A Fast Distance Between Histograms

Francesc Serratosa[1] and Alberto Sanfeliu[2]

[1] Universitat Rovira I Virgili, Dept. d'Enginyeria Informàtica i Matemàtiques, Spain
francesc.serratosa@.urv.net
[2] Universitat Politècnica de Catalunya, Institut de Robòtica i Informàtica Industrial, Spain
sanfeliu@iri.upc.es

Abstract. In this paper we present a new method for comparing histograms. Its main advantage is that it takes less time than previous methods.

The present distances between histograms are defined on a structure called signature, which is a lossless representation of histograms. Moreover, the type of the elements of the sets that the histograms represent are ordinal, nominal and modulo.

We show that the computational cost of these distances is $O(z')$ for the ordinal and nominal types and $O(z'^2)$ for the modulo one, where z' is the number of non-empty bins of the histograms. In the literature, the computational cost of the algorithms presented depends on the number of bins in the histograms. In most applications, the histograms are sparse, so considering only the non-empty bins dramatically reduces the time needed for comparison.

The distances we present in this paper are experimentally validated on image retrieval and the positioning of mobile robots through image recognition.

1 Introduction

A histogram of a set with respect a measurement represents the frequency of quantified values of that measurement in the samples. Finding the distance or similarity between histograms is important in pattern classification or clustering and image retrieval. Several measures of similarity between histograms have therefore been used in computer vision and pattern recognition.

Most of the distance measures in the literature (there is an interesting compilation in [1]) consider the overlap or intersection between two histograms as a function of the distance value but do not take into account the similarity in the non-overlapping parts of the two histograms. For this reason, Rubner presented in [2] a new definition of the distance measure between histograms that overcomes this problem of non-overlapping parts. Called Earth Mover's Distance, it is defined as the minimum amount of work that must be performed to transform one histogram into another by moving distribution mass. This author used the simplex algorithm. Later, Cha presented in [1] three algorithms for obtaining the distance between one-dimensional histograms that use the Earth Mover's Distance. These algorithms compute the distance between histograms when the type of measurements are *nominal*, *ordinal* and *modulo* in $O(z)$, $O(z)$ and $O(z^2)$, respectively, and where z the number of levels or bins.

Often, for specific set measurements, only a small fraction of the *bins* in a histogram contains significant information, i.e. most of the *bins* are empty. This is more frequent

when the dimensions of the element domain increase. In such cases, the methods that use histograms as fixed-sized structures are not very efficient. For this reason, Rubner [2] presented variable-size descriptions called *signatures*, which do not explicitly consider the empty bins.

If the statistical properties of the data are known *a priori*, the similarity measures can be improved by smoothing projections, as we can see in [3]. In [4] an algorithm was presented that used the *intersection function*, L_1 *norm*, L_2 *norm* and X^2 *test* to compute the distance between histograms. In [5], the authors performed image retrieval based on colour histograms. Because the distance measure between colours is computationally expensive, they presented a low dimensional and easy-to-compute distance measure and showed that this was a lower boundary for the colour-histogram distance measure. An exact histogram-matching algorithm was presented in [6]. The aim of this algorithm was to study how various image characteristics affect colour reproduction by perturbing them in a known way.

Given two histograms, it is often useful to define a quantitative measure of their dissimilarity in order to approximate perceptual dissimilarity as well as possible. We therefore believe that a good definition of the distance between histograms needs to consider the distance between the basic features of the elements of the set i.e. similar pairs of histograms defined from different basic features may obtain different distances between histograms. We call the distance between set elements the *ground distance*.

In this paper we present the distances between histograms whose computational cost depends only on the non-empty bins rather than, as in the algorithms in [1,2], on the total number of bins. The type of measurements are *nominal*, *ordinal* and *modulo* and the computational cost is $O(z')$, $O(z')$ and $O(z'^2)$, respectively, where z' is the number of non-empty bins in the histograms. In [7], we show that these distances are the same as the distances between the histograms in [1] but that the computational time for each comparison is lower when the histograms are large or sparse. We also depict the algorithms to compute them not shown here due to lack of space.

The next sections are organised as follows. In section 2 we define the histograms and signatures. In section 3 we present three possible types of measurements and their related distances. In section 4 we use these distances as ground distances when defining the distances between signatures. In section 6 we address image retrieval problem with the proposed distance measures. Finally, we conclude by stressing the advantage of using the distance between signatures.

2 Histograms and Signatures

In this section, we formally define histograms and signatures. We end this section with a simple example to show the representations of the histograms and signatures given a set of measurements.

2.1 Histogram Definition

Let x be a measurement that can have one of T values contained in the set $X=\{x_1,...x_T\}$. Consider a set of n elements whose measurements of the value of x are $A=\{a_1,...a_n\}$, where $a_t \in X$.

The histogram of the set A along measurement x is $H(x,A)$, which is an ordered list consisting of the number of occurrences of the discrete values of x among the a_t. As we are interested only in comparing the histograms and sets of the same measurement x, $H(A)$ will be used instead of $H(x,A)$ without loss of generality. If $H_i(A)$, $1 \le i \le T$, denotes the number of elements of A that have value x_i, then $H(A)=[H_1(A), ...,H_T(A)]$ where

$$H_i(A) = \sum_{t=1}^{n} C_{i,t}^{A} \quad \text{and} \quad C_{i,t}^{A} = \begin{cases} 1 & \text{if } a_t = x_i \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

The elements $H_i(A)$ are usually called *bins* of the histogram.

2.2 Signature Definition

Let $H(A)=[H_1(A), ...,H_T(A)]$ and $S(A)=[S_1(A), ...,S_z(A)]$ be the histogram and the signature of the set A, respectively. Each $S_k(A)$, $1 \le k \le z \le T$ comprises a pair of terms, $S_k(A)=\{w_k, m_k\}$. The first term, w_k, shows the relation between the signature $S(A)$ and the histogram $H(A)$. Therefore, if the $w_k=i$ then the second term, m_k, is the number of elements of A that have value x_i, i.e. $m_k=H_i(A)$ where $w_k<w_t \Leftrightarrow k<t$ and $m_k>0$.

The signature of a set is a lossless representation of its histogram in which the *bins* of the histogram whose value is 0 are not expressed implicitly. From the signature definition, we obtain the following expression,

$$H_{w_k}(A) = m_k \quad where \quad 1 \le k \le z \tag{2}$$

2.3 Extended Signature

The **extended signature** is one in which some empty bins have been added. That is, we allow $m_i=0$ for some bins. This is a useful structure for ensuring that, given a pair of signatures to be compared, the number of bins is the same and that each bin in both signatures represents the same bin in the histograms.

2.4 Example

In this section we show a pair of sets with their histogram and signature representations. This example is used to explain the distance measures in the next sections. Figure 1 shows the sets A and B and their histogram representations. Both sets have 10 elements between 1 and 8. The horizontal axis in the histograms represents the values of the elements and the vertical axis represents the number of elements with this value.

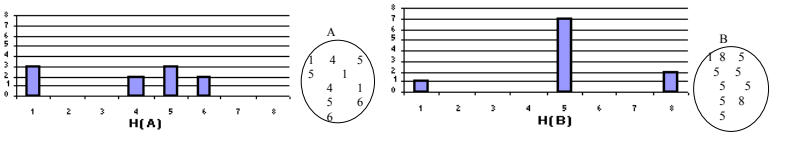

Fig. 1. Sets A and B and their histograms

Figure 2 shows the signature representation of sets *A* and *B*. The length of the signatures is 4 and 3, respectively. The vertical axis represents the number of elements of each bin and the horizontal axis represents the bins of the signature. Set A has 2 elements with a value of 6 since this value is represented by the bin 4 (W_4^A=6) and the value of the vertical axis is 2 at bin 4.

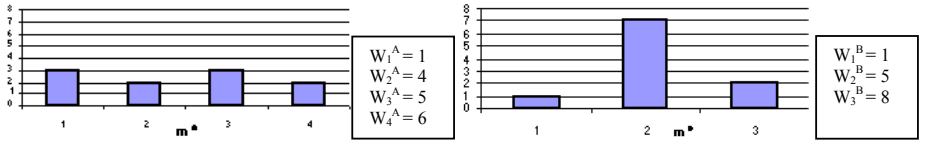

Fig. 2. Signature representation of the sets A and B

Figure 3 shows the extended signatures of the sets *A* and *B* with *5* bins. Note that the value that the extended signatures represents for each bin, w_i, is the same for both signatures.

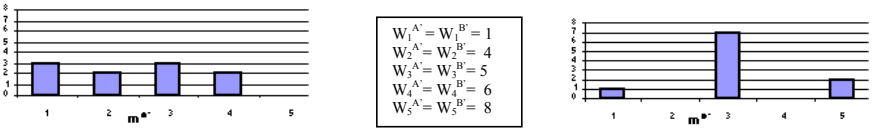

Fig. 3. Extended Signatures A' and B'. The number of elements m_i is represented graphically and the value of its elements is represented by w_i.

3 Type of Measurements and Distance Between Them

We consider three types of measurements, called nominal, ordinal and modulo. In a nominal measurement, each value of the measurement is a name and there is no relation, such as greater than or lower than, between them (e.g. the names of students). In an ordinal measurement, the values are ordered (e.g. the age of the students). Finally, in a modulo measurement, the values are ordered but they form a ring because of the arithmetic modulo operation (e.g. the angle in a circumference).

Corresponding to these three types of measurements, we define three measures of difference between two measurement levels $a \in X$ and $b \in X$, as follows:

a) Nominal distance:

$$d_{nom}(a,b) = \begin{cases} 0 & if \quad a = b \\ 1 & otherwise \end{cases} \tag{3}$$

The distance value between two nominal measurement values is either match or mismatch, which are mathematically represented by 0 or 1.

b) Ordinal distance:

$$d_{ord}(a,b)=|a-b| \tag{4}$$

The distance value between two ordinal measurement values is computed by the absolute difference of each element.

c) Modulo distance:

$$d_{\mathrm{mod}}(a,b)=\begin{cases} |a-b| & if\ |a-b|\le T/2 \\ T-|a-b| & otherwise \end{cases} \tag{5}$$

The distance value between two modulo measurement values is the interior difference of each element.

4 Distance Between Signatures

In this section, we present the nominal, ordinal and modulo distances between signatures. For the following definitions of the distances and for the algorithms section, we assume that the extended signatures of *S(A)* and *S(B)* are *S(A')* and *S(B')*, respectively, where $S_i(A')=\{w_i^{A'},m_i^{A'}\}$ and $S_i(B')=\{w_i^{B'},m_i^{B'}\}$. The number of bins of *S(A)* and *S(B)* is z^A and z^B and the number of bins of both extended signatures is *z'*.

4.1 Nominal Distance

The nominal distance between the histograms in [5] is the number of elements that do not overlap or intersect. We redefine this distance using signatures as follows,

$$D_{nom}(S(A),S(B))=\sum_{i=1}^{z'}\left|m_i^{A'}-m_i^{B'}\right| \tag{6}$$

4.2 Ordinal Distance

The ordinal distance between two histograms was presented in [6] as the minimum work needed to transform one histogram into another. Histogram *H(A)* can be transformed into histogram *H(B)* by moving elements to the left or to the right and the total number of all the necessary minimum movements is the distance between them. There are two operations. Suppose an element *a* that belongs to bin *i*. One operation is *move left (a)*. This result of this operation is that element *a* belongs to bin *i-1* and its cost is 1. This operation is impossible for the elements that belong to bin *1*. Another operation is *move right (a)*. Similarly, after this operation, *a* belongs to bin *i+1* and the cost is 1. The same restriction applies to the right-most bin. These operations are graphically represented by right-to-left arrows and left-to-right arrows. The total number of arrows is the distance value. This is the shortest movement and there is no other way to move elements in shorter steps and transform one histogram to the other. The distance between signatures is defined as follows,

$$D_{ord}(S(A),S(B))=\sum_{i=1}^{z'-1}\left[\left(w_{i+1}^{A'}-w_i^{A'}\right)\left|\sum_{j=1}^{i}\left(m_j^{A'}-m_j^{B'}\right)\right|\right] \tag{7}$$

The arrows do not have a constant size (or constant cost) but depend on the distance between bins. If element *a* belongs to bin *i*, the result of operation *move left (a)* is that the element *a* belongs to bin *i-1* and its cost is $w_i - w_{i-1}$. Similarly, after the operation *move right(a)*, the element *a* belongs to bin *i+1* and the cost is $w_{i+1} - w_i$. In equation (7), the number of arrows that go from bin *i* to bin *i+1* is described by the inner addition and the cost of these arrows is $w_{i+1} - w_i$.

4.3 Modulo Distance

One major difference in modulo type histograms or signatures is that the first bin and the last bin are considered to be adjacent to each other. It therefore forms a closed circle due to the nature of the data type. Transforming a modulo type histogram or signature into another while computing their distance should allow cells to move from the first bin to the last bin, or vice versa, at the cost of a single movement. Now, cells or blocks of earth can move from the first bin to the last bin with the operation *move left (1)* in the histogram case or *move left* (w_1) in the signature case. Similarly, blocks can move from the last bin to the first one with the operations *move right (T)* in the histogram case or *move right* ($w_{z'}$) in the signature case.

The cost of these operations is calculated as for the cost of the operations in the ordinal distance except for the movements of blocks from the first bin to the last or vice versa. For the distance between histograms, the cost, as in all the movements, is one. For the distance between signatures, the real distance between bins or the length of the arrows has to be considered. The cost of these movements is therefore the sum of three terms (see figure 4.a): (a) the cost from the last bin of the signature, $w_{z'}$, to the last bin of the histogram, *T*; (b) the cost from the last bin of the histogram, *T*, to the first bin of the histogram, *1*; (c) the cost from the first bin of the histogram, *1*, to the first bin of the signature, w_1. The costs are then calculated as the length of these terms. The cost of (a) is $T-w_{z'}$, the cost of (b) is *1* (similar to the cost between histograms) and the cost of (c) is w_1-1. Therefore, the final cost from the last bin to the first or vice versa between signatures is $w_1-w_{z'}+T$.

Example. Figure 4.b shows graphically the minimum arrows needed to get the modulo distance in (a) the histogram case and (b) the signature case. The distance is

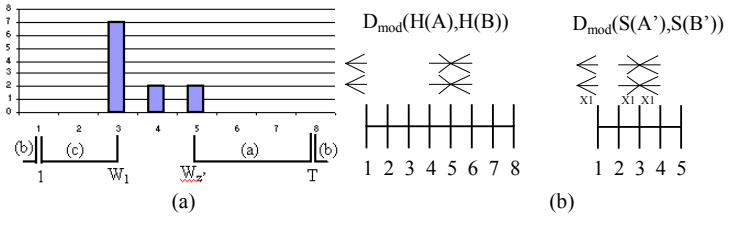

Fig. 4. (a) The three terms that need to be considered in order to compute the cost of moving blocks from the last bin to the first or vice versa in the modulo distance between signatures. (b) Arrow representation of the modulo distance in case of the histograms and signatures.

obtained as in the ordinal example except that the arrows from the first bin to the last are allowed or vice versa. The value of the distance between signatures is 2x1+2x1+2x1=6. In this signature representation, the cost of the two arrows that go from the first bin to the last bin is one. This is because w_1=1 (the first bin in the histogram representation) and w_5=8 (the last bin in the histogram representation, T=8). This cost is then 1-8-8=1.

Due to the previously explained modulo properties, we can transform one signature or histogram into another in several ways. In one of these ways, there is a minimum distance whose number of movements (or the cost of the arrows and the number of arrows) is the lowest. If there is a borderline between bins that has both directional arrows, they are cancelled out. These movements are redundant, so the distance cannot be obtained through this configuration of arrows. To find the minimum configuration of arrows, we can add a complete chain in the histogram or signature of the same directional arrows and the opposite arrows on the same border between bins are then cancelled out. The modulo distance between signatures is defined as

$$D_{\mathrm{mod}}(S(A),S(B)) = \min_{c}\left\{\sum_{i=1}^{z'-1}\left[\left(w_{i+1}^{A'}-w_{i}^{A'}\right)\left|c+\sum_{j=1}^{i}\left(m_{j}^{A'}-m_{j}^{B'}\right)\right|\right]+\left(w_{1}^{A'}-w_{z'}^{A'}+T\right)|c|\right\} \tag{8}$$

The cost of moving a block of earth from one bin to another is not 1 but the length of the arrows or the distance between the bins (as explained in the ordinal distance between signatures). The cost of the movement of blocks from the first bin to the last or vice versa is w_1-$w_{z'}$+T and the cost of the other movements is $w^{A'}{}_{i+1}$-$w^{A'}{}_{i}$. The term c represents the chains of left arrows or right arrows added to the current arrow representation. The absolute value of c at the end of the expression is the number of chains added to the current representation. It comes from the cost of the arrows from the last bin to the first or vice versa.

Example. Figure 5 shows five different transformations of signature *S(A)* to signature *S(B)* and their related costs. In the first transformation, one chain of right arrows is added (c=1). In the second transformation, no chains are added (c=0), so the cost is the same as the ordinal distance. In the third to the last transformations, 1, 2 and 3 chains of left arrows are added, respectively. We can see that the minimum cost is 6 and c=-2, the distance value is 6 for the modulo distance and 14 for the ordinal distance.

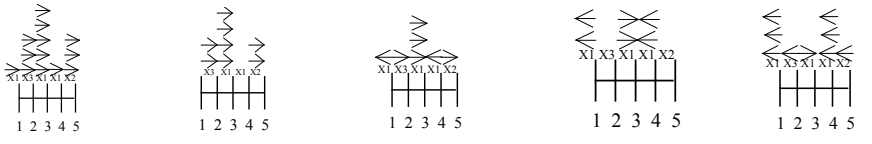

Fig. 5. Five different transformations of signature S(A) to signature S(B) with their related c and cost obtained

5 Experiment with Colour Images

To show the validity of our new method, we first tested the ordinal and modulo distances between histograms and between signatures. We used 1000 images (640 x 480 pixels) obtained from public databases. To validate the ordinal distance, we calculated the histograms from the illumination coordinate with 2^8 levels (table 1) and with 2^{16} levels (table 3). Also, to test the modulo distance, the histograms represented the hue coordinate with 2^8 levels (table 2) and with 2^{16} levels (table 4). Each table below shows the results of 5 different tests. In the first and second rows, the distance between histograms and signatures, respectively, are computed. In the other three rows, the distance between signatures is computed but, in order to reduce the length of the signature (and therefore increase the speed), the bins with fewer elements than 100, 200 or 300 in tables 1 and 2 and fewer elements than 1, 2 or 3 in tables 3 and 4 were removed. The first column shows the number of bins of the histogram (first cell) or signatures (the other four cells). The second column shows the increase in speed if we use signatures instead of histograms. It is calculated as the ratio between the run time of the histogram method and that of the signature method. The third column shows the average correctness. The last column shows the decrease in correctness as a result of using the signatures with filtered histograms, which is obtained as the ratio of the correctness of the histogram to the correctness of each filter.

Tables 1 to 4 show that our method is more useful when the number of levels increases, since the number of empty bins tends to increase. Moreover, the increase is greater when comparing the histograms of the hues, because the algorithm has a quadratic computational cost. Note that in the case of the first filter (third experiment in the tables), there is no decrease in correctness although the increase in speed is greater than with the signature method.

Table 1. Illumination 2^8 bins. Ordinal histogram.

	Length	Increase Speed	Correct.	Decrease in Correct.
Histo.	265	1	78%	1
Signa.	235	1.12	78%	1
Sig100	157	1.68	78%	1
Sig200	106	2.50	69%	0.88
Sig300	57	4.64	57%	0.73

Table 2. Hue 2^8 bins. Modulo histogram.

	Length	Increase Speed	Correct.	Decrease in Correct.
Histo.	265	1	86%	1
Signa.	215	1.23	86%	1
Sig100	131	2.02	85%	0.98
Sig200	95	2.78	73%	0.84
Sig300	45	5.88	65%	0.75

Table 3. Illumination 2^{16} bins. Ordinal histogram.

	Length	Increase Speed	Correct.	Decrease in Correct.
Histo.	65,536	1	81%	1
Signa.	245	267.49	81%	1
Sig. 1	115	569.87	81%	1
Sig. 2	87	753.28	67%	0.82
Sig. 3	32	2048.00	55%	0.67

Table 4. Hue 2^{16} bins. Modulo histogram.

	Length	Increase Speed	Correct.	Decrease in Correct.
Histo.	65,536	1	89%	1
Signa.	205	319.68	89%	1
Sig. 1	127	516.03	89%	1
Sig. 2	99	661.97	78%	0.87
Sig. 3	51	1285.01	69%	0.77

6 Conclusions

We have presented the nominal, ordinal and modulo distance between signatures. We have shown that signatures are a lossless representation of histograms and that computing the distances between signatures is the same as computing the distances between histograms but with a lower computational time. We have validated these new distances with a huge amount of real images and observed an important saving of time since most of the histograms are sparse. Moreover, when we applied filtering techniques to the histograms, the number of bins of the signatures decreased, so the run time of their comparison also decreased.

References

1. S.-H. Cha, S. N. Srihari, "On measuring the distance between histograms" Pattern Recognition 35, pp: 1355–1370, 2002.
2. Y. Rubner, C. Tomasi, and L. J. Guibas, "A Metric for Distributions with Applications to Image Databases" International Journal of Computer Vision 40 (2), pp: 99-121, 2000.
3. J.-K. Kamarainen, V. Kyrki, J. Llonen, H. Kälviäinen, "Improving similarity measures of histograms using smoothing projections", Pattern Recognition Letters 24, pp: 2009–2019, 2003.
4. F.-D. Jou, K.-Ch. Fan, Y.-L. Chang, "Efficient matching of large-size histograms", Pattern Recognition Letters 25, pp: 277–286, 2004.
5. J.Hafner, J.S. Sawhney, W. Equitz, M. Flicker & W. Niblack, "Efficient Colour Histogram Indexing for Quadratic Form Distance Functions", Trans. On Pattern Analysis and Machine Intelligence, 17 (7), pp: 729-735, 1995.
6. J. Morovic, J. Shaw & P-L. Sun, "A fast, non-iterative and exact histogram matching algorithm", Pattern Recognition Letters 23, pp:127–135, 2002.
7. F. Serratosa & A. Sanfeliu, "Signatures versus Histograms: Definitions, Distances and Algorithms", Submitted to Pattern recognition, 2005.