BIG DATA PROJECT REPORT

APPROXIMATE EARTH MOVER'S DISTANCE IN LINEAR TIME

D Sai Vineeth **K Mahidhar** **B Eshwanth**

CS14BTECH11011 **CS14BTECH11019** **CS14BTECH11012**

ABSTRACT

The earth mover's distance is an important perceptually meaningful metric for comparing histograms, but it suffers from high $O(N^3 \log N)$ computational complexity. The paper presents 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.

The paper proves that the resulting wavelet EMD metric is equivalent to EMD, i.e. the ratio of the two is bounded. The paper also provide estimates for the bounds. The weighted wavelet transform can be computed in linear time in the number of histogram bins.The paper experimentally show that wavelet EMD is a good approximation to EMD, has similar performance, but requires much less computation. Now we introduce the concept of EMD and go through the wavelet EMD and prove that wavelet EMD can be computed in linear time.

1. INTRODUCTION

HISTOGRAM DESCRIPTORS :

Histogram descriptors are a powerful representation for matching and recognition.They have been used extensively in vision applications like shape matching ,keypoint matching and 3D object recognition. Colour and texture histograms are also used for content based image retrieval. These descriptors are often compared using binwise dissimilarity measures like Euclidean norm.The popular SIFT descriptor is a gradient orientation location histogram. A similar histogram shifting will occur if the keypoint is not localized accurately.

EARTH MOVERS'S 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. 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.

WAVELET EMD :

The paper presents 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.The wavelet EMD metric can be computed in O(N) time.

In the primal form, the objective function is the total flow weighted ground distance between all bin pairs.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 function f assigned to each bin, subject to the constraint that any two bin potentials cannot differ by more than the ground distance.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 Holder continuity condition and is somewhat between
continuity and differentiability. The wavelet coefficients of a Holder 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 formula is:

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

Where , p is the n dimensional difference histogram and p_{lamda} are its wavelet coefficients.We call this as wavelet EMD.

2. RELATED WORK

2.1 EMD-L_1

- Various approximation algorithms have been suggested to speed up the computation of EMD
- EMD could be computed in $O(N^2)$ time if an L_1 ground distance is used instead of the usual Euclidean distance

2.2 DIFFUSION DISTANCE

- The diffusion distance is computed by constructing a Gaussian pyramid from the difference histogram and summing up the L_1 norms of the various levels
- Although this algorithm also computes EMD in O(N), it is not an approximation to the EMD and it is a different method unlike our algorithm

3. THEORY

Auxiliary Wavelet Domain Problem

$$\text{Maximize } \mathbf{p}^t\mathbf{f} = \sum_k p_k f_k + \sum_\lambda p_\lambda f_\lambda$$

Where,

$$|f_k| \leq C_0 \quad \text{and} \quad |f_\lambda| \leq C_1 2^{-j(s+n/2)}$$

Since we use orthonormal wavelets that preserve inner products, the wavelet problem has the same objective function as the initial dual problem

If f belongs to Holder Class, then C_h<C i.e

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

Changing the initial constraint from $C_h(f) < 1$ to $C_h(f) < C$, scales the optimal value by C.

According to Meyer's theorem, $a_{12}C_1 \leq C \leq a_{21}C_0 + a_{22}C_1$
Where, constants a_{12}, a_{21} and a_{22} depend on wavelet and s(0<s<1).

They set $C_0 = 0$ because it gives the tightest bounds and setting constant C_1 to 1,they get the simple distance measure

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

4. IMPLEMENTATION

- We used the SIMPLIcity image database of 10 classes with 100 images each.
- We split Wavelet EMD computation into two parts:
- First,the histogram descriptor is converted into wavelet domain and its coefficients are scaled according to the following equation.

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

- The wavelet EMD distance between two descriptors is the L1 distance between these coefficients.

We calculated Wavelet Emd for Daubechies, Symlet, Gaussian wavelets and compared the bound ratio (checks closeness of wavelet EMD to Actual Emd value).

Since in wavelet transform,we superimpose the wavelet on the histogram across different scales and shifts.We calculated wavelet coefficients across different scales parallely which is a improvement.

5. RESULTS :

- Various graphs are plotted showing the performance of the wavelet EMD algorithm with the actual EMD algorithm.

- It can be seen that the EMD value calculated by wavelet EMD algorithm is in the vicinity of the actual EMD value along with the advantage of taking linear time.

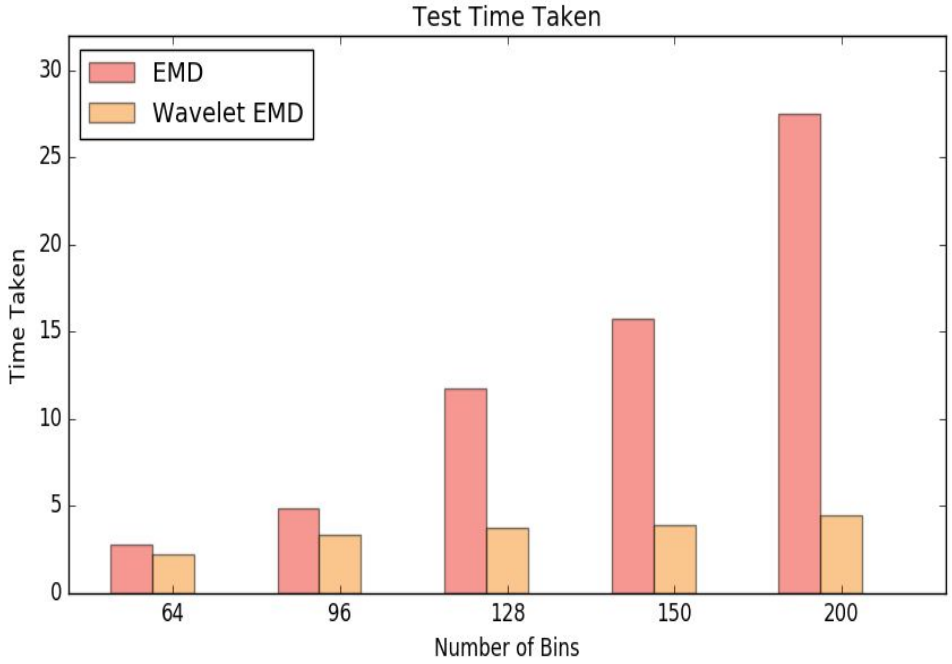
- Also, different wavelets were used to see which one performs better for our dataset.

COMPARING IMAGES :

Comparison	Actual EMD	Wavelet EMD
1st and 2nd image	476.8125	519.85625
1st and 3rd image	1497.0	1965.8

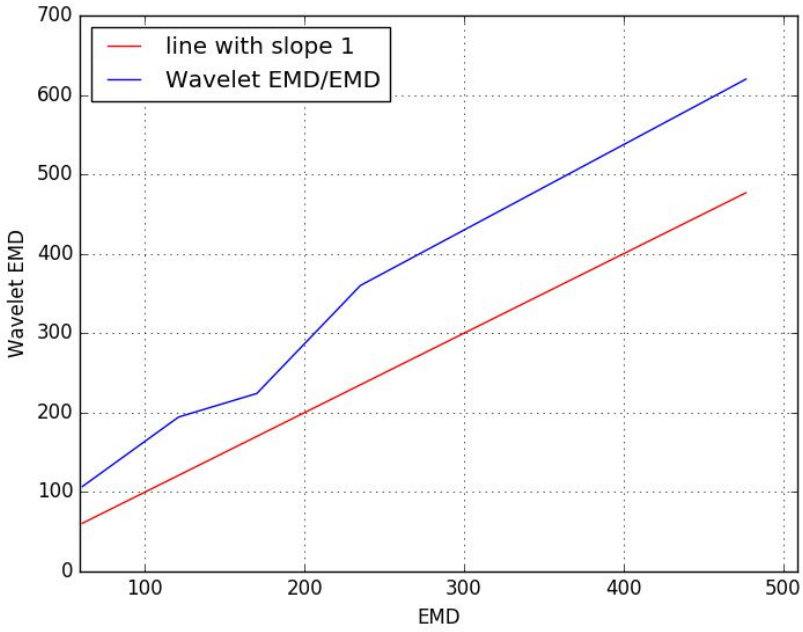
We can observe that first and second images are equivalent and have less EMD and Wavelet EMD values compared to value between first and third which are not similar.So it is clear that wavelet EMD is following same trend as EMD and has nearly equal value.

Wavelet Emd Performance - Time Complexity



From the graph, it can be inferred that along with finding a good approximation for the value of EMD, the wavelet EMD algorithm takes less time compared to the actual EMD algorithm.We can also observe that with increase in number of bins , time difference between EMD and wavelet EMD is exponentially increasing.

Wavelet Emd Value vs Actual Emd Value



We can see that ratio of Wavelet EMD/Actual EMD is close to 1. So the wavelet EMD value is a good approximation to the actual EMD value as inferred from the graph.

Wavelets Performance

Wavelet	**Bound ratio**
db3	2.45
sym5	1.2
gaus3	1.8

Lower value of the ratio for sym5 means that the EMD value is more bounded.

Every image is a different type of signal so some wavelets might capture info from some kind of images efficiently.

6. IMPROVEMENTS

- **Parallelization:**
  - The calculation of the wavelet coefficients by varying scales and shifts could be parallelized which could reduce the time complexity of the algorithm.
  - The multiprocessing module in python was used to parallelize the computation of wavelet coefficients.
  - Parallelizing code was found to reduce the time complexity by a significant extent.
  - For 1000 images time taken to run the code without parallelization = 7493.81 sec
  - For the same 1000 images time taken to run the code with our parallelization code = 2217.42 sec
  - So we got a significant change in the time taken to run wavelet EMD for large dataset.

- **Color correlogram:**
  - One of the limitations of histograms is that they do not take into account the spatial positions of the pixels. As a result, the EMD distance of an image and its flipped version will be zero. But this is weird considering that the two images are different.
  - A color correlogram of an image is a table indexed by color pairs, where the k-th entry for <i, j> specifies the probability of finding a pixel of color j at a distance k from a pixel of color i in the image.

- Such an image feature turns out to be robust in tolerating large changes in appearance of the same scene caused by changes in viewing positions, changes in the background scene, partial occlusions, camera zoom that causes radical changes in shape.
- So, color correlograms could be used instead of histograms for the same application which could negate this flaw.

7. Conclusion And Future Work

The paper has introduced a new method to approximate the earth mover's distance between two histograms using weighted wavelet transform coefficients of the difference histogram. They 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.We can also extend this work and use for image recognition which will be benefited with fast EMD computation.