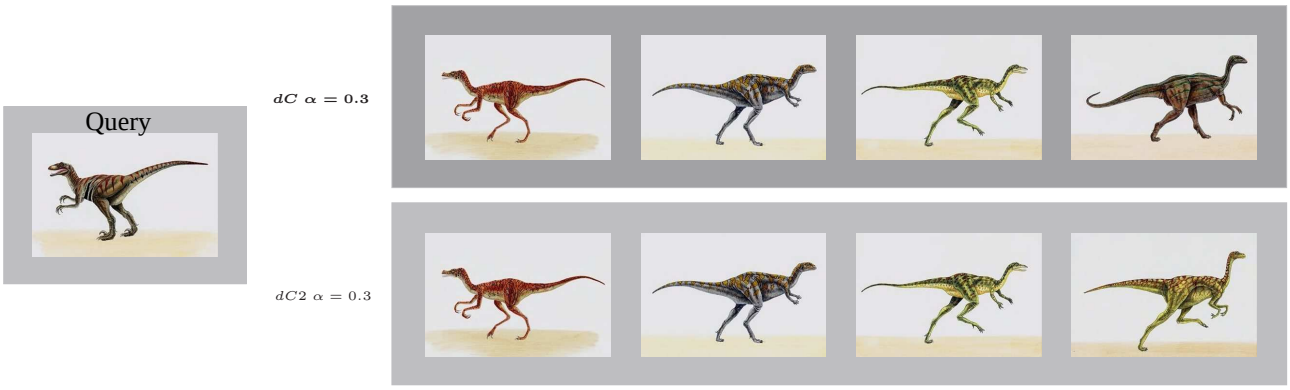


Figure 21.

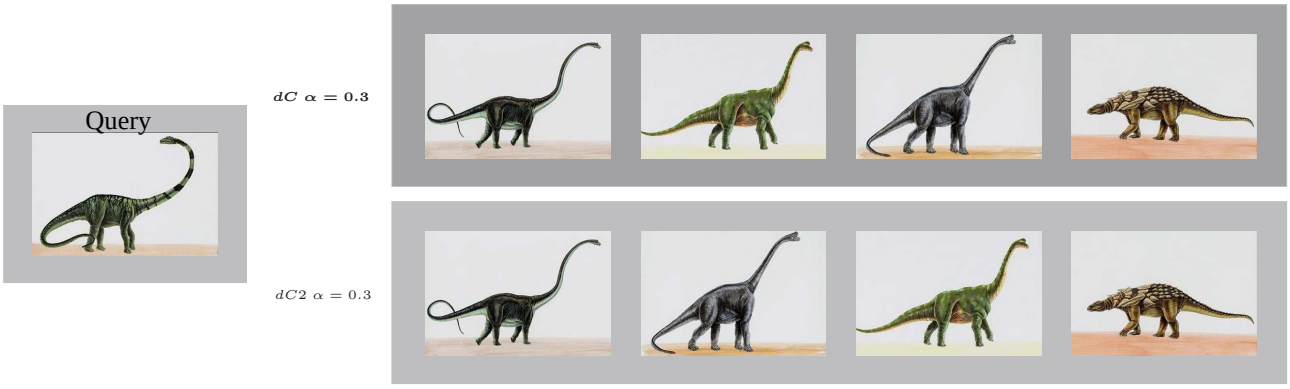


Figure 22.

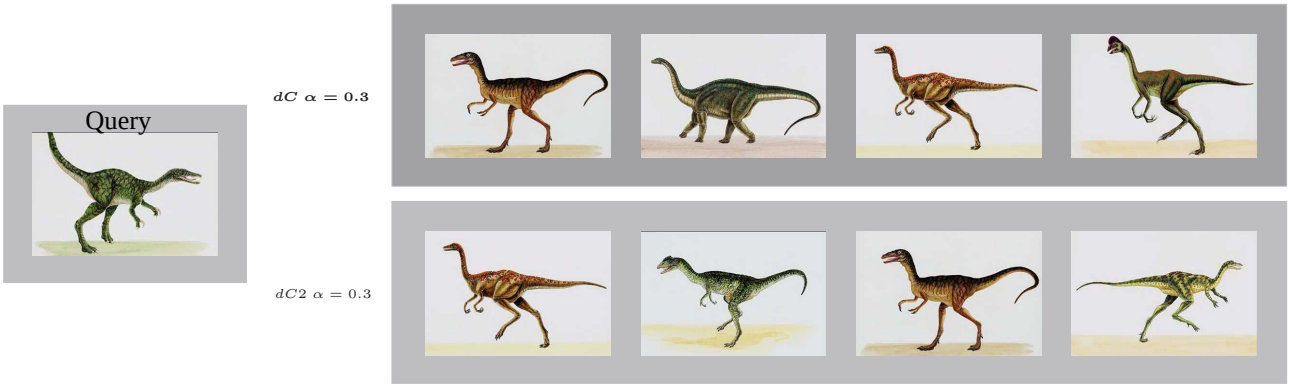


Figure 23.

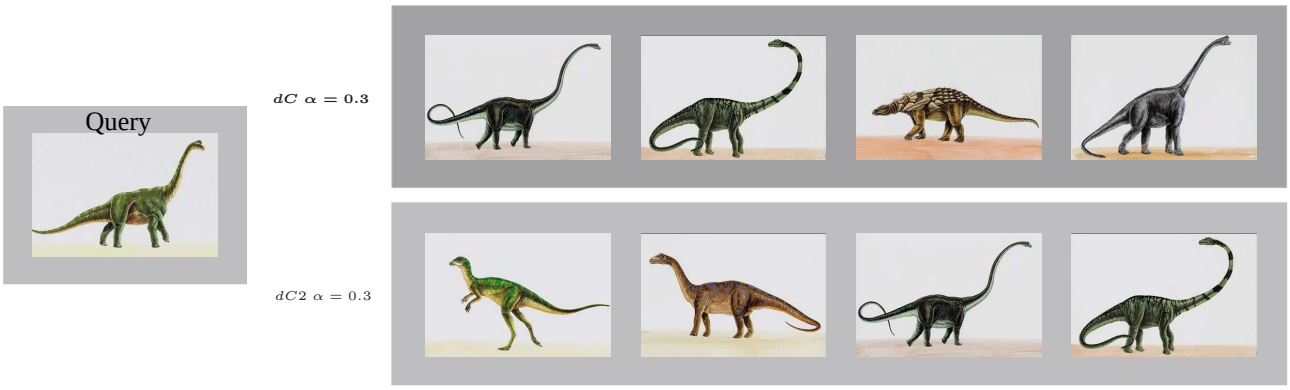


Figure 24.

Figure 25.

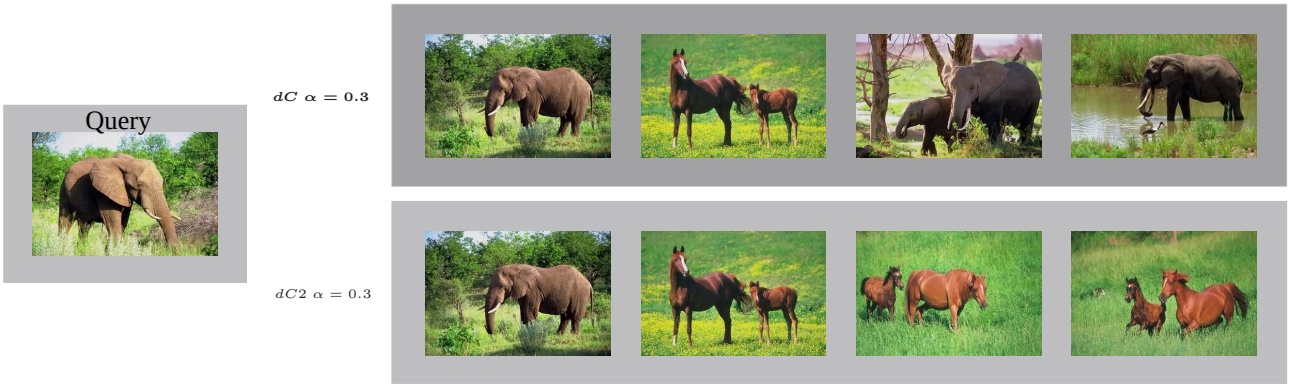


Figure 26.

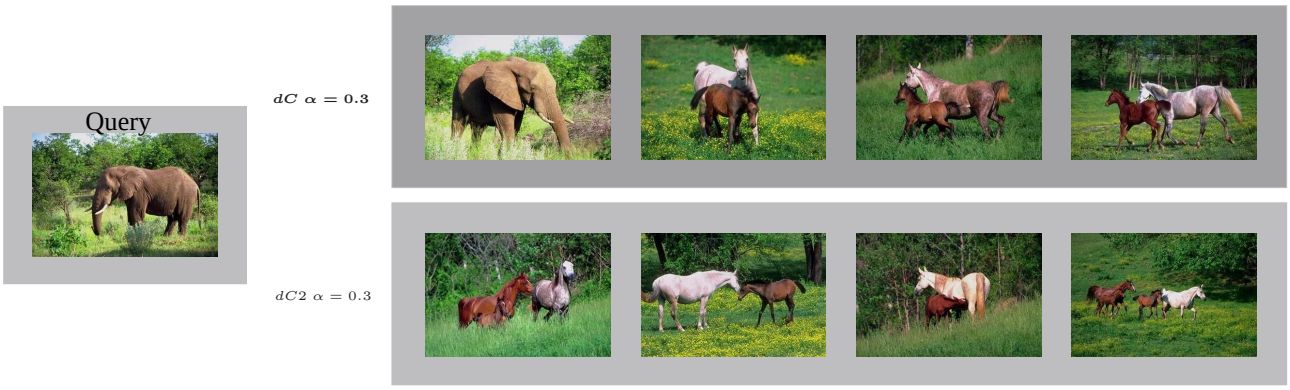


Figure 27.

Figure 28.

Figure 29.

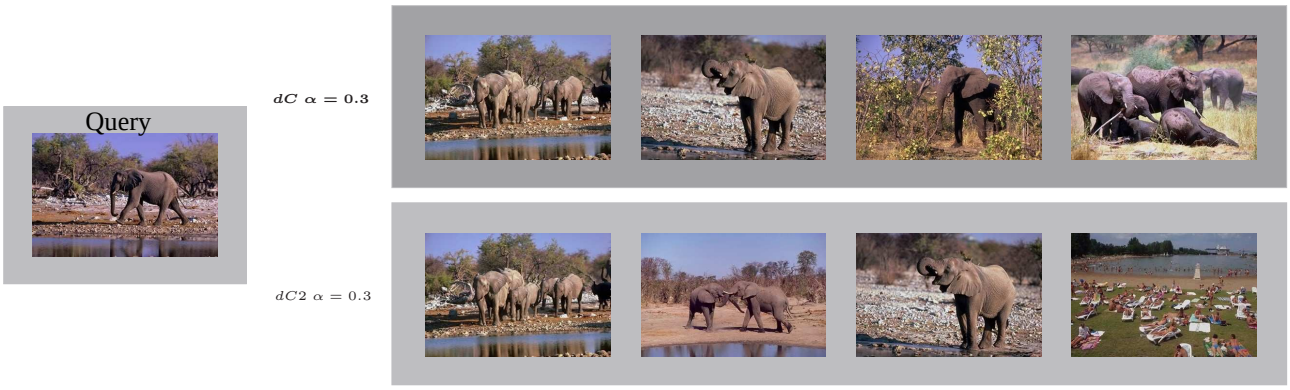


Figure 30.

Figure 31.

Figure 32.

Figure 33.

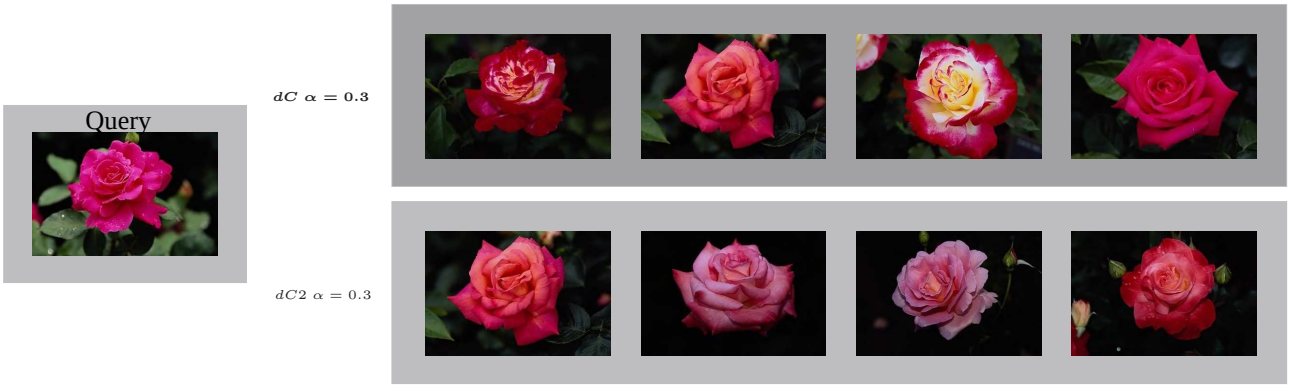


Figure 34.

Figure 35.

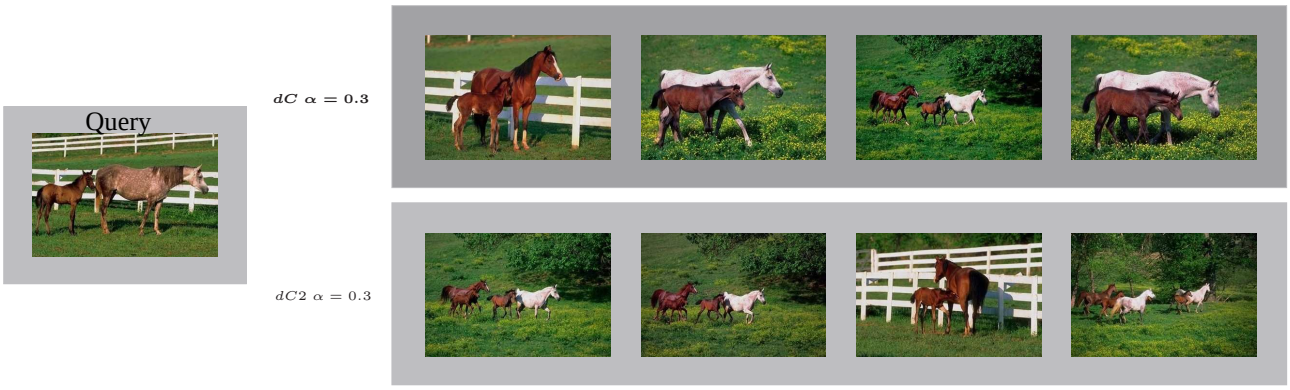


Figure 36.

Figure 37.

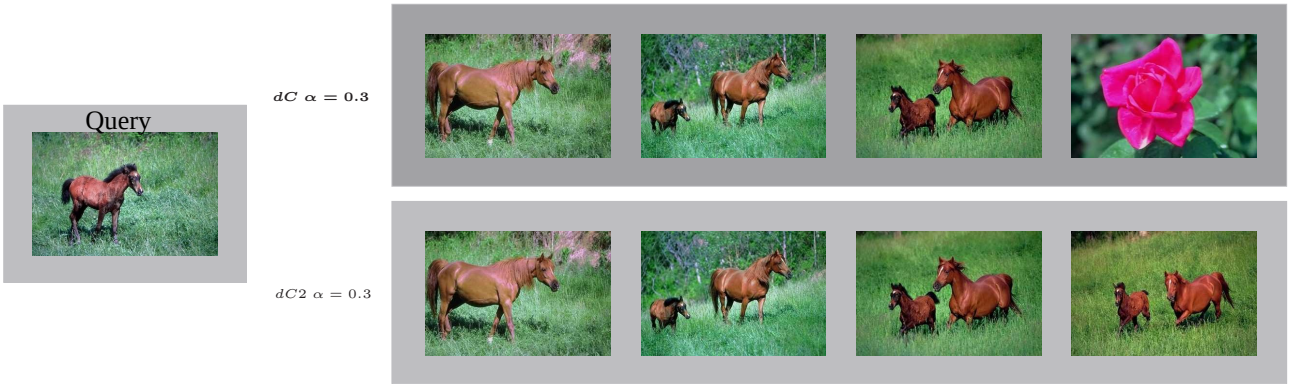


Figure 38.

Figure 39.

Figure 40.

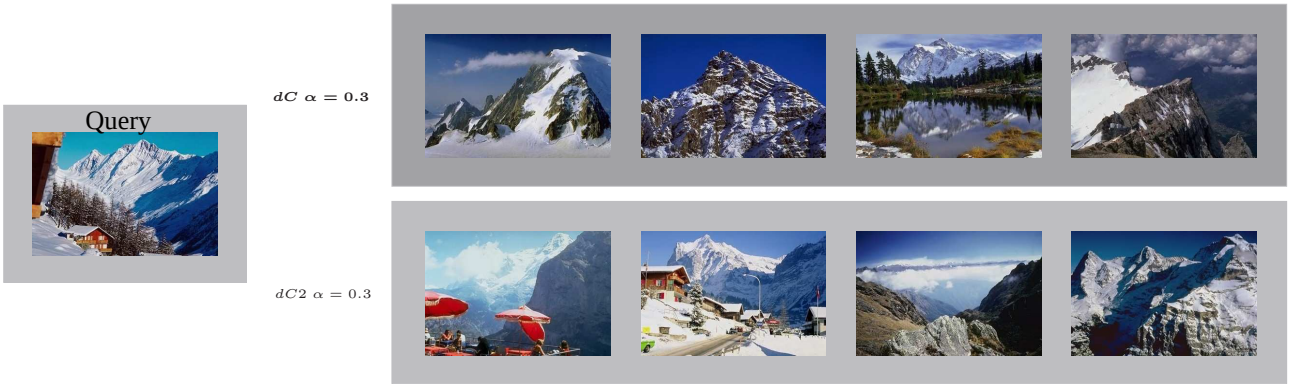


Figure 41.

Figure 42.

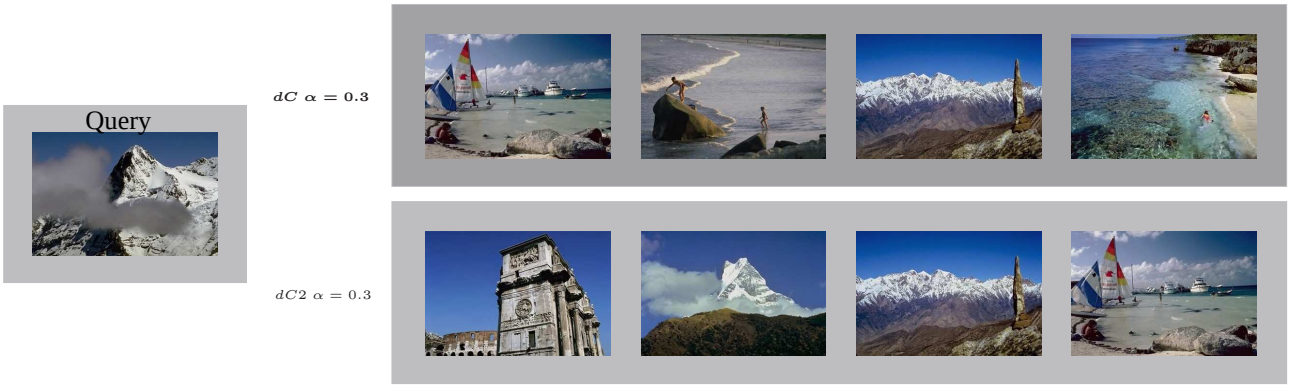


Figure 43.

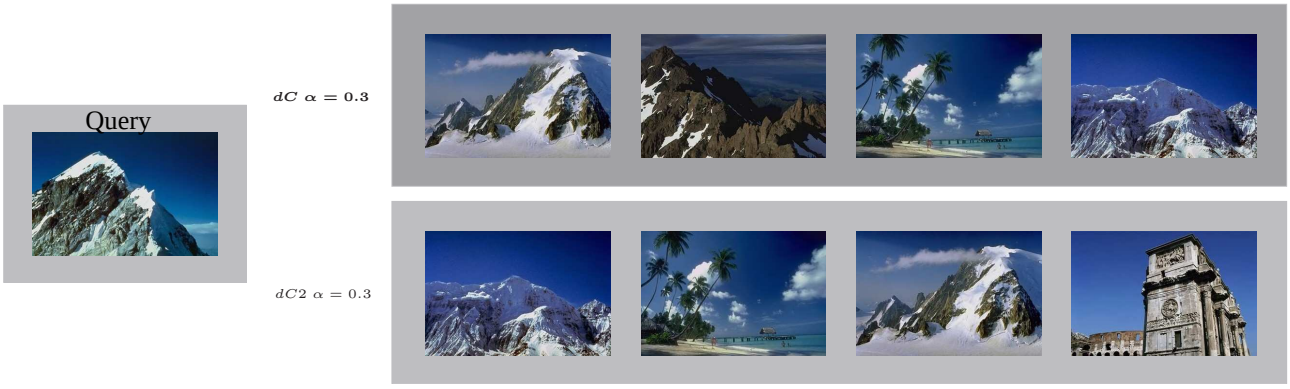


Figure 44.

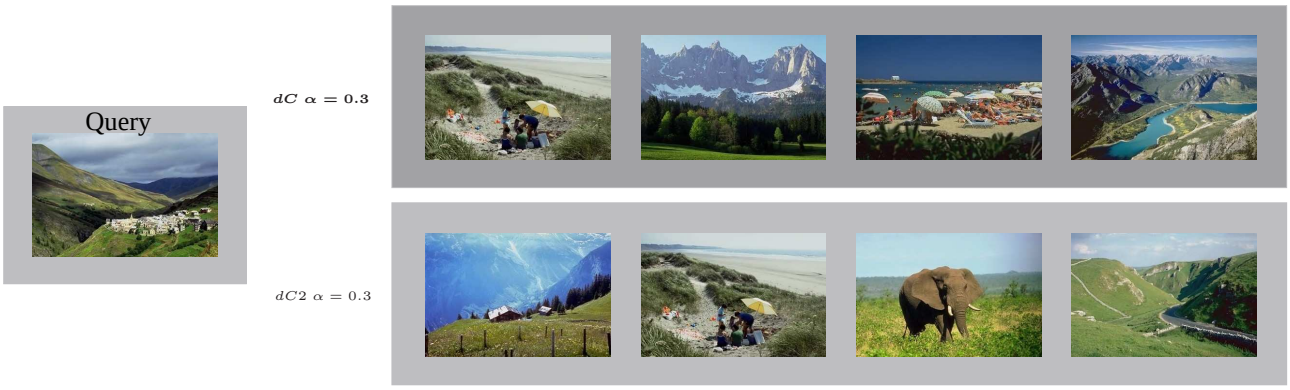


Figure 45.

Figure 46.

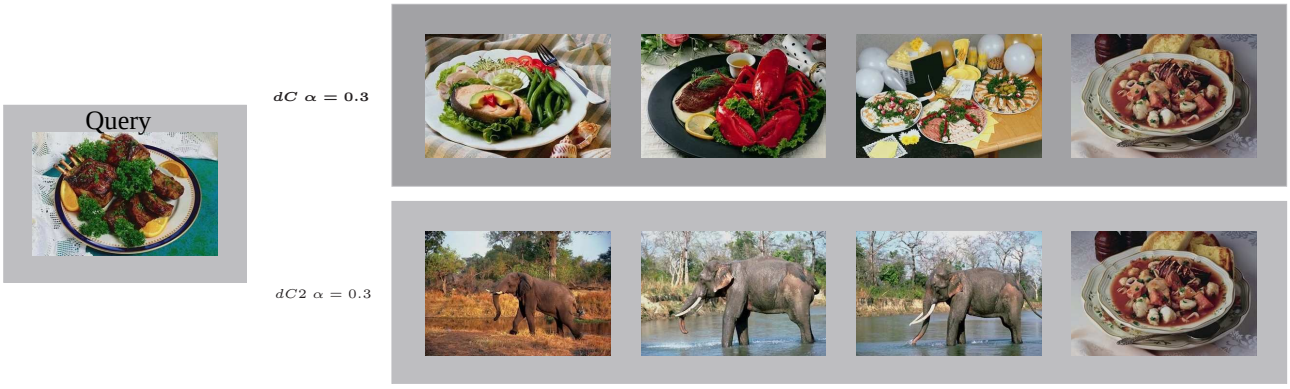


Figure 47.

Figure 48.

Figure 49.

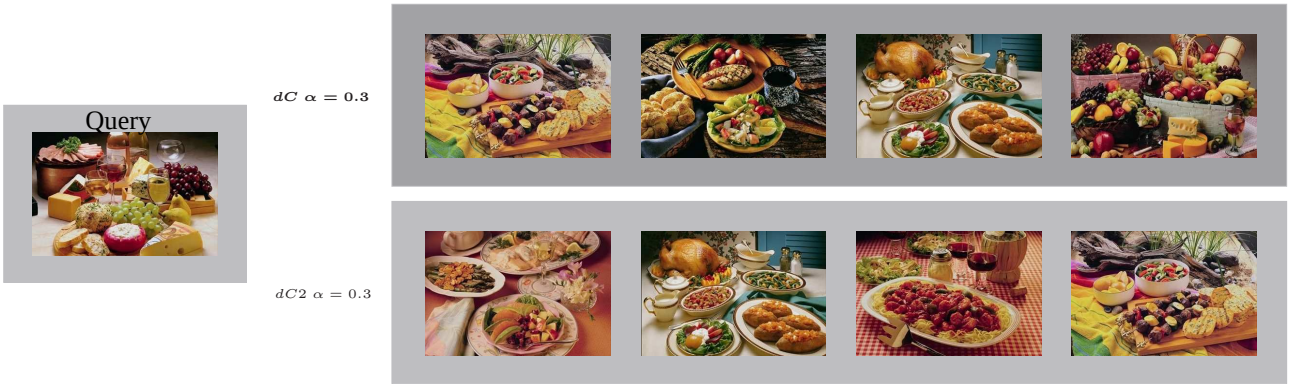


Figure 50.

References

[1] O. Pele and M. Werman. Fast and robust earth mover's distances. In *ICCV*, 2009. 1

Fast and Robust Earth Mover's Distances Additional Results - Comparing SIFT/CSIFT with All Distances

Ofir Pele
The Hebrew University of Jerusalem
ofirpele@cs.huji.ac.il

Michael Werman
The Hebrew University of Jerusalem
werman@cs.huji.ac.il

This document contains additional results for image retrieval that were omitted from the paper - "Fast and Robust Earth Mover's Distances" [3] due to lack of space. The results are for image retrieval as described in the paper. The results are for all pairs of descriptors and distance measures.

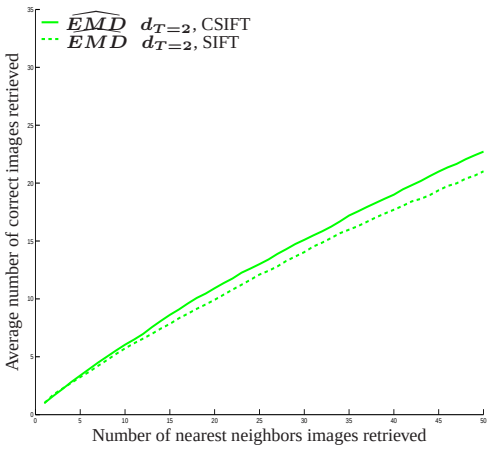


Figure 1.

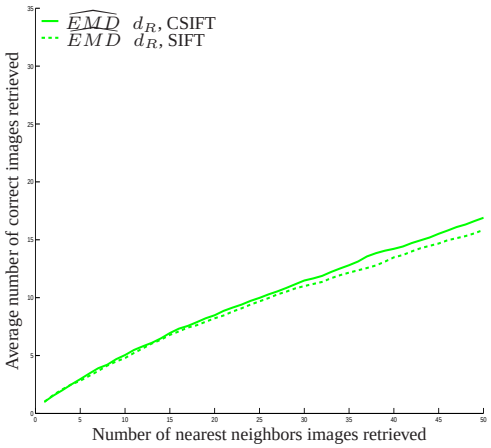


Figure 2.

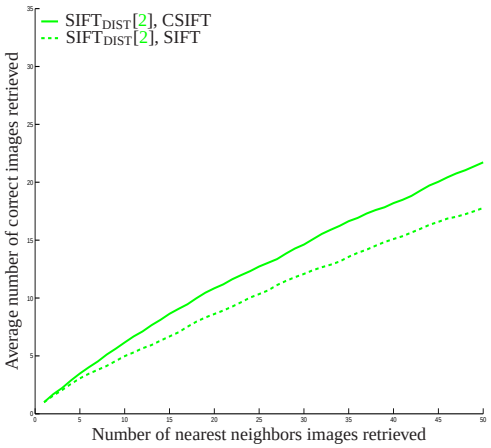


Figure 3.

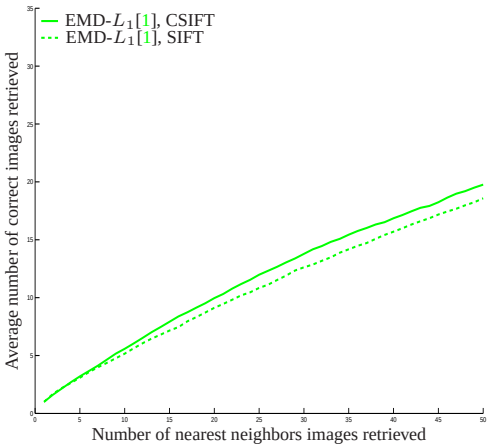


Figure 4.

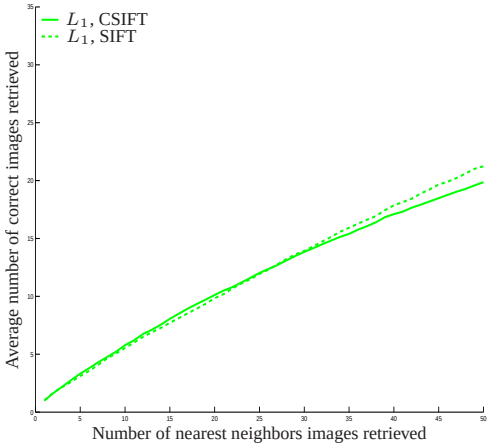


Figure 5.

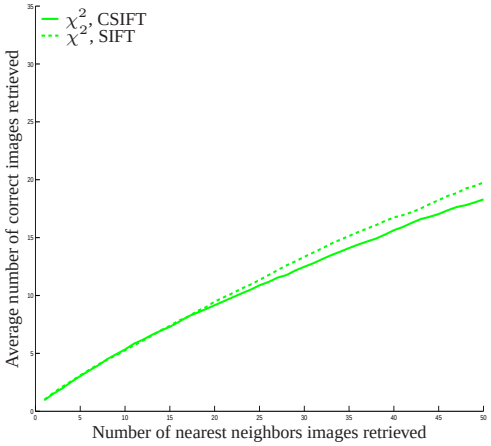


Figure 6.

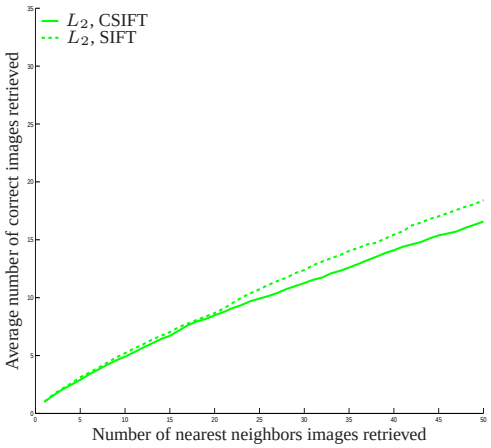


Figure 7.

References

[1] H. Ling and K. Okada. An Efficient Earth Mover's Distance Algorithm for Robust Histogram Comparison. *PAMI*, 2007. 1

[2] O. Pele and M. Werman. A linear time histogram metric for improved sift matching. In *ECCV*, 2008. 1

[3] O. Pele and M. Werman. Fast and robust earth mover's distances. In *ICCV*, 2009. 1

Chapter 3

The Quadratic-Chi Histogram Distance Family - Appendices

The Quadratic-Chi Histogram Distance Family - Appendices

Ofir Pele and Michael Werman

School of Computer Science
The Hebrew University of Jerusalem
{ofirpele,werman}@cs.huji.ac.il

1 Introduction

This document contains the appendices for the paper "The Quadratic-Chi Histogram Distance Family" [1], proofs and additional results. In section 2 we prove that all Quadratic-Chi histogram distances are continuous. In section 3 we prove that EMD, $\widehat{\text{EMD}}$ and all Quadratic-Chi histogram distances are *Similarity-Matrix-Quantization-Invariant*. In section 4 we present additional shape classification results. In section 5 we compare experimentally color image distances with and without spatial pruning. In section 6 we compare experimentally distances which resembles QC distances, but are either not *Similarity-Matrix-Quantization-Invariant* or not *Sparseness-Invariant*. In section 7 we compare the results for all the pairs of descriptors (SIFT/CSIFT) and distance measures used in the image retrieval experiments.

2 Quadratic-Chi Histogram Distance Continuity Proof

In this section we prove that a Quadratic-Chi (QC) histogram distance is continuous. We first recall its definition.

Let P and Q be two non-negative bounded histograms. That is, $P,Q \in [0,U]^N$, where the bound U can be any finite number. Let A be a non-negative symmetric bounded bin-similarity matrix such that each diagonal element is bigger or equal to every other element in its row (this demand is weaker than being a strongly dominant matrix) and to 1. That is, $A \in [0,U]^N \times [0,U]^N$ and $\forall i,j\ A_{ii} \geq A_{ij}$ and $\forall i\ A_{ii} \geq 1$. Let $0 \leq m < 1$ be the normalization factor. A Quadratic-Chi (QC) histogram distance is defined as:

$$\text{QC}_m^A(P,Q) = \sqrt{\sum_{ij} \left(\frac{(P_i - Q_i)}{\left(\sum_c (P_c + Q_c) A_{ci}\right)^m} \right) \left(\frac{(P_j - Q_j)}{\left(\sum_c (P_c + Q_c) A_{cj}\right)^m} \right) A_{ij}} \quad (1)$$

Theorem 1. *Each addend denominator:* $(\sum_c (P_c + Q_c) A_{ci})^m (\sum_c (P_c + Q_c) A_{cj})^m$*, is zero only if the addend numerator:* $(P_i - Q_i)(P_j - Q_j) A_{ij}$*, is zero.*

Proof.

$$(\sum_c (P_c+Q_c)A_{ci})^m(\sum_c (P_c+Q_c)A_{cj})^m = 0 \quad \Rightarrow \quad (2)$$
$$((P_i+Q_i)A_{ii})^m((P_j+Q_j)A_{jj})^m = 0 \quad \Rightarrow \quad (3)$$
$$(P_i=0 \wedge Q_i=0)\vee(P_j=0 \wedge Q_j=0) \quad \Rightarrow \quad (4)$$
$$((P_i-Q_i)=0)\vee((P_j-Q_j)=0) \quad \Rightarrow \quad (5)$$
$$(P_i-Q_i)(P_j-Q_j)A_{ij}=0 \quad (6)$$

Eqs. 3,4 and 6 are true as everything is non-negative and $A_{ii} \geq 1$.

We defined $\frac{0}{0}=0$. So, in order to prove continuity, we need to prove that when QC addend's denominator tends to zero, the whole addend tends to zero (this is the only point where discontinuity can occur).

Theorem 2.

$$(\sum_c (P_c+Q_c)A_{ci})^m(\sum_c (P_c+Q_c)A_{cj})^m \to 0 \Rightarrow \left(\frac{(P_i-Q_i)}{(\sum_c (P_c+Q_c)A_{ci})^m}\right)\left(\frac{(P_j-Q_j)}{(\sum_c (P_c+Q_c)A_{cj})^m}\right)A_{ij} \to 0 \quad (7)$$

Proof.

$$(\sum_c (P_c+Q_c)A_{ci})^m(\sum_c (P_c+Q_c)A_{cj})^m \to 0 \quad \Rightarrow \quad (8)$$
$$((P_i+Q_i)A_{ii})^m((P_j+Q_j)A_{jj})^m \to 0 \quad \Rightarrow \quad (9)$$
$$((P_i+Q_i)^m \to 0)\vee((P_j+Q_j)^m \to 0) \quad \Rightarrow \quad (10)$$
$$\big((P_i \to 0) \wedge (Q_i \to 0)\big)\vee\big((P_j \to 0) \wedge (Q_j \to 0)\big) \quad \Rightarrow \quad (11)$$
$$\left(\frac{(P_i-Q_i)}{((P_i+Q_i)A_{ii})^m}\right)\left(\frac{(P_j-Q_j)}{((P_j+Q_j)A_{jj})^m}\right)A_{ij} \to 0 \quad \Rightarrow \quad (12)$$
$$\left(\frac{(P_i-Q_i)}{(\sum_c (P_c+Q_c)A_{ci})^m}\right)\left(\frac{(P_j-Q_j)}{(\sum_c (P_c+Q_c)A_{cj})^m}\right)A_{ij} \to 0 \quad \Rightarrow \quad (13)$$

Eq. 9 is true as everything is non-negative and $m \geq 0$. Eq. 10 is true as if the whole product tends to zero, at least one of its multiplicands should tend to zero and $A_{ii} \geq 1$ and $A_{jj} \geq 1$. Eq. 11 is true as everything is non-negative and $m \geq 0$. Eq. 12 is true as $|(P_i-Q_i)| \leq |(P_i+Q_i)|A_{ii}$ and $|(P_j-Q_j)| \leq |(P_j+Q_j)|A_{jj}$ and $0 \leq m < 1$ (note that the strictly smaller than one is required here) and everything is bounded. Eq. 13 finishes the proof and is true as:

$$\left|\left(\frac{(P_i-Q_i)}{(\sum_c (P_c+Q_c)A_{ci})^m}\right)\right| < \left|\left(\frac{(P_i-Q_i)}{((P_i+Q_i)A_{ii})^m}\right)\right| \wedge \quad (14)$$
$$\left|\left(\frac{(P_j-Q_j)}{(\sum_c (P_c+Q_c)A_{cj})^m}\right)\right| < \left|\left(\frac{(P_j-Q_j)}{((P_j+Q_j)A_{jj})^m}\right)\right| \quad (15)$$

3 EMD, $\widehat{\text{EMD}}$ and Quadratic-Chi Histogram Distances *Similarity-Matrix-Quantization-Invariance* Proof

In this section we prove that EMD, $\widehat{\text{EMD}}$ and all Quadratic-Chi histogram distances are *Similarity-Matrix-Quantization-Invariant*. We recall *Similarity-Matrix-Quantization-Invariance* definition:

Definition 1. *Let $\mathcal{D}$ be a cross-bin histogram distance between two histograms P and Q and let A be the bin-similarity/distance matrix. We assume P, Q and A are non-negative and that A is symmetric. Let $A_{k,:}$ be the kth row of A. Let $V = [V_1, \ldots, V_N]$ be a non-negative vector and $0 \leq \alpha \leq 1$. We define $V^{\alpha,k,b} = [\ldots, \alpha V_k, \ldots, V_b + (1-\alpha)V_k, \ldots]$. That is, $V^{\alpha,k,b}$ is a transformation of V where $(1-\alpha)V_k$ mass has moved from bin k to bin b. We define $\mathcal{D}$ to be* Similarity-Matrix-Quantization-Invariant *if:*

$$A_{k,:} = A_{b,:} \;\Rightarrow\; \forall \;\; 0 \leq \alpha \leq 1,\; 0 \leq \beta \leq 1 \quad \mathcal{D}^A(P,Q) = \mathcal{D}^A(P^{\alpha,k,b}, Q^{\beta,k,b}) \tag{16}$$

If $\mathcal{D}$ is symmetric then Eq. 16 is equivalent to:

$$A_{k,:} = A_{b,:} \;\Rightarrow\; \forall \;\; 0 \leq \alpha \leq 1 \quad \mathcal{D}^A(P,Q) = \mathcal{D}^A(P^{\alpha,k,b}, Q) \tag{17}$$

EMD [2] and $\widehat{\text{EMD}}$ [3] distances are *Similarity-Matrix-Quantization-Invariant*. Since both are min-cost-flow problems [4], taking mass from one source/sink and moving it to another one, where both sources/sinks are connected to exactly the same sinks/sources, with exactly the same costs, will not change the min-cost solution.

We begin the proof that a Quadratic-Chi (QC) histogram distance (Eq. 1) is *Similarity-Matrix-Quantization-Invariant* with a lemma:

$$(\sum_c (P_c^{\alpha,k,b} + Q_c)A_{ci})^m = (\sum_c (P_c + Q_c)A_{ci})^m \tag{18}$$

Proof.

$$\sum_c (P_c^{\alpha,k,b} + Q_c)A_{ci} = \tag{19}$$

$$(\sum_{c \neq k,b} (P_c^{\alpha,k,b} + Q_c)A_{ci}) + ((P_k^{\alpha,k,b} + Q_k)A_{ki}) + ((P_b^{\alpha,k,b} + Q_b)A_{bi}) = \tag{20}$$

$$(\sum_{c \neq k,b} (P_c + Q_c)A_{ci}) + ((\alpha P_k + Q_k)A_{ki}) + ((P_b + (1-\alpha)P_k + Q_b)A_{bi}) = \tag{21}$$

$$(\sum_{c \neq k,b} (P_c + Q_c)A_{ci}) + ((\alpha P_k + Q_k)A_{ki}) + ((P_b + (1-\alpha)P_k + Q_b)A_{ki}) = \tag{22}$$

$$(\sum_{c \neq k,b} (P_c + Q_c)A_{ci}) + ((P_k + Q_k + P_b + Q_b)A_{ki}) = \tag{23}$$

$$(\sum_{c \neq k,b} (P_c + Q_c)A_{ci}) + ((P_k + Q_k)A_{ki}) + ((P_b + Q_b)A_{bi}) = \tag{24}$$

$$\sum_c (P_c + Q_c)A_{ci} \tag{25}$$

$$\Updownarrow \tag{26}$$

$$(\sum_c (P_c^{\alpha,k,b} + Q_c)A_{ci})^m = (\sum_c (P_c + Q_c)A_{ci})^m \tag{27}$$

In the proof we use the definitions and in Eqs. 22,24 we use $A_{ki} = A_{bi}$.

We now prove that Quadratic-Chi histogram distance is *Similarity-Matrix-Quantization-Invariant* using Eq. 17. That is we prove that:

$$A_{k,:} = A_{b,:} \;\Rightarrow\; \forall\; 0 \leq \alpha \leq 1 \quad QC_m^A(P,Q) = QC_m^A(P^{\alpha,k,b},Q) \tag{28}$$

We will prove that $(QC_m^A(P,Q))^2 = (QC_m^A(P^{\alpha,k,b},Q))^2$ which directly implies Eq. 28.

We start by separating $(QC_m^A(P^{\alpha,k,b},Q))^2$ addends into three disjoint and complementary types. The first type is when both i and j are different from k and b. For these addends the equality is direct from the definition. The second type is when both i and j are either k or b. For this type the proof is:

$$\sum_{i=(k\vee b),j=(k\vee b)} \frac{(P_i^{\alpha,k,b}-Q_i)(P_j^{\alpha,k,b}-Q_j)A_{ij}}{\left(\sum_c(P_c^{\alpha,k,b}+Q_c)A_{ci}\right)^m\left(\sum_c(P_c^{\alpha,k,b}+Q_c)A_{cj}\right)^m} = \tag{29}$$

$$\sum_{i=(k\vee b),j=(k\vee b)} \frac{(P_i^{\alpha,k,b}-Q_i)(P_j^{\alpha,k,b}-Q_j)A_{ij}}{\left(\sum_c(P_c+Q_c)A_{ci}\right)^m\left(\sum_c(P_c+Q_c)A_{cj}\right)^m} = \tag{30}$$

$$\frac{(P_k^{\alpha,k,b}-Q_k)(P_k^{\alpha,k,b}-Q_k)A_{kk}}{\left(\sum_c(P_c+Q_c)A_{ck}\right)^{2m}} + \frac{(P_k^{\alpha,k,b}-Q_k)(P_b^{\alpha,k,b}-Q_b)A_{kb}}{\left(\sum_c(P_c+Q_c)A_{ck}\right)^m\left(\sum_c(P_c+Q_c)A_{cb}\right)^m} +$$

$$\frac{(P_b^{\alpha,k,b}-Q_b)(P_k^{\alpha,k,b}-Q_k)A_{bk}}{\left(\sum_c(P_c+Q_c)A_{cb}\right)^m\left(\sum_c(P_c+Q_c)A_{ck}\right)^m} + \frac{(P_b^{\alpha,k,b}-Q_b)(P_b^{\alpha,b,b}-Q_b)A_{bb}}{\left(\sum_c(P_c+Q_c)A_{cb}\right)^{2m}} = \tag{31}$$

$$\frac{\left((\alpha P_k-Q_k)^2+2(\alpha P_k-Q_k)((P_b+(1-\alpha)P_k)-Q_b)+((P_b+(1-\alpha)P_k)-Q_b)^2\right)A_{kk}}{\left(\sum_c(P_c+Q_c)A_{ck}\right)^{2m}} = \tag{32}$$

$$\frac{\left((\alpha P_k-Q_k)+((P_b+(1-\alpha)P_k)-Q_b)\right)^2A_{kk}}{\left(\sum_c(P_c+Q_c)A_{ck}\right)^{2m}} = \tag{33}$$

$$\frac{\left((P_k-Q_k)+(P_b-Q_b)\right)^2A_{kk}}{\left(\sum_c(P_c+Q_c)A_{ck}\right)^{2m}} = \tag{34}$$

$$\frac{\left((P_k-Q_k)^2+2(P_k-Q_k)(P_b-Q_b)+(P_b-Q_b)^2\right)A_{kk}}{\left(\sum_c(P_c+Q_c)A_{ck}\right)^{2m}} = \tag{35}$$

$$\sum_{i=(k\vee b),j=(k\vee b)} \frac{(P_i-Q_i)(P_j-Q_j)A_{ij}}{\left(\sum_c(P_c+Q_c)A_{ci}\right)^m\left(\sum_c(P_c+Q_c)A_{cj}\right)^m} \tag{36}$$

Eq. 30 is based on the lemma in Eq. 18. Eqs. 32,36 are true as $A_{ck} = A_{cb}$ and $A_{kk} = A_{kb} = A_{bb} = A_{bk}$.

The third type of addends is when only i equals b or k or only j equals b or k. These sub-cases are symmetric, so it is enough to prove equality for only one of them:

$$\sum_{i=(k\vee b),j\neq(k\vee b)} \frac{(P_i^{\alpha,k,b}-Q_i)(P_j^{\alpha,k,b}-Q_j)A_{ij}}{\left(\sum_c(P_c^{\alpha,k,b}+Q_c)A_{ci}\right)^m\left(\sum_c(P_c^{\alpha,k,b}+Q_c)A_{cj}\right)^m} = \tag{37}$$

$$\sum_{i=(k\vee b),j\neq(k\vee b)} \frac{(P_i^{\alpha,k,b}-Q_i)(P_j-Q_j)A_{ij}}{\left(\sum_c(P_c+Q_c)A_{ci}\right)^m\left(\sum_c(P_c+Q_c)A_{cj}\right)^m} = \tag{38}$$

$$\sum_{j\neq(k\vee b)} \left(\frac{(P_k^{\alpha,k,b}-Q_k)(P_j-Q_j)A_{kj}}{\left(\sum_c(P_c+Q_c)A_{ck}\right)^m\left(\sum_c(P_c+Q_c)A_{cj}\right)^m} + \frac{(P_b^{\alpha,k,b}-Q_b)(P_j-Q_j)A_{bj}}{\left(\sum_c(P_c+Q_c)A_{cb}\right)^m\left(\sum_c(P_c+Q_c)A_{cj}\right)^m} \right) = \tag{39}$$

$$\sum_{j\neq(k\vee b)} \left(\frac{(\alpha P_k-Q_k+(P_b+(1-\alpha)P_k)-Q_b)(P_j-Q_j)A_{kj}}{\left(\sum_c(P_c+Q_c)A_{ck}\right)^m\left(\sum_c(P_c+Q_c)A_{cj}\right)^m} \right) = \tag{40}$$

$$\sum_{j\neq(k\vee b)} \left(\frac{((P_k-Q_k)+(P_b-Q_b))(P_j-Q_j)A_{kj}}{\left(\sum_c(P_c+Q_c)A_{ck}\right)^m\left(\sum_c(P_c+Q_c)A_{cj}\right)^m} \right) = \tag{41}$$

$$\sum_{i=(k\vee b),j\neq(k\vee b)} \frac{(P_i-Q_i)(P_j-Q_j)A_{ij}}{\left(\sum_c(P_c+Q_c)A_{ci}\right)^m\left(\sum_c(P_c+Q_c)A_{cj}\right)^m} \tag{42}$$

Eq. 38 is based on the lemma in Eq. 18. Eqs. 40,42 are true as $A_{kj} = A_{bj}$ and $A_{ck} = A_{cb}$.

Since we covered the three disjoint and complementary cases we have proved that a Quadratic-Chi histogram distance is *Similarity-Matrix-Quantization-Invariant*.

4 Additional Shape Classification Results

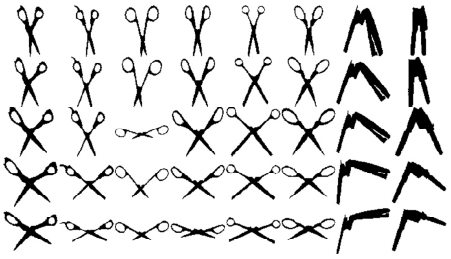
This section presents shape classification results using the same framework as Ling et al. [5,6,7]. We first present results for the articulated shape data set [5,7]. Then we present results for the Kimia-216 data set [8].

4.1 Articulated Data Set

The articulated shape data set [5,7] has 40 images from 8 different objects. Each object has 5 images articulated to different degrees, see Fig. 1(a). The dataset is very challenging because of the similarity between different objects (especially the scissors).

We used Belongie et al.'s Shape Context (SC) [9] and Ling and Jacobs' Inner Distance Shape Context (IDSC) [5].

To evaluate results, for each image, the 4 most similar matches are chosen from other images in the dataset. The retrieval result is summarized as the number of 1st,



(a)

Results using SC[9]

	Top 1	Top 2	Top 3	Top 4	AUC%
QCN$^{1-\frac{\text{dsc}T=2}{2}}$	20	6	10	5	0.307
QCN^{I}	18	7	5	8	0.278
QCS$^{1-\frac{\text{dsc}T=2}{2}}$	23	11	11	4	0.378
QCS^{I}	19	9	9	6	0.318
χ^2	25	13	13	7	0.430
$QF^{1-\frac{\text{dsc}T=2}{2}}$	25	16	10	7	0.438
L_2	25	15	8	7	0.420
$\widehat{\text{EMD}}_1^{\text{dsc}T=2}$	23	13	11	7	0.400
L_1	20	10	9	5	0.333
SIFT$_{\text{DIST}}$[10]	20	9	6	9	0.320
EMD-L_1[6]	15	10	6	10	0.280
Diffusion[11]	19	10	7	6	0.315
Bhattacharyya[12]	25	14	9	9	0.422
KL[13]	12	8	4	9	0.223
JS[14]	25	16	9	7	0.432

(b)

Results using IDSC[5]

	Top 1	Top 2	Top 3	Top 4	AUC%
QCN$^{1-\frac{\text{dsc}T=2}{2}}$	39	38	38	34	0.950
QCN^{I}	40	37	36	33	0.940
QCS$^{1-\frac{\text{dsc}T=2}{2}}$	39	35	38	28	0.912
QCS^{I}	40	34	37	27	0.907
χ^2	40	36	36	21	0.902
$QF^{1-\frac{\text{dsc}T=2}{2}}$	40	34	39	19	0.897
L_2	39	35	35	18	0.873
$\widehat{\text{EMD}}_1^{\text{dsc}T=2}$	39	36	35	27	0.902
L_1	39	35	35	25	0.890
SIFT$_{\text{DIST}}$[10]	38	37	27	22	0.848
EMD-L_1[6]	39	35	38	30	0.917
Diffusion[11]	39	35	34	23	0.880
Bhattacharyya[12]	40	37	32	23	0.895
KL[13]	40	38	36	29	0.938
JS[14]	40	35	37	21	0.900

(c)

Fig. 1. (a) Articulated shape database. This dataset contains 40 images from 8 objects. Each column contains five images of the same object with different articulation.
(b) Shape classification results using Belongie et al.'s Shape Context (SC)[9]. All distances performance is poor.
(c) Shape classification results using Ling and Jacobs' Inner Distance Shape Context (IDSC)[5]. QCN$^{1-\frac{\text{dsc}T=2}{2}}$ outperformed all other distances. QCN^{I} performance was excellent, which shows the importance of the normalization factor. All cross-bin distances outperformed their bin-by-bin versions. Again, χ^2 and QF improved upon L_2. QCN and QCS which are mathematically sound combinations of χ^2 and QF outperformed both.

2nd, 3rd and 4th most similar matches that come from the correct object. Tables (b) and (c) in Fig. 1 show the retrieval results for SC and IDSC respectively. Using SC descriptors, the performance was poor on this dataset for all kinds of distances. Using IDSC descriptors, the QCN$^{1-\frac{\mathrm{dsc}_{T=2}}{2}}$ histogram distance outperformed all the other distances. QCN^{I} performance was excellent, which shows the importance of the normalization factor. All cross-bin distances outperformed their bin-by-bin versions. χ^2 and QF improved upon L_2. QCN and QCS which are mathematically sound combinations of χ^2 and QF outperformed both.

4.2 Kimia-216 Data Set

The Kimia-216 shape data set [8] has 216 images grouped into 18 classes with 12 images in each class, see Fig. 2(a).

We tested for shape classification with Belongie et al. Shape Context (SC) [9]. We do not present results for Ling and Jacob's IDSC [5], as the code that computes IDSC [7] crashes on this data set (segmentation fault).

To evaluate results, for each image, the 11 most similar matches are chosen from other images in the dataset (as there are 12 images in each class). The retrieval result is summarized as the number of 1st, 2nd, ..., 11th and most similar matches that come from the correct object. Table (b) in Fig. 2(b) show the retrieval results. Again, the QCN$^{1-\frac{\mathrm{dsc}_{T=2}}{2}}$ histogram distance outperformed all the other distances. Again, χ^2 and QF improved upon L_2 and QCN and QCS outperformed both.

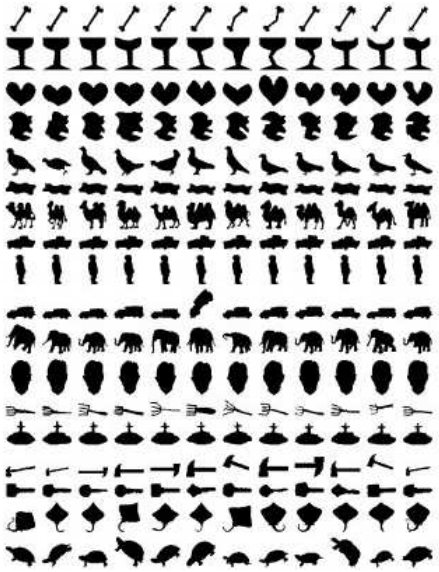
5 Comparison of Color Image Distances With and Without Spatial Pruning

In this section we compare experimentally the color image distances that have spatial pruning (those used in [1]) to distances without such a pruning (those used in [3]). Results are presented in Fig. 3. Pruning slightly improved accuracy for QF and slightly reduced accuracy for QCN, QCS and $\widehat{\mathrm{EMD}}$. However, pruning considerably reduced running time (see Table 1 in page 15).

6 Comparison to *Non-Similarity-Matrix-Quantization-Invariant* and *Non-Sparseness-Invariant* Quadratic-Chi-like Histogram Distances

In this section we compare experimentally distances which resembles QC distances, but are either not *Similarity-Matrix-Quantization-Invariant* or not *Sparseness-Invariant*. The comparison shows that these properties considerably boost performance, especially for sparse color histograms.

As in the QC definition, let P and Q be two non-negative bounded histograms. That is, $P, Q \in [0, U]^N$. Let A be a non-negative symmetric bounded bin-similarity matrix such that each diagonal element is bigger or equal to every other element in its



(a)

Results using SC[9]

	Top 1	Top 2	Top 3	Top 4	Top 5	Top 6	Top 7	Top 8	Top 9	Top 10	Top 11	AUC%
QCN$^{1-\frac{\text{dsc}_{T=2}}{2}}$	215	216	215	216	214	214	207	207	200	192	173	0.981
QCN^{I}	215	208	212	201	198	198	190	177	164	150	127	0.920
QCS$^{1-\frac{\text{dsc}_{T=2}}{2}}$	216	215	215	212	212	205	204	205	193	184	155	0.969
QCS^{I}	215	215	214	214	215	208	209	205	197	190	167	0.976
χ^2	216	215	212	211	210	203	199	199	192	177	156	0.960
$QF^{1-\frac{\text{dsc}_{T=2}}{2}}$	214	211	207	206	197	190	183	183	167	162	146	0.920
L_2	213	210	203	204	195	189	183	178	167	149	139	0.910
$\widehat{\text{EMD}}_1^{\text{dsc}_{T=2}}$	215	215	215	215	214	207	207	202	201	187	162	0.974
L_1	215	215	215	214	213	209	211	207	199	192	170	0.978
SIFT$_{\text{DIST}}$[10]	215	215	215	215	215	211	210	207	201	189	162	0.979
EMD-L_1[6]	215	215	212	213	212	208	204	203	197	184	161	0.969
Diffusion[11]	215	215	214	214	214	214	208	204	200	196	174	0.979
Bhattacharyya[12]	216	215	213	212	212	206	203	206	193	177	159	0.967
KL[13]	212	203	200	199	195	186	187	179	166	151	130	0.899
JS[14]	216	215	212	210	208	205	200	197	189	182	152	0.959

(b)

Fig. 2. (a) Kimia-216 shape database. This dataset has 216 images from 18 classes. Each row contains 12 images of the same class.
(b) Shape classification results using the Shape Context (SC)[9]. We do not present results for Ling and Jacob's IDSC [5], as the code that computes IDSC [7] crashed (segmentation fault). Again, the QCN$^{1-\frac{\text{dsc}_{T=2}}{2}}$ histogram distance outperformed all other distances. Again, χ^2 and QF improved upon L_2. QCN and QCS which are mathematically sound combinations of χ^2 and QF outperformed both.

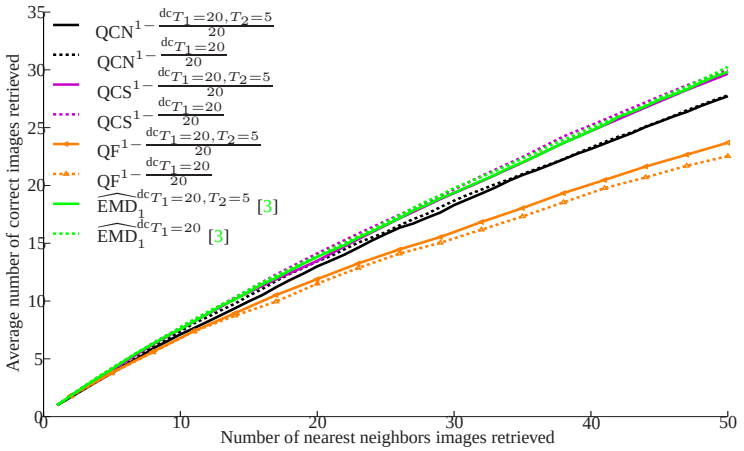


Fig. 3. Results for image retrieval comparing distances which are pruning spatially far away pixels (painted with solid lines) to those that are not (painted with broken lines). Pruning slightly improved accuracy for QF and slightly reduced accuracy for QCN, QCS and $\widehat{\text{EMD}}$. However, pruning considerably reduced running time (see Table 1 in page 15).

row (this demand is weaker than being a strongly dominant matrix) and to 1. That is, $A \in [0,U]^N \times [0,U]^N$ and $\forall i,j\ A_{ii} \geq A_{ij}$ and $\forall i\ A_{ii} \geq 1$. Let $0 \leq m < 1$ be the normalization factor. A QC-like histogram distance which is not *Sparseness-Invariant* is defined as:

$$\text{NSI-QC}_m^A(P,Q) = \sqrt{\sum_{ij}\left(\frac{\left(\sum_c P_c A_{ci} - \sum_c Q_c A_{ci}\right)}{\left(\sum_c (P_c+Q_c)A_{ci}\right)^m}\right)\left(\frac{\left(\sum_c P_c A_{cj} - \sum_c Q_c A_{cj}\right)}{\left(\sum_c (P_c+Q_c)A_{cj}\right)^m}\right)A_{ij}} \tag{43}$$

A QC-like histogram distance which is not *Similarity-Matrix-Quantization-Invariant* is defined as:

$$\text{NQI-QC}_m^A(P,Q) = \sqrt{\sum_{ij}\left(\frac{(P_i - Q_i)}{(P_i+Q_i)^m}\right)\left(\frac{(P_j - Q_j)}{(P_j+Q_j)^m}\right)A_{ij}} \tag{44}$$

If I is the identity matrix then:

$$\text{QC}_m^I(P,Q) = \text{NSI-QC}_m^I(P,Q) = \text{NQI-QC}_m^I(P,Q) \tag{45}$$

For any bin-similarity A we get the following equality:

$$\text{NQI-QC}_m^I(AP,AQ) = \text{NSI-QC}_m^I(P,Q) \tag{46}$$

We present results for image retrieval comparing QCN, NSI-QCN(NSI-QC$_{0.9}$), NQI-QCN(NQI-QC$_{0.9}$), QCS, NSI-QCS(NSI-QC$_{0.5}$) and NQI-QCS(NQI-QC$_{0.5}$) in Fig. 4(a) for SIFT-like descriptors (CSIFT as the CSIFT descriptor, compared to SIFT was better for all distance measures, see Fig. 7) and in Fig. 4(b) for small L*a*b* images with the pruning distance used in the paper [1] and in 4(c) for small L*a*b* images with a non-pruning distance as was used in [3]. The results show that *Similarity-Matrix-Quantization-Invariance* and *Sparseness-Invariance* considerably boost performance. The gain in performance is large using the L*a*b* images, which are sparse 5-dimensional histograms (x, y, L*, a*, b*). That is, the query histogram was always: $[1,\ldots,1,0,\ldots,0]$ and each image being compared to the query was represented by the histogram: $[0,\ldots,0,1,\ldots,1]$. That is, the dimension of the space is $32\times 48\times 256^3$ where only 32×48 entries are non-zero in each histograms. The bin-similarity matrix is computed for each pair of images. For these histograms:

$$\text{NQI-QCN}^A(P,Q)=\text{NQI-QCS}^A(P,Q)=\text{QF}^A(P,Q) \tag{47}$$

The time complexity of all the distances tested in this section is linear. In practice (see Table 1), QCS distances are faster to compute than QCN distances. In addition, NQI-QC distances are the fastest to compute, QC distances second, and NSI-QC distances have the longest running time. However, the differences between the running times are minor as all distances have linear time complexity.

7 Comparing SIFT/CSIFT with All Distances for Image Retrieval

In this section we present results for image retrieval as described in the paper. The results are for all pairs of descriptors (SIFT/CSIFT) and distance measures and given in Figs. 5, 6,7.

References

1. Pele, O., Werman, M.: The quadratic-chi histogram distance family. In: ECCV. (2010) 1, 7, 10
2. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. IJCV (2000) 3
3. Pele, O., Werman, M.: Fast and robust earth mover's distances. In: ICCV. (2009) 3, 7, 9, 10, 13
4. Ahuja, R., Magnanti, T., Orlin, J.: Network flows: theory, algorithms, and applications. (1993) 3
5. Ling, H., Jacobs, D.: Shape classification using the inner-distance. PAMI (2007) 5, 6, 7, 8
6. Ling, H., Okada, K.: An Efficient Earth Mover's Distance Algorithm for Robust Histogram Comparison. PAMI (2007) 5, 6, 8, 13
7. Ling, H.: Articulated shape benchmark and idsc code. http://www.ist.temple.edu/~hbling/code/inner-dist-articu-distrbution.zip (2010) 5, 7, 8
8. Sebastian, T., Kimia, B.: Curves vs. skeletons in object recognition. SP (2005) 5, 7
9. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. PAMI (2002) 5, 6, 7, 8

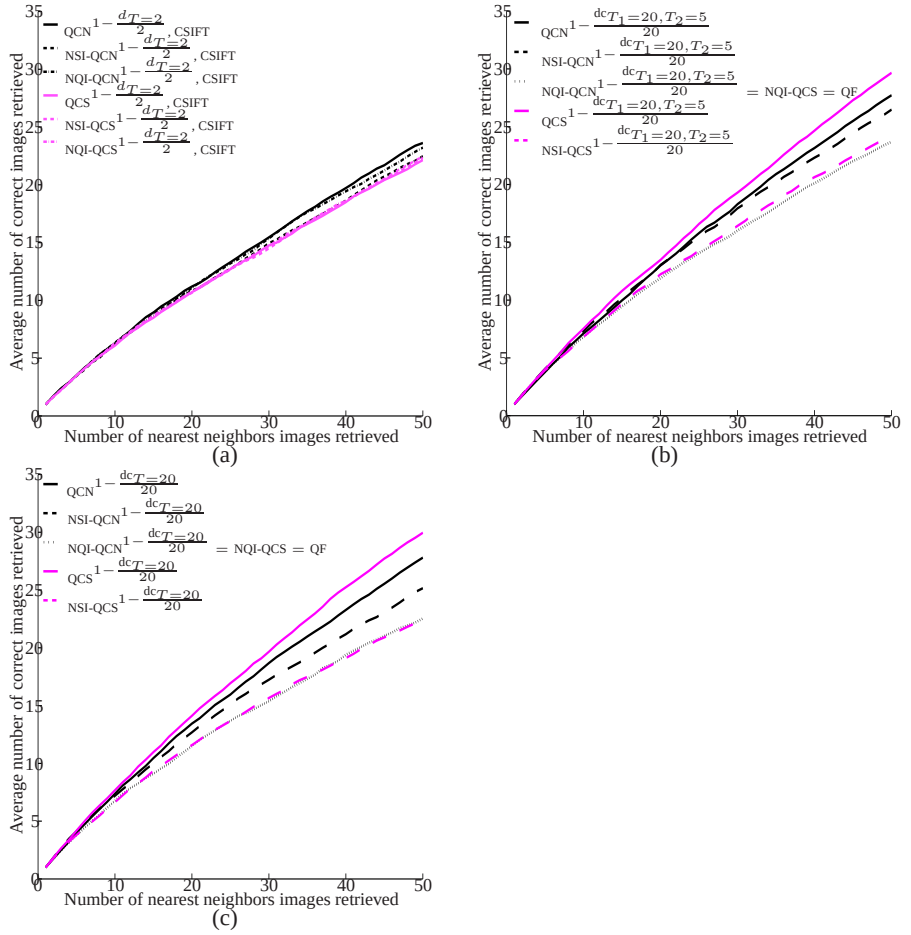


Fig. 4. Results for image retrieval comparing QC which is *Sparseness-Invariant* and *Similarity-Matrix-Quantization-Invariant* to NSI-QC and to NQI-QC which are not *Sparseness-Invariant* and not *Similarity-Matrix-Quantization-Invariant* respectively.
(a) **SIFT-like descriptors.** For each distance measure, we present the descriptor (SIFT/CSIFT) with which it performed best. The results for all the pairs of descriptors and distance measures can be found in Figs. 5, 6 and 7. It can be seen that QCN which is both *Sparseness-Invariant* and *Similarity-Matrix-Quantization-Invariant* has the best performance. However, the differences in performance are small.
(b) **L*a*b* images with pruning and (c) without pruning** Here, the *Sparseness-Invariant* and *Similarity-Matrix-Quantization-Invariant* properties are very important. This is probably because L*a*b* images are represented as very sparse histograms in a very high dimensional space (the dimension of the space is $32 \times 48 \times 256^3$ where only 32×48 entries are non-zero in each histograms). Comparing to (b) we can also note that performance using this representation is better than those obtained with the SIFT-like descriptor.

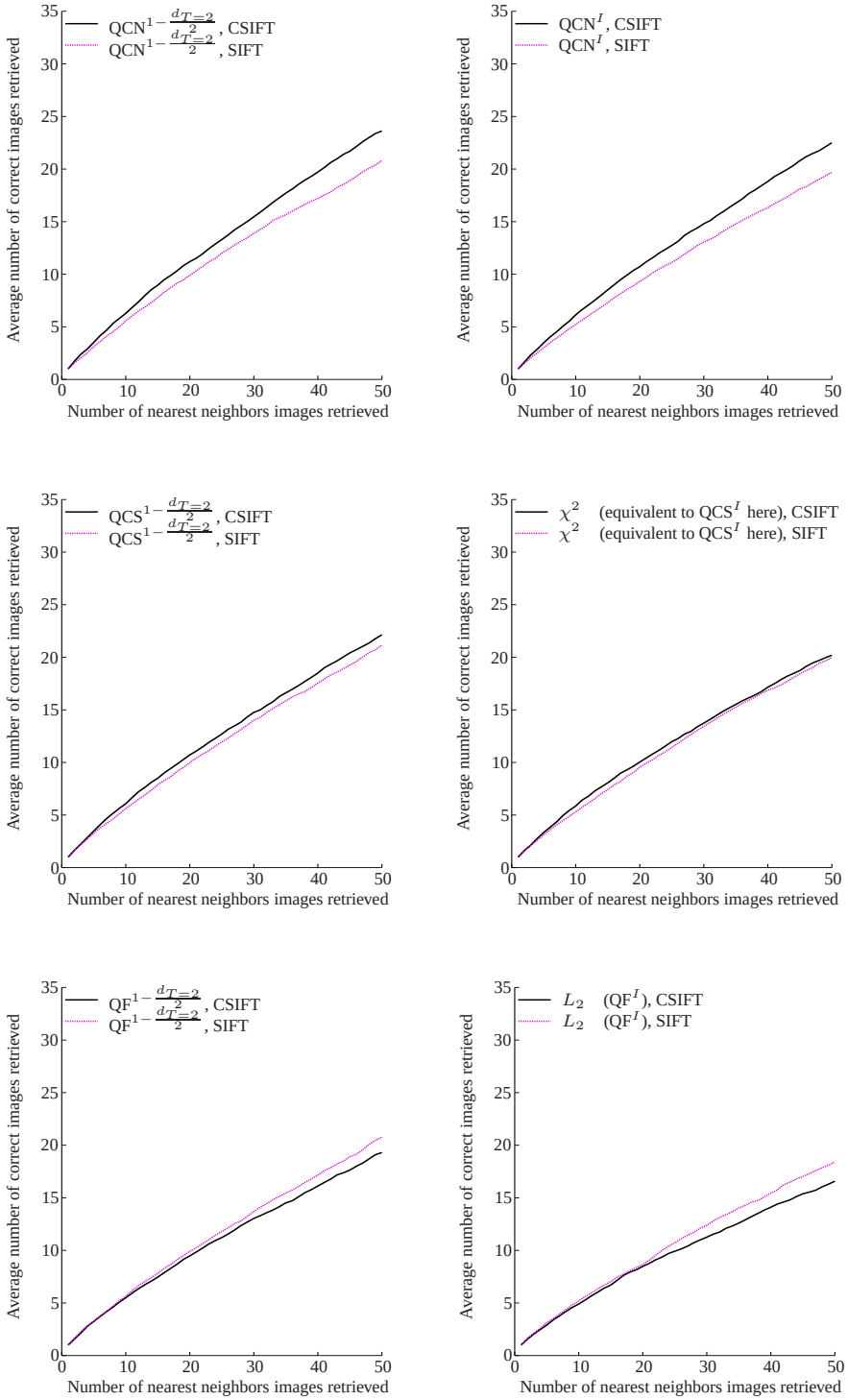


Fig. 5. Comparison of SIFT with CSIFT for all distances, part 1.

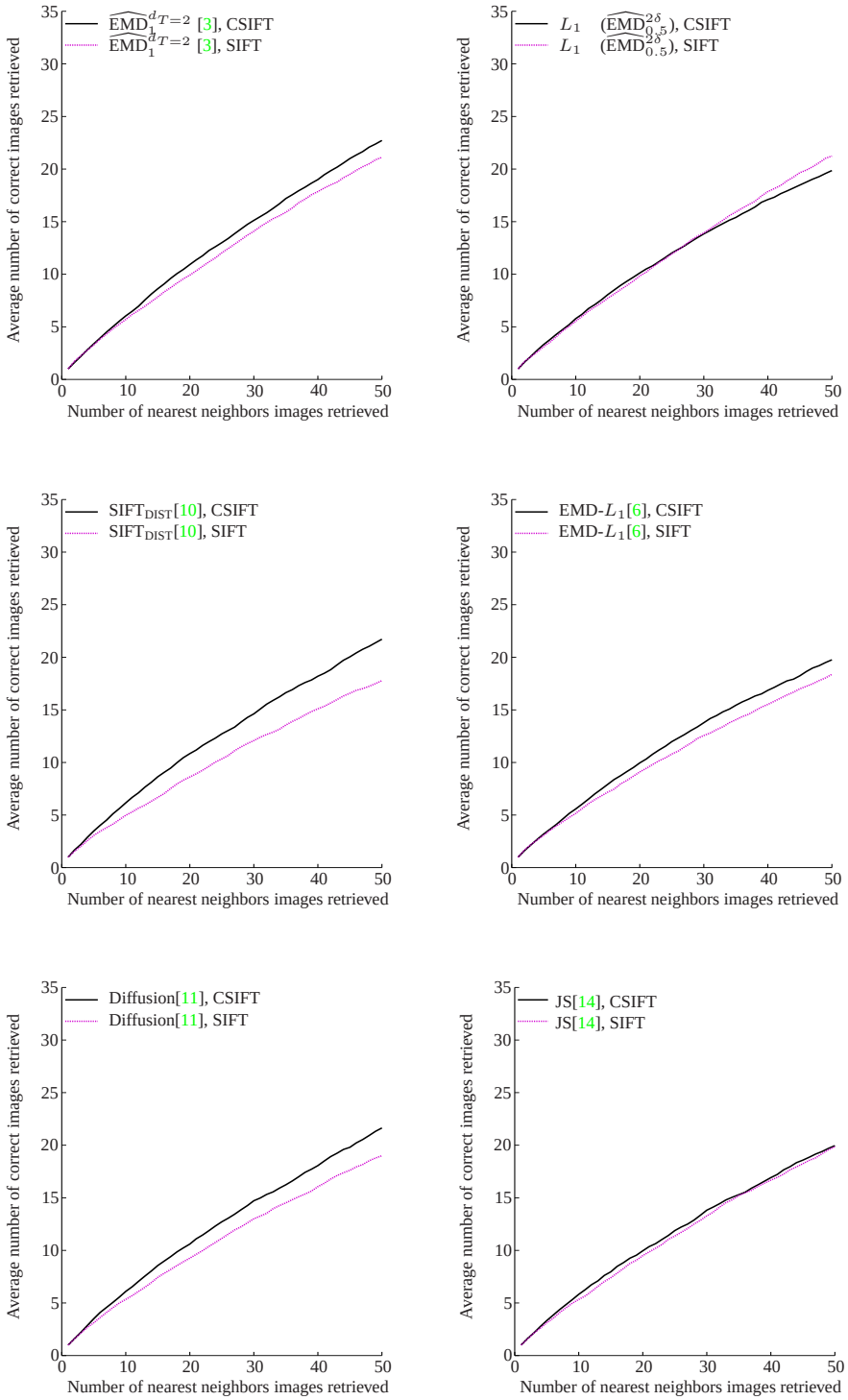


Fig. 6. Comparison of SIFT with CSIFT for all distances, part 2

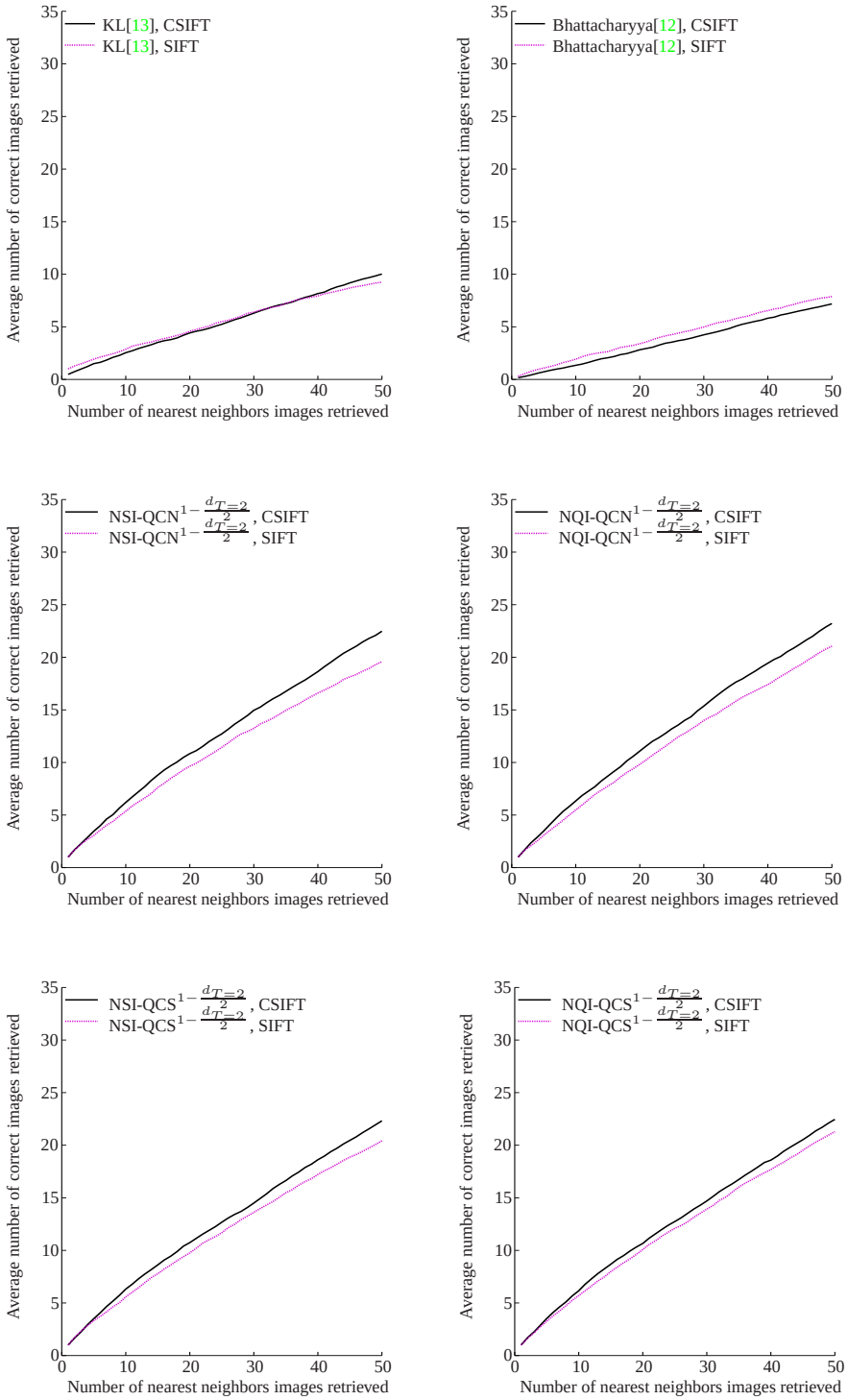


Fig. 7. Comparison of SIFT with CSIFT for all distances, part 3

Descriptor	QCN^{A_2}	NSI-QCN^{A_2}	NQI-QCN^{A_2}	QCS^{A_2}	NSI-QCS^{A_2}	NQI-QCS^{A_2}
(SIFT)	**0.15**	**0.18**	**0.14**	**0.07**	**0.1**	**0.06**

Descriptor	QCN$^{A_{20,5}}$	NSI-QCN$^{A_{20,5}}$	QF$^{A_{20,5}}$	QCS$^{A_{20,5}}$	NSI-QCS$^{A_{20,5}}$
(L*a*b*)	**20** (370)	**21** (371)	**11** (361)	**19** (369)	**18** (368)

Descriptor	QCN$^{A_{20}}$	NSI-QCN$^{A_{20}}$	QF$^{A_{20}}$	QCS$^{A_{20}}$	NSI-QCS$^{A_{20}}$
(L*a*b*)	**20** (3020)	**21** (3021)	**11** (3011)	**19** (3019)	**18** (3018)

Table 1. (SIFT) 384-dimensional SIFT-like descriptors matching time (in *milliseconds*). The distances from left to right are the same as the distances in Fig. 4 (a) from up to down.
(L*a*b*) 32×48 L*a*b* images matching time (in *milliseconds*) using a distance which prunes spatially far away pixels (middle row) and a distance which does not (bottom row). On the middle row, the distances from left to right are the same as the distances in Fig. 4 (b). On the bottom row, the distances from left to right are the same as the distances in Fig. 4 (c). In parentheses is the time it takes to compute the distance and the bin-similarity matrix as it cannot be computed offline.
To summaries results, QCS distances are faster to compute than QCN distances. Also, NQI-QC distances are the fastest to compute, QC distances second, and NSI-QC distances have the longest running time. However, the differences between the running times are minor as all distances have linear time complexity. Finally, pruning spatially far away pixels considerably reduced running time, almost without reducing accuracy (see Fig. 3).

10. Pele, O., Werman, M.: A linear time histogram metric for improved sift matching. In: ECCV. (2008) 6, 8, 13
11. Ling, H., Okada, K.: Diffusion distance for histogram comparison. In: CVPR. (2006) 6, 8, 13
12. Bhattacharyya, A.: On a measure of divergence between two statistical populations defined by their probability distributions. BCMS (1943) 6, 8, 14
13. Kullback, S., Leibler, R.: On information and sufficiency. AMS (1951) 6, 8, 14
14. Lin, J.: Divergence measures based on the Shannon entropy. IT (1991) 6, 8, 13