SHAPE-BASED FEATURES FOR IMAGE HASHING

Li Weng, Bart Preneel

Katholieke Universiteit Leuven
Department of Electrical Engineering, IBBT
Kasteelpark Arenberg 10, 3001 Heverlee, Belgium

ABSTRACT

Perceptual hashing is a solution for identification and authentication of multimedia content. The key of this technique is the extraction of proper features. In this paper, two features are proposed for natural image hashing. They are based on the description of shapes, in terms of contours and regions. The contour-based feature is formed by edge detection. The region-based feature is formed by the angular radial transform. Simulation results show that they have good robustness and discriminability. Compared to some other features, better ROC performance is achieved.

Index Terms— Perceptual, hash, image, fingerprint

1. INTRODUCTION

Identification and authentication of multimedia content are challenging issues. The main difficulty behind is that the same content can exist in different forms. For example, an image might undergo compression and format conversion during its distribution; it might be scaled; it might be smoothed, etc. These operations result in various binary representations of almost the same content. Consequently, it is difficult to track certain content from different sources. From a multimedia point of view, very often it is useful to consider these different versions as a whole rather than individually. One possible solution to meet this requirement is perceptual hashing, or multimedia fingerprinting. The principle of this technique is to extract robust features (fingerprints) from multimedia data. Robust features are those that do not vary if the content is not significantly changed. These features are resistant to incidental signal processing. In a perceptual hash algorithm, they are converted into a compact hash value. The problem of content identification can be solved by hash comparison. Since the hash value is dominated by the content, it can also be used for content authentication. The content can be judged as authentic if its perceptual hash value is not significantly changed.

A perceptual hash algorithm includes a hash generation part and a hash comparison part. The hash generation part normally consists of three stages: feature extraction, feature clustering and randomization [1]. Feature extraction is the stage where robust features are extracted. Feature clustering is a compression stage that can make the hash more compact. Randomization is a procedure that makes the hash value dependent on a secret key. It adds to the algorithm a security mechanism and also improves discriminability. The hash comparison part adopts a distance metric to measure the similarity between hash values, e.g., the Euclidean distance and the Hamming distance. Since hash values from similar contents tend to have a small distance, a threshold is set to judge if two hash values correspond to similar contents.

The performance of a perceptual hash algorithm is mainly characterized by robustness and discriminability. Robustness is the ability of a hash value to resist incidental distortion; discriminability means the uniqueness of a hash value. A good perceptual hash algorithm should have both strong robustness and good discriminability. Among these stages, feature extraction is the most important part, since the nature of extracted features essentially decides the performance.

Features can be global or local. Global features are extracted from a global perspective, e.g., from a whole image. Local features are extracted locally, e.g., from an image block. The advantage of global features is that they are compact and robust. However, they are only sensitive to global modification. For applications interested in local modification, local features are more suitable. The disadvantage of local features is that they may result in a longer hash value with less robustness. On the other hand, they achieve better discriminability. Features can also be distinguished by the type of distortion that they can resist. Generally speaking, there are geometric distortion and non-geometric distortion. Geometric distortion means those manipulations that might compromise the “synchronization” during hash generation, such as rotation, cropping, translation, etc. Non-geometric distortion can be interpreted as noise addition where noise might come from compression, Gaussian noise, filtering, etc. Currently, fea-
tures based on different principles have been proposed for images. For example, in [1], Swaminathan et al. use the Fourier-Mellin transform; in [2], Monga and Evans use feature points; in [3], Lefebvre et al. use the Radon transform.

In this work, two shape-based local features are proposed for perceptual hashing of natural images. The rationale behind is that the content of a natural image is mainly characterized by structural information, i.e., shapes. Therefore, features extracted from shapes are likely to well represent the content. Since shapes consist of surrounding contours and inside regions, one of the proposed features corresponds to contour information, and the other corresponds to region information. The contour-based feature is formed by edge detection. The region-based feature is formed by the angular radial transform. They mainly aim at non-geometric distortion. Their performance has been tested by simulation. The results show that they have strong robustness and good discriminability. Hypothesis testing indicates that high detection rates can be achieved at low false positive rates. Compared to some other features, better ROC performance is achieved.

The rest of the work is organized as follows: Section 2 describes the two features in detail; Section 3 examines their performance by simulation and shows the results; Section 4 concludes the work.

2. FEATURES FROM SHAPE INFORMATION

In this section, two features are proposed for natural image hashing. They are based on shape information in terms of contours and regions. In order to show their effectiveness, two algorithms are defined. Both algorithms are block-based. They extract local features from image blocks. At the beginning of each algorithm, some common pre-processing is applied. An image is first converted to gray scale and re-sized to a canonical size $512 \times 384$. It is then filtered by a Gaussian filter and a median filter to remove slight noise. After that, histogram equalization is performed to stabilize the contrast. The algorithm details are described below.

The first feature is extracted from image contours. First, a Sobel edge detector is applied. The result is a binary image. This image is divided into $64 \times 64$ non-overlapping blocks. A hash value is generated for each block. Since incidental distortion should not introduce a significant change in the hash value, the hash value is formed by noting the approximate distribution of edge points. Edge points are those points where discontinuities occur, detected by the edge detector. Each block is further divided into $32 \times 32$ sub-blocks with 50% overlapping, so that there are nine sub-blocks in each block. The pixel values of each binary sub-block are added up to derive nine values. They represent the edge density distribution of a block. Next, the mean of these nine values is computed. Each of the nine values is compared with the mean. A “1” bit is assigned to a sub-block if the corresponding value is greater than the mean; otherwise a “0” bit is assigned. In this way, each block is represented by a nine bit hash. Hash values from all blocks are concatenated to form a final hash of 432 bits.

The second feature is extracted by the angular radial transform (ART) [4]. This transform is adopted by the MPEG-7 standard as a method to produce a region-based shape descriptor [5]. The ART is a complex two-dimensional orthogonal transform, defined on a unit disk in polar coordinates. ART coefficients $F_{nm}$ are computed using the formula:

$$F_{nm} = \langle V_{nm}(\rho, \theta), f(\rho, \theta) \rangle$$

$$= \int_{0}^{2\pi} \int_{0}^{1} V_{nm}^*(\rho, \theta), f(\rho, \theta) \rho d\rho d\theta$$

where $f(\rho, \theta)$ is an image intensity function in polar coordinates and $V_{nm}(\rho, \theta)$ is the ART basis function of order $n$ and $m$. The basis functions are defined as

$$V_{nm}(\rho, \theta) = \frac{1}{2\pi} \exp(jm\theta)R_n(\rho),$$

$$R_n(\rho) = \begin{cases} 1, & n = 0 \\ 2\cos(n\rho), & n \neq 0 \end{cases}$$

The transform coefficients can efficiently represent shapes, and are robust to noise. In the simulation later, 24 basis functions are used, with $m \in \{0, 1, \cdots, 7\}$ and $n \in \{0, 1, 2\}$. An image is first binarized. It is then divided into $128 \times 128$ non-overlapping blocks. Each block is transformed by ART. The phases of the coefficients are kept. Each phase is quantized by a 2-bit uniform quantizer with Gray coding. They are concatenated to make a hash of 576 bits.

The distance metric used for hash comparison is the normalized Hamming distance, or the bit error rate (BER), defined as

$$d_{xy} = \frac{1}{N} \sum_{i=0}^{N-1} |x_i - y_i|$$

where $x$ and $y$ are two binary vectors of length $N$.

3. SIMULATION RESULTS

The performance of the proposed features has been examined by simulation, with respect to robustness and discriminability. The results are shown in this section. A set of 432 different natural images is used in the simulation. The image content consists of several types: architecture, landscape, sculpture, objects, humanoid, vehicle. Each type includes 72 images of three canonical sizes: $1600 \times 1200$, $1024 \times 768$, and $640 \times 480$. A perceptual hash value is generated from each image.

In order to test the robustness against incidental distortion, a set of legitimate operations are defined in Table 1. They are applied to each image to produce 20 distorted versions. Since the proposed features target at non-geometric distortion, there are not many geometric operations. For each distorted image, a hash value is computed. It is compared with
the original hash, and the BER is noted. The average BERs are shown in Table 2 for all the operations. The ART-based method exhibits BERs lower than 0.1 for almost all distortions. The contour-based method performs slightly worse, except for Gaussian noise. In common, both methods give larger BERs when the distortion is geometric. Besides, the contour-based method is also more vulnerable to salt and pepper noise, scaling, and strong JPEG compression, since they seriously affect the contour detection. The largest BERs for the two methods are 0.1635 and 0.23, given by rotation and scaling, respectively. Assuming that the BER between hash values of different contents should be around 0.5, the results are mostly acceptable.

The above assumption is verified by testing the discriminability. The average distance between two hash values of different contents is computed. Since there are 432 images and each has 20 distorted versions, in total there are \((432 \times 21^2) = 41055336\) pairs of different contents. The BER between each pair is computed and noted. The average BER for the ART-based method is 0.4406. For the contour-based method, the average BER is 0.4737. Therefore, the assumption holds. The contour-based method shows slightly better discriminability. This is because contours are more characteristic than regions. Note that the results are obtained without randomization, so that they reveal the “natural” performance of the proposed features. Better discriminability might be achieved by applying extra randomization. For example, Johnson and Ramchandran use “dithering” [6]. Swaminathan et al. allocate random weights to features [1].

Considering robustness and discriminability together, the overall performance is evaluated by hypothesis testing. When comparing a pair of hash values, two hypotheses \(H_0\) and \(H_1\) are defined as: \(H_0\) – they correspond to different contents; \(H_1\) – they correspond to similar contents. The performance can be given by the false positive rate \(P_f\) and the detection rate \(P_d\). They are defined as

\[
P_f = \text{Probability}\{d < T | H_0 \text{ is true}\} \quad (6)
\]

\[
P_d = \text{Probability}\{d < T | H_1 \text{ is true}\} \quad , \quad (7)
\]

where \(d\) is the BER and \(T\) is the threshold. By choosing different values for \(T\), different combinations of \(P_f\) and \(P_d\) are derived. They correspond to the receiver operating characteristic (ROC) curves, shown in Figure 1. These curves are computed from all the \((432 \times 21) = 41146056\) pairs of hash values. Since \(P_d\) increases with \(P_f\), it is preferred to have a high \(P_d\) while keeping \(P_f\) low. Both curves generally look steep. Some typical values are listed below.

<table>
<thead>
<tr>
<th>(P_f)</th>
<th>(P_d) (ART/ Contour)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0001</td>
<td>0.9658 0.8866</td>
</tr>
<tr>
<td>0.001</td>
<td>0.9921 0.9416</td>
</tr>
<tr>
<td>0.01</td>
<td>0.9952 0.9751</td>
</tr>
<tr>
<td>0.1</td>
<td>0.9939 0.9918</td>
</tr>
</tbody>
</table>

The ART-based method performs better than the contour-based method. For comparison, the ROC curves for other two features are also simulated and shown in Figure 1. The first one (720 bits) is the edge histogram proposed by Roy and Sun [7]; the other one (1536 bits) is based on fourth-order moments proposed by Yu et al. [8], but in order to make the hash size comparable to others, instead of quantization, each DCT coefficient is compared with the median value to derive a bit. The curves show that the proposed features both outperform the existing ones.

### Table 1. A set of legitimate distortions.

<table>
<thead>
<tr>
<th>Distortion name</th>
<th>Parameter setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rotation</td>
<td>Angle: 1°, 2°</td>
</tr>
<tr>
<td>Gaussian noise</td>
<td>Standard deviation: 50, 75</td>
</tr>
<tr>
<td>Central cropping</td>
<td>Percentage: 1%, 2%</td>
</tr>
<tr>
<td>JPEG compression</td>
<td>Quality factor: 10, 5</td>
</tr>
<tr>
<td>Scaling</td>
<td>Ratio: 0.2, 0.1</td>
</tr>
<tr>
<td>Median filter</td>
<td>Window size: 5 × 5, 7 × 7</td>
</tr>
<tr>
<td>Gaussian filter</td>
<td>Window size: 5 × 5, 7 × 7</td>
</tr>
<tr>
<td>Sharpening</td>
<td>Strength: 0.2, 0.5, 1</td>
</tr>
<tr>
<td>Salt &amp; pepper removal</td>
<td>Noise density: 0.05, 0.1 (^1)</td>
</tr>
<tr>
<td>Random row/column removal</td>
<td>No. of row/column(s): 10, 20</td>
</tr>
</tbody>
</table>

\(^1\)These are Matlab function parameters.
4. CONCLUSION AND DISCUSSION

In this work, two features are proposed for perceptual hashing of natural images. One is based on edge detection. The other is based on the angular radial transform. Their effectiveness is demonstrated by two algorithms. Both algorithms extract local features from image blocks. A hash value is composed by concatenating hash vectors from all blocks. The size of a hash is less than 600 bits. The robustness and discriminability of the proposed features have been evaluated. Simulation shows that the proposed features can well resist some typical non-geometric distortion and slight geometric distortion; the discriminability is also satisfactory. The overall performance is shown by the ROC curves. They indicate that high detection rates can be achieved at low false positive rates, and they outperform some existing features. Therefore, the proposed features are proper candidates for image hashing.

Some issues might be interesting for future investigation. Intuitively, if the proposed features can be used in a combined way, better performance might be achieved. Currently, they cannot cope with strong geometric distortion. This drawback significantly limits the robustness. The proposed algorithms also lack a randomization stage. Extending to key-based versions will facilitate applications that require higher security. In the ART-based algorithm, images are converted to black & white; coefficient phases are used instead of magnitudes. The performance may be different if gray images or magnitudes are used. For the contour-based algorithm, the performance can be affected by the underlying edge detector, thus the selection of a suitable edge detector is important.

5. REFERENCES


