Abstract—Recently, with the increasing number of internet users, detecting non legal copy of video sequence is of outstanding importance. This paper provides a new approach for content based video copy detection (CBVCD) which is invariant to rotation and flipping attacks. The proposed scheme is based on binary statistical image features (BSIF) descriptor using a new ring decomposition. The ring partition is particularly suitable for rotation/flipping attacks that affect the video frames. In fact, the visual content of each ring is kept constant when the video frames are rotated or flipped. The proposed VCD system was evaluated under TRECVID 2009 database and compared to others algorithms based on local binary pattern (LBP), local phase quantization (LPQ) or histogram of oriented gradient (HOG) descriptors. The experimental results demonstrated that the proposed descriptor is effective for all the attacks that can affect a video sequence and particularly in the case of rotation and flipping attacks.

Keywords—Content video protection, video copy detection, ring decomposition, BSIF descriptor.

I. INTRODUCTION

With the rapid development of the Internet, and the availability of video capturing devices such as digital cameras, and smart phones, the size of digital video collection is increasing rapidly. Efficient video searching, browsing and retrieval tools are required by users from various domains, including remote sensing, fashion, crime prevention, publishing, medicine, architecture, etc. For this purpose, many general purpose video retrieval systems have been developed. There are two well known frameworks: watermarking and content-based techniques.

During the past two decades, digital watermarking has been played an important role for securing data transmission over the Internet network. It has been applied for all data format such as images, audio and video [1]. Unfortunately, this technology presents some failures and drawbacks. When a watermark is inserted it can introduce some degradations to digital multimedia. Moreover, an already online video can not be recovered for inserting a logo. However, content based copy detection can be considered as an attractive alternative or a complementary application to watermarking technology.

Extracting content based video features not only allows us to quickly find video copies in large databases, but also can ensure content security of digital video. Generally, a short string which is a video feature is extracted and used in video retrieval [2], video authentication [3], video watermarking [4], video forgery detection [5] or video indexing [6].

There are many video copy detection algorithms successfully designed in last decade. However, from an applied context, there are still some limitations in video copy detection. For example, video rotation, flipping and scaling attacks are useful operations. In the postprocessing of video sequences, rotation, flipping or scaling have been frequently exploited to manipulate video data in order to claim copyright. However, most the proposed schemes for CBVCD are sensitive to rotation or flipping attacks [7].

In this paper, we propose an efficient robust video copy detection to meet both rotation and flipping robustness and good discriminative capability. This method is based on a ring decomposition and a BSIF descriptor. The main contribution in this work is the introduction of a feature extraction descriptor for video sequence which is achieved by a ring decomposition. This partition is invariant to video rotation and flipping, and then makes our algorithm resistant to these attacks. The rest of this paper is organized as follows; section II briefly discusses the related works. A detailed description of the proposed system is presented in Section III. The simulation results against different attacks including rotation and flipping are presented in section IV. Concluding remarks are drawn in Section V.

II. RELATED WORK

Many researchers have devoted themselves to developing CBVCD algorithms. In terms of the underlying techniques, the existing CBVCD algorithms can be roughly classified into two categories as follows, global descriptors and local descriptors techniques. However, the most important here is to show the current trend in designing CBVCD systems.

An approach for CBVCD in large videos archives dataset, namely ‘ViCopT’ for Video Copy Tracking is proposed in [8]. It is based on the estimation of points of interest movement and on the assignment of a label to these points behavior. To assign the labels, the author proposed to use a heuristic thresholds or a clustering algorithm. This indexing method has some attractive properties, it gives a rich, compact and generic descriptor, while labels of behavior provide a high level description of the video content. The work in [9] details a CBVCD approach based on a combination of global descriptors, an
automatic weighting algorithm, a pivot-based index structure, an approximate similarity search, and a voting algorithm for copy detection and localization, however it shows low performance against rotation and flipping attacks. In [10], a VCD system based on speeded up robust feature (SURF) and ordinal measure (OM) is proposed. Since SURF is an invariant feature based on scale space theory, the local feature is extracted by SURF in a frame-by-frame manner. Each frame is divided into 4 × 4 blocks, and each block is transformed by Hilbert-order rasterization to count the number of SURF points. Unfortunately, the execution time of this approach cannot be appreciated. In [11], a video retrieval approach using local non-negative matrix factorization (LNMF) is presented. This method exploits the convergence of factorized matrix extracted from original and transformed video sequence to design a new way to detect video copies. During this work, a LNMF based shot detection method is proposed for constructing a video identification framework and a LNMF based identification approach using Hausdorff distance is introduced, based on a two-stage search process.

In [12], a VCD framework is presented, based on fingerprints which are generated from the sub-bands of wavelet decomposed intensity image and localized intensity gradient histograms of low sub-band. The fingerprints reflect considerable discriminating capability and robustness against the attacks. Moreover, a robust sequence matching technique based on multivariate Wald-Wolfowitz test is designed. In [13], a motion vector-based CBVCD method was proposed, it uses the motion vectors extracted from frame sequences as a signature. However, when consecutive image frames are used, the resulting motion vectors are not descriptive enough because most vectors are either too small or they appear to scatter in all directions. The idea here is to calculate motion vectors in a lower frame rate than the actual frame rate of the video to overcome this problem. As a result, large vectors are obtained and they represent a given video in a robust manner. In [14], a novel frame-level descriptor for video is proposed. As a first step, the frames are partitioned into certain rings. In the second step, the Histogram of Oriented Gradient (HOG) and the Relative Mean Intensity (RMI) are extracted and considered as the feature vectors. Finally, the fusion of these two features is used to represent the global descriptor. However, this approach has been tested on small database that uses few attacks.

In [15], the authors have proposed an interesting CBVCD system based on fusion of BSIF and relative mean intensity. The proposed approach illustrates good performance against different attacks. However, its performance can not be appreciated when rotation and flipping attacks are considered. To improve the robustness against these attacks in this paper, we propose to use a ring decomposition that will act more effectively against rotation and flipping attacks.

III. PROPOSED SYSTEM

Our video copy detection system consists of the following steps, as shown in Fig. 1. After video decomposition, all the frames are undergo a pre-processing process, where borders are removed, and then each frame is transformed to BSIF domain. In the second step, the BSIF image is partitioned into different rings, which are then used to construct BSIF histograms. In the third step, resulting histograms are concatenated, and video histogram is finally formed by BSIF coefficients. In the following sections, we will present in detail the pre-processing process, the BSIF descriptor, and the flow chart of ring decomposition to better combat rotation and flipping attacks.

A. Pre-processing

Border is a common manipulation made in a video sequences. For each frame, we are interested in the significant content without borders. Besides, the intensities of the border are useless in frame analysis. We adopt a simple method, which removes the first few lines of each direction (left, right, top, bottom) whose sum of intensity is less than a threshold (20% of the maximum in this paper). Fig. 2 shows an example of two frames from TRECVID 2009 database after border removal.
Fig. 2. Examples of frames border removal. (top) original frames, (bottom) frames after border removal (samples frames from TRECVID 2009 dataset)

B. Binarised Static Image Features

BSIF descriptor represents each pixel by a binary code. These binary codes are constructed by learning a set of basis vectors from natural images using independent component analysis (ICA) and an efficient scalar quantization scheme [16]. The ICA is used to represent the data as a linear transformation of some latent independent components. Let $p$ denote the pixel grey values in an image patch concatenated into a vector. Using ICA, $p$ can be represented using a feature matrix $H$ as:

$$ p = H \cdot s $$  

where $s$ is a random vector and $H$ is constant that is the same for all different image patches. An approximation to $H$ up to a multiplicative constant can be retrieved without explicitly knowing the latent vector $s$, when a large enough number of training samples. Estimating $H$ is similar to determine the matrix $F$ which produces $s$ as the output of a number of linear filters as:

$$ s = F \cdot p $$  

where $F$ is considered as a filter applied to the pixels in $p$. The samples of a single patch are gathered into $z = (z_1, ..., z_N)$. $z_i$ is used to represent linear transformations of the independent components $s_i$. This is observed by multiplying both sides of Eq. (4) by the matrix performing the preprocessing and obtain:

$$ z = \mathcal{N} \cdot s $$  

where matrix $\mathcal{N}$ is the multiplication of $F$ by the preprocessing transformation matrix, $V$, which is used for whitening and dimensionality reduction. Here, for a matrix $U$ to be invertible, the number of independent components should be chosen in a way that it equals the number of variables produced after the whitening transformation. Under this condition, the system in Eq. (4) would be invertible in a unique way, producing the latent vector $r$ as a linear function of $z$ as:

$$ r = U \cdot z $$

where matrix $U$ represents the inverse of matrix $\mathcal{N}$. The filter matrix $F$ in Eq. (2) can then be obtained by multiplying the linear transformations $U$ and $V$.

$$ F = U \cdot V $$  

Consequently, the independent components $r_i$ of vector $r$ are obtained as:

$$ r = U \cdot V \cdot p $$

Finally, a useful post-processing step is binarising $r_i$ by thresholding at zero to produce the binarised features $b_i$ as:

$$ b_i = \begin{cases} 1 & r_i > 0 \\ 0 & \text{otherwise} \end{cases} $$

lastly, the binarised features $b_i$ form the BSIF image.

C. Ring decomposition

To overcome the invariance to rotation and flipping attacks seen in [15], we propose to replace the decomposition blocks within frame with circular one. This makes the detector invariant to rotation and flipping. Fig. 3 shows a central part of a video frame, and the corresponding part of the rotated video frame, where image center is considered as origin of coordinates. Obviously, visual contents in the corresponding rings of Figs. 3(a) and 3(b) are kept unchanged after frame rotation. The process is as follows: each BSIF image of the video sequence is divided into different rings and from each ring a BSIF histogram is extracted. The obtained histograms are than concatenated to form a vector which represent the frame content. Finally, all frames vectors are concatenated to form secondary frame, invariant to rotation. Fig. 4 is a representative diagram of proposed descriptor, where (a) is the original frame, (b) is the BSIF frame divided into four rings, (c) is the secondary matrix or frame formed by histograms of these rings and (d) is the global histogram which represent the frame content. Detailed steps for ring decomposition are as follows [18]:

$$ A = \pi r_n^2 $$

where $A$ is the area of inscribed circle,

$$ \mu_A = \left[A/n\right] $$
where \( n \) is the total number of rings with a radius \( r_n \) of the outmost circle within the image and \( \mu_A \) is the average area of each ring. So, \( r_1 \) can be computed by

\[
r_1 = \sqrt{\frac{\mu_A}{\pi}}
\]

(10)

Thus, other radius \( r_k \) \((k = 2, 3, ..., n - 1)\) can be obtained by the following equation:

\[
r_k = \sqrt{\frac{\mu_A + \pi r_{k-1}^2}{\pi}}
\]

(11)

If we consider \( p(x, y) \) to be the value of the pixel in the \( y \)th row and the \( x \)th column of the image \((1 \leq x, y \leq m)\), we suggest that \((x_c, y_c)\) are the coordinates of the image center. Thus, \( x_c = m/2 + 0.5 \) and \( y_c = m/2 + 0.5 \) if \( m \) is an even number. Otherwise, \( x_c = (m+1)/2 \) and \( y_c = (m+1)/2 \). Thus, the distance between \( p(x, y) \) and the image center \((x_c, y_c)\) can be measured by the euclidean distance as follows:

\[
d_{x,y} = \sqrt{(x-x_c)^2 + (y-y_c)^2}
\]

(12)

After obtaining the circle radius and pixel distances, the pixel values are grouped into \( n \) sets as follows:

\[
R_1 = \{p(x,y) \mid d_{x,y} \leq r_1\}
\]

(13)

\[
R_k = \{p(x,y) \mid r_{k-1} < d_{x,y} \leq r_k\} \quad (k = 2, 3, ..., n)
\]

(14)

Later, a sorted vector \( u_k \) is made in ascending order by reorganizing the elements of \( R_k \) \((k = 1; 2; ..., n)\). This operation guarantees that \( u_k \) is unrelated to rotation or flipping. As pixel coordinates are discrete, the pixel number of each set is not always equal to \( \mu_A \). Since, the pixels of each ring are expected to form a column of a new matrix, \( u_k \) is then mapped to a new vector \( v_k \) sized \( \mu_A \times 1 \) by linear interpolation. Thus, the new matrix \( V \) is obtained by arranging these new vectors as follows:

\[
V = [v_1, v_2, v_3, ..., v_n]
\]

(15)

As \( v_k \) is unrelated to rotation/flipping, \( V \) is also invariant to these operations. Except the rotation/flipping-invariant merit, the new matrix has another advantage that it has fewer columns than the original image. The feature vector that represents the video sequence can be finally extracted from the \( V \) matrix by using BSIF descriptor.

D. Matching process

Video matching plays an important role in the CBVCD. In this works we choose Chi-square distance as the similarity metric to match two descriptors \( D_1 \) and \( D_2 \), which represent the original and query video descriptors, respectively. In the matching process, a minimization process is employed. For an input query video, we find the clips with the minimal distance (maximal similarity) between descriptors in each source video. Then, we select the one with the lowest distance in the source. This distance is computed by the following equation:

\[
d^2 = \left( \frac{(D_1 - D_2)^2}{D_2} \right)
\]

(16)

where \( D_1 \) represents the descriptor extracted from the original video database, \( D_2 \) is the descriptor extracted from the transformed query video. It is worth noting, that minimum distance value signifies an exact match with the query.

IV. EXPERIMENTAL RESULTS

For the experiments, we used the video data from the video database of TRECVID 2009 Copy Detection task [16]. It includes web, TV archives and movies, and cover documentaries, movies, sport events, TV shows and cartoons. 30 videos are used to construct a new reference video database, 300 new query videos with different copy attacks are copies of the 30 videos (original). The format of video is MPEG-1 with 352 × 288 pixels and 25 fps.

To prove the correctness of the proposed circular decomposition, we start the experiments by extracting the global BSIF histogram using blocks decomposition as it was proposed in [15]. The deterioration in performance under different rotation and flipping is depicted in Table. I. The obtained results demonstrate that by increasing the rotation angle, the detection rate will be decreased since the difference between original blocks and rotated blocks will be increased.
TABLE I
DETECTION RATE OF [15] FOR ROTATION AND FLIPPING ATTACKS

<table>
<thead>
<tr>
<th>Attacks</th>
<th>Detection rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rotation (5°)</td>
<td>96.29</td>
</tr>
<tr>
<td>Rotation (10°)</td>
<td>93.72</td>
</tr>
<tr>
<td>Rotation (30°)</td>
<td>81.52</td>
</tr>
<tr>
<td>Rotation (90°)</td>
<td>77.11</td>
</tr>
<tr>
<td>Flipping</td>
<td>80.05</td>
</tr>
</tbody>
</table>

Fig. 5. Example of transformations: (a) original frame, (b) picture-in-picture Type 1, (c) insertion of pattern, (d) strong re-encoding, (e) change of gamma, (f) letterbox, (g) white noise, (h) crop, (i) shift, (j) caption/text insertion, (k) flip, and (l) picture-in-picture Type 2.

To evaluate the robustness of the proposed scheme based on ring decomposition, several transformations are constructed including rotation, flipping, AVC/H.264 compression, logo embedding, gamma correction, video in video, etc. Fig. 5 summarizes the different attacks used in the experimental study. It gives example frames of some transformations used in the evaluation, this example is extracted from TRECVID 2009 database.

To show advantages of the proposed CBVCD system based on the new rings decomposition, we have firstly tested the variation of the matching quality of the proposed approach when the number of the ring \( n \) used to extract the video features is varied. We have set \( n \) to be 4/8/12/16. Table II illustrates the matching results for rotation and flipping attacks. It is observed that, for a fixed rank, the whole CBVCD performances will be improved when the number of rings increases. In fact, the number of rings is equal to column number of the global descriptor. Fewer columns will lead to fewer features in the final descriptor, which will inevitably hurt the discriminative capability.

Secondly, we have compared the proposed CBVCD approach with other CBVCD systems which are based on some state-of-the-art descriptors, e.g: the local binary pattern [19], the local phase quantization [20], and the histograms of oriented gradients [21]. Each of these techniques extract the local features from non-overlapping blocks within the input image. Figs. 6 and 7 show the ROC curves comparison of different descriptor under rotation and flipping attacks. It is observed that, the proposed descriptor outperforms the others.

As a second comparison step, we compare our system to two previous systems described in [15] and [22]. The work in [22] is based on a major incline-based fast alignment method to find potential alignment positions between the compared videos. This approach is selected for its low complexity and its robustness for several attacks. Table III gives a comparison of simulation results of the proposed approach against other state of the art systems under the TRECVID 2009 corpus.

These experiments show that the proposed descriptor is effective and efficient for almost the attacks used in the TRECVID 2009 database. It is clear that this descriptor is

<table>
<thead>
<tr>
<th>Ring number ((n))</th>
<th>Attacks</th>
<th>Rotation %</th>
<th>Flipping %</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td></td>
<td>87.33</td>
<td>91</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>89.67</td>
<td>93.85</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>95</td>
<td>97</td>
</tr>
<tr>
<td>16</td>
<td></td>
<td>95.07</td>
<td>97.16</td>
</tr>
</tbody>
</table>

Fig. 6. ROC curve comparisons among different descriptors under rotation attack

Fig. 7. ROC curve comparisons among different descriptors under flipping attack
TABLE III

<table>
<thead>
<tr>
<th>#</th>
<th>Transformation Description</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Transf.: blur, gamma change, ...etc</td>
<td>93</td>
<td>96.6</td>
<td>95</td>
</tr>
<tr>
<td>2</td>
<td>Transf.: crop, shift, contrast, ...etc</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>Insertion of pattern</td>
<td>97</td>
<td>99</td>
<td>99</td>
</tr>
<tr>
<td>4</td>
<td>Strong re-encoding</td>
<td>77</td>
<td>32</td>
<td>03</td>
</tr>
<tr>
<td>5</td>
<td>Change of gamma</td>
<td>100</td>
<td>69</td>
<td>32</td>
</tr>
<tr>
<td>6</td>
<td>03 transf.: blur, gamma change, ...etc</td>
<td>97</td>
<td>93</td>
<td>93</td>
</tr>
<tr>
<td>7</td>
<td>03 transf.: crop, shift, contrast, ...etc</td>
<td>100</td>
<td>100</td>
<td>98</td>
</tr>
<tr>
<td>8</td>
<td>05 transf. chosen from T2 - T8</td>
<td>100</td>
<td>100</td>
<td>95</td>
</tr>
<tr>
<td>9</td>
<td>Rotation (90°)</td>
<td>82</td>
<td>72</td>
<td>72</td>
</tr>
<tr>
<td>10</td>
<td>Flipping</td>
<td>85</td>
<td>80</td>
<td>97</td>
</tr>
</tbody>
</table>

TABLE IV

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ren et al.</td>
<td>137</td>
</tr>
<tr>
<td>Jiang et al.</td>
<td>69</td>
</tr>
<tr>
<td>BSIF-square</td>
<td>32</td>
</tr>
<tr>
<td>BSIF-ring</td>
<td>45</td>
</tr>
</tbody>
</table>

more effective to rotation/flipping attacks, although it kept good performance against other attacks. Furthermore, the ring decomposition adds a slight complexity in comparison with the square decomposition, as it is shown in Table IV.

V. CONCLUSIONS

In this paper, we have proposed a robust CBVCD based on ring decomposition, which is robust against image rotation and flipping and has a desirable discriminate capability. A key component of our CBVCD is the construction of new matrix based on ring decomposition, which is used to represent the video content. This matrix is rotation/flip-invariant, and therefore makes the video content resilient to rotation/flip. Experiments have been conducted on TRECVID 2009 database to validate our CBVCD performances, and showed that our CBVCD system is robust against content-preserving manipulations, such as frame rotation, flipping, AVC/H.264 compression, watermark embedding, Gaussian low-pass filtering, gamma correction, brightness adjustment, contrast adjustment, and image scaling, and reaches good discrimination. The ROC curve comparisons have indicated that the proposed CBVCD show better robustness against rotation/flip/ing than some well-known features extraction descriptor for content base copy detection, such as LBP, LPQ and HOG. In the future, our research will focus on extending spatial descriptor by adding temporal component to improve the performance of proposed CBVCD approach.

REFERENCES