ON THE EARTH MOVER'S DISTANCE AS A HISTOGRAM SIMILARITY METRIC FOR IMAGE RETRIEVAL

*Zhenghua Yu and Gunawan Herman**

National ICT Australia, Sydney
{Zhenghua.Yu, Gunawan.Herman}@nicta.com.au

ABSTRACT

In this paper, the performance of the Earth Mover's Distance (EMD) vs χ^2 distance as histogram similarity metrics for image retrieval is evaluated experimentally. Ground truth is generated in two ways: through a novel approach of extracting frames from video shots and through random sampling. Quantitative evaluations of the retrieval performance with regular partitioning, clustered and adaptive binning histograms and with various number of bins are given. Meanwhile the influence of ground distance functions to the performance of the EMD is also studied. The findings of the paper may constitute guidelines in adopting the EMD.

1. INTRODUCTION

In content based image retrieval systems, histogram is one of the most frequently used methods to represent features such as color and texture. A color histogram partitions a color space (eg RGB) into a number of bins and represents an image by the distribution of pixel values in those bins. Depending on the partitioning method, there exist in general two types of histograms: fixed binning and adaptive binning. Fixed binning histograms can be further classified into regular partitioning and clustered histograms. See [1] for more detailed definition of different histograms.

The Earth Mover's Distance (EMD) was introduced in [2] as a histogram similarity metric for image retrieval. Given two distributions of points, the EMD measures the minimal cost to transform one distribution to the other. Compared to traditional similarity metrics such as Minkowski distance and χ^2 distance, it possesses several unique features, including the exploitation of ground distance between histogram bins; the support of adaptive binning; and the support of partial matches. Since its introduction, the EMD has been adopted in several image retrieval systems[4, 5].

The performance of the EMD was evaluated in [2, 3] using color and texture features. However in this paper we will focus on color only, as color is the most widely used feature in image retrieval systems. In [2], two color feature based image retrieval experiments were conducted with the main difference in ground truth generation: random sampling and annotation. The experiments suffer from some drawbacks, such as relatively small scale of test (94 input images in random sampling, and 2 query classes in annotation), no thorough investigation with various number of bins, and sometimes inconsistent results (EMD vs Jeffrey divergence and χ^2 across the two queries in annotation test). In [3], experiments were designed similar to [2] and the results were largely similar. The authors concluded that EMD performed very well for the small sample sizes, χ^2 performed better for the larger sample sizes (especially for texture). In another paper[1], EMD performed poorly (worse than L_2) in image retrieval test. The authors noticed that image size of 6144 pixels was used, far larger than the sample sizes used in [3].

Meanwhile as the EMD exploits ground distance between bins, the ground distance function may have critical influence to the performance of the EMD. However there is little study of this influence in literature so far.

Considering the shortcomings in the previous experiments and the sometimes inconsistent results, it is warranted to conduct another independent evaluation of the performance of the EMD in this paper, with two primary purposes: to thoroughly benchmark the performance of the EMD vs another popular metric, and to study the influence of the EMD ground distance to its performance.

The paper is organized as follows: the experiment methodology is described in Section 2, and the experimental results are presented in Section 3. Given the page limit, we only present the main experimental results in this paper.

2. EXPERIMENT METHODOLOGY

To evaluate the performance of histogram similarity metrics experimentally, there exist several difficulties. As pointed out in [3], how to generate ground truth is arguably the hardest problem. Besides, the test settings and parameters should cover major variations that could affect the performance of different metrics. We basically conducted image

*National ICT Australia is funded through the Australian Government's *Backing Australia's Ability* initiative, in part through the Australian Research Council.

retrieval tests similar to [2, 3], however many new test settings and more classes were included to expose the performance of the metrics tested. We also developed a novel ground truth generation method.

The performance of the EMD was benchmarked against several other metrics in [2, 3]. However given the large number of possible parameter combinations and the computational complexity involved, it is not practical to conduct large scale tests using all available metrics and report in this paper. On the other hand, the relative performances of these traditional metrics are largely known and the main purpose of this paper is to evaluate the EMD only. Therefore it was decided to benchmark the EMD against χ^2 distance only. χ^2 was selected based on its high performance in a previous evaluation [3], its low computational complexity and its sound statistical interpretation. The code of EMD was downloaded from the author's website[1].

Two types of experiments are conducted in the paper, with the main difference in the method to generate the ground truth, namely shot segmentation and random sampling. The Hue, Saturation and Value (HSV) color space is used, as it is the most natural color space and has been frequently adopted in image retrieval systems [6].

2.1. Image retrieval with shot segmentation ground truth

In [2, 3], image retrieval experiments were conducted with the image classification ground truth collected through a time consuming process of manually annotating each image. A different way of generating ground truth was adopted in this paper. Color histogram based image similarity metric has been a major part of video shot segmentation systems. Consequently histogram similarity metrics' performance can be evaluated in the context of shot segmentation. Each shot consists of several similar frames so frames from the same shot can be regarded as within the same class. Compared to manually annotating each image, there is very little subjectiveness in shot segmentation. In this paper, we randomly selected sequences from four different commercial TV channels. The content included news and soccer. The sequences were manually segmented into 300 shots. 10 frames (CIF size) from each shot were then selected to form a 3000-images test database. 32 classes (ie 320 images) were randomly selected as the query input. Each of the 320 images was compared to all images in the database using the EMD and χ^2. The precision among the most similar 10 returned images for each query was calculated. A positive sample was the one from the same shot as the query input. Average precision was then calculated over all queries.

Several ground distance functions were experimented for the EMD, including the geometric distance between two colors $dist_0$ [6] and some other distance functions. Given two HSV colors (h_1, s_1, v_1) and (h_2, s_2, v_2),

$$\begin{aligned}
dist_0 &= 1/\sqrt{5}[(v_1 - v_2)^2 + \\
&\quad (s_1\cos(2\pi h_1) - s_2\cos(2\pi h_2))^2 + \\
&\quad (s_1\sin(2\pi h_1) - s_2\sin(2\pi h_2))^2]^{\frac{1}{2}} \quad (1)\\
dist_A &= 1 - exp(-(dist_0)^{\frac{1}{2}} \times 4) \quad (2)\\
dist_B &= 1 - exp(-dist_0/\sigma) \quad (3)\\
dist_C &= 1 - exp(-dist_0/(2\sigma)) \quad (4)\\
dist_D &= 1 - exp(-2dist_0/\sigma) \quad (5)\\
dist_E &= 1 - exp(-dist_0) \quad (6)\\
dist_F &= 1 - exp(-(dist_0)^{\frac{1}{2}} \times 2) \quad (7)\\
dist_G &= 1 - exp(-(dist_0)^{\frac{1}{2}} \times 8) \quad (8)
\end{aligned}$$

where σ is the standard deviation of all color features in the database. $dist_B$ is the function suggested in [2].

2.2. Image retrieval with random sampling ground truth

We adopted the method of generating ground truth through random sampling [2], however the number of input images was increased from 94 to 310. Images from two databases (http://www.cs.washington.edu/research/imagedatabase/ and http://wang.ist.psu.edu/docs/related/) were combined and manually classified into 31 different classes, covering a wide range of content. 10 images were randomly selected from each class. From each image, 16 samples were taken, each containing N randomly chosen disjoint sets of pixels. N was a variable whose value was 8 in most experiments. At this point, there were $310 \times 16 = 4960$ image samples in total. All the 16 samples taken from the same image can be regarded as originating from the same image/distribution and very similar to each other, hence forming a class. Each of the samples was then represented by a color histogram, where two binning methods were applied: clustered and adaptive binning. In both cases K-means was used as the clustering method to partition the color space. Image retrieval experiments were carried out with a leave-one-out procedure, ie, every image sample was used as input to retrieve similar images from the 4960 samples database. The similarity between samples was determined by the distance (EMD or χ^2) between the respective histograms. The number of retrieved images were varied and the respective average recall (over all samples) was calculated.

3. RESULTS AND ANALYSIS

3.1. Shot segmentation ground truth test results

The following notation is adopted to describe test cases, eg, $emd\ adaptive\ 8\ dist_A$ represents EMD, adaptive binning (via K-means) with 8 bins, and $dist_A$ distance function, while $chi\ regular\ 128\ 4:4:8$ represents χ^2, regular partitioning with 128 bins (4, 4, 8 bins for S, V, and H each).

[1] http://www.cs.duke.edu/∼tomasi/software/emd.htm

Table 1. Shot segmentation ground truth test results

Bins	Test Case	Precision
8	$emd\ regular\ 8\ 2:2:2\ dist_A$	0.670000
	$emd\ regular\ 8\ 2:2:2\ dist_B$	0.673125
	$emd\ adaptive\ 8\ dist_A$	0.712500
	$chi\ regular\ 8\ 2:2:2$	0.713437
	$emd\ adaptive\ 8\ dist_B$	0.737500
128	$emd\ regular\ 128\ 4:4:8\ dist_B$	0.812188
	$emd\ adaptive\ 128\ dist_B$	0.820312
	$emd\ regular\ 128\ 4:4:8\ dist_A$	0.833125
	$emd\ adaptive\ 128\ dist_A$	0.841875
	$chi\ regular\ 128\ 4:4:8$	0.852500
256	$emd\ regular\ 256\ 4:4:16\ dist_0$	0.803750
	$emd\ regular\ 256\ 4:4:16\ dist_E$	0.807188
	$emd\ regular\ 256\ 4:4:16\ dist_C$	0.810938
	$emd\ regular\ 256\ 4:4:16\ dist_B$	0.815938
	$emd\ regular\ 256\ 4:4:16\ dist_D$	0.826563
	$emd\ regular\ 256\ 4:4:16\ dist_F$	0.838125
	$emd\ regular\ 256\ 4:4:16\ dist_A$	0.840313
	$emd\ regular\ 256\ 4:4:16\ dist_G$	0.844688
	$chi\ regular\ 256\ 4:4:16$	0.859687

The image retrieval results are reported in Table 1. The following main conclusions may be drawn from the table:

1. Using the same regular binning, χ^2 always has better precision than the EMD with any ground distance function.

2. EMD adaptive binning at 8 bins is not as good as any test case at 128 or 256 bins. This is different from [2].

3. EMD adaptive binning has higher precision than χ^2 at small number of bins (8), but worse than χ^2 at large number of bins (128). This can be attributed to the quality of adaptive binning which makes remarkable difference in approximating the color distribution at small number of bins.

4. The influence of ground distance functions to the performance of the EMD was fully evaluated at 256 bins. Functions of the family $dist_A$, $dist_G$ and $dist_F$ perform the best. For all distance functions, we plot the ground distances as a function of $dist_0$ in Fig 1. By cross-checking the test results and Fig 1, it may be concluded that the ground distance should increase rapidly as the difference in color increases.

5. In other EMD tests (8 bins and 128 bins), $dist_A$ is better than $dist_B$ with an exception at 8 bins. Similar results can be observed in Section 3.2.

3.2. Random sampling ground truth test results

In random sampling tests of [2, 3], 8 pixels per image sample were used. However 8 is very small so the binning method may have critical influence to the performance, and it is too harsh to use regular partitioning with so little bins. Therefore the EMD and the χ^2 were both tested with clustered binning. The EMD also used adaptive binning method. Two ground distance functions were tested. $dist_A$ was selected given its superior performance in section 3.1, and $dist_B$ was adopted because it was used in [2]. The following notation was adopted in this section to describe different

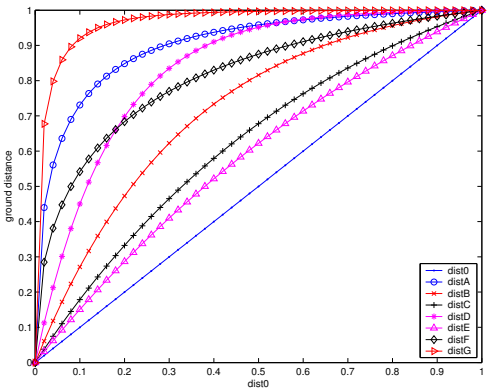


Fig. 1. Ground distances as a function of $dist_0$

test cases, eg $chi\ clustered\ 8\ 32$ means χ^2 distance, clustered binning, 8 pixels per sample, and 32 histogram bins.

χ^2 with various number of histogram bins In this test, the same clustered binning partition was applied to all samples. Each sample contained 8 pixels only. The number of histogram bins was varied between 8 and 512. The experiment results are shown in Fig. 2(a). From the diagram, it is shown that the recall increases as the number of bins increases from 8 with peak performance at 128 bins, and then decreases as the number of bins increases further. The phenomenon can be explained as follows. Since each sample contains only 8 pixels, with small number of histogram bins, most pixels in one sample are likely to belong to the same bin. On the other hand, with large number of histogram bins and only 8 pixels, it is unlikely that two samples will have any pixel in identical bins. Therefore in both cases many samples may have the same distance, which greatly reduces the ability of χ^2 to identify samples from the same class.

Given its superior performance, 128 histogram bins will be used in all χ^2 8 pixels per sample tests below.

EMD and χ^2 at 8 pixels and 8 bins In this test 8 pixels per sample and 8 histogram bins were used. The EMD was tested with both clustered and adaptive histogram bins and with two ground distance functions - $dist_A$ and $dist_B$. From Fig. 2(b), there is little surprise to see that two EMD adaptive cases perform the best, because with adaptive histogram, 8 pixels and 8 bins, each pixel will be assigned to the bin whose mean value is exactly the pixel value, so there is no binning effect at all. Meanwhile with the same number of pixels, $chi\ clustered\ 8\ 128$ is close to the average recall of EMD adaptive 8 bins and even better than EMD at large number of retrieved samples. When both the EMD and χ^2 adopt clustered 8-bins histograms, χ^2 is slightly better than the EMD. Both $dist_A$ and $dist_B$ achieve almost identical performance, so only $dist_A$ curve is shown in the figure.

EMD and χ^2 at 8 pixels and 4 bins In the above test, the EMD does not suffer from binning effect at all, which is not realistic in practical applications. When the same 8 pixels per sample condition is adopted, EMD with adaptive 4

bins histogram is worse than χ^2 with clustered 128-bins histogram (see Fig. 2(c)). In this case, $chi\ clustered\ 8\ 128$ has more average used number of bins (6.3) than $EMD\ adaptive\ 8\ 4$ (average used number of bins 4.0), however the clustering quality of the EMD is much better than χ^2 because it is adaptive to each individual sample (it has 17316 distinct bins in total vs 64 for χ^2). Meanwhile EMD adaptive 4 bins is significantly better than any clustered 8 bins method.

EMD and χ^2 at 128 pixels per sample As shown in Fig. 2(d), when 128 pixels are taken for each sample, at 8 bins per sample, EMD with adaptive histograms ($dist_B$) is better than χ^2, and χ^2 is better than the EMD with the same clustered 8 bins histograms. Secondly, using the same sample pixels, χ^2 with more bins (128) is better than adaptive histogram/EMD with 8 bins.

3.3. A summary of the major findings of the paper

The main findings of the paper can be summarized below (in the context of image retrieval with color features):

1. EMD with adaptive histogram has significantly better precision/recall than EMD and χ^2 with regular partitioning or clustered histogram at the same number of bins when the number of bins is small.

2. The performance of χ^2 varies according to the number of bins used and there may exist optimal number of bins.

3. The performance of the EMD with adaptive binning and small number of bins may be matched or exceeded by χ^2 with more number of bins.

4. EMD with either regular partitioning or clustered histogram is inferior to χ^2 using the same histogram.

5. EMD ground distance functions which increase sharply as a function of the geometric distance between two colors usually have better retrieval performance.

6. The main results are consistent across the two tests and the small/large number of pixels per image conditions.

4. REFERENCES

[1] W. K. Leow and R. Li, "The analysis and applications of adaptive-binning color histograms," *Comput Vis Image Underst*, vol 94, no 1-3, pp 67-91, 2004.

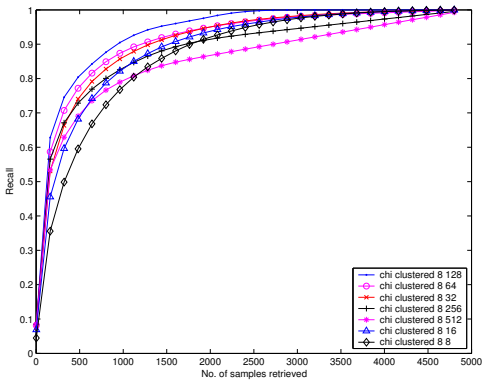
[2] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int Jour of Comp Vision*, vol 40, no 2, pp 99-121, 2000.

[3] Y. Rubner et al, "Empirical evaluation of dissimilarity measures for color and texture," *Comput Vis Image Underst*, vol 84, no 1, pp 25-43, 2001.

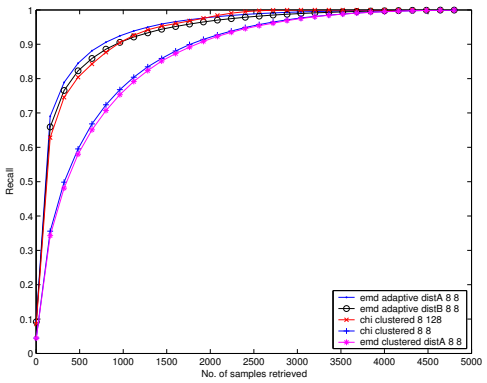
[4] F. Jing et al, "An efficient and effective region-based image retrieval framework," *IEEE Trans Image Proc*, vol 13, no 5, pp 699-709, 2004.

[5] S. Wang, L.-T. Chia and D. Rajan, "Efficient image retrieval using MPEG-7 descriptors," *ICIP*, vol 3, pp 14-17, 2003.

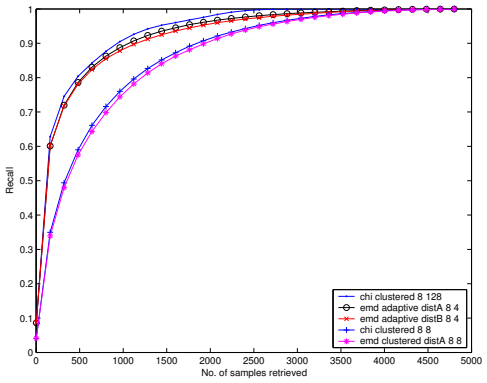
[6] J. R.Smith and S.-F. Chang, "VisualSeek: A fully automated content based image query system," *ACM Multimedia*, pp 87-98, 1996.



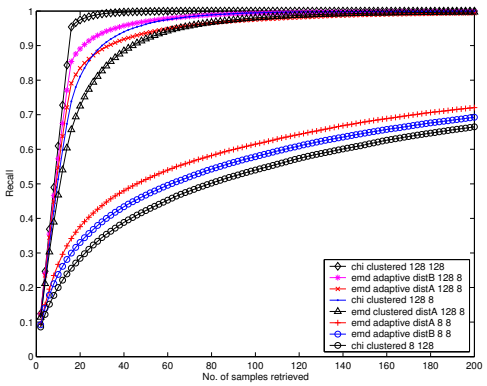
(a) χ^2 with various number of histogram bins



(b) EMD and χ^2 at 8 pixels and 8 bins



(c) EMD and χ^2 at 8 pixels and 4 bins



(d) EMD and χ^2 at 128 pixels per sample

Fig. 2. Random sampling test results.