Approximate earth mover's distance in linear time

Sameer Shirdhonkar and David W. Jacobs
Center for Automation Research, University of Maryland, College Park, USA
{ssameer,djacobs}@umiacs.umd.edu

Abstract

The earth mover's distance (EMD) [16] 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.

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 [12] and 3D object recognition [7]. Colour and texture histograms [16] are also used for content based image retrieval. These descriptors are often compared using binwise dissimilarity measures like Euclidean norm 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 have large values. The popular SIFT descriptor [12] 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. For perceptually meaningful ground distances, EMD agrees with perceptual dissimilarity better than other measures [16]. EMD has been successfully used for image retrieval by comparing colour and texture histograms [16], contour matching [3], image registration [2] and pattern matching in medical images [5]. Ling and Okada [11] report improved performance when comparing various histogram descriptors with EMD over the χ^2 statistic and the L_2 norm. 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 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

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

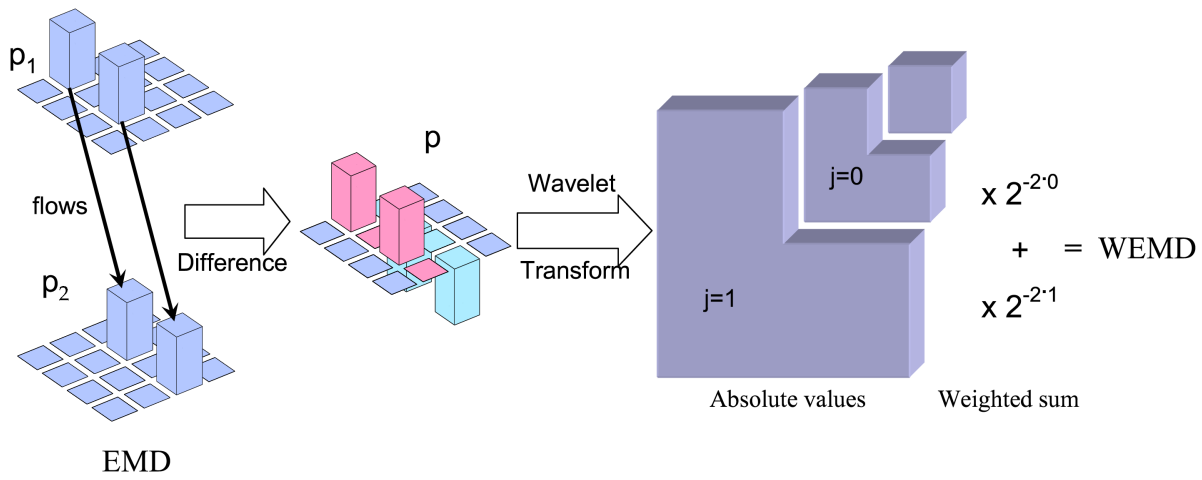

Figure 1. Computation of wavelet EMD

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 [15] is the corresponding problem for continuous distributions. Both problems admit duals with the same optimal value. The important insight in our approximation 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, the dual EMD is given by :

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

subject to the constraint that any two bin potentials cannot differ by more than 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 a (non-vertical) straight line at any point. This is a Hölder 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 has the 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 wavelet 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. For a single wavelet, the mass to be moved is proportional to the volume of $|\psi_j(x)|$, i.e. to $2^{-jn/2}$. The distance travelled is proportional to the span of the wavelet 2^{-j} (According to Meyer's [14] convention, a wavelet at scale j is the mother wavelet squeezed 2^j times.) The sum of all distances is an approximation to EMD and is thus given 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

EMD for signatures [16]	Discrete EMD for histograms	Continuous EMD for distributions
Signatures $f(i;1), f(i,2)$ In general, $\sum_i f(i;1) \neq \sum_i f(i;2)$	Histograms $f(i;1), f(i,2)$ $\sum_i f(i;1) = \sum_i f(i;2) = 1$ Difference $f(i) := f(i;1) - f(i;2)$	Distributions $p_1(x), p_2(x)$ $\int p_1(x)dx = \int p_2(x)dx = 1$ Difference $p(x) := p_1(x) - p_2(x)$
Ground distance $d_{ij} \geq 0$ Flow (from bin i to bin j) $g_{ij} \geq 0$	Ground distance $d_{ij} \geq 0$ Flow (from bin i to bin j) $g_{ij} \geq 0$ Potential π_i	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

approximation to the actual distance.

2. Related Work

The earth movers distance was introduced in vision by Werman *et al.* [19], though they did not use this name. Rubner *et al.* [16] 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 ([8] 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 ([8] 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 [11] 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 [6] 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 linear time as well, our algorithm has deterministic error bounds. We will also show empirically that our algorithm is more accurate.

The diffusion distance introduced by Ling and Okada in [10] 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 [5] 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 [15] first formulation of the problem as a linear program and the study of its duality in 1942. Minimal l_1 metric, Kantorovich metric [15], Wasserstein distance and Mallows distance [9] are equivalent formulations of EMD and are computed in the same way. 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. Note that discrete EMD is a special case of continuous EMD since a discrete distribution can be represented as a set of delta functions. Consequently, our results for continuous EMD are valid for discrete EMD with histograms as well. The problem is simpler for histograms than for signatures. 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 [15], 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 $S \subset \mathbb{R}^n$. 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. The Kantorovich-Rubinstein transshipment problem (KRP) is to find

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

where the infimum is over all joint probability densities 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 *Kantorovich–Rubinstein theorem* states that the problem admits the dual representation :

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

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

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

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

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 ([14] 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 [14] Chapter 2) as:

$$f(x) = \sum_k f_k \phi(x-k) + \sum_\lambda f_\lambda \psi_\lambda(x) \tag{7}$$

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

$$\psi_\lambda(x) := 2^{nj/2} \psi^\epsilon(2^j x - k) \tag{8}$$

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

$$f_k = \int f(x) \bar{\phi}(x-k) dx, \quad k \in \mathbb{Z}^n \tag{9}$$

$$f_\lambda = \int f(x) \bar{\psi}_\lambda(x) dx, \quad \lambda \in \Lambda, \quad j \geq 0 \tag{10}$$

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

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 Hölder class $C^s(\mathbb{R}^n)$ if the following supremum exists and is finite :

$$C_H(f) := \sup_{x \neq y} \frac{|f(x) - f(y)|}{\|x-y\|^s} \tag{11}$$

We can now state the constraint (6) simply as

$$C_H(f) < 1 \tag{12}$$

The following theorem from Meyer ([14] 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*

$$\begin{aligned} &|f_k| \leq C_0, \quad k \in \mathbb{Z}^n \quad \textit{and} \\ &|f_\lambda| \leq C_1 2^{-j(n/2+s)}, \quad \lambda \in \Lambda_j, \quad j \geq 0 \end{aligned} \tag{13}$$

for some constants C_0 and C_1.

A little modification to the proof of this theorem (see [17]) gives the following lemma:

Lemma 1. *For $0 < s < 1$, if the wavelet series coefficients of the function f are bounded as in (13), then $f \in C^s$ with $C_H(f) < C$ such that*

$$a_{12}(\psi; s)C_1 \leq C \leq a_{21}(\psi; s)C_0 + a_{22}(\psi; s)C_1 \tag{14}$$

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

$$C_H(f) := \sup_{|x-y|\geq 1} \frac{|f(x) - f(y)|}{\|x - y\|^s}, \tag{15}$$

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

The constants a_{12}, a_{21} and a_{22} are estimated in the technical report [17]. 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 $C_1 > 0$,*

$$\hat{\mu}_c = C_0 \sum_k |p_k| + C_1 \sum_\lambda 2^{-j(s+n/2)} |p_\lambda| \tag{16}$$

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

$$C_L \hat{\mu}_c \leq \dot{\mu}_c \leq C_U \hat{\mu}_c \tag{17}$$

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

Proof. Consider the auxiliary wavelet domain problem :

$$\begin{aligned} &\text{Maximize } \mathbf{p}^t\mathbf{f} = \sum_k p_k f_k + \sum_\lambda p_\lambda f_\lambda \\ &\text{subject to} \quad |f_k| \leq C_0 \quad \text{and} \quad |f_\lambda| \leq C_1 2^{-j(s+n/2)} \end{aligned} \tag{18}$$

$\mathbf{p}$ and $\mathbf{f}$ are coefficient vectors of p_λ and f_μ. It is easy to see that $\hat{\mu}_c$ in (16) is the solution of this problem. We need to show that the ratio of the optimal values of the two problems are bounded by two constants C_L and C_U. Since we use orthonormal wavelets that preserve inner products, the wavelet problem (18) 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 (18) will allow functions with $C_H(f) < C$, where C is bounded by the limits in (14). So, all functions with $C_H(f)$ less than the lower bound above 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 (14). This is equivalent to (17) with

$$\begin{aligned} C_L &= a_{12}(\psi; s)C_1 \quad \text{and} \\ C_U &= a_{21}(\psi; s)C_0 + a_{22}(\psi; s)C_1. \end{aligned} \tag{19}$$

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 (19) are valid again and we have the required equivalence. □

A similar but more complex result holds for biorthogonal wavelets as well. See [17] for details. We set $C_0 = 0$ because this gives us the tightest bounds in (14). Setting the constant C_1 to 1, we get the simple distance measure :

$$d(p)_{wemd} = \sum_\lambda |p_\lambda| 2^{-j(s+n/2)} \tag{20}$$

$$\text{The bounds ratio } \frac{C_U}{C_L} = \frac{a_{22}(\psi; s)}{a_{12}(\psi; s)} \tag{21}$$

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.

4. Experiments

First, in section (4.1), we will discuss some implementation issues that affect the accuracy and other aspects of

wavelet EMD. In section (4.2), we will describe how to choose appropriate wavelets. Finally, in section (4.3), We will describe experiments that demonstrate that the wavelet EMD behaves very similar to EMD, but can be computed much faster.

4.1. 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 (Manhattan) distance between these coefficients. We should note the following points while computing wavelet EMD :

1. Initialization: The standard Mallat filter bank algorithm ([13] 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, but not if we want to use wavelet coefficients to represent signal properties like Hölder continuity. We use the wavelet transform initialization method (algorithm 2) of Zhang, Tian and Peng [20]. 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 sparsity: Most wavelet transform 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.

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.

4. Histogram dimensionality: Although the computation time increases linearly with the number of bins N, it grows exponentially with the histogram dimension n. This method may become impractical for more than 4–5 dimensions. Further, any sparsity in the high dimensional histogram may be lost when computing the wavelet transform leading to increased space requirement. We restrict our experiments to histograms of dimensionality 1, 2 and 3 only.

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

4.2. 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 theoretical maximum error estimates (C_U/C_L) for some common wavelets in 1D. These estimates [17] are computed through combinatorial optimization and are hard to compute for higher dimensions. 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 using wavelet EMD error on random histogram pairs. Since uniformly random histogram pairs tend to have EMD concentrated in a small range, we instead generated only the first histogram randomly. The second histogram was obtained by changing this at random bins by random amounts. The number of changed bins and the maximum allowed change at a bin was gradually increased. These random histogram pairs have well distributed EMDs. Wavelet EMD was scaled to make $Mean(WEMD/EMD) = 1$. The estimated 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. Table 3 shows these two quantities and the computation time in MATLAB R2007a on an

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 histogram pairs for various wavelets

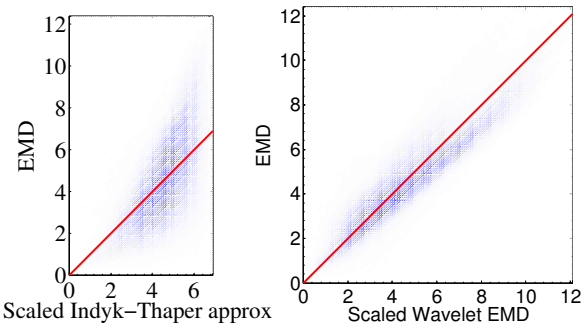

Figure 2. EMD approximations with Wavelet EMD using order 3 Coiflets is better than with Indyk and Thaper's [6] method. Darker shades indicate greater point density. The diagonal red (dark) line indicates points of zero error.

Intel Xeon 3GHz PC for a set of 100 random 16×16 histogram pairs. The time can be improved if optimized wavelet transform implementations are used. We observed that Coiflets of order 3 and symmetric Daubechies wavelets of order 5 had low errors. We use order 3 coiflets in our experiments.

4.3. Image retrieval: colour histograms

We tested wavelet EMD on content based image retrieval using colour histograms, since this is the most recognized application of EMD. We used the SIMPLIcity image database [18] of 10 image classes with 100 images each. We show that wavelet EMD provides a better approximation to EMD than other EMD approximation methods in terms of distance values as well as retrieval performance for colour histograms. 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. To reduce EMD computation time, each histogram was clustered to 64 clusters before computing EMD, similar to [16]. No clustering was done while computing approximations, since it was not necessary.

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

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.

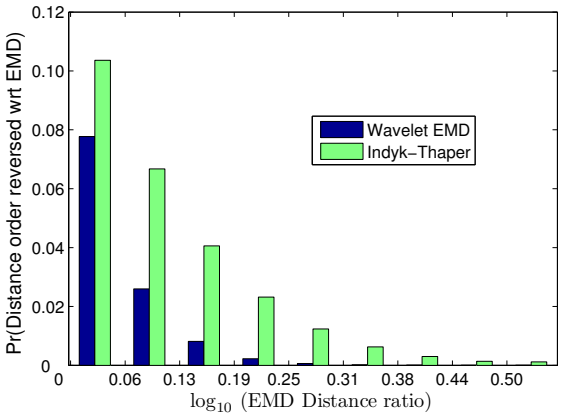

Figure 3. Wavelet EMD is less likely to disagree with EMD about ordering of histogram distances than Indyk-Thaper.

The scatter plots in figure 2 compare the wavelet EMD approximation with that of Indyk and Thaper [6] for distances computed between all colour histogram pairs in the dataset. 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. 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. Note that the normalized RMS error is 18% for wavelet EMD compared to 43% for Indyk-Thaper. Although wavelet EMD needs more preprocessing time than the other two methods, the actual comparison is very fast. For nearest neighbour searches, comparison time is far more important than preprocessing time. For example, in our 1000 image database, it will take $63.92s$ and $22.51s$ to retrieve the image most similar to a query image using EMD and the Indyk-Thaper approximation respectively. Using wavelet EMD, the closest image can be

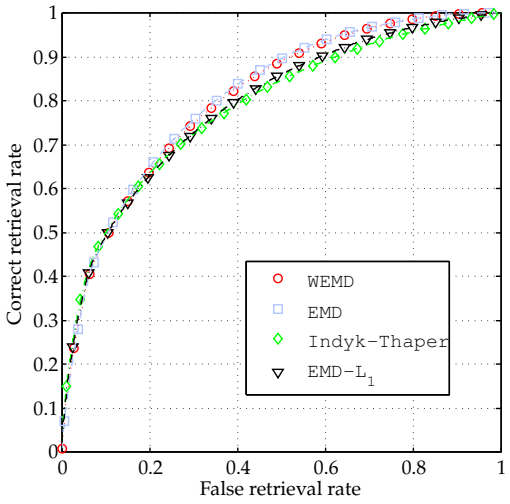

Figure 4. Colour histograms for content based image retrieval: wavelet EMD performance compared to other EMD methods

retrieved in merely $2.46s$.

Another method of measuring 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 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.

5. 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.

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.

[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.

[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.

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

[5] 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.

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

[7] 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.

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

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

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

[11] 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.

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

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

[14] Y. Meyer. *Wavelets and Operators, Vol 1*. Cambridge university press, 1992.

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

[16] 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.

[17] S. Shirdhonkar and D. W. Jacobs. Approximate earth mover's distance in linear time. Technical report, University of Maryland, College Park, MD, USA, 2008.

[18] 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.

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

[20] 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.