Signatures versus Histograms: Definitions, Distances and Algorithms

Francesc Serratosa[1] & 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. The aim of this paper is to present a new method to compare histograms. The main advantage is that there is an important time-complexity reduction respect the methods presented before. This reduction is statistically and analytically demonstrated in the paper.
The distances between histograms that we present 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, being z' the number of non-empty bins of the histograms. The computational cost of the algorithms presented in the literature depends on the number of bins of the histograms. In most of the applications, the obtained histograms are sparse, then considering only the non-empty bins makes the time consuming of the comparison drastically decrease.
The distances and algorithms presented in this paper are experimentally validated on the comparison of images obtained from public databases and positioning of mobile robots through the recognition of indoor scenes (captured in a learning stage).

1. Introduction

A histogram of a set with respect to a measurement represents the frequency of quantified values of that measurement among the samples. Finding the distance or similarity between histograms is an important issue in pattern classification or clustering and image retrieval. For this reason, a number of measures of similarity between histograms have been proposed and used in computer vision and pattern recognition. Protein classification is one of the common histogram applications [9]. Moreover, if the ordering of the elements in the sample is unimportant, the histogram obtained from this set is a lossless representation of it and can be reconstructed from its histogram. Then, we can compute the distance between sets in an efficient way by computing the distance between their histograms.
The probabilistic approaches use histograms based on the fact that the histogram of a measurement provides the basis for an empirical estimate of the probability density function [1]. Computing the distance between probability density functions can be regarded as the same as computing the Bayes probability. This is equivalent to measuring the overlap between probability density functions as the distance. The *B-distance* [2], proposed by Kailath, measures the distance between populations. It is a value between 0 and 1 and provides bounds on the Bayes misclassification probability. An approach closely related to the *B-distance* was proposed by Matusita [3]. Finally, Kullback generalised the concept of probabilistic uncertainty or

"entropy" and introduced the *K-L-distance* measure [1,4] that is the minimum cross entropy.

Most of the distance measures presented in the literature (there is an interesting compilation in [5]) consider the overlap or intersection between two histograms as a function of the distance value but they do not take into account the similarity on the non-overlapping parts of the two histograms. For this reason, Rubner presented in [6] a new definition of the distance measure between histograms that overcomes this non-overlapping parts problem. It was called Earth Mover's Distance and it is defined as the minimum amount of work that must be performed to transform one histogram into the other one by moving distribution mass. They used the simplex algorithm [8] to compute the distance measure and the method presented in [7] to search a good initialisation. Latter, Cha presented in [5] three algorithms to obtain the distance between one-dimensional histograms that use the Earth Mover's Distance. These algorithms computed the distance between histograms when the type of measurements where *nominal*, *ordinal* and *modulo* in $O(z)$, $O(z)$ and $O(z^2)$ respectively, being z the number of levels or bins.

Often, for specific set measurements, only a small fraction of the *bins* in a histogram contain significant information, that is, most of the *bins* are empty. This is more frequent when the dimensions of the element domain increase. In that cases, the methods that use histograms as fixed-sized structures obtain poor efficiency. For this reason, Rubner [6] presented the variable-size descriptions called *signatures*. In that representations, the empty bins where not explicitly considered.

If the statistical properties of the data are a priori known, the similarity measures can be improved by the smoothing projections as it was shown in [10]. Moreover, these projections can be applicable for reduction of the dimensionality of the data and also to represent sparse data in a more tight form in the projection subspace.

Given two histograms, it is often useful to define a quantitative measure of their dissimilarity with the intent of approximating perceptual dissimilarity as well as possible. To that aim, we consider that a good definition of a distance between histograms needs to take into consideration a distance between the basic features of the elements of the set. That is, similar pairs of histograms defined from different basic features may obtain different distance value between histograms. We call the distance between set elements the *ground distance*.

In [12], they performed image retrieval based on colour histograms. Do to the distance measure between colours is computationally expensive, they presented a low dimensional and easy to compute distance measure. They show that this measure is a lower bound on the colour-histogram distance measure.

An exact histogram-matching algorithm was presented in [13]. The aim of this algorithm was to study the influence of various image characteristics on colour reproduction by perturbing them in a known way. Furthermore, this perturbation would be done in a way whereby a set of heterogeneous images would be the starting point and this set would be transformed so as to make their histogram the same for all the images. The aim of the algorithm was not the comparison of histograms but to arrive to a transformation look up table and transform the target image according to it.

It was presented in [11] an algorithm to compute the distance between histograms that used the *intersection function*, L_1 *norm*, L_2 *norm* and X^2 *test*. The main feature of this algorithm was that the input was a built histogram (obtained from the target image) and another image. Then, it was not necessary to build the histogram of the image of the database to compute the distance between histograms.

Finally, the applications of the references commented before use the histograms as global information of images. Histograms can also be used in structural pattern recognition. For instance, Serratosa defined the Function-Described Graphs [14], which is structure that represents a cluster of Attributed Graphs in which there is a probability density function in each node of the structure described by a histogram. Thus, to compare clusters (that is, to compare Function-Described Graphs), it is needed a distance between histograms to compare each of their nodes. Latter, the same authors defined the Second-Order Random Graphs [15]. This structure represents also a cluster of Attributed Graphs but there is much amount of information since there is a joint probability in each node described by a 2-dimensional histogram. The computational cost of comparing graphs is exponential respect the number of nodes in the worst case. There are some efficient algorithms that obtain sub-optimal distances in polynomial cost respect the number of nodes [16]. For this reason, it is important to reduce the time consuming comparing their nodes.
In this paper, we present the algorithms to compute the distances between histograms that the computational cost depends only on the non-empty bins instead of the number of bins as it is in the algorithms presented in [5,6]. The type of measurements where *nominal*, *ordinal* and *modulo* and the computational cost where $O(z')$, $O(z')$ and $O(z'^2)$ respectively, being z' the number of non-empty bins of the histograms. We show that these distances obtain the same value than the distances between histograms presented in [5] although the computational time for each comparison decreases when the histograms have a large size or they are sparse. Furthermore, we suppose that we do not have a priori probabilistic information of the histograms. For this reason, the methods presented in [10] are not useful.
The subsequent sections are constructed as follows. First, we define the histograms and signatures. Then in section 3 we present three possible types of measurements and their related distances. These distances will be used in the next section to define the distances between signatures. In section 5, we depict the basic algorithms to compute the distances between signatures. In section 6 we validate our algorithms on two different scenarios. The histograms to be compared are obtained from images obtained from databases and indoor scenes, respectively. Finally, we conclude with emphasis of the advantage of using the distance between signatures and using the proposed algorithms.

2. Histograms & Signatures

In this section, we formally give a definition of histograms and signatures. The section finishes 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 which 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

$$H_i(A) = \sum_{t=1}^{n} C_{i,t}^{A} \tag{1}$$

and the individual costs are defined as

$$C_{i,t}^{A} = \begin{cases} 1 & if\ a_t = x_i \\ 0 & otherwise \end{cases} \tag{2}$$

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 $S_k(A)$, $1 \le k \le z \le T$ is composed by 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 $H(A)$. Thus, if the $w_k=i$ then the second term, m_k, is the number of elements of A that have value x_i, that is, $m_k=H_i(A)$ 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 that has value 0 are not expressed implicitly. From the signature definition, we obtain the following expression,

$$H_{w_k}(A) = m_k \quad where \quad 1 \le k \le z \tag{3}$$

2.3. Extended Signature

The **extended signature** is a signature in which the minimum number of empty bins have been added to assure that, given a pair of signatures to be compared, the number of bins is the same. Moreover, 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 and values are contained from 1 to 8. Horizontal axis in the histograms represents the values of the elements and the vertical axis represents the number of elements that have this value, that is m_i. Empty bins are the ones that $m_i=0$.

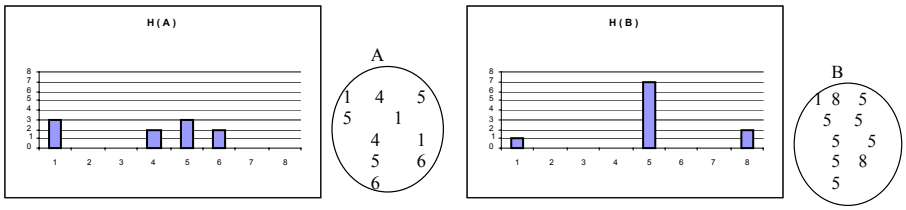


Figure 1. Sets *A* and *B* and its histograms.

Figure 2 shows the signature representation of the 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. The set A has 2 elements with value 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. In the signature representation there is not any empty bin.

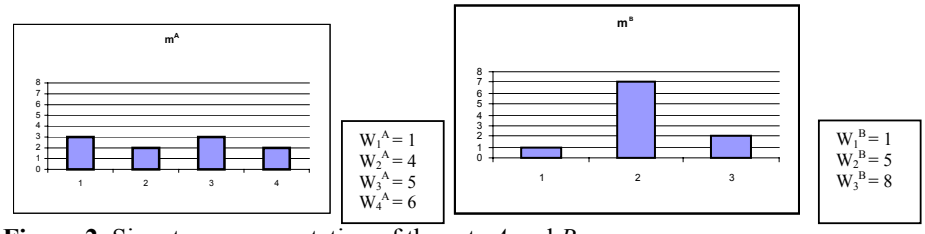


Figure 2. Signature representation of the sets *A* and *B*.

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. Moreover, in *A'* and *B'*, one and two empty bins have been added, respectively.

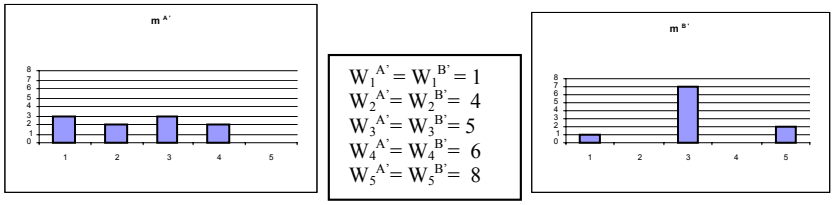


Figure 3. Extended Signatures *A'* and *B'*. The number of elements m_i are represented graphically and the value of its elements is represented by w_i.

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 an there is not any relation between them such as great than or lower than (e.g. the names of the students). In an ordinal measurement, the values are ordered (e.g. the age of the students). Finally, in the modulo measurement, measurement values are ordered but form a ring due to the arithmetic modulo operation (e.g. the angle in a circumference). Corresponding to the three types of measurements mentioned before, we define three measures of difference between two measurement levels $a \in X$ and $b \in X$ as follows:

a) Nominal distance:

$$d_{nom}(a,b)=\begin{cases}0 & if \quad a=b\\ 1 & otherwise\end{cases} \tag{4}$$

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

b) Ordinal distance:

$$d_{ord}(a,b)=|a-b| \tag{5}$$

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

c) Modulo distance:

$$d_{\mathrm{mod}}(a,b)=\begin{cases}|a-b| & if\ |a-b|\le T/2\\ T-|a-b| & otherwise\end{cases} \tag{6}$$

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

Metric Property. The three measures in equations (4)-(6) satisfy the following necessary properties of a metric:

a) Reflexivity: d(a,b)=0.

b) Non-negativity: d(a,b)≥0.

c) Commutativity: d(a,b)= d(b,a).

d) Triangle inequality: d(a,c)≤ d(a,b)+ d(b,c).

Proof. Since they are straightforward facts, we omit the proofs. Moreover, the proof of the triangle inequality for the modulo distance is depicted in [5].

4. Distance between Signatures

The aim of this section is to present the new distances between signatures. To do so and for each type of elements (nominal, ordinal and modulo), we first show de definition of the distance between histograms and then we move on the new definitions of the distance between signatures. The algorithms used to obtain the extended signatures and the three distances are described in the algorithms section.

For the following definitions of the distances and also 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 histograms presented in [5] is the number of elements that do not overlap or intersect. It is defined as follows,

$$D_{nom}(H(A),H(B))=\sum_{i=1}^{z}\left|H_i(A)-H_i(B)\right| \quad (7)$$

We define this distance through their extended signatures as follows,

$$D_{nom}(S(A),S(B))=\sum_{i=1}^{z'}\left|m_i^{A'}-m_i^{B'}\right| \quad (8)$$

Theorem 1. *The nominal-distance value between signatures is the same than the nominal-distance value between histograms.*

Proof theorem 1. The bins in the histograms that are not represented explicitly in the signatures are the ones that in both histograms are empty, $H_i(A) = H_i(B) = 0$. Then, the addition of these bins does not affect on the distance value.

Example. We consider the extended signatures *A'* and *B'* shown in figure 3. The nominal distance is defined as the addition of the difference between the number of elements. In this case it is 2+2+4+2+2=12.

4.2. Ordinal Distance

The ordinal distance between two histograms was presented in [6] as the minimum of work needed to transform one histogram to the other. Histogram *H(A)* can be transformed into histogram *H(B)* by moving elements to left or right and the total of all necessary minimum movements is the distance between them. There are two operations. Suppose an element *a* that belong to the bin *i*. One operation is *move left (a)*. This operation results that the element *a* belong to bin *i-1* and the cost to do so is 1. This operation is impossible to the elements that belong to the bin *1*. Another operation is *move right (a)*. Similarly, after the operation, *a* belongs to the 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. Figure 4 shows the arrows needed to transform (a) histogram *H(A)* to histogram *H(B)* and (b) the extended signature *S(A')* to *S(B')*. The total number of arrows is the distance value. It 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 histograms was defined in [5] as follows,

$$D_{ord}(H(A),H(B))=\sum_{i=1}^{T-1}\left|\sum_{j=1}^{i}\left(H_j(A)-H_j(B)\right)\right| \quad (9)$$

There is a slight difference between equation (9) and the equation of the distance presented in [5]. They calculated the case *i=T*, we have not considered it in equation (9) since the addition of all the arrows is always 0 when the sets have the same number of elements.

We have defined our new distance between signatures similarly to the distance between histograms as follows,

$$D_{ord}(S(A),S(B)) = \sum_{i=1}^{z'-1}\left[\left(w_{i+1}^{A'} - w_i^{A'}\right)\left|\sum_{j=1}^{i}\left(m_j^{A'} - m_j^{B'}\right)\right|\right] \qquad (10)$$

The main difference is that we have to take into consideration that the difference between bins is not constant. In equation (9), the number of arrows that goes from bin *i* to bin *i+1* is described by $\left|\sum_{j=1}^{i}(H_j(A) - H_j(B))\right|$ and the cost of one arrow (or the operation *move right* or *move left*) is 1 as described before. Our arrows have not a constant size (or constant cost) but they depend on the distance between bins. If element *a* belongs to the bin *i*, the operation *move left (a)* results that the element *a* belong to bin *i-1* and the cost to do so is $w_i - w_{i-1}$. Similarly, after the operation *move right(a)*, the element *a* belongs to the bin *i+1* and the cost is $w_{i+1} - w_i$. In equation (10), the number of arrows that goes from bin *i* to bin *i+1* is described by $\left|\sum_{j=1}^{i}\left(m_j^{A'} - m_j^{B'}\right)\right|$ and the cost of these arrows is $w_{i+1} - w_i$.

In the extreme case in which the signature and the histogram have equal number of bins, all the arrows have length 1 do to $w_i - w_{i-1} = 1$ and we obtain similar expressions in both distances.

Example. Figure 4 shows the graphic representation of the arrows in the histogram distance (a) and in the signature distance (b). They represent the minimum necessary movements. In the case of the distance between histograms, the distance is the number of arrows. But in the signature case, the distance is the number of arrows multiplied by the length of the arrows (shown under the arrows). For instance, in the first arrows, the length is 3 since $w_1^{A'}$-w_2^{A}=4-1=3. The distance value between signatures is 3x2+1x4+2x2=14, which is the number of arrows in the histogram distance.

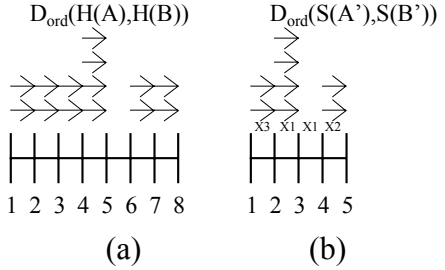


Figure 4. Arrow representation of the ordinal distance using (a) histograms and (b) signatures.

Theorem 2. *The ordinal-distance value between signatures is the same than the ordinal-distance value between histograms.*

The following lemma makes easier the demonstration of the theorem. First, suppose that the relation between the bins of the extended signatures and histograms is $w_k=i$ and $w_{k+1}=p$ being $p>i$.

Lemma 1. The accumulative addition of the difference between histograms is the same than the accumulative addition of the difference between extended signatures when $w_k=i$.

$$\sum_{j=1}^{i}\left(H_j(A)-H_j(B)\right)=\sum_{j=1}^{k}\left(m_j^{A'}-m_j^{B'}\right) \qquad (11)$$

Proof lemma 1. This is a straight-forward fact since the terms that $H_j(A)= H_j(B)=0$ are not considered in the extended signatures and also $H_j(A)- H_j(B)=0$.

Proof theorem 2.
By definition of the extended signatures we have,

$$\left.\begin{array}{l} H_i(A)\neq 0 \quad or \quad H_i(B)\neq 0 \\ H_{i+1}(A)=0 \quad and \quad H_{i+1}(B)=0 \\ \ldots \\ H_{p-1}(A)=0 \quad and \quad H_{p-1}(B)=0 \\ H_p(A)\neq 0 \quad or \quad H_p(B)\neq 0 \end{array}\right\} \quad (p-i-1)\equiv(w_{k+1}-w_k-1) \quad files \qquad (12)$$

then, it is easy to see that $\sum_{j=1}^{i}\left(H_j(A)-H_j(B)\right)=\sum_{j=1}^{p}\left(H_j(A)-H_j(B)\right)$. So, if we add the absolute value of these terms as follows, $\sum_{t=i}^{p}\left|\sum_{j=1}^{t}\left(H_j(A)-H_j(B)\right)\right|$, we get that this expression is similar to $(p-i)\left|\sum_{j=1}^{i}\left(H_j(A)-H_j(B)\right)\right|$. If we substitute $(p-i)$ by $(w_{k+1}-w_k)$ and we use equation (11) we arrive to the following expression $\sum_{t=i}^{p}\left|\sum_{j=1}^{t}\left(H_j(A)-H_j(B)\right)\right|=\left(w_{k+1}^{A'}-w_k^{A'}\right)\left|\sum_{j=1}^{k}\left(m_j^{A'}-m_j^{B'}\right)\right|$. This expression is true for all the bins, so we obtain equation (10) by adding all the terms.

4.3. Modulo Distance

One major difference in a modulo type histograms or signatures is that the first bin and the last bin are considered to be adjacent to each other, and hence, it forms a closed circle, due to the nature of the data type. Transforming a modulo type histogram or signature to another while computing their distance should allow cells to move from the first bin to the last one or vice versa at a cost of a single movement. Now, cells or blocks of earth can move from the fist 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 are calculated similarly to the cost of the operations in the ordinal distance except for the movements of blocks from the first bin to the last one or viceversa. In the case of the distance between histograms, the cost is one, as in all the movements. In the case of the distance between signatures, it has to be considered the real distance between bins or the length of the arrows. Thus, the cost of these movements are the addition of three terms (see figure 5). (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. Then, the costs are calculated as the lenght of these terms. The cost of (a) is T-$w_{z'}$, the cost of (b) is 1 (similarly 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 one or viceversa between signatures is w_1-$w_{z'}$+T.

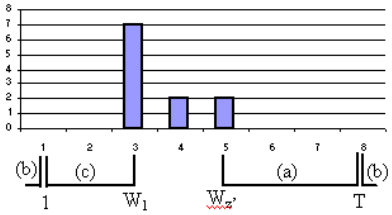


Figure 5. The three terms that have to be considered to compute the cost of moving blocks from the last bin to the first one or viceversa in the modulo distance between signatures.

Example. Figure 6 shows graphically the minimum arrows necessary to get the modulo distance in (a) the histogram case and (b) the signature case. The distance is obtained similarly to the ordinal example except that arrows from the first bin to the last one are allowed or vice versa. The value of the distance between signatures is 2x1+2x1+2x1=6. In ths signature representation, the cost of the two arrows that go from the first bin to the last bin is one. This is do to the fact that w_1=1 (first bin in the histogram representation) and w_5=8 (last bin in the histogram representation, T=8). Then this cost is 1-8-8=1.

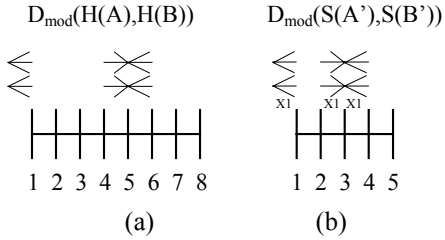
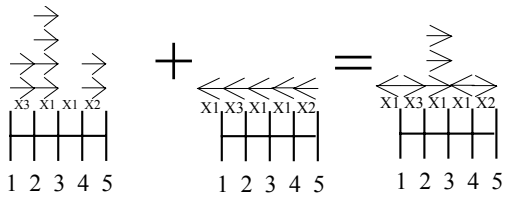


Figure 6. Arrow representation of the modulo distance using (a) histograms and (b) signatures.

Due to the modulo properties explained before, we can transform one signature or histogram into another one in several ways. Among these ways, there exists 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 border line between bins that has both directional arrows, they are cancelled out. These movements are redundant and so the distance cannot be obtained through this configuration of arrows. To find the

minimum cofiguration of arrows, we can add a complete chain in the histogram or signature of same directional arrows, then the opposite arrows on the same border between bins are cancelled out. Figure 7 shows the operation of adding a chain of left arrows to an arrow representation. The cost of the first representation is 3x2+ 1x4+ 1x0+ 2x2= 14 and the cost of the last representation is 1x1+ 3x1+ 1x3+ 1x1+ 2x1= 10.



(a) (b) (c)

Figure 7. (a) Arrow representation of the modulo distance between signatures. (b) Addition of a chain of left arrows. (c) The final arrow representation.

An algorithm to compute the modulo distance between histograms was presented in [5] although it was not described the mathematical expression of the distance. We propose here a new expression for the modulo distance that their algorithm calculates,

$$D_{\mathrm{mod}}\left(H(A),H(B)\right)=\min_{c}\left\{\sum_{i=1}^{T-1}\left[\left|c+\sum_{j=1}^{i}\left(H_j(A)-H_j(B)\right)\right|\right]+|c|\right\} \quad (13)$$

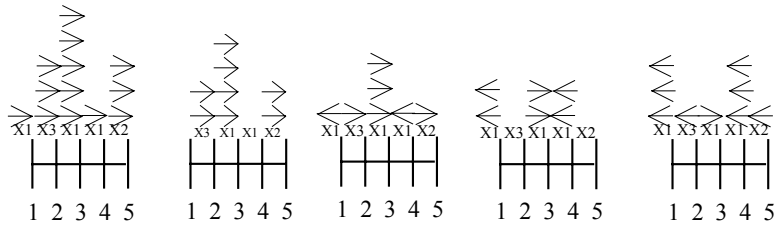
where *c* represents the chains of left arrows or right arrows added to the current arrow representation. The absolute value of *c* at the end of the expression is the number of chains added to the current representation. It comes from the cost of the arrows from the last bin to the first one or vice versa.

The modulo distance between signatures is defined similarly as follows,

$$D_{\mathrm{mod}}\left(S(A),S(B)\right)=\min_{c}\left\{\sum_{i=1}^{z'-1}\left[\left(w_{i+1}^{A'}-w_i^{A'}\right)\left|c+\sum_{j=1}^{i}\left(m_j^{A'}-m_j^{B'}\right)\right|\right]+\left(w_1^{A'}-w_{z'}^{A'}+T\right)|c|\right\} \quad (14)$$

This expression is similar to the one for the histograms. The main difference is that the cost of moving a block of earth from one bin to another one is not 1 but it is the length of the arrows or the distance between the bins (as it was explained in the ordinal distance between signatures). The cost of the movement of blocks from the first bin to the last one or viceversa is w_1-$w_{z'}$+T and the costs of the other movements is $w^{A'}{}_{i+1}$-$w^{A'}{}_{i}$.

Example. Figure 8 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 are added (c=1). In the second one, no chains are added (c=0), thus the cost is the same than the ordinal distance. In the third to the last ones, 1, 2 and 3 chains of left arrows are added, respectively. We can see that the minimum cost is 6 and it is the case that c=-2, then the distance value is 6 for the modulo distance and 14 for the ordinal distance.



c=1	c=0	c=-1	c=-2	c=-3
cost=22	cost=14	cost=10	cost=6	cost=12

Figure 8. Five different transformations of signature *S(A)* to the signature *S(B)* with their related *c* and the obtained cost.

Theorem 3. *The ordinal-distance value between signatures is the same than the ordinal-distance value between histograms.*

Proof theorem 3. The proof that both distances are the same is very similar to the one for the ordinal distance. We assume the situation of equation (12), then, we have that $c+\sum_{j=1}^{i}\left(H_j(A)-H_j(B)\right)=c+\sum_{j=1}^{p}\left(H_j(A)-H_j(B)\right)$. So, if we add the absolute value of these terms as follows, $\sum_{t=i}^{p}\left|c+\sum_{j=1}^{t}\left(H_j(A)-H_j(B)\right)\right|$, we get that this expression is similar to $(p-i)\left|c+\sum_{j=1}^{i}\left(H_j(A)-H_j(B)\right)\right|$. If we substitute $(p-i)$ by $(w_{k+1}-w_k)$ and we use equation (11) we arrive to the following expression $\sum_{t=i}^{p}\left|c+\sum_{j=1}^{t}\left(H_j(A)-H_j(B)\right)\right|=\left(w_{k+1}^{A'}-w_k^{A'}\right)\left|c+\sum_{j=1}^{k}\left(m_j^{A'}-m_j^{B'}\right)\right|$. This expression is true for all the bins, so we obtain equation (14) by adding all the terms.

5. Algorithms

We present the pseudo code of 4 algorithms. The first one extends two signatures, which is the first step to compute the distances between signatures. The other algorithms compute the distance between signatures.

5.1. Extended Signatures

Given two signatures, the process *Extended_Signature* obtains two minimum extended signatures and the number of bins of both extended signatures. The two extended signatures have the same number of bins but each one have the same information than the original signature. To do so, some bins with the null value have to be added.

```
{S(A'),S(B'),z'} = Extended_Signature {S(A),S(B)}
1. i=0 j=0 z'=0
2. while(i<z^A or j<z^B)
3.   if(w_i^A < w_j^B or j = z^B)
4.       w_z'^A' = w_z'^B' = w_i^A
```

```
5.       m_z'^A' = m_i^A   m_z'^B' = 0
6.       i++   z'++
7.    else if(w_i^A > w_j^B or i = z^A)
8.       w_z'^A' = w_z'^B' = w_j^B
9.       m_z'^A' = 0   m_z'^B' = m_j^B
10.      j++   z'++
11.   else // w_i^A = w_j^B and i < z^A and j < z^B
12.      w_z'^A' = w_z'^B' = w_i^A
13.      m_z'^A' = m_i^A   m_z'^B' = m_j^B
14.      i++   j++ z'++
```

Correctness of the procedure
The aim of the algorithm is to fill the extended signatures with the values of both signatures taking into consideration the order of the positions of the bins. We can discern into three different cases. In the first one, (lines 3 – 6), the extended signature *A'* is filled with information and *B'* with an empty bin. This is because the order of the bin in the signature *A* is smaller than the one in *B* or because there are no more bins in *B*. In the second one, (lines 7 - 10), we have the inverse situation. And in the last case, (lines 11 – 14), both bins in the signatures are non-empty and so their extended signatures are filled with the same value. The worst-case time complexity of this procedure is *O(z)*, being *z* the length of both histograms. This is the case when the intersection of the signatures is null and the union of them has not any non-empty bins. Then, the execution of the procedure never goes through lines 11 – 14 and the extended signatures have *z* bins. The best-case time complexity appears when both signatures and also the union of them have the same number of bins.

5.2. Nominal Distance

The process *Nominal_Distance* obtains the value of the nominal distance of two signatures.

```
D_nom = Nominal_Distance {S(A),S(B)}
{S(A'),S(B'),z'} = Extended_Signature {S(A),S(B)}
1.  D_nom = 0
2.  for (i = 1 to z')
3.     D_nom += abs(m_i^A' - m_i^B')
```

Correctness of the procedure
Since it is a straight-forward fact, we omit the proof. The time complexity of this procedure is *O(z')*, being *z'* the number of bins of the extended signatures. The worst case appears when the length of the extended signatures is the length of the histograms, *z'=z*.

5.3. Ordinal Distance

The process *Ordinal_Distance* obtains the value of the ordinal distance of two signatures.

```
D_ord = Ordinal_Distance {S(A),S(B)}
{S(A'),S(B'),z'} = Extended_Signature {S(A),S(B)}
1. D_ord = 0 p = 0
2. for (i = 1 to z'-1)
```

```
3.   p += m_i^A' - m_i^B'
4.   D_nom += (w_i+1^A' - w_i^A') * abs(p)
```

The algorithm computes, for each bin, the sum of the product of two terms. The first one is the length of each arrow (distance between the i^{th} bin and the $i+1^{th}$), represented by $(w_{i+1}^{A'} - w_i^{A'})$ and the second one is the number of arrows between the bins, represented by the absolute value of *p*.

Correctness of the procedure

The following lemma is crucial for the demonstration of the correctness of the algorithm. First, suppose that we have successfully constructed the arrow representation of the histograms such that the distance is the minimum.

Lemma 2. Let the variable of the algorithm *p* at step *i*, denote the number of arrows from the bin *i* to the bin *i+1* of the extended signatures. It is positive if arrows are heading to right or negative otherwise. The algorithm computes *p* as follows,

$$p = \sum_{j=1}^{i} \left(m_i^{A'} - m_i^{B'} \right) \tag{15}$$

Proof lemma 2. Consider two extended sub-signatures, $S_{1..i}(A')$ and $S_{1..i}(B')$ where bins are *1* to *i*. After transforming, population of $S_{1..i}(A')+p$ must be equal to that of $S_{1..i}(B')$. If $p \neq \sum_{j=1}^{i} m_i^{A'} - \sum_{j=1}^{i} m_i^{B'}$ then there is no way to transform $S_{1..i}(A')$ to $S_{1..i}(B')+p$. By contradiction, equation (15) holds.

Theorem 4. *The procedure Ordinal_Distance correctly finds the minimum distance between two signatures*.

Proof theorem 4. Note that, given a pair of signatures, the distribution of the arrows is the only variable of the distance since the length of the arrows is a constant. As equation (15) is true for all levels and it is the only way to transform one sub-signature to another one, the distance has to be obtained as this distribution of arrows. Therefore, the distance is obtained by $\sum_{j=1}^{i} \left[\left(m_i^{A'} - m_i^{B'} \right) | p | \right]$. This is equivalent to the equation of the distance (14) if *p* is substituted using equation (15).
The time complexity of this procedure is $O(z')$ as in the nominal case.

5.4. Modulo Distance

The process *Modulo_Distance* obtains the modulo distance of two signatures.

```
D_mod = Modulo_Distance {S(A),S(B)}
{S(A'),S(B'),z'} = Extended_Signature {S(A),S(B)}
1.  D_mod = 0 p[0] = m_0^A' - m_0^B'
2.  for (i = 2 to z') p[i] = m_i^A' - m_i^B' + p[i-1]
3.  for (i = 1 to z-1') D_mod += (w_i+1^A' - w_i^A') * abs(p[i])
4.  do
```

```
5.      D2=0
6.      c = min positive {p[i] for 1≤i≤z'}
7.      Temp[i]=p[i]-c for 1≤i≤z'
8.      for (i = 1 to z'-1) D2 += (w_{i+1}^{A'} - w_i^{A'}) * abs(Temp[i])
9.      if (D_mod > D2) D_mod = D2
10.     p[i]= Temp [i] for 1≤i≤z'
11. while(D_mod > D2)
12. do
13.     D2=0
14.     c = max negative {p[i] for 1≤i≤z'}
15.     Temp[i]=p[i]-c for 1≤i≤z'
16.     for (i = 1 to z'-1) D2 += (w_{i+1}^{A'} - w_i^{A'}) * abs(Temp[i])
17.     if (D_mod > D2) D_mod = D2
18.     p[i]= Temp [i] for 1≤i≤z'
19. while(D_mod > D2)
```

Correctness of the procedure

The arrow representation of minimum distance can be achieved from any arbitrary valid arrow representation by combination of two basic operations: Increasing the chains of right arrows (when the value of *c* is positive) or increasing the chains of left arrows (when the value of *c* is negative). The distance value can increase infinitely but there exists only one minima among valid representations. In order to reach to the minima, first the algorithms tests for increasing positively *c* if whether it gives higher or lower distance value. If the distance reduces, keep applying the operations until no more reduction occurs. Then, the algorithms does the same operations but increasing negatively *c*. With these two actions, the algorithm guarantees that all possible combinations of correct representations of arrows are tested.

The procedure runs in $O(z'^2)$ time. The lines 1 to 3 obtains the ordinal distance. In the lines 4 – 11 chains of right arrows are added to the current arrow representation until there is no more reduction to the total number of arrows. This increment is considered in the algorithm by the variable *c*. Next, chains of left arrows are added in the similar manner (lines 12 – 19).

6. Validation of the method and algorithms

The method and algorithms presented in this paper are applied on histograms, independently on the kind of the original set from which they have been obtained, i.e. images [20], discretized probability-density functions [14],… The only condition to use our method is to know the type of elements of the original set: ordinal, nominal or modulo.

In this paper, we validate our method on the comparison of images. We use the distance between histograms as a metric to compare images. It is important to note that we are not interested in the sophisticated techniques of image retrieval, i.e. [17,18,19]. We show that, using our algorithms, the classification of images through their histograms is really fast and keeps a high ratio of correctness. Some image retrieval techniques could be applied using our technique as the distance between images.

In the next two sections, we first experiment on images obtained from databases and second on indoor scenes. In both experiments, we show that there is an important reduction of the run time when the signature distance is used respect the histogram distance, although the ratio of recognition does not decrease.

6.1. Experiment with colour images

To show the validity of our new method, we have 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 calculate the histograms from the illumination coordinate with 2^8 levels (table 1) and with 2^{16} levels (table 3). And also, to test the modulo distance, the histograms represent the hue coordinate with 2^8 levels (table 2) and with 2^{16} levels (table 4). Each of the tables below shows the results of 5 different tests. In the first and second files of the tables, the distance where computed between histograms and signatures, respectively. In the other three, the distance was computed between signatures but, with the aim of reducing the length of the signature (and so to increase the speed), the bins that had less elements than 100, 200 or 300 in tables 1 and 2 and less elements than 1, 2 or 3 in tables 3 and 4 where removed. The first column is the number of bins of the histogram (first cell) or signatures (the other four cells). The second column represents the increase of speed if we use signatures respect histograms. It is calculated as the ratio between the run time of the histogram method and the signature method. The third column is the average correctness. The last column represents the decrease of correctness due to using the signatures with filtered histograms. It is obtained as the ratio of the correctness of the histogram by the correctness of each filter.

	Length	Increase Speed	Correct.	Decrease Correct.
Histo.	265	1	78%	1
Signa.	235	1.12	78%	1
Signa. 100	157	1.68	78%	1
Signa. 200	106	2.50	69%	0.88
Signa. 300	57	4.64	57%	0.73

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

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

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

	Length	Increase Speed	Correct.	Decrease Correct.
Histo.	265	1	86%	1
Signa.	215	1.23	86%	1
Signa. 100	131	2.02	85%	0.98
Signa. 200	95	2.78	73%	0.84
Signa. 300	45	5.88	65%	0.75

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

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

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

Tables 1 to 4 show us that our method is much useful when the number of levels increases since the number of empty bins tends to increase. Moreover, while comparing the histogram of the hues, the increase is much important do to 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 the correctness although there is much increase in the speed respect the signature method (second experiment).

6.2. Experiment on indoor scenes

Signatures have also been compared with histograms using indoor scenes. These scenes were used for robot positioning. In the learning stage, the robot is guided through the offices and corridors while the exact position is introduced and the robot captures the scenes. In the recognition stage, the robot assumes to be in the position that captures the most similar scene obtained in the learning stage. The main advantage of this technique (also reported in [21,22]) is that any mechanic method is not needed. Moreover, image retrieval techniques are neither used, since the new scene is not compared on all the scenes of the database, but only on the ones that are known to be near of the supposed position of the robot. Finally, if the scene is not recognised, the last position of the robot is assumed to be the present one and another image is captured. In [22], a same robot-positioning method was used. The main difference is that they used structural information of the image and they needed to segment the image. For this reason, the ratio of recognition was supposed to be higher but also the computational time. The main advantage of our method is that the image has not to be processed, only the histogram of it is needed.

Figure 9 shows 8 different scenes. From each scene, we have taken 3 images with a slight difference of the position. Furthermore, we show the histogram of the luminance and the hue of these images. We have used the histograms of the luminance to test the cardinal and ordinal distance ant the other histograms to test the modulo distance. Note that there is a similarity between the three histograms of the same scene and also that the hue histograms are much sparse than the luminance ones.

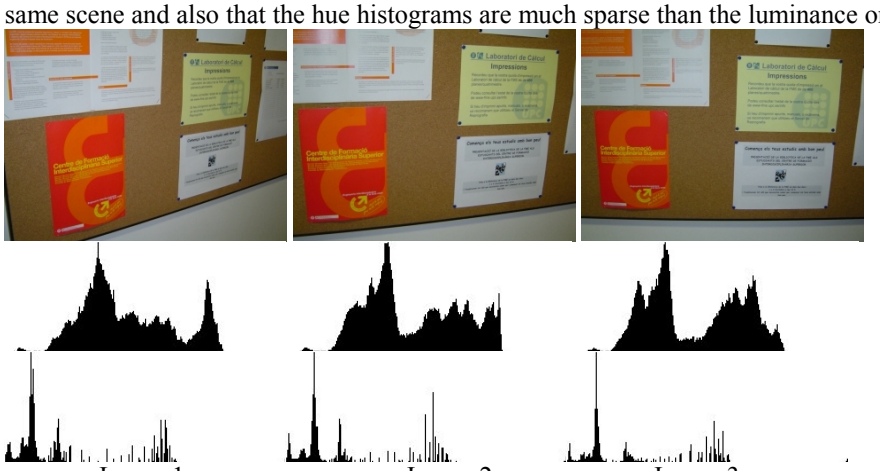


Image 1 Image 2 Image 3

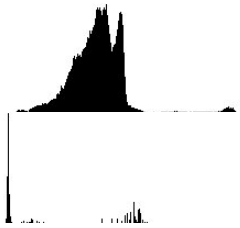


Image 4

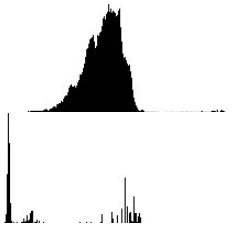


Image 5

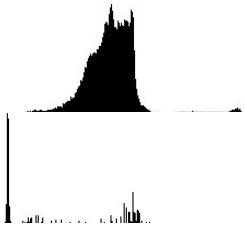


Image 6

Image 7

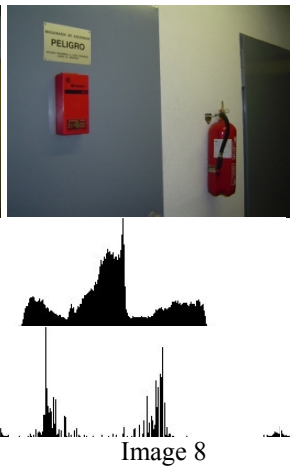


Image 8

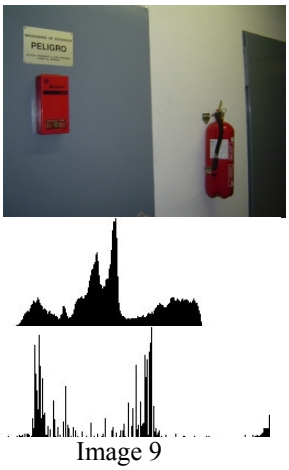


Image 9

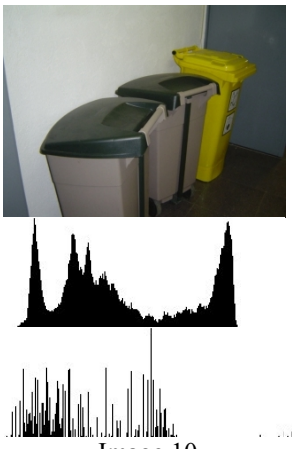


Image 10

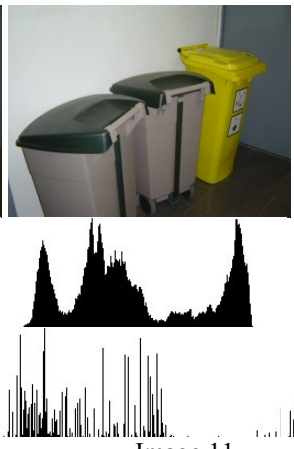


Image 11

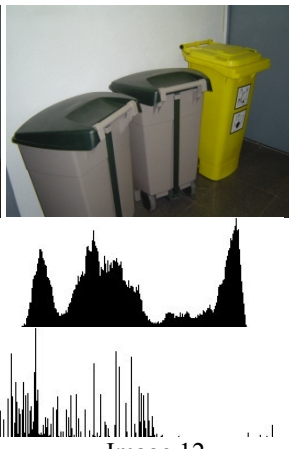


Image 12

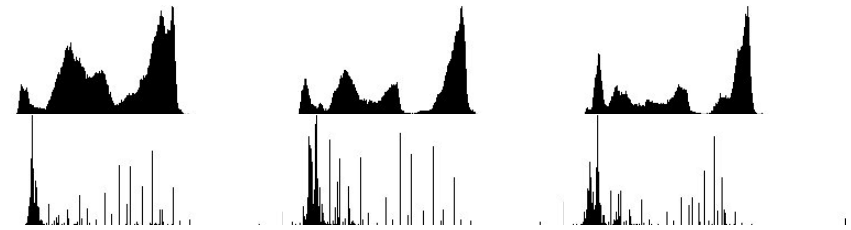


Image 13 Image 14 Image 15

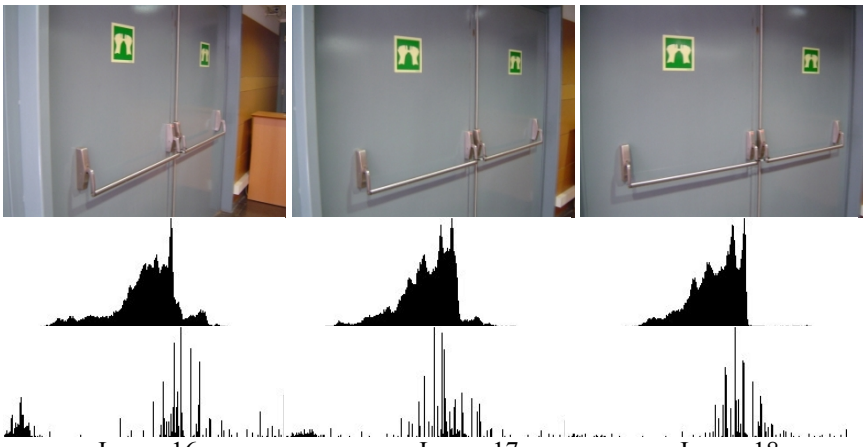


Image 16 Image 17 Image 18

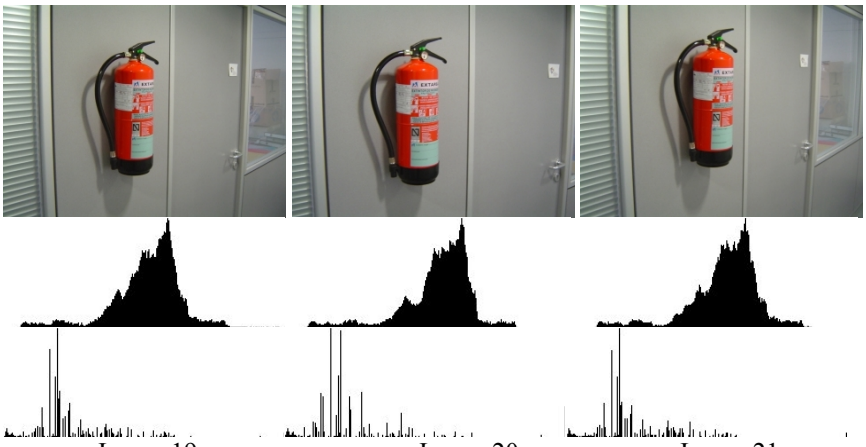


Image 19 Image 20 Image 21

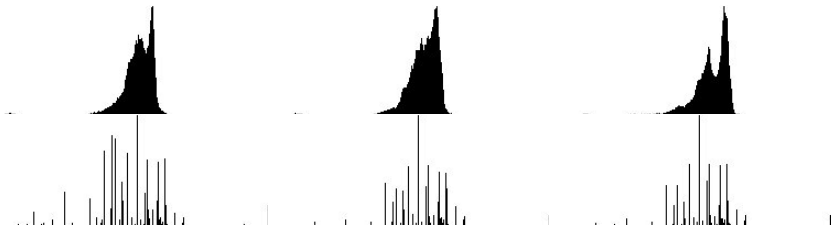


Image 22 Image 23 Image 24

Figure 9. Images from indoor scenes. Below each image, its luminance and hue histogram.

Table 5 shows the mismatched images and the ratio of recognition using the cardinal and ordinal distances for the luminance histogram and the modulo distance for the hue histogram. We consider a mismatched image if the three images that obtained smaller distance are not from its scene. As it is expected, the hue histograms obtain better results. Nevertheless, there is not a big difference these images are not much saturated and so there is little information on the hue. This is the reason because the hue histograms are sparse.

Histogram	Distance	Mismatched images	Recognition
Luminance	Cardinal	1,3,8,12,16,17,18	70.8%
Luminance	Ordinal	3,16,19	87.5%
Hue	Modulo	15,17	91.6%

Table 5. Mismatched images and ratio of recognition using luminance and hue histograms and the three distances: Cardinal, Ordinal and Modulo.

Table 6 shows the run time and ratio of recognition obtained from three experiments. The first one (a), the results where computed using the cardinal distance on the luminance histograms. The second one (b), the cardinal distance was changed by the ordinal distance. And the third experiment (c), the modulo distance was computed on the hue histograms. From each experiment, we obtained the run time and the ratio of recognition in four cases. 1: Comparing histograms. 2: Comparing signatures. 3: Comparing filtered signatures. The threshold of the filter was situated as much higher as possible, when that the ratio of recognition began to decrease. 4: The same as the third case but with a higher threshold of the filter. In all the cases, the run time was normalised such that the run time of the histogram in the first case was 100.
It is interesting to realize the decrease on the run time in the case of the modulo distance when filter a is applied. There is a decrease from 526 to 78.

	Card. Dist.				**Ord. Dist.**				**Mod. Dist.**			
	Histo.	Sign.			Histo.		Sign.		Histo.		Sign.	
		No filter	Filter a	Filter b		No filter	Filter a	Filter b		No filter	Filter a	Filter b
Run time	100	85	51	24	105	87	45	21	526	233	78	69
% Recog.	70.8	70.8	70.5	55.3	87.5	87.5	87.3	75.2	91.6	91.6	91.4	87.2
	(a) Luminance				(b) Luminance				(c) Hue			

Table 6. Run time and ratio of recognition obtained from three experiments on the luminance histograms (a) and (b) and on the hue histograms (c).

7. Conclusions and future work

We have presented the nominal, ordinal and modulo distance between signatures and the algorithms used to compute them. We have shown that signatures are a lossless representation of histograms and that computing the distance between signatures is the same than between histograms but with a lower computational time. We have validated these new algorithms with a huge amount of real images and we have realised that there is an important time saving do to most of the histograms are sparse. Moreover, if we apply filtering techniques on the histograms, the number of bins of the signatures reduces and so the run time of their comparison.

Albeit the signatures and histograms that we dealt with in this paper are one-dimensional, it can be useful in many applications the comparison between multi-dimensional histograms. The only difference in our equations of the distances would be the definition of the ground distance (nominal, ordinal or modulo). Nevertheless, defining the algorithms to compute the distance between multi-dimensional signatures is non-trivial because the increase of possible assignments. We leave the design of fast algorithms to compute these distances as open problems.

References

1. R.O. Duda, P.E. Hart & D.G. Stork, *Pattern Classification*, 2nd edition, Wiley, New York, 2000.
2. T. Kailath, "The divergence and bhattacharyya distance measures in signal selection", *IEEE Transactions Community Technol*. COM-15, 1, pp:52-60, 1967.
3. K. Matusita, "Decision rules, based on the distance, for problems of fit, two samples and estimation", *Annals Mathematic Statistics*, 26, pp: 631-640, 1955.
4. J.E. Shore & R.M. Gray, "Minimum cross-entropy pattern classification and cluster analysis", Transactions on Pattern Analysis and Machine Intelligence, 4 (1), pp: 11-17, 1982.
5. S.-H. Cha, S. N. Srihari, "On measuring the distance between histograms" *Pattern Recognition* 35, pp: 1355–1370, 2002.
6. 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.
7. E. J. Russell. "Extension of Dantzig's algorithm to finding an initial near-optimal basis for the transportation problem", *Operations Research*, 17, pp: 187-191, 1969.
8. *Numerical Recipes in C: The Art of Scientific Computing*, ISBN 0-521-43108-5.
9. Y-P Nieh & K.Y.J. Zhang, "A two-dimensional histogram-matchinig method for protein phase refinement and extension", Biological Crystallography, 55, pp:1893-1900, 1999.
10. 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.
11. F.-D. Jou, K.-Ch. Fan, Y.-L. Chang, "Efficient matching of large-size histograms", *Pattern Recognition Letters* 25, pp: 277–286, 2004.
12. 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.

13. J. Morovic, J. Shaw & P-L. Sun, "A fast, non-iterative and exact histogram matching algorithm", *Pattern Recognition Letters* 23, pp:127–135, 2002.
14. F. Serratosa, R. Alquézar y A. Sanfeliu, "Function-Described Graphs for modeling objects represented by attributed graphs", *Pattern Recognition*, 36 (3), pp. 781-798, 2003.
15. A. Sanfeliu, F. Serratosa & R. Alquézar, "Second-Order Random Graphs for modeling sets of Attributed Graphs and their application to object learning and recognition", *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 18, (3), pp: 375-396, 2004.
16. F. Serratosa, R. Alquézar and A. Sanfeliu (a), "Efficient algorithms for matching attributed graphs and function-described graphs", *Proceedings 15th International Conference on Pattern Recognition*, *ICPR'2000, Barcelona, Spain*, 2000.
17. Wei Jiang, Guihua Er, Qionghai Dai and Jinwei Gu, "Hidden annotation for image retrieval with long-term relevance feedback learning", *Pattern Recognition*, Vol. 38, (11), 2005, pp. 2007-2021.
18. Ke Lu and Xiaofei He, "Image retrieval based on incremental subspace learning", *Pattern Recognition*, Vol. 38, (11), 2005, pp. 2047-2054.
19. Y. Peng-Yeng, B: Bhanu, Ch. Kuang-Cheng, A. Dong; "Integrating Relevance Feedback Techniques for Image Retrieval Using Reinforcement Learning", *Pattern Analysis and Machine Intelligence*, Vol. 27 (10), 2005, pp.1536 – 1551.
20. M. Pi, M.K. Mandal, A. Basu, "Image retrieval based on histogram of fractal parameters", *Multimedia, IEEE Transactions on*, Vol. 7 (4), 2005, pp. 597 – 605.
21. G. Wells, Ch. Venaille & C. Torras, "Vision-based robot positioning using neural networks", *Image and Vision Computing*, Vol 14 (10), 1996, pp. 715-732.
22. E.Staffeti, A.Grau, F.Serratosa & A.Sanfeliu, "Object and Image indexing based on Region Connection Calculus and Oriented Matroids theory", *Discrete Applied Mathematics*, Vol. 147, (2-3), 2005 pp: 345-361.