Silhouettes Based Evaluation of the Effectiveness in Image Retrieval

A. AMATO\(^1\)-T. DELVECCHIO\(^1\) - V.DI LECCE\(^2\)*

\(^1\) Engineering Faculty of Politecnico di Bari – Viale del Turismo n.8 – 74100 Taranto - ITALY
\(^2\) DEE- Politecnico di Bari – Via Re David 200 – 70121 Bari - ITALY

Abstract: New tools for combining queries in a useful way are needed. The rise of multimedia data repositories is changing the way we look at the process of searching a particular piece of information, so we will be forced to change many of our usual concepts about database organization. Goal of this paper is to evaluate the retrieval performances of same well known methods, using the silhouettes based Jain’s approach. The low level features selected, typically used in automatic indexing, are angular spectrum, Gabor, Hough, and Tamura. The tests includes both natural and synthetic images.

Key-words: Retrieval, similarity, Tversky, silhouette.

1 Introduction

The importance of similarity in everyday life is that we share the world in categories, classes, groups, situations of identification, comparing them with the old and already experienced ones.

This concept kept the attention of psychologists in the last twenty years, till to the recent sophisticated mathematical developments: mathematical definition of image features and distance model suitable in human similarity modelling.

The last trend is to realise similarity retrieval in complex multimedia data [1], matching won’t be expressive enough to answer the needs of database systems, in which the main operation is similarity assessment. Although the matching process can be sophisticated, there will be always a subset of database “matching” and another one “matching not” the queries: there can be considerable difficulties to determine if a given item matches or not but, at least ideally, the match and no-match sets form a partition of the whole database. Matching is all that result about a process of reasoning on the input data, and on inferences based on structured data encoded in high level symbols. In this approach, the existence of a target image in the database is not postulated, but we order the images according to similarity with the query, given a fixed similarity criterion [2]. The consequences is the fact that the answer to any query is the whole database.

In images retrieval, matching method is used to determine if two different images are both sides (aspects) of the same thing. Matching is the main and fundamental operation in a database. It means to compare an item with the query, and to decide if this one satisfies the query or not.

The main differences between match based and similarity searches could be described in this way: the result of a match based search is a partition of the database in the set of the images matching the query and the set of images that do not, the result of a similarity search is a permutation of the whole database (in particular, a sorting with the respect to the similarity criterion).

Searching for a reasonable similarity measure, the most obvious comparison to do and to look for is human similarity assessment. In fact, when a user makes a sketch, or selects a prototype image, looking for something similar, he has in is mind his own ideas of similarity. Consequence is that the similarity used by database has to be as similar as possible to human similarity, if we want a satisfied result of searches.

In this paper after a brief presentation of Similarity Theory well discussed in [3], the b/w silhouettes parametrization and the under test retrieval methods are presented. Finally the results of tests are presented in terms of automatic and human retrieval.

2 Similarity Theory: stimuli, signatures and distance models

Measurement is the process of assigning numbers to objects according to a set of rules. This process serves to describe and organise phenomena and to test results of the phenomena model under investigation.

In the automatic similarity evaluation it needs to correlate two stimuli (images) in the perceptual space (human) with the distance from two points representing the signatures (extracted from images) in the signature space, where signatures and distance represent the mathematical model of human similarity under test.
A large number of researchers uses Euclidean distance model assuming that the psychological space is a metric space. In [4] it is highlight that is more fair to use the Euclidean distance among the numerical representation of selected image features (often called, as in this paper, signature).

The fig. 1 shows the correspondence between images $\rightarrow$ stimuli $\rightarrow$ signatures $\rightarrow$ distance $\rightarrow$ (computable similarity).

In the papers [5,6] Di Lecce and Guerriero have investigated the relationship between images and computable similarity\(^1\) using human panel and the the shot\(^2\) definition: in the proposed scheme the chain stimuli $\rightarrow$ signatures $\rightarrow$ distance has been considered as a black box.

An in-depth study is in [3] where a concise presentation is in the following. Stimuli are points in perceptual space, a similarity between the stimulus Ap and Bp are one of the cardinal points of cognitive theories of similarity. Various hypothesis are been investigated. In the purely psychological approach based on multidimensional representation an image is represented as a point in a multidimensional space [7]. The position of the points is typically derived from Scaling Techniques such as Multidimensional Scaling (MDS) [8]. The MSD is a non-metrical space, and the goal is to find a projection space in which the interpoint distance are monotonically related to a human panel response about (di-)similarity.

In the purely computational approach a natural scene is represented by a collection of values (signature) explicitly derived from the 2-D image containing basic image low-level features, comparable to retinal-brain sensitivity, like shape, colours, patterns.

The first step is to evaluate that distance concept from Aa and Ba (in the artificial similarity space, \(d(Aa,Ba)\) of fig. 1) is related to distance from Ap and Bp (in the stimuli or perceptual space, \(d(Ap,Bp)\)), while \(d(Ap,Bp)=f[d(Aa,ba)]\). The second step is in triangle inequality verify, introducing a third point C:

\[
\text{If } d(Ap,Bp)+d(Bp,Cp) \geq d(Ap,Cp) \text{ then } d(Aa,Ba)+d(Ba,Ca) \geq d(Aa,Ca)
\]

The function \(d\) is frequently Euclidean metric or city-block metric according to the Fred Attneave work[9].

---

\(^{1}\) It is the distance between the signatures that represents the images.

\(^{2}\) A shot is a video sequence of contiguous frames generated during a continuous operation and therefore represent a continuous action in time or space. Frames extracted from a shot are different but have the same semantic content.
Relevant contributions are offered by Thurstone and Shepard, Tversky, Ashby and Perrin and many others in the last years, mainly oriented to distance axioms verify. Really interesting researches are focused on fuzzy and neural models, often used in faces recognition and identification [10], but it seems the high complexity of the neural or fuzzy networks restricts the practicability at specific low-complexity cases.

For these reasons it is selected a class of images containing a b/w silhouettes in which are recognisable the Tversky [11] similarity proposition, also known as “Contrast Model”: first, if dissimilarity is to be represented as a metric distance, it must follow the three metric axioms of minimality, symmetry, and the triangle inequality, but data contrary to these axioms have arisen in various experimental situations with humans; second, few stimuli differ from each other in only a few continuous dimensions such as size and color. Tversky’s “Contrast Model” systematise this feature approach. A central assumption of the model is that the similarity of object A to object B is a function of the features common to A and B ("A and B"), those in A but not in B (symbolized "A-B") and those in B but not in A ("B-A"). Based on this and several other assumptions, Tversky derived the following relationship:

\[ S(A,B) = xf(A \text{ and } B) - yf(A-B) - zf(B-F) \]

where S is an interval scale of similarity, f is a function of salience of the various features considered, and x, y and z are parameters that provide for differences in focus on the different components. Note especially that similarity is not just a function of common features, but depends also on features that are unique to each object. This formulation makes principled sense of several characteristics of similarity data that contradict the metric assumptions discussed above. The most troubling is probably asymmetry. This often goes along with task asymmetry; for example, "how similar is A to B" may give a different answer than "how similar is B to A".

3 Silhouettes generation

According to the previous section it has been generate a class of images containing a b/w silhouettes. Fig. 2 shown a silhouette used in the test. The silhouette is parameterised by ratio between height/width of the central rectangle (body) and number of fingers for each side.

Name of image is coded as "Ahn_ar_n_dnl1.gif", where:

- A prefix of images
- h ratio between width/height*10
- nu finger number on upper side
- nr finger number on right side
- nd finger number on down side
- nl finger number on left side

Figure 2: the silhouettes a68020 used as stimulus in retrieval tests.

4 Algorithm and distance model

For our comparison we adopted 7 signature, based on the following algorithms.

4.1 Angular spectrum

The image visual properties are mainly related to the largest image components, and among their features the shape, the texture and the orientation play a major role. In many cases shapes can also be defined in terms of presence and distribution of oriented sub-components. A thick object has most of its lines arranged along its primary direction, while a thin elongated object has a peak in the line orientation distribution. Therefore the orientation of objects within an image is a key attribute in the definition of the similarity with other images.

Following this assumption, we have defined a metric for image classification based on orientation in the two-dimensional space, that is quantified by signatures composed of angular spectra of image components [12].

We approach the problem of finding the distribution of image lines direction by analysing its
Fourier transform. Some pre-processing is applied before computing the Fourier transform: a 2D Hamming function transforms the signal describing the image into a periodic signal, that can be represented by a finite number of frequency components. The visual effect of applying a Hamming function is comparable to a low pass filter whose effect is negligible in the center of the image, and increases towards the borders. The borders reduce their contribution to the image frequency spectrum, focusing the analysis on the central part of the image.

The low frequencies correspond to large components within the image, which are usually relevant for interpreting the image and can be assumed to be most frequently localized in central areas (e.g., foreground objects) or spanning all the image (e.g., recurrent shapes, landscapes). Their contribution to the image is almost completely preserved.

High frequencies correspond to small details and fine-grain textures, a reduction of these components is equivalent to focusing on the foreground component located in the image central area, and ignoring the details of the peripheral contour.

For each image we generate a signature formed by 7*10 values. The polar representation of the image spectrum is split in 10 sectors, in each sector 7 harmonics are evaluated and recorded in a histogram. In order to evaluate the distance between two angular spectrum signatures, we use a weighted Euclidean distance:

$$d_{w}(n,m) = w_{n,m} \cdot d(n,m)$$

where the weight $w_{n,m}$ depends on the distance between the largest term of the reference image signature and the largest term in the first band (terms 1:10) of the examined signature.

### 4.2 Hough transform

The Hough transform [13] involves the transformation of a line in Cartesian coordinate space to a point in polar coordinate space.

A straight line can be parametrically described as:

$$\rho = x \cos(\theta) + y \sin(\theta)$$

where $\rho$ is the normal distance of the line from the origin and $\theta$ is the angle with the x axis.

The Hough transform of a line is simply a point of coordinate $(\rho, \theta)$ in the polar domain.

Duda and Hart [14] have proposed the Hough transform technique for line and curve detection in binary images.

In our test we used this techniques to find the main direction of the edges of the images. The Canny [15] algorithm is used to detect the edge of the image, for each edge point the Hough transform is computed and the results are recorded in a histogram. After all data points are transformed, the histogram can be examined to find the main line’s direction. This histogram, organized in 18 bins, is used as the image signature.

For the distance evaluation in accordance with [16] we used the Euclidean distance.

### 4.3 Gabor

Gabor functions, because of their good approximation to the receptive fields of cortical cells, are often used in models of image representation in the visual cortex [17-18-19].

The Gabor function, or Gaussian wavelet [19], are complex exponential function modulated by Gaussian function. These functions performing a local Fourier analysis can be used for sampling the space-frequency domain.

If the gaussian envelope have circular symmetry, and the complex exponential has zero phase, the filter expression, tuned to the frequency $f_0$ with orientation $\theta_0$, and centred at the origin $(x_0=0, y_0=0)$ is:

$$g_{0,0,f_0,\theta_0}(x, y) = \exp(-\frac{a}{2} (x^2 + y^2)) \cdot \exp(i2\pi f_0(x \cdot \cos \theta_0 + y \cdot \sin \theta_0))$$

where $a$ determine the spatial frequency bandwidth. Changing the central frequency, the bandwidth and the orientation, a set of filters is obtained. Each filter provides information contained in a particular direction in the image.

In our evaluation we used a set of 12 filters, obtained changing $\theta_0$ with a step of 45° from 0° to 135° and sampling the frequency domain in 4 octaves. The energies of the filters output form the image signature.

Being these energies a periodic function of $\theta$ with period $\pi$, a rotation invariant signature can be generated by their Fourier coefficient [20] since a rotation of the image corresponds to a translation of the periodic function and Fourier transform is invariant to translation.

The filtering operation was performed in the space domain by the efficient implementation
presented in [21] and the distance between two signatures is evaluated by the Euclidean distance.

4.4 Pattern directionality

This method, based on the histogram of local edge probabilities against directional angle [21], utilizes the fact that gradient is a vector, so it has both magnitude and direction.

In the discrete case, the magnitude $|\Delta G|$ and the local edge direction $\theta$ are approximately as follows:

4) $|\Delta G| = \frac{(|\Delta_H| + |\Delta_V|)}{2}$

5) $\theta = \tan^{-1}\left(\frac{\Delta_V}{\Delta_H}\right) + \pi / 2$

where $\Delta_H$ and $\Delta_V$ represent the 3*3 Sobel operators.

$\theta$ is a real number $(0 \leq \theta \leq \pi)$ measured counterclockwise so the horizontal direction is zero.

The method is applied to the 25 windows extracted from the image.

The histogram, obtained quantizing $\theta$ and counting the points with magnitude $|\Delta G|$ greater than a non-critical threshold $t$, forms the signature.

4.5 Local pattern directionality

Texture is an important feature of a visible surface where repetition of fundamental pattern occurs.

Typical texture features are contrast, color, uniformity, coarseness, roughness, frequency, density, and directionality. The spatial relationship of different textures can be considered semantically related to the image, and then used in image classification and retrieval.

Tamura [21] proved that six basic textural features (coarseness, contrast, directionality, line-likeness, regularity, and roughness) are related to psychological measurements for human subjects.

It is known that a modified set of the Tamura features has been used in the QBIC project [22].

We select two Tamura textural features and two derived ones for the presented evaluation.

For all the local features the similarity is evaluated by the Euclidean distance.

The technique described in 4.4. is applied to 25 windows extracted from the image. The position of the maximum value of each histogram, obtained quantizing $\theta$ and counting the points with magnitude $|\Delta G|$ greater than a non-critical threshold $t$, forms a signature component.

4.6 Roughness

For each window the roughness term is obtained by adding the contrast term $F_{con}$ and the coarseness term $F_{crs}$. Tamura define:

6) $F_{con} = \sigma * \left(\frac{\sigma^2}{\mu_4}\right)^{\frac{1}{4}}$

where $\sigma$ represents the standard deviation of the window pixel, $\mu_4$ is the fourth moment about the mean, and $\sigma^2$ is the variance.

While the essence of the $F_{crs}$ determination is to pick a large size as best when coarse texture is present even though microtexture is also present but to pick a small size when only fine texture is present (a complete discussion and the mathematical expression of $F_{crs}$ can be found in the Tamura paper at the reference [22]).

4.7 Local horizontal/vertical prevalence

This technique can be derived from a syntactic patterns approach [24] using the Tamura directionality concept applied only to horizontal and vertical directions.

This signature is formed by 50 terms, the first 25 are used to report the intensity of prevalent horizontal components in the windows, the last 25 terms are used for the intensity of vertical components, if prevalent. If no prevalent direction is found both terms are 0.

The distance between these signatures is the sum of two terms, the horizontal and the vertical components, each computed by the Euclidean distance.

5 Retrieval performance comparison

Retrieval performances are measured in terms of efficiency and effectiveness. Typically efficiency is concerned with how much work is performed to answer a query.

Due to the aim of this research, specifically to identify the sensibility of test methods using silhouettes, the results are presented as retrieved images (fig. 3, each method are used to retrieve 10 images organised in a row) and in table form (tab. 1, each method are used to retrieve 10 images organised in a column). Tab.2 resume the retrieval performance evaluated by a human panel (50 students of our Faculty).
Each student expressed his opinion (a grade from 0 to 10) on each retrieved sequence. The table can be considered as human or ground normalized truth.

In the case of natural images [5] to evaluate the effectiveness of retrieval method (feature and distance model) we have applied a measure proposed by Mether [25].

The effectiveness measure is defined as follow:

\[
\text{Effectiveness} = \begin{cases} 
\frac{n}{N} & \text{if } T > N \\
\frac{n}{T} & \text{if } T \leq N 
\end{cases}
\]

where:
- \(T\) is the number of images required by the user in the query,
- \(N\) is the total number of relevant images in the database (known \textit{a priori}),
- \(n\) is the number of relevant images retrieved.

In the table 3 the results are summarized.

6 Conclusion

Searching for a reasonable similarity measure, the most obvious comparison to do and to look for is human similarity assessment.

In this paper after a brief presentation of Similarity Theory proposed by Jain and Santini in [3] a test is implemented to compare an algorithm silhouettes based approach with a human evaluation of similarity. In the test different image basic features are used.

The distance used is the Euclidean model. It is well known that many others distance model can be used.

Our idea is to perform additional test to identify a weighing function suitable in tuning results of algorithmic sorting with human expectation.


