A Linear Time Histogram Metric for Improved SIFT Matching

Ofir Pele[1] and Michael Werman[1]

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

Abstract. We present a new metric between histograms such as SIFT descriptors and a linear time algorithm for its computation. It is common practice to use the L_2 metric for comparing SIFT descriptors. This practice assumes that SIFT bins are aligned, an assumption which is often not correct due to quantization, distortion, occlusion etc.
In this paper we present a new Earth Mover's Distance (EMD) variant. We show that it is a metric (unlike the original EMD [1] which is a metric only for normalized histograms). Moreover, it is a natural extension of the L_1 metric. Second, we propose a linear time algorithm for the computation of the EMD variant, with a robust ground distance for oriented gradients. Finally, extensive experimental results on the Mikolajczyk and Schmid dataset [2] show that our method outperforms state of the art distances.

1 Introduction

Histograms of oriented gradient descriptors [3–6] are ubiquitous tools in numerous computer vision tasks. One of the most successful is the Scale Invariant Feature Transform (SIFT) [3]. In a recent performance evaluation [2] the SIFT descriptor was shown to outperform other local descriptors. The SIFT descriptor has proven to be successful in applications such as object recognition [3, 7–9], object class detection [10–12], image retrieval [13–16], robot localization [17], building panoramas [18] and image classification [19].

It is common practice to use the L_2 metric for comparing SIFT descriptors. This practice assumes that the histogram domains are aligned. However this assumption is violated through quantization, shape deformation, detector localization errors, etc. Although the SIFT algorithm has steps that reduce the effect of quantization, this is still a liability, as can be seen by the fact that increasing the number of orientation bins negatively affects performance [3].

The Earth Mover's Distance (EMD) [1] is a cross-bin distance that addresses this alignment problem. EMD is defined as the minimal cost that must be paid to transform one histogram into the other, where there is a "ground distance" between the basic features that are aggregated into the histogram. There are two main problems with the EMD. First, it is not a metric between non-normalized histograms. Second, for a general ground distance, it has a high run time.

In this paper we present an Earth Mover's Distance (EMD) variant. We show that it is a metric, if it is used with a metric ground distance. Second, we present a linear time algorithm for the computation of the EMD variant, with a robust ground distance for oriented gradients. Finally, we present experimental results for SIFT matching on the Mikolajczyk and Schmid dataset [2] showing that our method outperforms state of the art distances such as L_2, EMD-L_1 [20], diffusion distance [21] and EMD_{MOD} [22, 23].

This paper is organized as follows. Section 2 is an overview of previous work. Section 3 introduces the new EMD variant. Section 4 introduces the new SIFT metric and the linear time algorithm for its computation. Section 5 describes the experimental setup and Section 6 presents the results. Finally, conclusions are drawn in Section 7.

2 Previous Work

Early work using cross-bin distances for histogram comparison can be found in [24, 25, 22, 26]. Shen and Wong [24] proposed to unfold two integer histograms, sort them and then compute the L_1 distance between the unfolded histograms. To compute the modulo matching distance between cyclic histograms they proposed taking the minimum from all cyclic permutations. This distance is equivalent to EMD between two normalized histograms. Werman et al. [25] showed that this distance is equal to the L_1 distance between the cumulative histograms. They also proved that matching two cyclic histograms by examining only cyclic permutations is in effect optimal. Cha and Srihari [27] rediscovered these algorithms and described a character writer identification application. Werman et al. [22] proposed an $O(M \log M)$ algorithm for finding a minimal matching between two sets of M points on a circle. The algorithm can be adapted to compute the EMD between two N-bin, normalized histograms with time complexity $O(N)$ (Appendix A in [23]).

Peleg et al. [26] suggested using the EMD for grayscale images and using linear programming to compute it. Rubner et al. [1] suggested using EMD for color and texture images. They computed the EMD using a specific linear programming algorithm - the transportation simplex. The algorithm worst case time complexity is exponential. Practical run time was shown to be super-cubic ($\Omega(N^3) \cap O(N^4)$). Interior-point algorithms with time complexity $O(N^3 \log N)$ can also be used. All of these algorithms have high computational cost.

Indyk and Thaper [28] proposed approximating the EMD by embedding it into an Euclidean space. Embedding time complexity is $O(Nd \log \Delta)$, where N is the feature set size, d is the feature space dimension and Δ is the diameter of the union of the two feature sets.

Recently Ling and Okada proposed general cross-bin distances for histogram descriptors. The first is EMD-L_1[20]; *i.e.* EMD with L_1 as the ground distance. To execute the EMD-L_1 computation, they propose a tree-based algorithm, Tree-EMD. Tree-EMD exploits the fact that a basic feasible solution of the simplex algorithm-based solver forms a spanning tree when the EMD-L_1 is modeled as

a network flow optimization problem. The worst case time complexity is exponential. Empirically, they show that this new algorithm has an average time complexity $O(N^2)$. Ling and Okada also proposed the diffusion distance [21]. They defined the difference between two histograms to be a temperature field. The diffusion distance was derived as the sum of dissimilarities over scales. The algorithm run time is linear.

For a comprehensive review of EMD and its applications in computer vision we refer the reader to Ling and Okada's paper [20].

3 The new EMD variant - $\widehat{EMD}$

This section introduces $\widehat{EMD}$, a new Earth Mover's Distance variant. We show that it is a metric (unlike the original EMD [1] which is a metric only for normalized histograms). Moreover, it is a natural extension of the L_1 metric.

The Earth Mover's Distance (EMD) [1] is defined as the minimal cost that must be paid to transform one histogram into the other, where there is a "ground distance" between the basic features that are aggregated into the histogram.

Given two histograms P, Q the EMD as defined by Rubner et al. [1] is:

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

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

where $\{f_{ij}\}$ denotes the flows. Each f_{ij} represents the amount transported from the ith supply to the jth demand. We call d_{ij} the *ground distance* between bin i and bin j in the histograms.

We propose $\widehat{EMD}$:

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

Note that for two probability histograms (*i.e.* total mass equal to one) EMD and $\widehat{EMD}$ are equivalent. However, if the masses are not equal, $\widehat{EMD}$ adds one supplier or demander such that the masses on both sides becomes equal. The ground distance between this supplier or demander to all other demanders or suppliers respectively is set to be α times the maximum ground distance. In addition, the $\widehat{EMD}$ is not normalized by the total flow.

Note that $\widehat{EMD}$ with $\alpha = 0.5$ and with the Kroncker δ ground distance multiplied by two ($d_{ij} = 0$ if $i = j$, 2 otherwise) is equal to the L_1 metric.

If $\alpha \ge 0.5$ and the ground distance is a metric, $\widehat{EMD}$ is a metric (unlike the original EMD [1] which is a metric only for normalized histograms). A proof

is given in Appendix B [23]. Being a metric can lead to more efficient data structures and search algorithms.

We now give two examples of when the usage of $\widehat{EMD}$ is more appropriate (both of which are the case for the SIFT descriptors). The first is when the total mass of the histograms is important. For example, let $P = (1, 0)$, $Q = (0, 1)$, $P' = (9, 0)$, $Q' = (0, 9)$. Using L_1 as a ground distance and $\alpha = 1$, $EMD(P, Q) = 1 = EMD(P', Q')$, while $\widehat{EMD}(P, Q) = 1 < 9 = \widehat{EMD}(P', Q')$. The second is when the difference in total mass between histograms is a distinctive cue. For example, let $P = (1, 0)$, $Q = (1, 7)$. Using L_1 as a ground distance and $\alpha = 1$, $EMD(P, Q) = 0$, while $\widehat{EMD}(P, Q) = 7$.

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

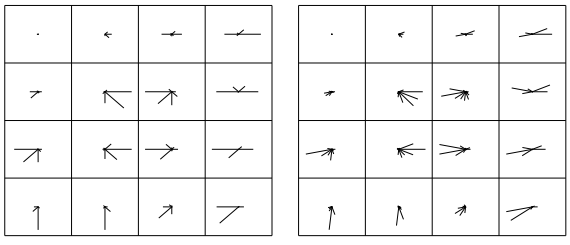
This section introduces SIFT$_{\text{DIST}}$, a new metric between SIFT descriptors. It is common practice to use the L_2 metric for comparing SIFT descriptors. This practice assumes that the SIFT histograms are aligned, so that a bin in one histogram is only compared to the corresponding bin in the other histogram. This is often not the case, due to quantization, distortion, occlusion, etc. Our distance has three instead of two matching costs: zero-cost for exact corresponding bins, one-cost for neighboring bins and two-cost for farther bins and for the extra mass. Thus the metric is robust to small errors and outliers.

The section first defines SIFT$_{\text{DIST}}$ and then presents a linear time algorithm for its computation.

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

This section first describes the SIFT descriptor. Second, it proposes Thresholded Modulo Earth Mover's Distance (EMD$_{\text{TMOD}}$), an EMD variant for oriented gradient histograms. Finally it defines the SIFT$_{\text{DIST}}$.

The SIFT descriptor [3] is a $M \times M \times N$ histogram. Each of the $M \times M$ spatial cells contains an N-bin histogram of oriented gradients. See Fig. 1 for a visualization of SIFT descriptors.



(a) (b)

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

Let $A = \{0, \ldots, N-1\}$ be N points, equally spaced, on a circle. The modulo L_1 distance for two points $i, j \in A$ is:

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

The thresholded modulo L_1 distance is defined as:

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

EMD_{TMOD} is defined as $\widehat{EMD}$ (Eq. 3) with ground distance $D_{TDMO}(i,j)$ and $\alpha = 1$. D_{TDMO} is a metric. It follows from Appendix B [23] that EMD_{TMOD} is also a metric. Using EMD_{TMOD} the transportation cost to the two nearby bins is 1, while for farther bins and for the extra mass it is 2 (see Fig. 2). For example, for $N = 16$ we assume that all differences larger than $22.5°$ are caused by outliers and should be assigned the same transportation cost.

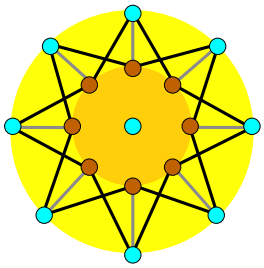


Fig. 2. The flow network of EMD_{TMOD} for $N = 8$. The brown vertexes on the inner circle are the bins of the "supply" histogram, P. The cyan vertexes on the outer circle are the bins of the "demand" histogram, Q. We assume without loss of generality that $\sum_i P_i \geq \sum_j Q_j$; thus we add one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

The $\text{SIFT}_{\text{DIST}}$ between two SIFT descriptors is defined as the sum over the EMD_{TMOD} between all the $M \times M$ oriented gradient histograms. $\text{SIFT}_{\text{DIST}}$ is also an EMD, where edges between spatial bins have an infinite cost.

$\widehat{EMD}$ (Eq. 3) has two advantages over Rubner's EMD (Eq. 1) for comparing SIFT descriptors. First, the difference in total gradient magnitude between SIFT spatial cells is an important distinctive cue. Using Rubner's definition this cue is ignored. Second, $\widehat{EMD}$ is a metric even for non-normalized histograms.

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

As $\text{SIFT}_{\text{DIST}}$ is a sum of EMD_{TMOD} solutions, we present a linear time algorithm for EMD_{TMOD}.

Like all other Earth Mover's Distances, EMD_{TMOD} can be solved by a max-flow-min-cost algorithm. Each bin i in the first histogram is connected to: bin i in the second histogram with a *zero-cost* edge, two nearby bins with *one-cost* edges and to all other bins with *two-cost* edges. See Fig. 2 for an illustration of this flow network.

The algorithm starts by saturating all *zero-cost* edges. As the ground distance (Eq. 5) obeys the triangle inequality, this step does not change the minimum cost solution [22]. Note that after the first step finishes from each of the supplier-demander pairs that were connected with a *zero-cost edge*, either the supplier is empty or the demander is full.

To minimize the cost after the first step, the algorithm needs to maximize the flow through the *one-cost* edges; *i.e.* the problem becomes a max-flow problem on a graph with N vertexes and at most N edges, where the maximum path length is 1 as flow goes only from supply to demand (see Fig. 2). This step starts by checking whether all vertexes are degree 2, if yes we remove an arbitrary edge. Second, we traverse the suppliers' vertexes clockwise, twice. For each degree one vertex we saturate its edge. If we did not remove an edge at the beginning of this step, there will be no augmenting paths. If an edge was removed, the algorithm flows through all augmenting paths. All these paths can be found by returning the edge and expanding from it.

The algorithm finishes by flowing through all *two-cost* edges, which is equivalent to multiplying the maximum of the total remaining supply and demand by two and adding it to the distance.

5 Experimental Setup

We evaluate $\text{SIFT}_{\text{DIST}}$ using a test protocol similar to that of Mikolajczyk and Schmid [2]. The dataset was downloaded from [29]. The test data contain eight folders, each with six images with different geometric and photometric transformations and for different scene types. Fig. 3 shows the first and the third image from each folder. Six image transformations are evaluated: viewpoint change, scale and rotation, image blur, light change and JPEG compression. The images are either of planar scenes or the camera position was fixed during acquisition. The images are, therefore, always related by a homography. The ground truth homographies are supplied with the dataset. For further details about the dataset we refer the reader to [2].

The evaluation criterion is based on the number of correct and false matches obtained for an image pair. The match definition depends on the matching strategy. The matching strategy we use is a symmetric version of Lowe's ratio matching [3]. Let $a \in A$ and $b \in B$ be two descriptors and $n(a, A)$ and $n(b, B)$ be their spatial neighbors; that is, all descriptors in A and B respectively such that the ratio of the intersection and union of their regions with a and b respectively is larger than 0.5. a and b are matched, if all the following conditions hold: $a = \arg\min_{a' \in A} D(a', b)$, $a_2 = \arg\min_{a' \in A \setminus n(a,A)} D(a', b)$, $b = \arg\min_{b' \in B} D(a, b')$, $b_2 = \arg\min_{b' \in B \setminus n(b,B)} D(a, b')$ and $\min\left(\frac{D(a_2,b)}{D(a,b)}, \frac{D(a,b_2)}{D(a,b)}\right) \geq R$. R value is varied to obtain the curves. Note that for $R = 1$ the matching strategy is the symmetric nearest neighbor strategy. The technique of not using overly close regions as a second best match was used by Forssén and Lowe [30].

The match correctness is determined by the *overlap error* [2]. It measures how well regions A and B correspond under a known homography H, and is defined

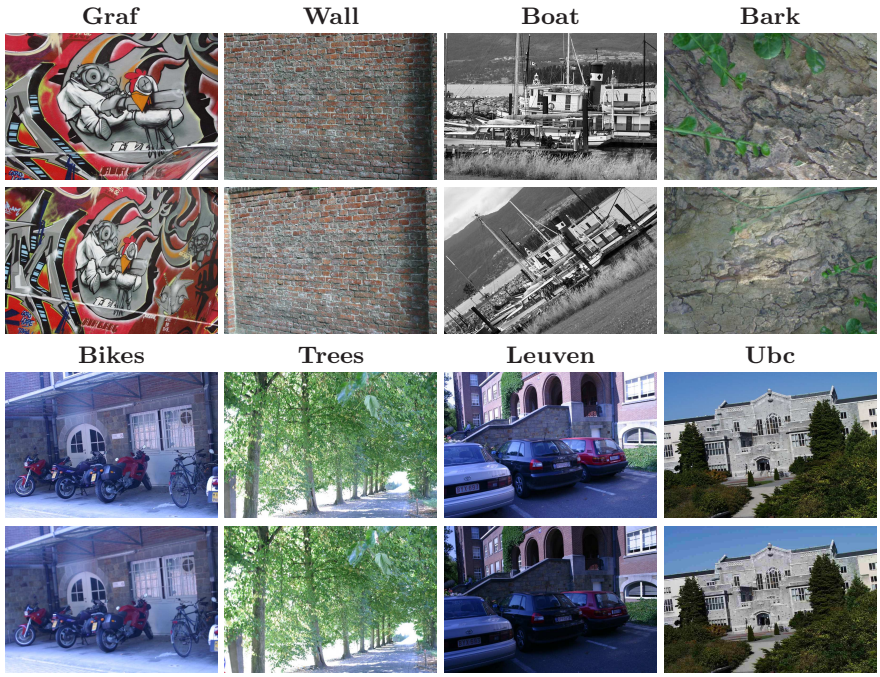


Fig. 3. Examples of test images (left): **Graf** (viewpoint change, structured scene), **Wall** (viewpoint change, textured scene), **Boat** (scale change + image rotation, structured scene), **Bark** (scale change + image rotation, textured scene), **Bikes** (image blur, structured scene), **Trees**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

by the ratio of the intersection and union of the regions: $\epsilon_S = 1 - \frac{A \bigcap H^T B H}{A \bigcup H^T B H}$. As in [2] a match is assumed to be correct if $\epsilon_S < 0.5$. The correspondence number (possible correct matches) is the maximum matching size in the correct match, bi-partite graph. The results are presented with *recall* versus 1 - *precision*: $recall = \frac{\#\text{correct matches}}{\#\text{correspondences}}$, $1 - precision = \frac{\#\text{false matches}}{\#\text{all matches}}$.

In all experiments we used Vedaldi's SIFT detector and descriptor implementation [31]. All parameters were set to the defaults except the oriented gradient bins number, where we also tested our method with a SIFT descriptor having 16 oriented gradient bins (see Fig. 1 (b) in page 4).

6 Results

In this section, we present and discuss the experimental results. The performance of $\text{SIFT}_{\text{DIST}}$ is compared to that of L_2, EMD-L_1 [20], diffusion distance [21]

and EMD_{MOD}[1] [22, 23] for viewpoint change, scale and rotation, image blur, light change and JPEG compression. The matching was done between SIFT descriptors with eight and sixteen orientation bins (see Fig. 1 in page 4). Finally, we present run time results.

Figs. 4,5,6,7 are 1-precision vs. recall graphs. Due to space constraints we present graphs for the matching of the first to the third and fifth images from each folder in the Mikolajczyk and Schmid dataset [29]. Results for the rest of the data are similar and are in [32].

In all of the experiments $\text{SIFT}_{\text{DIST}}$ computed between SIFT descriptors with sixteen oriented gradient bins (SIFT-16) outperforms all other methods.
$\text{SIFT}_{\text{DIST}}$ computed between the original SIFT descriptors with eight oriented gradient bins (SIFT-8) is usually the second. Also, using $\text{SIFT}_{\text{DIST}}$ consistently produces results with greater precision and recall for the symmetric nearest neighbor matching (the rightmost point of each curve).

Increasing the number of orientation bins from eight to sixteen decreases the performance of L_2 and increases the performance of $\text{SIFT}_{\text{DIST}}$. This can be explained by the $\text{SIFT}_{\text{DIST}}$ robustness to quantization errors.

Fig. 4 shows matching results for viewpoint change on structured (**Graf**) and textured (**Wall**) scenes . $\text{SIFT}_{\text{DIST}}$ has the highest ranking. As in [2] performance is better on the textured scene. Performance decreases with viewpoint change (compare **Graf-3** to **Graf-5** and **Wall-3** to **Wall-5**), while the distance ranking remains the same. For large viewpoint change, performance is poor (see **Graf-5**) and the resulting graphs are not smooth. This can be explained by the fact that the SIFT detector and descriptor are not affine invariant.

Fig. 5 shows matching results for similarity transformation on structured (**Boat**) and textured (**Bark**) scenes. $\text{SIFT}_{\text{DIST}}$ outperforms all other distances. As in [2] performance is better on the textured scene.

Fig. 6 shows matching results for image blur on structured (**Bikes**) and textured (**Trees**) scenes. $\text{SIFT}_{\text{DIST}}$ has the highest ranking. The SIFT descriptor is affected by image blur. A similar observation was made by Mikolajczyk and Schmid [2].

Fig. 7 - **Leuven** shows matching results for light change. $\text{SIFT}_{\text{DIST}}$ obtains the best matching score. The performance decreases with lack of light (compare **Leuven-3** to **Leuven-5**), although not drastically.

Fig. 7 - **Ubc** shows matching results for JPEG compression. $\text{SIFT}_{\text{DIST}}$ outperforms all other distances. Performance decreases with compression level (compare **Ubc-3** to **Ubc-5**).

Table 1 present run time results. All runs were conducted on a Dual-Core AMD Opteron 2.6GHz processor. The table contains the run time in seconds of each distance computation between two sets of 1000 SIFT descriptors with eight and sixteen oriented gradient bins. Note that we measured the run time of $(L_2)^2$ and not L_2 as computing the root does not change the order of elements and is time consuming. $\text{SIFT}_{\text{DIST}}$ is the fastest cross-bin distance.

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

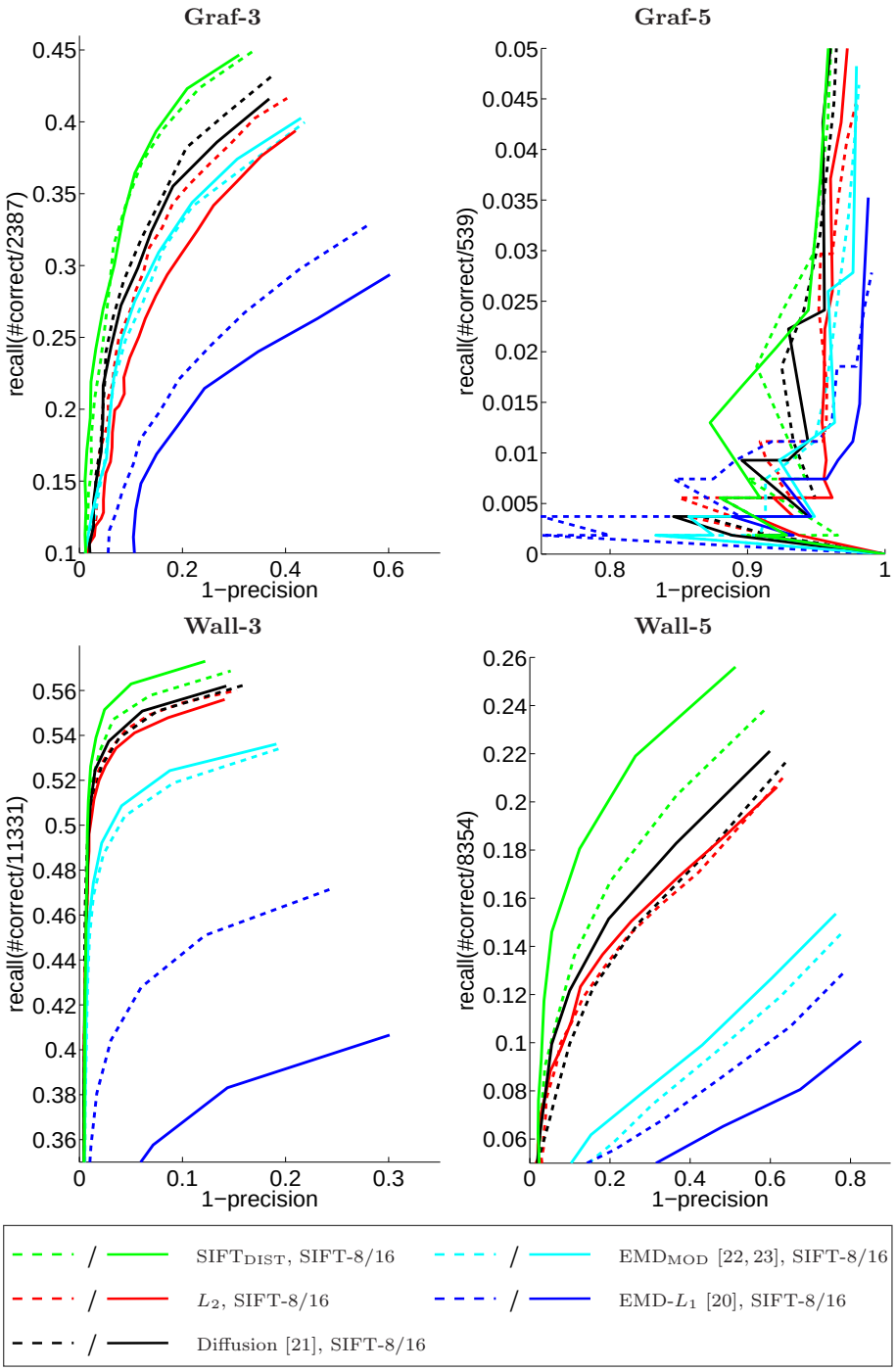


Fig. 4. Results on the Mikolajczyk and Schmid dataset [2]. Should be viewed in color.

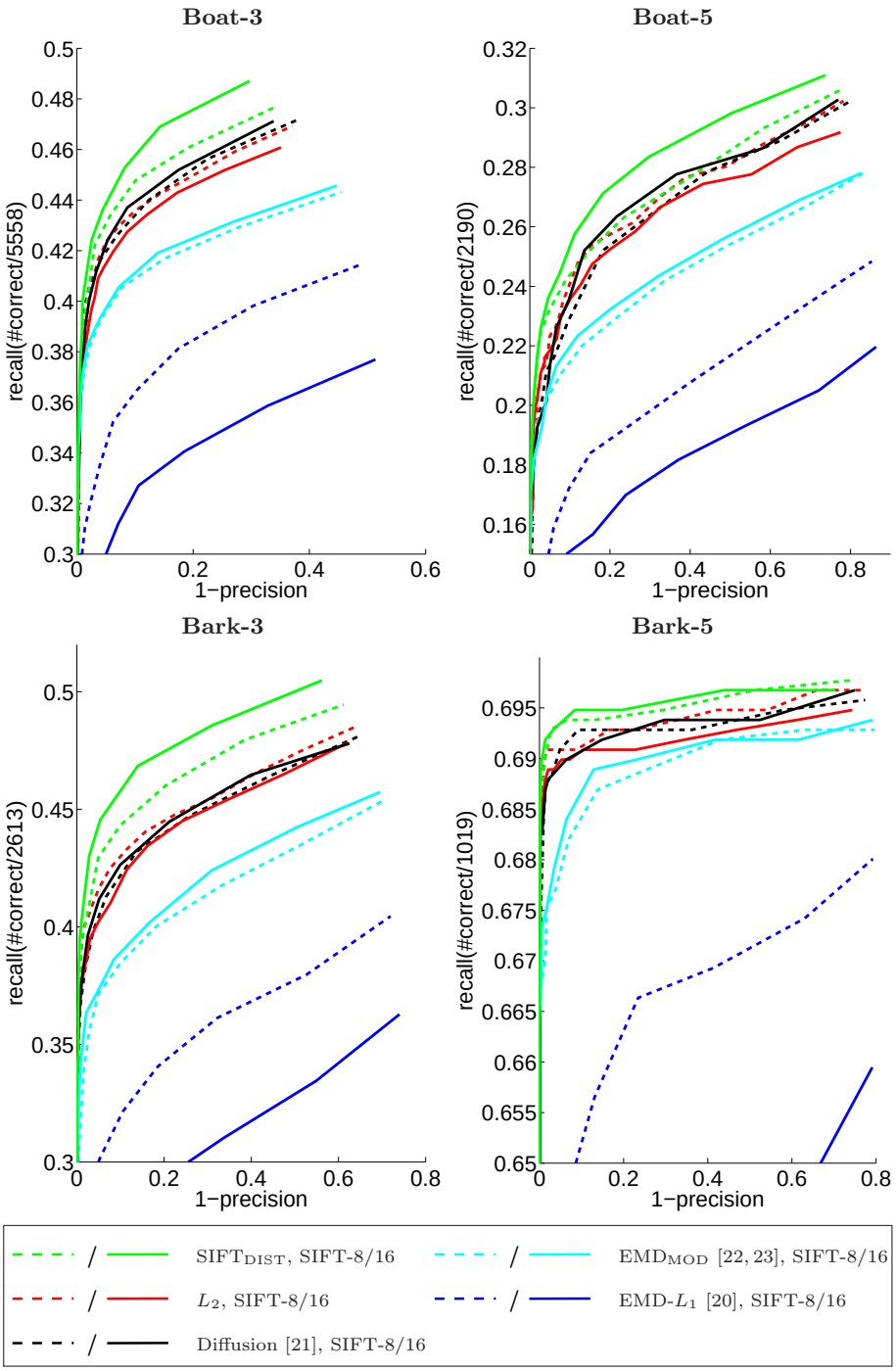


Fig. 5. Results on the Mikolajczyk and Schmid dataset [2]. Should be viewed in color.

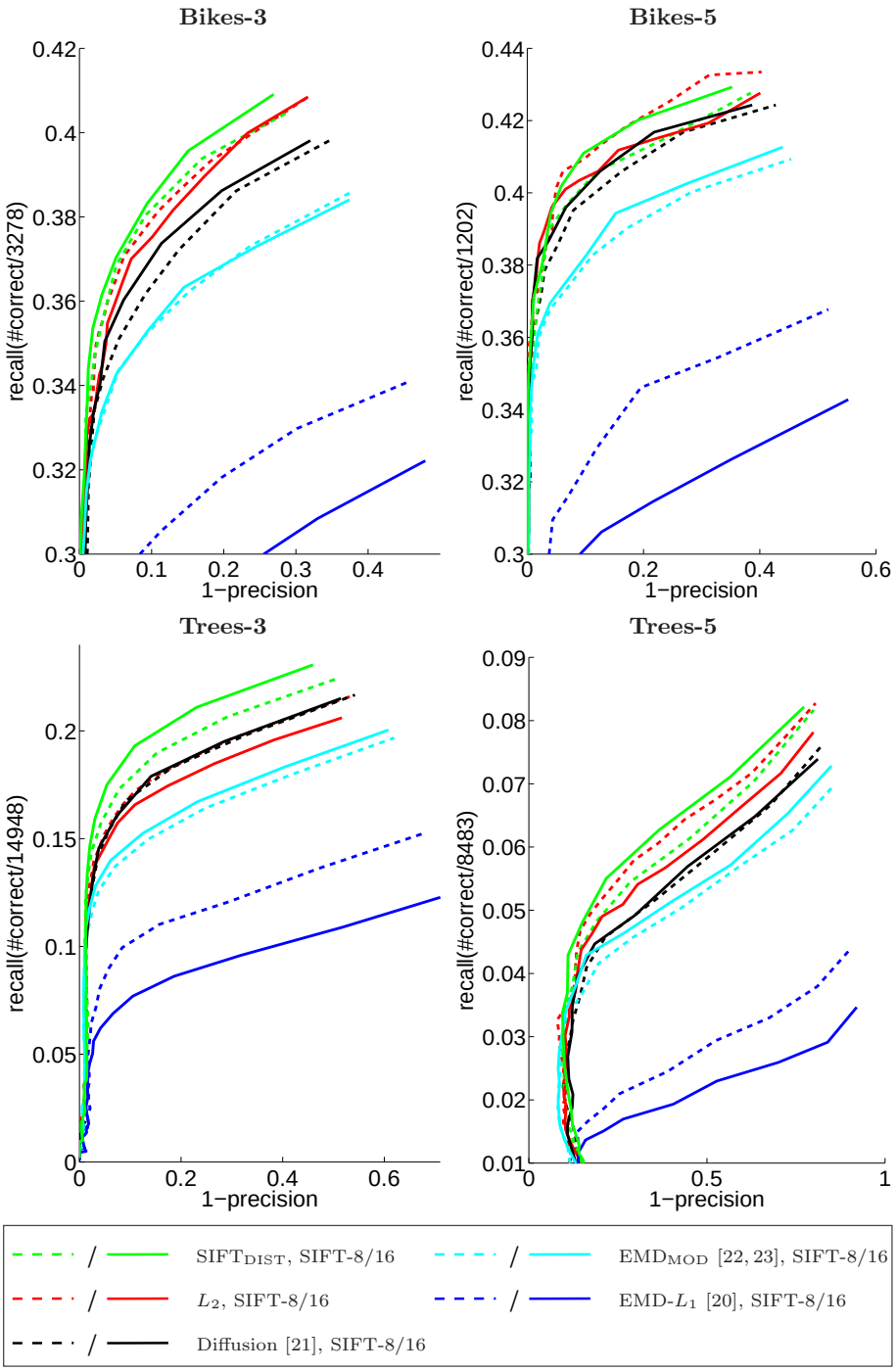


Fig. 6. Results on the Mikolajczyk and Schmid dataset [2]. Should be viewed in color.

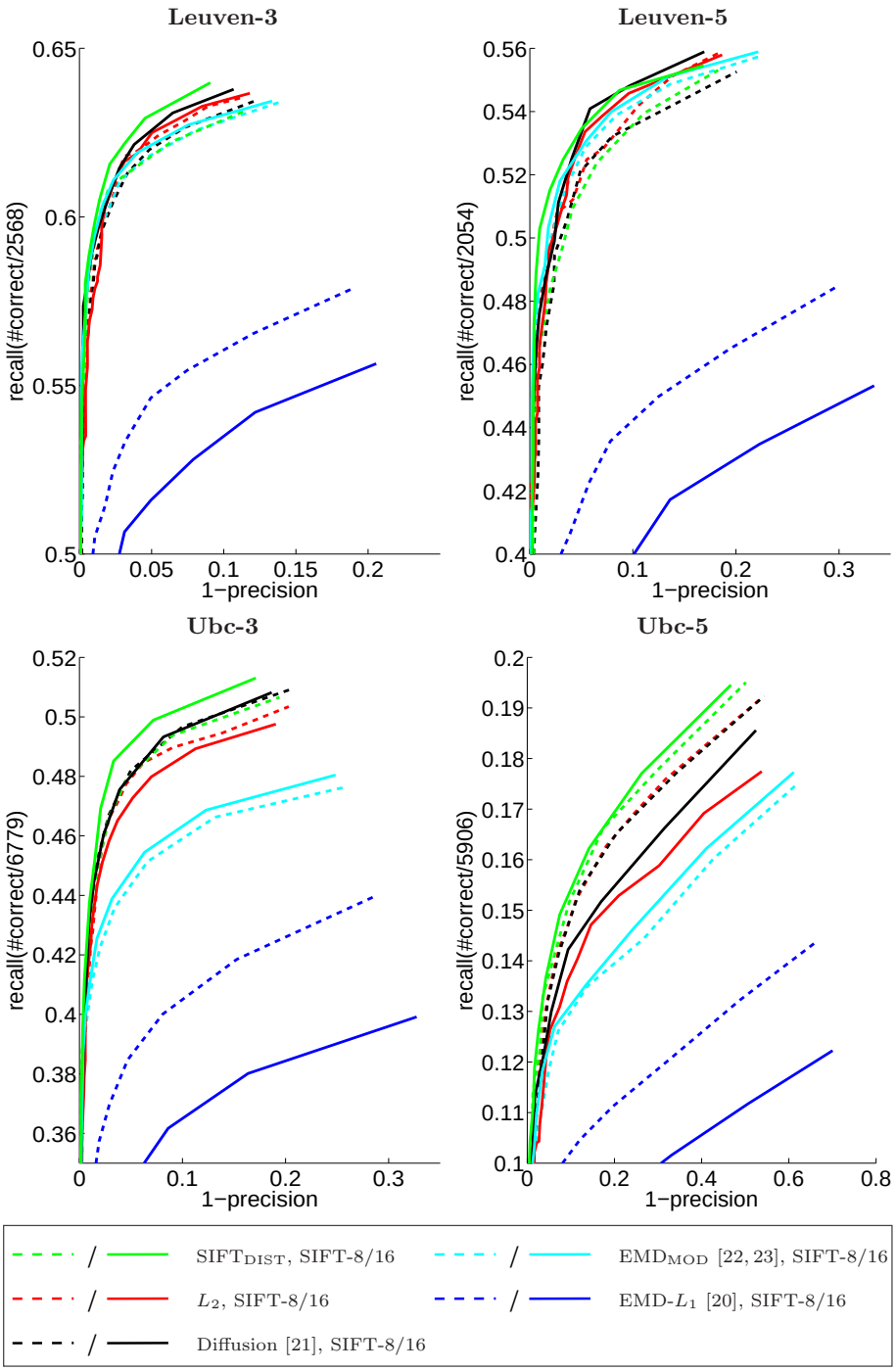


Fig. 7. Results on the Mikolajczyk and Schmid dataset [2]. Should be viewed in color.

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

Table 1. Run time results in seconds of 10^6 distance computations

7 Conclusions

We presented a new cross-bin metric between histograms and a linear time algorithm for its computation. Extensive experimental results for SIFT matching on the Mikolajczyk and Schmid dataset [2] showed that our method outperforms state of the art distances.

The speed can be further improved using techniques such as Bayesian sequential hypothesis testing [33], sub linear indexing [34] and approximate nearest neighbor [35, 36]. The new cross-bin histogram metric may also be useful for other histograms, either cyclic (*e.g.* hue in color images) or non-cyclic (*e.g.* intensity in grayscale images). The project homepage, including code (C++ and Matlab wrappers) is at: `http://www.cs.huji.ac.il/~ofirpele/SiftDist`

References

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

13. Sivic, J., Zisserman, A.: Video Google: a text retrieval approach to object matching in videos. In: ICCV. (2003) 1470–1477
14. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: CVPR. (2007)
15. Snavely, N., Seitz, S., Szeliski, R.: Photo tourism: exploring photo collections in 3D. ACM Transactions on Graphics (TOG) **25**(3) (2006) 835–846
16. Sivic, J., Everingham, M., Zisserman, A.: Person Spotting: Video Shot Retrieval for Face Sets. In: CIVR, Springer (2005)
17. Se, S., Lowe, D., Little, J.: Local and global localization for mobile robots using visuallandmarks. In: IROS. Volume 1. (2001)
18. Brown, M., Lowe, D.: Recognising panoramas. In: ICCV. Volume 1. (2003) 3
19. Nowak, E., Jurie, F., Triggs, B.: Sampling strategies for bag-of-features image classification. In: ECCV. Volume 4., Springer (2006) 490–503
20. Ling, H., Okada, K.: An Efficient Earth Mover's Distance Algorithm for Robust Histogram Comparison. IEEE Trans. Pattern Analysis and Machine Intelligence **29**(5) (2007) 840–853
21. Ling, H., Okada, K.: Diffusion distance for histogram comparison. In: CVPR. Volume 1. (2006) 246–253
22. Werman, M., Peleg, S., Melter, R., Kong, T.: Bipartite graph matching for points on a line or a circle. Journal of Algorithms **7**(2) (1986) 277–284
23. `http://www.cs.huji.ac.il/~ofirpele/publications/ECCV2008.pdf`
24. Shen, H., Wong, A.: Generalized texture representation and metric. Computer vision, graphics, and image processing **23**(2) (1983) 187–206
25. Werman, M., Peleg, S., Rosenfeld, A.: A distance metric for multidimensional histograms. Computer Vision, Graphics, and Image Processing **32**(3) (1985)
26. Peleg, S., Werman, M., Rom, H.: A unified approach to the change of resolution: Space and gray-level. IEEE Trans. Pattern Analysis and Machine Intelligence. **11**(7) (1989) 739–742
27. Cha, S., Srihari, S.: On measuring the distance between histograms. Pattern Recognition **35**(6) (2002) 1355–1370
28. Indyk, P., Thaper, N.: Fast image retrieval via embeddings. In: 3rd International Workshop on Statistical and Computational Theories of Vision. (Oct 2003)
29. `http://www.robots.ox.ac.uk/~vgg/research/affine/index.html`
30. Forssén, P., Lowe, D.: Shape Descriptors for Maximally Stable Extremal Regions. In: ICCV. (2007) 1–8
31. `http://vision.ucla.edu/~vedaldi/code/sift/sift.html`
32. `http://www.cs.huji.ac.il/~ofirpele/publications/ECCV2008addRes.pdf`
33. Pele, O., Werman, M.: Robust real time pattern matching using bayesian sequential hypothesis testing. IEEE Trans. Pattern Analysis and Machine Intelligence. **30**(8) (2008) 1427–1443
34. Obdrzalek, S., Matas, J.: Sub-linear indexing for large scale object recognition. In: BMVC. Volume 1. (2005) 1–10
35. Arya, S., Mount, D., Netanyahu, N., Silverman, R., Wu, A.: An optimal algorithm for approximate nearest neighbor searching fixed dimensions. Journal of the ACM (JACM) **45**(6) (1998) 891–923
36. Beis, J., Lowe, D.: Shape indexing using approximate nearest-neighbour search in high-dimensional spaces. In: CVPR. (1997) 1000–1006
37. Cabrelli, C., Molter, U.: A linear time algorithm for a matching problem on the circle. Information Processing Letters **66**(3) (1998) 161–164

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$EMD(\tilde{A},\tilde{B})+EMD(\tilde{B},\tilde{C}) \geq EMD(\tilde{A},\tilde{C}) \tag{17}$$
$$\Downarrow (S \geq 0) \tag{18}$$
$$(EMD(\tilde{A},\tilde{B}) \times S)+(EMD(\tilde{B},\tilde{C}) \times S) \geq (EMD(\tilde{A},\tilde{C}) \times S) \tag{19}$$
$$\Downarrow (\text{Eqs. 14, 15, 16}) \tag{20}$$
$$\widehat{EMD}_\alpha(A,B)+\widehat{EMD}_\alpha(B,C) \geq \widehat{EMD}_\alpha(A,C) \tag{21}$$