Binary Data Comparison using Similarity Indices and Principal Components Analysis

Nouhoun Kane, Khalid Aznag, Ahmed El Oirrak and Mohammed Kaddioui
Computer Science Department, Faculty of Sciences Semlalia, Cadi Ayyad University, Morocco

Abstract: This work is a study of binary data, especially binary images sources of information widely used. In general, comparing two binary images that represent the same content is not always easy because an image can undergo transformations like Translation, rotation, affinity, resolution change and scale or change in appearance. In this paper, we will try to solve the translation and rotation problems. For translation case, the similarity indices are used between the image rows or blocks. In the case of rotation, first the coordinate’s contours are extracted, then we compute the covariance matrix used in the Principal Components Analysis (PCA) and the corresponding eigen values which are invariant to this type of movement. We also, compare our approach having complexity $O(M+N)$ to Hausdorff Distance (HD) that has complexity of $O(M \times N)$ for an $M \times N$ image. In our approach, HD is used only to compare distance between 1D signatures.

Keywords: Binary images, covariance matrix, similarity index, HD.

Received August 31, 2013; accepted April 20, 2014

1. Introduction

To compare two images, we must first represent these images through effective descriptors. These descriptors represent general information on color, texture and shapes of the extracted image [24]. Their choice determines the effectiveness of the method and is a delicate step of indexing [23]. The color histogram is widely used as an indexing descriptor space [9].

Other characterizations of contours are possible such as fourier coefficients, eccentricity, euler number and moment invariants [10, 11, 27]. But, some methods like SIFT and SURF [4, 18] are not suitable for binary features.

On the other hand, characterizations based on the autocorrelation function are used for textured images [19]. However, textures are useful only for textured images which are a special case.

Discriminating descriptors once extracted are arranged to form the signature of the image. Signatures are then used to compare images [20]. This comparison must prove the degree of similarity between images. There are two ways to structure information to form a signature: Global signature and local or partial signature. The histogram (Figure 2) [3] is an example of global signature and techniques used in this paper are also global.

This paper is organized as follows: Section 2 describes the problem of translation between binary images and how we can solve it by calculating a similarity measure between rows of the image. Section 3 presents the rotation problem in a picture and it turned out that the covariance matrix of contour coordinates is an invariant measure according to rotation. In section 4, we compare our proposed approach to Hausdorff based approach.

2. Translation and Similarity Index

Binary data is one of the most common representations of patterns and similarity measures between these types of data is essential in many problems such as clustering, classification, etc. Since, jaccard proposed a similarity measure to classify ecological species in 1901, many similarity measures and distances have been proposed in various areas. Implementing appropriate measures has for result more accurate data analysis.

For example, Jaccard similarity measure has been used for classification of ecological species [13]. Binary similarity measures were then applied in biology, anthropology, taxonomy, image retrieval, text retrieval, geology and chemistry [21, 25]. Recently they were actively used to solve identification of fingerprints, iris pictures problems and recognition of manuscripts characters [6, 7, 8]. Many articles discuss their properties and characteristics [12, 15].

The Simple Matching Coefficient (SMC) is a simple similarity index and the base of the most of other indices.

2.1. Simple Matching Coefficient

Given two objects $A$ and $B$, each with $n$ binary attributes, the SMC coefficient is a useful measure of the overlap that $A$ and $B$ share with their attributes. Each attribute of $A$ and $B$ can either be 0 or 1. The total number of each combination of attributes for both $A$ and $B$ are specified as follow:

- $f_{11}$: Represents the total number of attributes where $A$ and $B$ both have a value of 1.
- $f_{00}$: Represents the total number of attributes where the attribute of $A$ is 0 and the attribute of $B$ is 1.

...
• \( f_{10} \): Represents the total number of attributes where the attribute of \( A \) is 1 and the attribute of \( B \) is 0.
• \( f_{00} \): Represents the total number of attributes where \( A \) and \( B \) both have a value of 0.

Each attribute must fall into one of these categories, meaning that \( f_{11} + f_{00} + f_{10} + f_{01} = n \).

This index is defined by:

\[
SMC = \frac{f_{11} + f_{00}}{f_{11} + f_{01} + f_{00} + f_{10}}
\]

(1)

Where, \(| \text{matching attributes}| = f_{11} + f_{00} \) represents pixels that are matched (values are 0 or 1) and \(| \text{attributes}| = f_{10} + f_{11} + f_{00} \) represents the number of pixels on the two rows or vectors to be matched. Thus, we can rewrite the SMC index as follows:

\[
SMC^* = \frac{f_{11} + f_{00}}{f_{01} + f_{10} + f_{11} + f_{00}}
\]

(2)

2.2. Similarity between image Rows or Columns using Jaccard Index

Principle used for the jaccard index is similar to SMC except that no account is considered for pixels with values equal to 0.

\[
\text{jaccard} = \frac{| \text{matching present attributes}|}{| \text{attributes values present}|}
\]

(3)

\[
| \text{matching present attributes}| = f_{11}
\]

\mid | \text{attributes values present}| = f_{01} + f_{10} + f_{11}

In this work we chose to work with Jaccard index (Figures 1 and 2) using Equation 4 because images backgrounds are black. Technique used is as follows:

• We fix a row on image (row containing the center of gravity of the object i.e., white pixels).
• An index of similarity between each row of image and the fixed row is calculated.
• Signature image is formed by the values of the index.

Note:

• For images with white backgrounds, jaccard coefficient must be changed because the black color is now the objects within the image, so we have:

\[
\text{Jaccard} = \frac{f_{00}}{f_{01} + f_{10} + f_{00}}
\]

(4)

• We can also work with image columns instead of rows.

2.3. Similarity between Image Blocks

In this section, we propose to compute similarity indices between image blocks. Blocks generally represent the neighbourhood of a pixel, this neighbourhood is a rich source of local information, which is not the case for rows that are source of global information.

The steps of proposed technique are as following: Cutting the image into blocks, for two successive blocks we calculate the similarity index. This work is repeated for all blocks of the image in opposite direction. Two successive blocks are defined as follow:

\[
\begin{bmatrix}
I(i - 1, j - 1) & I(i - 1, j) & I(i - 1, j + 1) \\
I(i, j - 1) & I(i, j) & I(i, j + 1) \\
I(i + 1, j - 1) & I(i + 1, j) & I(i + 1, j + 1)
\end{bmatrix}
\]

and

\[
\begin{bmatrix}
I(i - 1, j + 2) & I(i - 1, j + 3) & I(i - 1, j + 4) \\
I(i, j + 2) & I(i, j + 3) & I(i, j + 4) \\
I(i + 1, j + 2) & I(i + 1, j + 3) & I(i + 1, j + 4)
\end{bmatrix}
\]

With \( i = 2: M - 4 \) and \( j = 2: M - 4 \). These two blocks are \( 8 \times 8 \) neighbourhoods for pixel \((i, j)\) and pixel \((i, j + 3)\). Thus, jaccard index is now defined between two matrices of size 3×3.

In Figure 3, we show 3 images on which we apply similarity blocks to extract signatures shown in Figure 4. The vector of all similarity indices is a signature for the image (Figure 4).

Figure 3. The three images used to test similarity between image blocks: Images in Figures 3-a and 3-b are close to each other while the third one in Figure 3-c is utterly different.

Figure 4. Signatures for images in Figure 3: Both first signatures (a and b) are similar while the third one (c) is different.
In Table 1, a Hausdorff Distance (HD) is computed between signatures of the three images. Let \( J_1, J_2 \), and \( J_3 \) denote vectors signatures for images in Figures 3-a, 3-b and 3-c respectively. Then:

<table>
<thead>
<tr>
<th>HD(( J_1, J_2 ))</th>
<th>HD(( J_1, J_3 ))</th>
<th>HD(( J_2, J_3 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>22.52</td>
<td>23.50</td>
</tr>
</tbody>
</table>

We use the discrete HD defined by

\[
HD(A, B) = \max_{a \in A, b \in B} \| a - b \|.
\]

Where, \( h(a, b) = \max \min |a|, |b| \).

We can also use Dynamic Time Warping (DTW) [16] or Chamfer Distance [2].

### 3. Rotation and PCA

PCA has its source in an article published in 1901 by Karl Pearson [14, 26, 28]. Here, we give a brief description of PCA principle (Suppose \( x_1, x_2, ..., x_M \) are \( N \times 1 \) vectors):

- **Step 1**: \( \bar{x} = \frac{1}{M} \sum_{i=1}^{M} x_i \)
- **Step 2**: Subtract the mean:
  \( \Phi_i = x_i - \bar{x} \)
- **Step 3**: Form the matrix:
  \( A = [\Phi_1 \Phi_2 \cdots \Phi_M] \)
  \((N \times M)\) matrix, then compute:
  \[
  C = \frac{1}{M} \sum_{n=1}^{M} \Phi_n \Phi_n^T = AA^T
  \]
  (sample covariance matrix, \( N \times N \), characterizes the scatter of the data)
- **Step 4**: Compute the eigen values of
  \( C: \lambda_1 > \lambda_2 > ... > \lambda_N \)
- **Step 5**: Compute the eigen vectors of:
  \( C: u_1, u_2, ..., u_N \)

Since \( C \) is symmetric, \( u_1, u_2, ..., u_N \) from a basic, (i.e., any vector \( x \) or actually \( x - \bar{x} \), can be written as a linear combination of the eigen vectors):

\[
x - \bar{x} = b_1 u_1 + b_2 u_2 + \cdots + b_N u_N = \sum_{i=1}^{N} b_i u_i
\]

Thus, the Principal Component Analysis (PCA) uses a matrix indicating the degree of similarity between variables and then compute projection matrix for variables in the new space. This symmetric matrix, which includes the variance of variables on the diagonal and elsewhere is called covariance matrix.

Covariance measures the degree of independence of two variables.

Under the action of rotation contours coordinates of the second image are related to coordinates contours of the first image by:

\[
\begin{pmatrix}
\hat{x} \\
\hat{y}
\end{pmatrix} = \begin{pmatrix}
\cos \theta & \sin \theta \\
-\sin \theta & \cos \theta
\end{pmatrix}
\begin{pmatrix}
x \\
y
\end{pmatrix}
\]

(5)

For all points

\[
\begin{pmatrix}
x \\
y
\end{pmatrix} = \begin{pmatrix}
x_i \\
y_i
\end{pmatrix}
\]

Where:

\[
\begin{pmatrix}
\hat{x} \\
\hat{y}
\end{pmatrix}: \text{Denotes contours vectors of second.}
\]

The analytical writing is given by:

\[
\begin{align*}
\hat{x} &= \cos \theta x_i + \sin \theta y_i \\
\hat{y} &= -\sin \theta x_i + \cos \theta y_i
\end{align*}
\]

Let \( M \) denotes:

\[
M = \begin{pmatrix}
x_i \\
y_i
\end{pmatrix}
\]

And

\[
\hat{M} = \begin{pmatrix}
\hat{x}_i \\
\hat{y}_i
\end{pmatrix}
\]

The covariance matrix for transformed image is:

\[
V = MM^T = \begin{pmatrix}
\hat{x}_i & \hat{y}_i
\end{pmatrix}
\begin{pmatrix}
\hat{x}_i \\
\hat{y}_i
\end{pmatrix}
\]

(6)

For \( i=1...n \) and \( j=1...n \) so, we have:

\[
\hat{x}_i \hat{x}_j + \hat{y}_i \hat{y}_j = \left( \cos \theta x_i + \sin \theta y_i \right) \left( \cos \theta x_j + \sin \theta y_j \right) + \\
\left( -\sin \theta x_i + \cos \theta y_i \right) \left( -\sin \theta x_j + \cos \theta y_j \right) + \\
\cos^2 \theta x_i x_j + \sin^2 \theta x_i x_j + \cos^2 \theta y_i y_j + \\
sin^2 \theta y_i y_j + \cos \theta \sin \theta y_i x_j - \cos \theta \sin \theta x_i y_j
\]

\[
= x_i x_j + y_i y_j
\]

Which proves invariance under rotation.

### 3.1. Experiments 1

In the following experiment (Figure 5), we apply this technique to eliminate rotation and extract two invariant values for each image. Those values are the two largest Eigen values of the covariance matrix (Table 2).
Figure 5. Two images used in experiments and their corresponding edges.

Table 2. The two largest Eigen values of the covariance matrix for contours.

<table>
<thead>
<tr>
<th>Two Eigen Values for $V$</th>
<th>Two Eigen Values for $V'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0287</td>
<td>0.0390</td>
</tr>
<tr>
<td>3.5376</td>
<td>3.4347</td>
</tr>
</tbody>
</table>

3.2. Experiments 2

For synthetic contours (Figures 6 and 7), we obtain a perfect result, the following 2D curve was created using the parametric equation:

\[
\begin{align*}
  x(t) &= 2 \cos(t) \\
  y(t) &= \sin(t) + \frac{1}{2} \sin(5t) \\
  t &\in [0, 2\pi]
\end{align*}
\]

(7)

Figure 6. Synthetic contour in Figure 6-a and the same contour in Figure 6-b after applying a synthetic rotation.

Figure 7. 2D display of the covariance matrix for contours.

Table 3. The two largest eigen values of covariance matrices in Figures 7-a and 7-b respectively.

<table>
<thead>
<tr>
<th>Synthetic Contour</th>
<th>Transformed Contour</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.25</td>
<td>6.25</td>
</tr>
<tr>
<td>20.40</td>
<td>20.40</td>
</tr>
</tbody>
</table>

To test our technique to noise resistance, noise with different percentage is added (Figure 8) and values in Table 4 does not change much compared to one illustrated in Table 3.

4. Comparison

In the work [1] an approach for binary image comparison without feature extraction was presented. It uses the windowed HD in a pixel adaptive way. They measure HD not between two full images, but between subimages extracted by a window. It is therefore necessary to define the extent of the HD in a window. This amounts to modify the definition of the overall measurement by introducing a restriction to sets of points at the window.

Let $A$ and $B$ be two bounded sets:

\[
HD_w(A, B) = \max(h_w(A, B), h_w(B, A))
\]

Where, $w$ denotes a window and:

\[
h_w(A, B) = \max_{a \in A, b \in B} \min \|a - b\|
\]

The size of the $w$ window must be fixed. This can be done by the user or automatically and globally or even automatically and locally according to the local environment.

This distance is not invariant to rotation and under translation (Figure 9). We have $HD_w(A, A+V) = |V|$. So, they did not take care of the situation where images might be rotated.

For two binary images, assuming that they have the same resolution and same orientation of object(s) in the images, the map of all local dissimilarity measures i.e., CDL (Figure 10) made on different places images constitute the proposed signature in this work.

Figure 9. Image A: A bold line on the left. Image B: A bold line to the right. Image C: A dots line on the right.
both image A and image B have the same signatures.

HD proved in this work that three non-nulls (different from zero) values.

(0) (pink), while signature for image C has (zero in this case), while signature for image C has three non-nulls (different from zero) values.

CDL (Figure 10) cannot tell us if images (Figure 9) are the same or not as it gives 2 lines (Figures 10-a and 10-b).

5. Conclusions

This work was the opportunity to expose problems related to automatic recognition of individual binary images. There are two main techniques for extracting a signature from binary images: The first one uses a similarity index between rows or blocks of the image, the second uses covariance matrix to eliminate the rotation transformation effect.

In the paper of Baudrier et al. [1] we can see that the \( HD_a \) is not invariant according to translation. We also, proved in this work that \( HD_a \) is invariant according to rotation using covariance matrix and its Eigen values.

In future work, as each row or column in binary image can be considered as 1D vector (so we have \( N \) or \( M \) binary vector), we can reuse algorithm such as BRIEF, ORB or BRISK [5, 17, 22] categorized as binary valued features and are reserved to binary dataset.

References


[18] Lowie G., “Distinctive Image Features from Scale-Invariant Keypoints,” the International


Ahmed El Oirrak He joined Cadi Ayyad University, Morocco, in 1999, first as an assistant professor, and received the Doctorate and Habilitation in signal processing from the Mohammed V University, Morocco, in 2001 and University Cadi Ayyad, Morocco, in 2010 respectively. He is presently a PH professor with the Faculty of Sciences of Marrakech Semlalia. His research interests include image processing, pattern recognition and their applications. He is the author of more than 20 publications.

Mohammed Kaddioui is a full professor of computer science at the Department of Computer Science, Faculty of Science Semlalia, Cadi Ayyad University, Morocco. His major field of study is information processing and management and computer graphics.

Nouhoun Kane is a PhD student at the Department of Computer Science, Faculty of Science Semlalia, Cady Ayyad University, Morocco. His current research interests are signal, text and image processing.

Khalid Aznag is a PhD student at the Department of Computer Science, Faculty of Science Semlalia, Cady Ayyad University, Morocco. His current research interests are 2D and 3D images and 2D curve.