CAR-TR-1025
CS-TR-4908
UMIACS-TR-2008-06

21 February 2008

Approximate earth mover's distance in linear time

Sameer Sheorey
David W. Jacobs

Center for Automation Research
University of Maryland
College Park, MD 20742-3411
(ssameer, djacobs)@umiacs.umd.edu

Abstract

The earth mover's distance (EMD) [21] is an important perceptually meaningful metric for comparing histograms, but it suffers from high ($O(n^3 \log n)$) computational complexity. We present a novel linear time algorithm for approximating the EMD for low dimensional histograms using the sum of absolute values of the weighted wavelet coefficients of the difference histogram. EMD computation is a special case of the Kantorovich-Rubinstein transshipment problem, and we exploit the Hölder continuity constraint in its dual form to convert it into a simple optimization problem with an explicit solution in the wavelet domain. We prove that the resulting wavelet EMD metric is equivalent to EMD, i.e. the ratio of the two is bounded. We also provide estimates for the bounds.

The weighted wavelet transform can be computed in time linear in the number of histogram bins, while the comparison is about as fast as for normal Euclidean distance or χ^2 statistic. We experimentally show that wavelet EMD is a good approximation to EMD, has similar performance, but requires much less computation.

Keywords: Fast earth mover's distance, wavelets, mass transportation problems, Kantorovich-Rubinstein metric.

1 Introduction

Histogram descriptors: Histogram descriptors are a powerful representation for matching and recognition. Their statistical nature gives them sufficient robustness while maintaining discriminative power. They have been used extensively in vision applications like shape matching [1], keypoint matching [17], texture analysis [13] and 3D object recognition [11]. Colour and texture histograms [21] are also used for content based image retrieval. These descriptors are often compared using binwise dissimilarity measures like Euclidean or other L_p norms or the χ^2 statistic. While these measures can be computed very fast and often give good results, they do not take into account all possible variations in the random variables whose distributions they compare. These unmodelled variations may lead to large measure values for changes in the distribution that are perceived to be small. For example, suppose we take two photos of a plain wall with strong and weak sunlight and compare their colour histograms. The histograms are shifted delta functions and have large binwise differences. Consequently, all of these measures will give large values. The popular SIFT descriptor [17] is a gradient orientation – location histogram. A similar histogram shifting will occur if the keypoint is not localized accurately.

Earth mover's distance: Crossbin distance measures take into account the fact that histograms are based in feature space and it is possible for histogram mass to *move* between bins in feature space. They penalize this movement according to the distance covered, called the *ground distance*. The earth mover's distance (EMD) is a natural and intuitive metric between histograms if we think of them as piles of sand sitting on the ground (feature space). Each grain of sand is an observed sample. To quantify the difference between two distributions, we can measure how far the grains of sand have to be moved so that the two distributions coincide exactly. *EMD is the minimal total ground distance travelled weighted by the amount of sand moved* (called *flow*). EMD makes sure that shifts in sample values are not penalized excessively. For the example of a shifted delta function, the EMD is simply the shift amount. EMD has been successfully used for image retrieval by comparing colour and texture histograms [21], contour matching [3], image registration [2], [6] and pattern matching in medical images [9], [8]. However, a major hurdle to using EMD is its $O(n^3 \log n)$ computational complexity (for an n-bin histogram).

Wavelet EMD: In this paper, we present a novel method for approximating the EMD for histograms p_1 and p_2 using a new metric on the weighted wavelet coefficients of the difference histogram. We show that this is equivalent to EMD, i.e. the ratio of EMD to wavelet EMD is always between two constants. Although our estimates for these constants are loose, we will show experimentally that our metric follows EMD closely and can be used instead without any significant performance difference. The wavelet EMD metric can be computed in $O(n)$ time.

EMD can be computed as the minimal value of a linear program. The Kantorovich-Rubinstein (KR) transshipment problem [20] is the corresponding problem for continuous distributions. Both problems admit duals with the same optimal value. The important insight in our algorithm is that the dual of the KR problem has a wavelet domain representation with a simple explicit solution.

In the primal form, the objective function is the total flow–weighted ground distance between all bin pairs. See table (1) for exact definitions. The flows must make up for the difference between the histograms at each corresponding bin. In the dual form, the optimization is over a potential f assigned to each bin. For a difference histogram $p := p_1 - p_2$, the dual EMD is

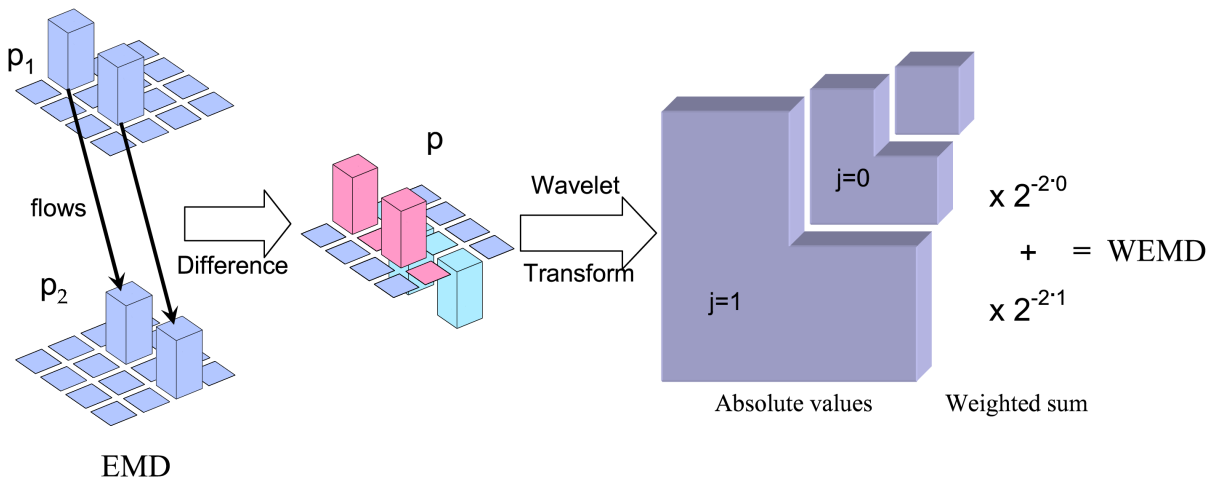

Figure 1: Computation of wavelet EMD

given by :

$$\text{Dual EMD} := \sup_f \int f(x)p(x)dx \tag{1}$$

subject to the constraint that the difference between two bin potentials is bounded by the ground distance $c(x, y) = \|x - y\|$, i.e. $f(x) - f(y) \leq \|x - y\|$. The objective function is the maximum inner product between the potential function and the difference histogram and is easily represented in the wavelet domain, since orthonormal wavelets preserve inner products. The constraint means that f cannot grow faster than some (non-vertical) straight line at any point. This is actually a Hölder (or Lipschitz) continuity condition and is somewhat between continuity and differentiability. The wavelet coefficients of a Hölder continuous function decay exponentially at fine scales, since fine scale wavelets represent rapid changes in the function. We thus have an equivalent constraint in the wavelet domain. The resulting optimization is trivial and gives an explicit solution :

$$d(p)_{wemd} := \sum_\lambda 2^{-j(1+n/2)}|p_\lambda| \tag{2}$$

p is the n dimensional difference histogram and p_λ are its wavelet coefficients. The index λ includes shifts and the scale j. We will call this the *wavelet EMD* between two histograms. This is clearly a metric. This is not exactly equal to the EMD since the Hölder continuity constraint can't be transformed exactly into the wavelet domain.

This surprising formula for approximating the EMD with wavelet coefficients of the difference histogram is the main contribution of this paper. By using appropriate wavelets, we can approximate EMD very well. Since the wavelet transform is a common linear time operation, we can compute this in time linear in the number of bins for uniform histograms. Figure (1) explains the EMD approximation algorithm in 2D.

Intuitively speaking, the wavelet transform splits up the difference histogram according to scale and location. Each wavelet coefficient represents an EMD subproblem that is solved separately. The sum of all distances is an approximation to EMD. This is a good approximation because wavelet transforms are well suited for separating local variations according to scale and

EMD for signatures [21]	**Discrete EMD for histograms**	**Continuous EMD for distributions**
Signatures $f(i;1), f(i,2)$ In general, $\sum_i f(i;1) \neq \sum_i f(i;2)$ Ground distance $d_{ij} \geq 0$ Flow (from bin i to bin j) $g_{ij} \geq 0$	Histograms $f(i;1), f(i,2)$ $\sum_i f(i;1) = \sum_i f(i;2) = 1$ Difference $f(i) := f(i;1) - f(i;2)$ Ground distance $d_{ij} \geq 0$ Flow (from bin i to bin j) $g_{ij} \geq 0$ Potential π_i	Distributions $p_1(x), p_2(x)$ $\int p_1(x)dx = \int p_2(x)dx = 1$ Difference $p(x) := p_1(x) - p_2(x)$ Cost function $c(x,y) \geq 0$ Joint distribution $q(x,y) \geq 0$ Potential $f(x)$
EMD $:= \min \frac{\sum_{ij} g_{ij} d_{ij}}{\sum_{ij} g_{ij}}$ s.t. $\sum_j g_{ij} \leq f(i;1), \sum_i g_{ij} \leq f(i;2)$, $\sum_{ij} g_{ij} = \min(\sum_i f(i;1), \sum_i f(i;2))$	EMD $:= \min \sum_{ij} g_{ij} d_{ij}$ s.t. $\sum_i g_{ik} - \sum_j g_{kj} = f(k)$	EMD $:= \inf \int c(x,y) q(x,y) dxdy$ s.t. $\int q(u,y)dy - \int q(x,u)dx = p(u)$
	Dual EMD $:= \max \sum_i \pi_i f(i)$ s.t. $\pi_i - \pi_j \leq d_{ij}$	Dual EMD $:= \sup \int f(x)p(x)dx$ s.t. $f(x) - f(y) \leq c(x,y)$

Table 1: Correspondence between EMD for signatures, discrete EMD and continuous EMD for probability distributions

position. For a single wavelet coefficient, the mass to be moved is proportional to $|p_\lambda| 2^{-jn/2}$, since this would be the wavelet coefficient if we use wavelets normalized by total mass, i.e. $\int |\psi_\lambda| = 1$. The distance travelled is proportional to the span of the wavelet 2^{-j} (According to Meyer's [19] convention, a wavelet at scale j is the mother wavelet squeezed 2^j times.) The total EMD is thus approximated by equation (2).

Approximation by scale and location separation is similar to the way packages are shipped over large distances. The total journey is broken into several hops – short and long. Short hops connect the source and destination to shipping hubs, while long hops connect the shipping hubs themselves. Packages from nearby towns merge at shipping hubs to travel together. Thus, the package journey is split into multiple scales, and the sum of the distances travelled is an approximation to the actual distance.

2 Related Work

The earth movers distance was introduced in vision by Werman *et al.* [23], though they did not use this name. Rubner *et al.* [21] extended this to comparing *signatures*: adaptive histograms of varying mass represented by weighted clusters. They computed the EMD using a linear program called *transportation simplex* and used it for content based image retrieval by comparing colour signatures. They obtained better performance than binwise measures. This method has an empirical time complexity between $O(n^3)$ and $O(n^4)$. EMD being a transportation problem, can also be modelled as a network flow problem ([12] chapter 9) in graph theory. The two histograms are represented by a single graph with a vertex for each bin and ground distances as the edge weights. The two histogram vertices act as sources and sinks respectively with bin contents as values. Computing EMD is now an *uncapacitated minimum cost flow problem* and can be solved by Orlin's algorithm ([12] section 9.5) in $O(n^3 \log n)$ time.

Various approximation algorithms have been suggested to speed up the computation of EMD. Ling and Okada [16] empirically showed that EMD could be computed in $O(N^2)$ time if an L_1 ground distance is used instead of the usual Euclidean distance. They used the EMD for comparing different histogram descriptors and noted improved performance compared to χ^2 and Euclidean distance.

Indyk and Thaper [10] use a randomized multiscale embedding of histograms into a space equipped with the l_1 norm. The multiscale hierarchy is obtained by a series of random shifting

and dyadic merging of bins. The histogram levels are weighted by powers of 2, with more weight at the coarser levels. They show that the l_1 norm computed in this space, averaged over all random shifts, is equivalent to the EMD. They do not prove this for individual random embeddings, and also do not estimate the constants that bound the ratio of this norm to EMD. They couple this with locality sensitive hashing for fast nearest neighbour image retrieval using colour signatures. Grauman and Darrell's pyramid match kernel [4] is based on this method. They use histogram intersection instead of l_1 distance at each level and inverted weights to obtain a similarity measure useful for matching partial histograms instead of a metric. Both these methods have a time complexity of $O(Tdm \log D)$ for d dimensional histograms with diameter D and m bins. The random embeddings are computed T times. Although these algorithms are fast, our algorithm gives deterministic error bounds. We will also show empirically that our algorithm is more accurate.

The diffusion distance introduced by Ling and Okada in [15] is computed by constructing a Gaussian pyramid from the difference histogram and summing up the L_1 norms of the various levels. Although this has some similarities with our algorithm, it is not an approximation to the EMD and may behave differently.

Holmes and Taylor [8], [9] use partial signature matching based on the EMD for identifying mammogram structures. They embed histograms into a learned Euclidean space to speed up computation.

The continuous EMD problem and its generalizations have a good basis in probability theory for comparing distributions and have been studied since Nobel prize winner L. V. Kantorovich's [20] first formulation of the problem as a linear program and the study of its duality in 1942. In this area, various equivalent formulations of EMD are minimal l_1 metric, Kantorovich metric [20], Wasserstein distance and Mallows distance [14]. General mass transportation problems have wide applications in mathematical economics, recursive stochastic equations for studying convergence of algorithms and stochastic differential equations.

3 Theory

The earth mover's distance is a metric between two probability distributions for metric ground distances. It is a special case of a class of optimization problems in applied probability theory called *mass transportation problems*. We will first look at the analogy between discrete and continuous EMD and state the dual form (section 3.1). Then, in section (3.2), we will describe how to convert the dual form into the wavelet domain. The wavelet domain dual problem has an explicit solution.

3.1 Continuous EMD and its dual

The wavelet domain connection of the EMD problem becomes clear only when we look at EMD for continuous distributions. Table (1) lists analogous terms between EMD for signatures and discrete and continuous versions of the EMD problem for distributions. The problem is simpler for histograms than for signatures because they must add up to 1. The objective function is simpler because the total flow $\sum_{ij} g_{ij} = 1$. The constraint is simpler as well and means that the flows must make up the difference between the two histograms. This is a mass conservation constraint. We will now formally state the continuous domain EMD problem [20], summarized in the third column of table (1).

Let P_1 and P_2 be probability distributions with densities p_1 and p_2 respectively, defined on a compact space $S \subset \mathbb{R}^n$. MOre generally, In general, we can consider P_1 and P_2 to be

non-negative *Borel* measures on S, i.e. $P_1, P_2 \in M_+(S)$. c is a continuous cost function on the Cartesian product space $S \times S$. Here, we will restrict c to be of the form $\|x-y\|^s$ with $0 < s \leq 1$. $s = 1$ gives us the usual Euclidean ground distance. Thus, c is always a norm. The Kantorovich-Rubinstein transshipment problem (KRP) is to find

$$\dot{\mu}_s = \inf_q \int ||x-y||^s \, q(x,y) dx dy \tag{3}$$

where the infimum is over all joint probability distributions Q with density q on $S \times S$. Q is analogous to flow in the discrete EMD problem and specifies how the source density p_1 is moved to the target density p_2. Thus the joint density q must satisfy the mass conservation constraint :

$$p_1(u) - p_2(u) = \int q(u,y) dy - \int q(x,u) dx \tag{4}$$

$p := p_1 - p_2$ is a difference density with the property that $\int p = 0$. The corresponding distribution P thus belongs to the class of Borel measures $M_0(S)$ with a total measure 0. The *Kantorovich–Rubinstein theorem* states that the problem admits the dual representation :

$$\dot{\mu}_s = \sup_f \int f(x)(p_1(x) - p_2(x)) dx \tag{5}$$

with the same optimal value. The supremum is over all bounded continuous functions f on S (called potentials) satisfying the order s Hölder continuity condition

$$f(x) - f(y) \leq ||x-y||^s \qquad \text{for all } x, y \in S \tag{6}$$

In the dual form, the EMD is the supremum of inner products of the difference density with a suitably smooth function.

Going back to the piles of sand, in the primal form, we try to find the flows q to convert p_1 into p_2 that move the sand by the least amount (3). In the dual form, we try to assign heights or potentials f to the various bins that will drive these flows. If we limit the change in the potentials by the ground distance (6), we can measure the total sand movement by the change in total height of the sand pile (5).

The potential f thus belongs to a *homogeneous Hölder space* of functions of order s denoted by $\dot{C}^s$. This is also referred to as a *homogeneous Lipschitz space* denoted by $\dot{\text{Lip}}s$. Hölder space membership is an indication of the global smoothness of a function. For $0 < s < 1$, a bounded, continuous function f belongs to the homogeneous Hölder class $\dot{C}^s(\mathbb{R}^n)$ if the following supremum exists and is finite :

$$C_H(f) := \sup_{x \neq y} \frac{|f(x) - f(y)|}{\|x-y\|^s} \tag{7}$$

This defines the Hölder seminorm of f. This is not a norm because it assigns zero length to all constants. We can now state the constraint (6) simply as

$$C_H(f) < 1 \tag{8}$$

The corresponding *inhomogeneous Hölder space* is equipped with the following norm :

$$\|f\|_{C^s} := \max\{C_H(f), \max_x |f(x)|\} \tag{9}$$

A simpler way of expressing the KR duality is in the form of the following Cauchy-Schwarz inequality :

$$|\langle f, p\rangle| := \left| \int f(x)p(x)dx \right| \leq \dot{\mu}_s(p) C_H(f) \tag{10}$$

The KR metric is a norm in the space of probability distributions $M_0(S)$ while C_H is a seminorm for the homogeneous Hölder space. The KR duality thus establishes an isometry between these two spaces, i.e. we can obtain the norm of a function in ones space as the maximum inner product with all functions of unit norm in the corresponding dual space.

3.2 EMD in the wavelet domain

Now we will look at expressing the dual problem in the wavelet domain. We can identify the various classes that a function belongs to by observing the rate of decay of its wavelet coefficients [19] (Chapter 6). For our application, we are interested in the wavelet characterization of Hölder spaces, since the potential f belongs to one. First we will explain some notation about the wavelet series representation of a function.

A function f in $\mathbb{R}^n$ can be expressed in terms of a wavelet series (Meyer [19] Chapter 2) as:

$$f(x) = \sum_k f_k \phi(x-k) + \sum_\lambda f_\lambda \psi_\lambda(x) \tag{11}$$

ϕ and ψ are the scaling function and wavelet respectively. k runs through all integer n–tuples and represents shifts, and $\lambda := (\epsilon, j, k)$. In n dimensions, we need $2^n - 1$ different wavelet functions which are indexed by ϵ. They are usually constructed by a tensor product of 1D wavelet functions along individual dimensions. For example, in 2D, we have horizontal ($\epsilon = 1$: $\psi(x)\phi(y)$), vertical ($\epsilon = 2$: $\phi(x)\psi(y)$) and diagonal ($\epsilon = 3$: $\psi(x)\psi(y)$) wavelets. j represents the scale and is a non-negative integer. Larger values of j mean finer scales with shorter wavelet functions. The set of all possible λ for a scale $j \geq 0$ is denoted by Λ_j and Λ is the union of all Λ_j. We thus have

$$\psi_\lambda(x) := 2^{nj/2} \psi^\epsilon(2^j x - k) \tag{12}$$

A wavelet ψ has regularity $r \in \mathbb{N}$ if it has derivatives up to order r and all of them (including ψ itself) have *fast decay*, i.e. they decay faster than any reciprocal polynomial for large x. For orthonormal wavelets, the coefficients can be computed as

$$f_k = \int f(x)\bar{\phi}(x-k)dx, \quad k \in \mathbb{Z}^n \tag{13}$$

$$f_\lambda = \int f(x)\bar{\psi}_\lambda(x)dx, \quad \lambda \in \Lambda, \quad j \geq 0 \tag{14}$$

$\bar{\phi}$ and $\bar{\psi}$ are complex conjugates of ϕ and ψ respectively.

The following theorem from Meyer ([19] section 6.4) can be used to characterize functions in $C^s(\mathbb{R}^n)$:

Theorem 1. *A function $f \in L^1_{loc}(\mathbb{R}^n)$, (i.e. $|f|$ is integrable over all compact subsets of $\mathbb{R}^n$) belongs to $C^s(\mathbb{R}^n)$ if and only if, in a wavelet decomposition of regularity $r \geq 1 > s$, the approximation coefficients f_k and detail coefficients f_λ satisfy*

$$\begin{aligned} &|f_k| \leq C_0, \quad k \in \mathbb{Z}^n \quad \textit{and} \\ &|f_\lambda| \leq C_1 2^{-j(n/2+s)}, \quad \lambda \in \Lambda_j, \quad j \geq 0 \end{aligned} \tag{15}$$

for some constants C_0 and C_1.

A little modification to the proof of this theorem (see Appendix A) gives the following lemma:

Lemma 1. *For $0 < s < 1$, if the wavelet series coefficients of the function f are bounded as in (15), then $f \in C^s$ with $C_H(f) < C$ such that*

$$a_{12}(\psi; s)C_1 \leq C \leq a_{21}(\psi; s)C_0 + a_{22}(\psi; s)C_1 \tag{16}$$

for some positive constants a_{12}, a_{21} and a_{22} that depend only on the wavelet and s. For discrete distributions, if we change the definition of $C_H(f)$ to

$$C_H(f) := \sup_{|x-y|\geq 1} \frac{|f(x) - f(y)|}{\|x - y\|^s}, \tag{17}$$

the same condition holds for $s = 1$ as well.

The constants a_{12}, a_{21} and a_{22} are estimated in Appendix A. Now we have all the ingredients necessary for our main result :

Theorem 2. *Consider the KR problem with the cost function $c(x, y) = \|x - y\|^s$, $s < 1$. Let p_k and p_λ be the wavelet transform coefficients (approximation and detail, respectively) of the difference density p generated by the orthonormal wavelet-scaling function pair ψ and ϕ with regularity $r \geq 1 > s$. Then for any non-negative constants C_0 and $C_1 > 0$,*

$$\hat{\mu}_s = C_0 \sum_k |p_k| + C_1 \sum_\lambda 2^{-j(s+n/2)} |p_\lambda| \tag{18}$$

is an equivalent metric to the KR metric $\dot{\mu}_s$; i.e. there exist positive constants C_L and C_U (depending only on the wavelet used) such that

$$C_L \hat{\mu}_s \leq \dot{\mu}_s \leq C_U \hat{\mu}_s \tag{19}$$

For discrete distributions, the same result holds for $s = 1$ as well.

Proof. Consider the auxiliary wavelet domain problem :

$$\begin{aligned} &\text{Maximize } \mathbf{p}^T\mathbf{f} = \sum_k p_k f_k + \sum_\lambda p_\lambda f_\lambda \\ \text{subject to} \quad &|f_k| \leq C_0 \quad \text{and} \quad |f_\lambda| \leq C_1 2^{-j(s+n/2)} \end{aligned} \tag{20}$$

$\mathbf{p}$ and $\mathbf{f}$ are coefficient vectors of p_λ and f_μ. It is easy to see that $\hat{\mu}_s$ in (18) is the solution of this problem. We need to show that the ratio of the optimal values of the KR problem and auxiliary wavelet problem are bounded by two constants C_L and C_U. Since we use orthonormal wavelets that preserve inner products, the wavelet problem (20) has the same objective function as the KR problem dual (5).

Note that changing the KR dual problem constraint $C_H(f) < 1$ to $C_H(f) < K$ for any $K > 0$ will simply have the effect of scaling the optimal value by K, since for every function f allowed by the original constraint, there is a corresponding function Kf allowed by the new constraint. Further, the constraints in the auxiliary problem (20) will allow functions with $C_H(f) < C$, where C is bounded by the limits in (16). So, all functions with $C_H(f)$ less than the lower bound in (16) are included by the constraint, and no function with $C_H(f)$ greater

than the upper bound are included. Consequently, the optimal value is scaled by a factor C that obeys the bounds in (16). This is equivalent to (19) with

$$\begin{aligned} C_L &= a_{12}(\psi; s)C_1 \quad \text{and} \\ C_U &= a_{21}(\psi; s)C_0 + a_{22}(\psi; s)C_1. \end{aligned} \tag{21}$$

The wavelet EMD metric is thus equivalent to EMD.

For discrete distributions, we can scale the domain so that the minimum distance between any two points is 1 or more. This scales the EMD by the same factor. Now the bounds (21) are valid again and we have the required equivalence. A similar but more complex result holds for biorthogonal wavelets as well. See Appendix B for details. □

We set $C_0 = 0$ because this gives us the tightest bounds in (16). Setting the constant C_1 to 1, we get the simple distance measure :

$$d(p)_{wemd} := \hat{\mu}_s\Big|_{C_0=0,C_1=1} = \sum_{\lambda} |p_\lambda| 2^{-j(s+n/2)} \tag{22}$$

$$\text{The bounds ratio } \frac{C_U}{C_L} = \frac{a_{22}(\psi; s)}{a_{12}(\psi; s)} \tag{23}$$

measures the maximum possible error. After scaling wavelet EMD suitably, the ratios WEMD/EMD and EMD/WEMD will always be less than the bounds ratio.

3.3 Why not the Fourier transform ?

At this point, it is clear that a wavelet representation can enable us to approximate EMD because of its effective characterization of the Hölder continuity of a function. Wavelets provide a tight (*if and only if*) characterization that enables us to construct an equivalent wavelet domain norm.

We know that a Fourier transform also characterizes Hölder or Lipschitz continuity. The two principal results concerning Fourier series and Lipschitz functions are ([18] Theorem 6.1 and []) :

$$\text{For } s > 0, \int_{\mathbb{R}} |\hat{f}(\omega)|(1 + |\omega|^s)d\omega < \infty \implies f \in C^s(\mathbb{R}) \implies |\hat{f}(\omega)| \leq \frac{K}{1 + |\omega|^s} \text{ for some } K > 0 \tag{24}$$

Neither of these two conditions gives a complete characterization of the Hölder space, the first includes extra functions that are not Hölder continuous while the second excludes some Hölder continuous functions. Hence, we cannot use the Fourier characterization for approximating EMD. Another orthonormal representation can be used instead of wavelets if it provides a tight characterization of Hölder continuity.

Intuitively speaking, the Fourier transform is not very good at localizing features or judging distances between them. So it cannot be used to measure distances between places of excess and deficit mass that computing the EMD requires.

4 EMD for partial histograms

For many applications, it may not be possible to gather enough data samples to construct a complete histogram. We are still required to compare two histograms constructed from a

different number of data samples. The trivial method of renormalizing the two histograms to the same number of samples is correct only if we can be sure that the measured data samples were picked uniformly from the histogram domain. This is rarely the case. For example, while constructing the colour histogram of an image patch, the colour values are sampled according to image content and will often be clustered together for nearby pixels. If part of the image patch is occluded, part of the histogram is likely to be missing.

We will first look at Kantorovich and Rubinstein original extension to deal with partial histograms. We will then look at how Rubner's signature EMD and examine its relation with the KR extension. Although this allows for a dual representation, it cannot be directly converted into the wavelet domain. Next we will look at Hanin's extension of the KR metric that surprisingly preserves our current wavelet domain algorithm.

4.1 Kantorovich-Rubinstein extension

Lets first take a look at how different quantities change when we have partial histograms. Our probability distributions p_1 and p_2 are now unnormalized non–negative *Borel* measures (i.e. they belong to the space $M_+(S)$). The difference distribution $p = p_1 - p_2$ now belongs to the space of general signed Borel measures $M(S)$ on S. We no longer have $\int_S p(x)dx = 0$, i.e. p need not belong to $M_0(S)$ anymore. As a result, the joint density $q(x, y)$ representing flows is also no longer normalized.

Kantorovich and Rubinstein's extension, referred to as the K-norm in [5], assigns a *waste* cost function $w(x)$ for unmatched mass left at the point x. Wasting extra mass is costlier than transporting it anywhere else

$$w(x) > \sup_{y \in S} d(x, y) \quad \forall x \in S \tag{25}$$

and it is not worthwhile to transfer mass to another position just to waste it there.

$$|w(x) - w(y)| \leq d(x, y) \quad \forall x, y \in S \tag{26}$$

The extended KR metric is a combination of the cost of transporting matching mass and wasting the rest.

$$\|p\|_{w,d} = \inf_{p_0 \in M_0(S)} \left\{ \|p_0\|_d + \int w(x)|p(x) - p_0(x)|dx \right\} \tag{27}$$

This extension reduces to the original KR norm in the case that $p \in M_0(S)$, since there is no waste

The KR extension may not be physically realistic in many cases since the waste cost depends on the size of the domain 25. Indeed, this extension cannot be used for unbounded domains at all. The next two extensions get around this limitation by ignoring the constraint 25.

4.2 Rubner's EMD for signatures

Rubner's original EMD formulation [21] for signatures deals with partial histograms by minimizing the ratio of movement work to total flow, i.e.

$$EMD_{\text{Rubner}} := \min_q \frac{\iint d(x, y)q(x, y)dxdy}{\iint q(x, y)dxdy} \tag{28}$$

$$\text{subject to} \quad \int q(x, y)dy \leq p_1(x), \quad \int q(x, y)dx \leq p_2(y) \tag{29}$$

$$\text{and} \quad \iint q(x, y)dxdy = \min \left\{ \int p_1(x)dx, \int p_2(y)dy \right\} \tag{30}$$

The flow is constrained so that all mass is transferred from the smaller distribution. Note that this imposes no penalty for the unmatched part of the larger histogram.

This can be simplified by normalizing p_1 and p_2 so that the smaller distribution has unit measure again. This scales the EMD by the normalizing value as well. Without loss of generality, we can assume that $\int p_2(x)dx = 1$ and $\int p_1(x)dx \geq 1$. This implies $\int p(x)dx \geq 0$. Since all mass from p_2 is to be transferred, we have $\iint q(x,y)dxdy = 1$, i.e. q is a proper joint probability density as before. Finally, the constraint $\int q(u,y)dy - \int q(x,u)du \leq p(u)$ means that overall, we cannot remove more mass from a bin at point u than the excess $p(u)$. The optimization will not allow transferring mass from a bin with mass deficit ($p(x) < 0$) to a bin with mass excess ($p(x) > 0$) since this increases the EMD without affecting the constraints. So, this single inequality actually implies these two constraints :

$$\begin{aligned} \int q(u,y)dy - \int q(x,u)du &= p_1(u) - p_2(u) \quad \text{if} \quad p_1(u) \leq p_2(u) \\ \int q(u,y)dy - \int q(x,u)du &\leq p_1(u) - p_2(u) \quad \text{otherwise} \end{aligned} \tag{31}$$

For metric costs, these two inequalities are equivalent to the two inequalities 29. 29 can be violated while satisfying 31 if q removes more mass from bin u than present and then puts it back again from other bins. Such a procedure can only increase EMD for metric cost functions.

We can write the simplified problem as :

$$\begin{aligned} EMD_{\text{simple}} &:= \inf_q \iint d(x,y)q(x,y)dxdy \\ \text{subject to} \quad & \int q(u,y)dy - \int q(x,u)du < p(u) \end{aligned} \tag{32}$$

Note that this is the same as the KR extension with a zero waste function.

This is rather similar to our original EMD problem, and the only difference is that we have an inequality constraint instead of equality constraint. We can show that this problem also admits a strong dual given by :

$$\begin{aligned} EMD_{simple,dual} &:= \sup_f \int f(x)p(x)dx \\ \text{subject to} \quad & f(x) - f(y) \leq d(x,y) \quad \text{and} \quad f(x) \leq 0 \end{aligned} \tag{33}$$

As before, we can easily convert the objective function and the Hölder continuity constraint into the wavelet domain. However, the extra constraint $f(x) \leq 0$ poses a serious problem. It does not have any direct wavelet domain conversion. Although there are indirect methods of ensuring negativity in the wavelet domain (for example, using the wavelet representation of convolution operators), they will not be able to give us a simple linear time algorithm. Now we will look at a different partial EMD formulation that allows us to continue using our current simple linear time algorithm.

4.3 Hanin's partial EMD formulation

Hanin [7] proposed a different extension to the Kantorovich–Rubinstein metric for partial histograms. Hanin's extension retains almost all the properties of the original KR metric. Although it is defined for any metric cost function, we will concentrate on the metric cost $d(x,y) := \|x-y\|^s, \quad 0 < s \leq 1$. It is defined as

$$\mu_s(p) := \inf_{p_0 \in M_0(S)} \{\dot{\mu}_s(p_0) + \text{Var}(p - p_0)\} \tag{34}$$

Here $\text{Var}(p) := \int |p(x)|dx$ is the total variation or L_1 norm. Note that the term total variation norm has different meanings in functional analysis and probability theory. Here we are using the probability theory meaning of the term. We can get Hanin's extension by setting the waste function $w(x) = 1$ in the KR extension. This is again a norm provided the cost function is a metric. If $\int p(x)dx = 0$ and D is the diameter of the support of p,

$$\mu_s(p) \leq \dot{\mu}_s(p) \leq \frac{1}{2}\max\{D, 2\}\mu_s(p) \tag{35}$$

So, Hanin's extension is in general equivalent to the KR metric. They are identical if $D \leq 2$. In fact, Hanin's extension behaves as if the distance metric was saturated at the value 2. The total variation cost of wasting positive and negative histogram masses of size δp would be $2\delta p$ while the transportation cost would be $d\delta p$. So, it is cheaper to waste histogram mass than move it a distance greater than 2 units. We can make sure that this does not happen and make Hanin's extension is identical to the KR metric by scaling the domain to a diameter 2 before computing EMD.

For our purposes, the most important property of Hanin's extension is that it preserves KR duality 10 in almost the same form.

$$\begin{aligned}\left|\int f(x)p(x)dx\right| &\leq \left|\int f(x)p_0(x)dx\right| + \left|\int f(x)(p(x) - p_0(x))dx\right| \\ &\leq C_H(f)\dot{\mu}_s(p_0) + \max_x |f(x)| \int |p(x) - p_0(x)|dx \\ &\leq ||f||_{C^s}\left[\dot{\mu}_s(p_0) + \text{Var}(p - p_0)\right]\end{aligned}$$

Using the definition 34,

$$|\langle f, p\rangle| \leq ||f||_{C^s}\, \mu_s(p) \tag{36}$$

Hanin [7] also shows that for any p, there exists an f such that equality is attained. This is identical to the original KR duality except that the potential function f now belongs to the corresponding *inhomogeneous Hölder space*, i.e. constants are now important. We can rephrase this duality relation as

$$\mu_s(p) = \sup_f \int f(x)p(x)dx \quad \text{subject to} \quad C_H(f) \leq 1 \quad \text{and} \quad \max|f(x)| \leq 1 \tag{37}$$

This clears our path for constructing a wavelet domain approximation. In fact, it is clear that both lemma 1 and theorem 2 are still valid. The WEMD approximation is still given by 18 as

$$\hat{\mu}_s = C_0 \sum_k |p_k| + C_1 \sum_\lambda 2^{-j(s+n/2)}|p_\lambda|$$

The only difference is that since we are using the inhomogeneous Hölder space, we can no longer set $C_0 = 0$. Very roughly, the ratio $\frac{C_0}{C_1}$ determines the relative weight given to the extra histogram mass in p. A higher ratio will give more weight to the total variation part of the norm.

5 Experiments

First, in section (5.1), we will discuss some implementation issues that affect the accuracy and other aspects of wavelet EMD. In section (5.2), we will describe how to choose appropriate wavelets. Finally, in section (5.3), We will describe experiments that demonstrate that the wavelet EMD behaves very similar to EMD, but can be computed much faster.

5.1 Some implementation notes

For applications that store computed histogram descriptors, we split the wavelet EMD computation into two parts. First, the histogram descriptor is converted into the wavelet domain and its coefficients are scaled according to equation (2). The wavelet EMD distance between two descriptors is now the L_1 distance between these coefficients. We should note the following points while computing wavelet EMD :

1. Initialization : The standard Mallat filter bank algorithm ([18] section 7.3.1) for computing the wavelet transform starts with fine level wavelet coefficients as input. We can use signal values as input if we want to reconstruct the signal again, as in compression or denoising. This does not work if we want to use wavelet coefficients to represent signal properties like Hölder continuity. We can approximate fine scale wavelet coefficients with signal values if we use *coiflets* ([18] section 7.2.3). Unfortunately, this is not accurate enough for our application. So, we use the wavelet transform initialization method (algorithm 2) of Zhang, Tian and Peng [24]. We assume that the histogram bin values are obtained from a block sampler.

2. Periodic and non-periodic histograms : For data like distance and intensity values, there are no samples outside the histogram limits and we use zero padding extension while computing the wavelet transform. Since angles are measured modulo 2π, angle dimensions are extended periodically. For example, SIFT descriptors are 3D histograms of gradient orientation with respect to location around the feature point. So, we should use periodic extension along the gradient orientation dimension and zero padding along the location dimensions.

3. Wavelet transform size : Zero padding increases the size of the wavelet transform. For each decomposition level, the histogram is padded with a vector of zeros about as long as the wavelet filter length. This is significant for multi-dimensional histograms that only have a few bins along each dimension. However, most of these coefficients are close to zero because the wavelet transform is a sparse representation. We can store the coefficients compactly as a sparse vector if we set small coefficients to zero. After weighting the coefficients, we keep the largest coefficients that contribute 95% to the total L_1 norm. The remaining are set to zero. The coefficients are then stacked to form a 1D sparse vector: the final descriptor representation. Descriptor comparison takes time linear in the number of non–zero coefficients. Although there may be about 1–5 times as many elements as in the original histogram, depending on its size and dimensionality, the required time is similar to that for χ^2 or Euclidean distance on similarly enlarged histograms.

4. Histogram dimensionality : Wavelet transform computation time increases exponentially ($O(2^n)$) with dimension using Mallat's fast wavelet transform (FWT) algorithm. On the other hand, the computation time for Swelden's lifting wavelet transform (LWT) algorithm does not depend significantly on the dimensionality. LWT reduces unnecessary computation by subsampling before filtering, unlike FWT. For long wavelet filters, LWT requires half as much time as FWT for each dimension, and hence the $O(2^n)$ factor is absent. So LWT is a far better choice for high dimensional histograms.

Another concern is the data storage requirements for high dimensional histograms. This is usually not a problem since high dimensional histograms tend to be sparse. However, the wavelet transform of a sparse array is quite likely to be non-sparse. At the cost of some accuracy, this problem can be mitigated by thresholding as described above.

Daubechies	C_U/C_L	Daub. symmetric	C_U/C_L
db3	6.33	sym3	6.33
db4	7.29	sym4	4.64
db5	9.92	sym5	6.01
db6	12.59	sym6	5.58
Coiflets	C_U/C_L	Ojanen	C_U/C_L
coif1	4.38	oj8	7.46
coif2	4.75	oj10	10.56
coif3	5.85	oj12	13.79

Table 2: Theoretical (loose) estimates for maximum error for various 1D wavelets. Ojanen wavelets have maximum smoothness for a given filter length. Coiflets have low error despite large support.

Next we will look at how to choose wavelets that approximate EMD well.

5.2 Which wavelets ?

The conditions of theorem (2) put some restrictions on the wavelets for which this works. We need wavelets with at least one derivative. This rules out the simple Haar wavelet.

We can try choosing the best possible wavelets by computing the bounds ratio C_U/C_L for $C_0 = 0, C_1 = 1$. Table (2) lists maximum error estimates (C_U/C_L) for some common wavelets in 1D. These estimates (see Appendix A) are computed through combinatorial optimization and are hard to compute for higher dimensions. Without explicit calculation, we cannot say how the bounds will change for a wavelet as the dimension increases. The estimate formulas do indicate that wavelets with small support and fast decay will have a high C_L. C_U will be low if the wavelet has a small absolute value maximum.

In higher dimensions, it is easier to choose wavelets empirically. We measured the error of wavelet EMD with respect to actual EMD for a set of 100 random 16×16 histogram pairs. Since uniform random histogram pairs tend to have EMD concentrated in a small range, we instead generated only one histogram randomly. The second histogram was obtained by changing this at random locations by random amounts. The number of locations as well as maximum allowed change at a location was gradually increased. These random histogram pairs have well distributed EMDs. Wavelet EMD was scaled to make its mean ratio with EMD 1. Table (3) shows the normalized RMS error and the observed bounds ratio C_U/C_L. The bounds ratio is the maximum of all the ratios WEMD/EMD and EMD/WEMD, while the normalized RMS error is the RMS deviation of the ratio WEMD/EMD from 1. The table also notes the time needed to compute wavelet EMD in MATLAB R2007a on an Intel Xeon 3GHz PC. This can be improved if optimized wavelet transform implementations are used. We observed that Coiflets of order 3 and symmetric Daubechies wavelets of order 5 produced good results. We use order 3 coiflets in our experiments.

5.3 Image retrieval: colour histograms

We tested wavelet EMD on content based image retrieval using colour histograms. We used the SIMPLIcity test database [22] that consists of 10 image classes with 100 images each. We will show that wavelet EMD provides a better approximation to EMD than other EMD approximation methods in terms of distance values as well as performance for colour histograms.

Wavelet	Normalized RMS error	Bounds ratio C_U/C_L	Time (ms)
db3	16%	1.91	28
db4	20%	2.45	36
db5	17%	1.98	43
db6	18%	1.93	49
sym3	16%	1.91	28
sym4	17%	2.18	31
sym5	**13%**	**1.50**	**34**
sym6	16%	2.00	44
coif1	16%	1.88	34
coif2	15%	1.85	45
coif3	**14%**	**1.87**	**74**
oj8	20%	2.44	37
oj10	18%	2.07	39
oj12	17%	1.82	43

Table 3: EMD approximation error for random 16×16 histograms for various wavelets

Method	Bounds ratio	Normalized RMS error	Preproc. time	Compare time
EMD	–	–	0.92 s	63 ms
Wavelet EMD	7.03	18%	2.35 s	0.11 ms
Indyk-Thaper	11.00	43%	0.51 s	22 ms

Table 4: Error and time requirements for 16x16x16 colour histograms. Preprocessing time includes colour space conversion, binning, clustering (EMD only) and weighted wavelet transform (WEMD). Indyk-Thaper random embedding is repeated 5 times.

We computed $16 \times 16 \times 16$ colour histograms in *Lab* colour space since Euclidean (ground) distances in this colour space are proportional to perceived colour differences. The histograms were clustered into 64 clusters each before computing EMD, but not for computing approximations.

The scatter plots in figure (2) compare the wavelet EMD approximation with that of Indyk and Thaper [10] for distances computed between these colour histograms. Both approximations are scaled to have a mean ratio with EMD of 1. The plot indicates that Wavelet EMD distances correlate better with EMD than Indyk and Thaper. Note that EMD and its approximations have a maximum value depending on the histogram size. The Indyk-Thaper scatter plot appears cut–off because its greater spread causes it to reach this limit faster. Table (4) shows the approximation errors and time requirements for EMD, wavelet EMD and Indyk and Thaper's method. Although wavelet EMD needs more preprocessing time than the other two methods, actual comparison is very fast.

Another method to measure approximation error, in the context of feature matching, is to measure the probability of distance order reversal, i.e. the probability that histogram p_1 is closer to histogram p_2 than to histogram p_3 according to EMD, but not according to an approximation. We expect this probability to decrease as p_3 moves farther away from p_1, compared to p_2, i.e. the ratio $EMD(p_1, p_3)/EMD(p_1, p_2)$ increases. Figure (3) shows that this probability starts lower and falls off faster for wavelet EMD than for Indyk and Thaper's approximation. We do not include EMD–L_1 in these comparisons because it uses a different ground distance.

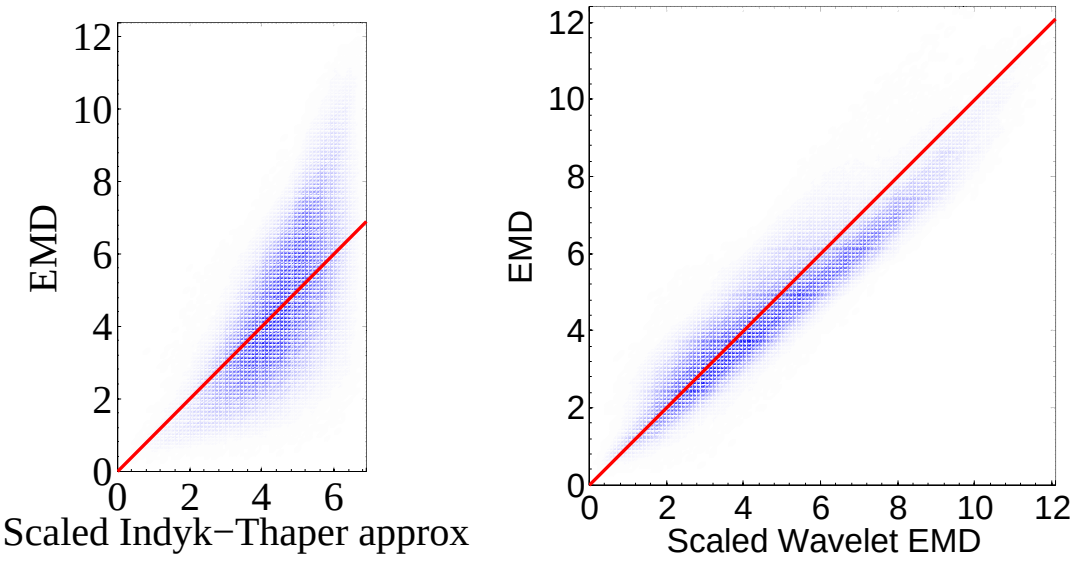

Figure 2: EMD approximations with Wavelet EMD using order 3 Coiflets is better than with Indyk and Thaper's [10] method. The red (dark) line indicates points of zero error.

Figure (4) shows ROC curves for EMD and its different approximation methods obtained from leave one out image retrieval experiments on this dataset. Wavelet EMD and EMD have almost the same performance, and this is better than EMD–L_1 and Indyk and Thaper's method.

6 Conclusion and future work

We have introduced a new method to approximate the earth mover's distance between two histograms using weighted wavelet transform coefficients of the difference histogram. We provide theoretical bounds to the maximum approximation error. Our experiments with colour histograms demonstrate that the wavelet EMD approximation preserves the performance of EMD while significantly reducing computation time.

In this paper, we have focussed our attention on approximation of EMD for full histograms. We would like to extend this to matching partial histograms as well. We also want to explore the use of different ground distances (different powers s) and other applications like image registration that can benefit from fast EMD computation.

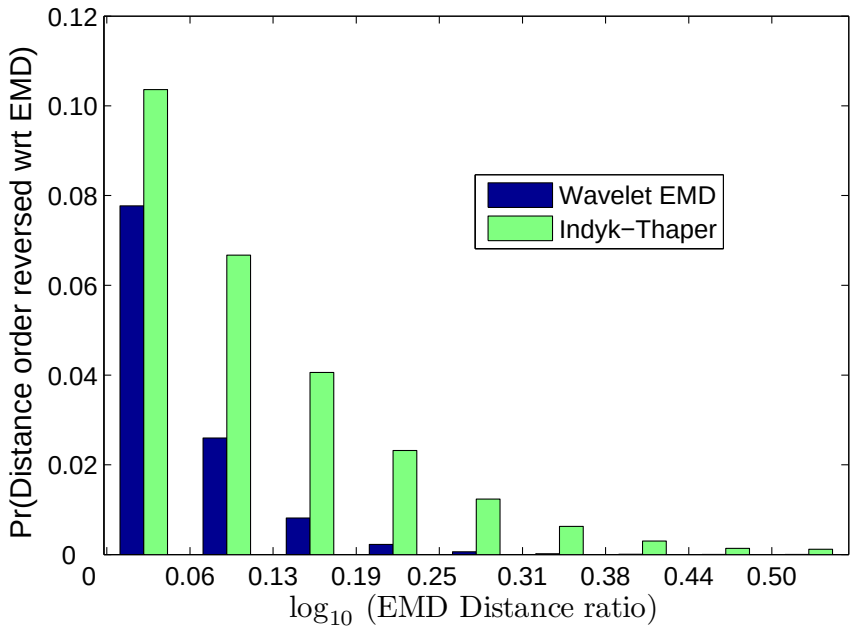

Figure 3: Wavelet EMD is less likely to disagree with EMD about ordering of histogram distances than Indyk-Thaper.

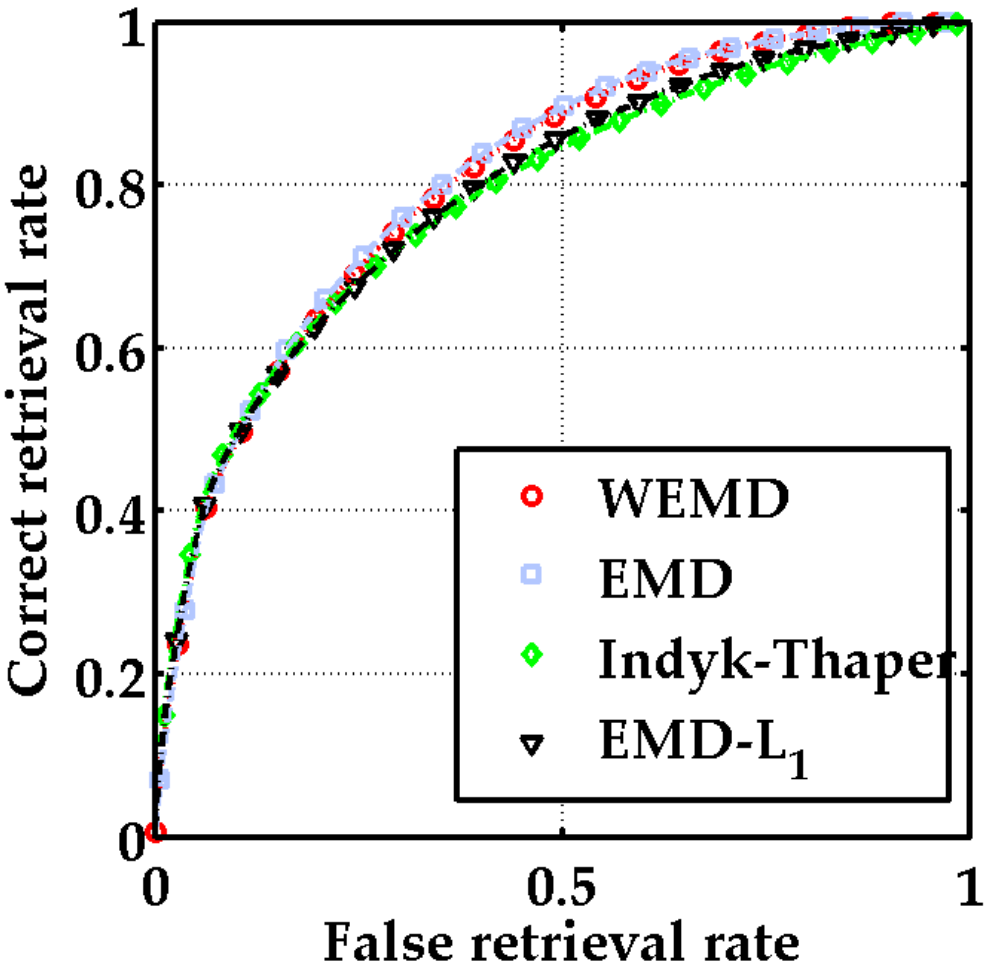

Figure 4: Colour histograms for content based image retrieval: wavelet EMD performance compared to other EMD methods

Appendix A Proof of Lemma (1)

Parts of this proof are adapted from Meyer ([19] section 6.4). We will start with the first inequality $a_{12}(\psi; s)C_1 \leq C$ in (16). The proof of this inequality corresponds to the proof of the *only if* part of theorem (1). For all functions $f \in C^s(\mathbb{R}^n), 0 < s \leq 1$ with the seminorm $C_H(f)$, we will compute bounds on their wavelet series coefficients. We will omit the dependence of C_H on f to simplify notation. Suppose that the wavelet coefficient bounds are actually attained. Using the definition of C_H, we can bound the values of $f(x)$ as :

$$|f(x) - f(k+r)| \leq C_H \|x - k - r\|^s \text{ for any } r \in \mathbb{R}^n, k \in \mathbb{Z}^n$$

Since the bounds are attained, we have

$$\begin{aligned}
C_0 &= \sup_k |f_k| = \sup_k \left| \int f(x)\phi(x-k)dx \right| \\
&= \sup_k \left| f(k+r) + \int (f(x) - f(k+r))\phi(x-k)dx \right| \qquad \left(\text{since} \int \phi(x-k)dx = 1 \right) \\
&\leq \sup_k |f(k+r)| + \int |f(x) - f(k+r)| \, |\phi(x-k)| \, dx \\
&\leq \sup_k |f(k+r)| + \int C_H \|x - k - r\|^s \, |\phi(x-k)| \, dx \\
&\leq ||f||_\infty + C_H \int \|x - r\|^s |\phi(x)| dx
\end{aligned}$$

Since this is true for all $r \in \mathbb{R}^n$,

$$C_0 \leq ||f||_\infty + C_H \inf_r \int \|x - r\|^s |\phi(x)| dx \tag{38}$$

If we define

$$a_{11}(\psi; s) := \frac{1}{\inf_r \int \|x - r\|^s |\phi(x)| dx}, \tag{39}$$

we can write this as

$$C_H \geq a_{11}(C_0 - ||f||_\infty) \tag{40}$$

Note that this constant depends only on the wavelet and s.

To compute a bound on C_1, we will first bound f_λ using the definition of C_H.

$$f(y) \leq f(2^{-j}(k+r)) + C_H \|y - 2^{-j}(k+r)\|^s$$

$$\begin{aligned}
|f_\lambda| &= \left| \int f(y)\psi_\lambda(y)dy \right| \\
&= \left| \int (f(y) - f(2^{-j}(k+r)))\psi_\lambda(y)dy \right| \\
&\leq \int \left| f(y) - f(2^{-j}(k+r)) \right| |\psi_\lambda(y)| \, dy \\
&\leq \int C_H \|y - 2^{-j}(k+r)\|^s \left| 2^{nj/2}\psi^\epsilon(2^j y - k) \right| dy
\end{aligned}$$

(using the n dimensional change of variables $x = 2^j y - k$, so $dx = 2^{nj} dy$)

$$\begin{aligned}
&= C_H \int \|x - r\|^s 2^{-js} 2^{-nj/2} \left|\psi^\epsilon(x)\right| dx \\
&= C_H 2^{-j(s+n/2)} \int \|x - r\|^s \left|\psi^\epsilon(x)\right| dx \\
\text{So } C_1 &= \sup_\lambda 2^{j(s+n/2)} \left|f_\lambda\right| \\
&\leq \sup_{j,\epsilon} 2^{j(s+n/2)} C_H 2^{-j(s+n/2)} \int \|x - r\|^s \left|\psi^\epsilon(x)\right| dx
\end{aligned} \tag{41}$$

Since this is true for all $r \in \mathbb{R}^n$,

$$C_1 \leq C_H \max_\epsilon \inf_r \int \|x - r\|^s \left|\psi^\epsilon(x)\right| dx \tag{42}$$

If we define

$$a_{12} := \frac{1}{\max_\epsilon \inf_r \int \|x - r\|^s \left|\psi^\epsilon(x)\right| dx}, \tag{43}$$

we can write this as

$$C_H \geq a_{12} C_1 \tag{44}$$

This constant too depends only on the wavelet and s.

From equations (40) and (44), we have

$$C_H \geq \max\left\{a_{11}(C_0 - ||f||_\infty), a_{12} C_1\right\} \tag{45}$$

If the bounds on the wavelet coefficients of f are not attained, we can instead say that

$$C_H \leq C \text{ such that } C \geq \max\left\{a_{11}(C_0 - ||f||_\infty), a_{12} C_1\right\} \tag{46}$$

Since its hard to know $||f||_\infty$ beforehand, we can simply use the looser bound (44),

$$C_H \leq C \text{ and } C \geq a_{12} C_1 \tag{47}$$

This is our first inequality.

Proving the second inequality is a bit more involved. This corresponds to the proof of the *if* part of theorem (1). We need to look at the converse problem: given a function defined by a wavelet series with approximation and detail coefficients bounded by C_0 and C_1 respectively, what is the corresponding bound on C_H ?

We start with the wavelet series of f

$$f(x) = \sum_k f_k \phi(x - k) + \sum_{\lambda \in \Lambda_j, j \geq 0} f_\lambda \psi_\lambda(x)$$

and split this into a Littlewood-Paley type series as

$$f(x) = \sum_{j >= -1} f_j(x) \tag{48}$$

$$\text{with } f_{-1}(x) = \sum_k f_k \phi(x - k) \tag{49}$$

$$\text{and } f_j(x) = \sum_{\lambda \in \Lambda_j} f_\lambda \psi_\lambda(x) \qquad \text{for } j \geq 0 \tag{50}$$

To begin with, we will establish some properties of the functions f_j. Consider the wavelet series $\Sigma\psi(x;\eta) := \sum_{k,\epsilon} \eta_{k,\epsilon}\psi^{(\epsilon)}(x-k)$ with $-1 \leq \eta_{k,\epsilon} \leq 1$. This is a convergent series because of the fast decay properties of wavelets. So,

$$||\Sigma\psi||_\infty := \sup_{x,\eta} |\Sigma\psi(x;\eta)| \tag{51}$$

is finite. This quantity can be computed for wavelets with compact support using combinatorial optimization if we note that the supremum will occur at $\eta_{k,\epsilon} \in \{-1,+1\}$. If we have $|f_\lambda| \leq C_1 2^{-j(s+n/2)}$, then

$$\begin{aligned} |f_j(x)| &\leq C_1 2^{-j(s+n/2)} 2^{nj/2} ||\Sigma\psi||_\infty \qquad \text{for all } x \\ \text{So, } ||f_j||_\infty &\leq C_1 ||\Sigma\psi||_\infty 2^{-js} \end{aligned} \tag{52}$$

With a similar argument, we get

$$||f_{-1}||_\infty \leq C_0 ||\Sigma\phi||_\infty \tag{53}$$

where $||\Sigma\phi||_\infty$ is defined similar to $||\Sigma\psi||_\infty$.

Now we can immediately bound $\|f\|_\infty$ as

$$\begin{aligned} \|f\|_\infty &\leq C_0 ||\Sigma\phi||_\infty + \sum_{j\geq 0} C_1 ||\Sigma\psi||_\infty 2^{-js} \\ \|f\|_\infty &\leq C_0 ||\Sigma\phi||_\infty + \frac{C_1}{1-2^{-s}} ||\Sigma\psi||_\infty \end{aligned} \tag{54}$$

Next, we will look at the first derivatives of the functions f_j. Since the wavelets have at least one derivative, we have for first derivatives with respect to all the components x_i $(i = 1,\ldots,n)$ of x :

$$\partial_{x_i} f_{-1}(x) = \sum_k f_k \partial_{x_i}\phi(x-k) \tag{55}$$

$$\begin{aligned} \partial_{x_i} f_j(x) &= \sum_{\lambda\in\Lambda_j} f_\lambda \partial_{x_i} 2^{nj/2}\psi^\epsilon(2^j x - k) \\ &= \sum_{\lambda\in\Lambda_j} f_\lambda 2^{j(n/2+1)} (\partial_{x_i}\psi^\epsilon)(2^j x - k) \end{aligned} \tag{56}$$

Again using the fast decay properties of wavelet derivatives, we can define the following convergent series and their absolute suprema :

$$\Sigma\phi^{(i)}(x;\eta) := \sum_k \eta_k \partial_{x_i}\phi(x-k) \qquad \|\Sigma\phi^{(i)}\|_\infty := \sup_{x,\eta} |\Sigma\phi^{(i)}(x;\eta)| \tag{57}$$

$$\Sigma\psi^{(i)}(x;\eta) := \sum_{k,\epsilon} \eta_{k,\epsilon} \partial_{x_i}\psi^{(\epsilon)}(x-k) \qquad \|\Sigma\psi^{(i)}\|_\infty := \sup_{x,\eta} |\Sigma\psi^{(i)}(x;\eta)| \tag{58}$$

Also, the Hölder space embedding $C^1 \subset C^s$ (every continuously differentiable function is Hölder continuous) for $s < 1$ implies that the series $\sum \phi(x;\eta) \in C^s$. We define

$$\left|\left|\sum \phi^s(x)\right|\right|_\infty := \sup_{x\neq y} \frac{\sum \phi(x;\eta) - \sum \phi(y;\eta)}{||x-y||^s} \tag{59}$$

Now we can bound the derivatives of f_j as :

$$\begin{aligned} |\partial_{x_i} f_j(x)| &\leq C_1 2^{-j(s+n/2)} 2^{j(n/2+1)} \|\Sigma\psi^{(i)}\|_\infty \qquad \text{for all } x \\ \text{So, } ||\partial_{x_i} f_j||_\infty &\leq C_1 \|\Sigma\psi^{(i)}\|_\infty 2^{-j(s-1)} \end{aligned} \tag{60}$$

Similarly, we also get

$$||\partial_{x_i} f_{-1}||_\infty \leq C_0 \|\Sigma\phi^{(i)}\|_\infty \tag{61}$$

Finally, we have everything we need to estimate C_H. Define $r_j(x;x_0) := f_j(x) - f_j(x_0)$ and $r(x;x_0) := f(x) - f(x_0) = \sum_j r_j(x;x_0)$, for any $x_0 \in \mathbb{R}^n$. Then, we need to find C_H s.t $|r(x;x_0)| \leq C_H \|x - x_0\|$. Let $m \in \mathbb{Z}$ be defined by $2^{-m} \leq \|x - x_0\| < 2.2^{-m}$. We can split the series for $r(x;x_0)$ as

$$r(x;x_0) = r_{-1}(x) + \sum_{j=0}^{m-1} r_j(x;x_0) + \sum_{j \geq m} r_j(x;x_0) \tag{62}$$

We have the following two cases :

Case 1: $\|x - x_0\| < 1$ so that $m > 0$
Starting with the last term of equation (62), we have :

$$\begin{aligned} \left| \sum_{j \geq m} r_j(x;x_0) \right| &\leq \sum_{j \geq m} |f_j(x)| + |f_j(x_0)| \\ &\leq \sum_{j \geq m} 2C_1 \|\Sigma\psi\|_\infty 2^{-js} \qquad \text{(from equation (52))} \\ &= 2C_1 \|\Sigma\psi\|_\infty \frac{2^{-ms}}{1 - 2^{-s}} \\ &\leq \frac{2C_1 \|\Sigma\psi\|_\infty}{1 - 2^{-s}} \|x - x_0\|^s \end{aligned} \tag{63}$$

This holds for $s = 1$ as well. To deal with the middle term of equation (62), we use the mean value theorem to bound each r_j.

$$\begin{aligned} |r_j(x;x_0)| &= \left| \sum_{k=1}^{n} (x_k - x_{0k}) \frac{\partial f_j}{\partial x_k}(x') \right| \qquad \text{(for some } x' \text{ between } x \text{ and } x_0) \\ &\leq \sum_i \|x - x_0\| \cdot \|\partial_{x_i} f_j\|_\infty \\ &\leq C_1 \sum_i \|\Sigma\psi^{(i)}\|_\infty 2^{j(1-s)} \|x - x_0\| \qquad \text{(from equation (60))} \end{aligned}$$

$$\begin{aligned} \text{So, } \left| \sum_{j=0}^{m-1} r_j(x;x_0) \right| &\leq C_1 \sum_i \|\Sigma\psi^{(i)}\|_\infty \sum_{j=0}^{m-1} 2^{j(1-s)} \|x - x_0\| \\ &= C_1 \sum_i \|\Sigma\psi^{(i)}\|_\infty \frac{2^{m(1-s)} - 1}{2^{1-s} - 1} \|x - x_0\| \end{aligned}$$

Now $2^{m-1} < \|x - x_0\|^{-1}$ implies $2^{m(1-s)}\|x - x_0\| < 2^{s-1}\|x - x_0\|^s$. So we get

$$\left|\sum_{j=0}^{m-1} r_j(x; x_0)\right| \leq \frac{C_1 \sum_i \|\Sigma\psi^{(i)}\|_\infty}{2^{(1-s)} - 1}(2^{s-1}\|x - x_0\|^s - \|x - x_0\|) \tag{64}$$

We cannot use this bound for $s = 1$. In that case, since we are adding up m terms with the same bound for each, we can use the fact that $m \leq 1 - \log_2\|x - x_0\|$ to get

$$\left|\sum_{j=0}^{m-1} r_j(x; x_0)\right| \leq C_1 \sum_i \|\Sigma\psi^{(i)}\|_\infty (1 - \log_2\|x - x_0\|)\|x - x_0\| \tag{65}$$

We can bound the first term of equation (62) using the Hölder norm bound from equation (59) as :

$$|r_{-1}(x)| \leq C_0 \left|\left|\sum \phi^s(x)\right|\right|_\infty \|x - x_0\|^s \tag{66}$$

Now we add the three terms from equations (66), (64), (63) to get

$$|r(x; x_0)| \leq \left(C_0 \left|\left|\sum \phi^s(x)\right|\right|_\infty + C_1 \frac{\sum_i \|\Sigma\psi^{(i)}\|_\infty}{2^{1-s}(2^{1-s} - 1)} + C_1 \frac{2\|\Sigma\psi\|_\infty}{1 - 2^{-s}}\right) \|x - x_0\|^s \tag{67}$$

For $s = 1$, we can add up everything to get

$$\begin{aligned} |r(x; x_0)| \leq\ & C_0 \sum_i \|\Sigma\phi^{(i)}\|_\infty \|x - x_0\| \\ & + \|\Sigma\psi^{(i)}\|_\infty (1 - \log_2\|x - x_0\|)\|x - x_0\| \\ & + 4C_1\|\Sigma\psi\|_\infty \|x - x_0\| \end{aligned} \tag{68}$$

The log term indicates that the wavelet coefficient decaying at the rate of $2^{-j(1+n/2)}$ is insufficient to restrict functions to the space C^1. Instead, this condition restricts functions to the Zygmund class $\Lambda_\star$, which includes some extra functions.

Case 2: $\|x - x_0\| \geq 1$ so that $m \leq 0$
The only change here is that the middle term disappears in equations (67) and (68).

Combining these two cases, for $s < 1$, we get the bound :

$$C_H \leq C_0 \left|\left|\sum \phi^s(x)\right|\right|_\infty + C_1 \frac{\sum_i \|\Sigma\psi^{(i)}\|_\infty}{2^{1-s}(2^{1-s} - 1)} + C_1 \frac{2\|\Sigma\psi\|_\infty}{1 - 2^{-s}} \tag{69}$$

If we define

$$a_{21}(\psi; s) := \left|\left|\sum \phi^s(x)\right|\right|_\infty \text{ and} \tag{70}$$

$$a_{22}(\psi; s) := \frac{\sum_i \|\Sigma\psi^{(i)}\|_\infty}{2^{1-s}(2^{1-s} - 1)} + \frac{2\|\Sigma\psi\|_\infty}{1 - 2^{-s}}, \tag{71}$$

we have the second inequality for $0 < s < 1$:

$$C_H \leq C \text{ and } C \leq a_{21}(\psi; s)C_0 + a_{22}(\psi; s)C_1 \tag{72}$$

If we restrict ourselves to case 2, i.e. $\|x - x_0\| \geq 1$, this inequality is still valid for $s = 1$ with the change that :

$$a_{22}(\psi; s) := \frac{2\|\Sigma\psi\|_\infty}{1 - 2^{-s}} \tag{73}$$

The bounds ratios in table (2) were calculated for 1D discrete distributions using this formula with $s = 1$.

From equations (72) and (47), we have the bounds in the lemma :

$$C_H \leq C \text{ and } a_{12}(\psi; s)C_1 \leq C \leq a_{21}(\psi; s)C_0 + a_{22}(\psi; s)C_1 \tag{74}$$

□

Appendix B WEMD with biorthogonal wavelets

Theorem (2) holds in a slightly changed form for biorthogonal wavelets as well. In the auxiliary wavelet domain problem (20), we can keep the constraint, but we have to change the objective function since biorthogonal wavelets don't preserve inner products. Since these wavelets are not orthonormal, the analysis (ϕ, ψ) and synthesis $(\tilde{\phi}, \tilde{\psi})$ scaling function and wavelet are different. They are related by the following biorthogonal relationship :

$$\int \phi(x-k)\tilde{\phi}(x-l) = \delta_{kl} \qquad \int \psi_\lambda(x)\tilde{\phi}(x-l) = 0$$

$$\int \psi_\lambda(x)\tilde{\psi}_\mu(x) = \delta_{\mu\lambda} \qquad \int \phi(x-k)\tilde{\psi}_\mu(x) = 0$$

The wavelet coefficients of a function in a biorthogonal wavelet series expansion are given by :

$$f_k = \int f(x)\phi(x-k)dx \qquad f_\lambda = \int f(x)\psi_\lambda(x)dx \tag{75}$$

and the function can be reconstructed as :

$$f(x) = \sum_k f_k\tilde{\phi}(x-k) + \sum_\lambda f_\lambda\tilde{\psi}_\lambda(x) \tag{76}$$

We can use equation (76) to compute the inner product of two functions.

$$\int f(x)p(x)dx = \int \left(\sum_k f_k\tilde{\phi}(x-k) + \sum_\lambda f_\lambda\tilde{\psi}_\lambda(x)\right)\left(\sum_l p_l\tilde{\phi}(x-l) + \sum_\mu p_\mu\tilde{\psi}_\mu(x)\right)dx$$

Let $\tilde{\theta}_\omega(x) := \tilde{\phi}(x-k)$ or $\tilde{\psi}_\lambda(x)$, i.e. the function $\tilde{\theta}$ represents either $\tilde{\phi}$ or $\tilde{\psi}$ and the index ω first runs over all k and then over all λ.

$$\begin{aligned} &= \sum_{\omega,\sigma} f_\omega p_\sigma \int \tilde{\theta}_\omega(x)\tilde{\theta}_\sigma(x)dx \\ &= \mathbf{f}^T\mathbf{U}\mathbf{p} \end{aligned} \tag{77}$$

where $\mathbf{f}$ and $\mathbf{p}$ are vectors of wavelet coefficients as before and

$$U_{\omega\sigma} := \int \tilde{\theta}_\omega(x)\tilde{\theta}_\sigma(x)dx \tag{78}$$

Thus the auxiliary wavelet domain problem now becomes :

$$\begin{aligned} &\text{Maximize } \mathbf{f}^T\mathbf{U}\mathbf{p} \\ &\text{subject to } |f_k| \le C_0 \quad \text{and} \quad |f_\lambda| \le C_1 2^{-j(s+n/2)} \end{aligned} \tag{79}$$

This is the same problem as before, except that we must change $\mathbf{p}$ to $\tilde{\mathbf{p}} := \mathbf{U}\mathbf{p}$. The solution is :

$$\hat{\mu}_s = C_0 \sum_k |\tilde{p}_k| + C_1 \sum_\lambda 2^{-j(s+n/2)}|\tilde{p}_\lambda| \tag{80}$$

If we set $C_0 = 0$ and $C_1 = 1$, we get the simplified formula :

$$d(p)_{wemd} := \sum_\lambda 2^{-j(s+n/2)}|\tilde{p}_\lambda| \tag{81}$$

Computing WEMD with biorthogonal wavelets will take a bit longer because we need to compute $\mathbf{U}\mathbf{p}$. This raises the overall complexity to $O(n^2)$, though we do not expect it to increase computation time significantly since matrix multiplication has much lower complexity constants than the fast wavelet transform. Although the matrix $\mathbf{U}$ is not sparse ($O(n)$ non-zeros), a lot of its elements are still zeros, and the rest can be precomputed and stored.

An advantage of using biorthogonal wavelets is that we can have wavelets with tighter bounds. The constant a_{12} depends on the analysis wavelet while a_{21} and a_{22} depend on the synthesis wavelet. Since biorthogonal wavelets offer more freedom in choosing these two, we can expect wavelets with lower bounds ratios

$$\frac{C_U}{C_L} = \frac{a_{22}(\tilde{\psi}; s)}{a_{12}(\psi; s)}. \tag{82}$$

References

[1] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape context. *IEEE Transactions on PAMI*, 24(4):509–522, Apr 2002. 1

[2] C. Chefd'hotel and G. Bousquet. Intensity-based image registration using EMD. In *Medical Imaging 2007: Image Proc. Proc. of the SPIE*, volume 6512, Mar. 2007. 1

[3] K. Grauman and T. Darrell. Fast contour matching using approximate earth mover's distance. In *IEEE Conference on CVPR*, volume 01, pages 220–227, 2004. 1

[4] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *IEEE ICCV*, pages 1458–1465, 2005. 4

[5] K. Guittet. Extended kantorovich norms : a tool for optimization. Technical Report 4402, INRIA, March 2002. 9

[6] S. Haker, L. Zhu, A. Tannenbaum, and S. Angenent. Optimal mass transport for registration and warping. *International Journal of Computer Vision*, 60:225–240, December 2004. 1

[7] L. Hanin. An extension of the kantorovich norm. *Contemporary Mathematics*, 226:113–130, 1997. 10, 11

[8] A. Holmes, C. Rose, and C. Taylor. Measuring similarity between pixel signatures. *Image and Vision Computing*, 20(4):239–248, April 2002. 1, 4

[9] A. Holmes, C. Rose, and C. Taylor. Transforming pixel signatures into an improved metric space. *Image and Vision Computing*, 20(9):701–707(7), August 2002. 1, 4

[10] P. Indyk and N. Thaper. Fast image retrieval via embeddings. In *3rd International Workshop on Statistical and Computational Theories of Vision (at ICCV)*, 2003. 3, 14, 15

[11] A. E. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3d scenes. *IEEE Transactions on PAMI*, 21(5):433–449, 1999. 1

[12] B. Korte and J. Vygen. *Combinatorial optimization: Theory and Algorithms*. Springer, 2000. 3

[13] S. Lazebnik, C. Schmid, and J. Ponce. A sparse texture representation using affine-invariant regions. *IEEE Transactions on PAMI*, 27(8):1265–1278, August 2005. 1

[14] E. Levina and P. Bickel. The earth movers distance is the mallows distance: Some insights from statistics. In *IEEE ICCV*, pages 251–256, 2001. 4

[15] H. Ling and K. Okada. Diffusion distance for histogram comparison. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 246–253, 2006. 4

[16] H. Ling and K. Okada. An efficient earth movers distance algorithm for robust histogram comparison. *IEEE Transactions on PAMI*, 29(5):840–853, May 2006. 3

[17] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 1

[18] S. Mallat. *A wavelet tour of signal processing*. Academic Press, second edition, 1998. 8, 12

[19] Y. Meyer. *Wavelets and Operators, Vol 1*. Cambridge university press, 1992. 3, 6, 18

[20] S. T. Rachev and L. Rüschendorf. *Mass Transportation Problems, Vol 1: Theory*. Springer, 1998. 1, 4

[21] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40:99–121, Nov 2000. 1, 3, 9

[22] J. Z. Wang, J. Li, and G. Wiederhold. Simplicity: Semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on PAMI*, 23(9):947–963, 2001. 13

[23] M. Werman, S. Peleg, and A. Rosenfeld. A distance metric for multidimensional histograms,. *Computer Vision, Graphics and Image Processing*, 32:328–336, December 1985. 3

[24] X.-P. Zhang, L.-S. Tian, and Y.-N. Peng. From the wavelet series to the discrete wavelet transform – The initialization. *IEEE Trans. on signal proc.*, 44(1):129–133, 1996. 12