Joint Color and Texture Descriptor Using Ring Decomposition for Robust Video Copy Detection in Large Databases

Yassine Himeur, Karima Ait Sadi
Centre de développement des Technologies Avancées.
Cité 20 Aout, Baba Hassen, 16303, Alger, Algérie.
Email: yhimeur@cdta.dz, aitsaadi@cdta.dz

Abstract—The growth of audiovisual content, and in particular video requires the creation of robust tools for detecting illegal copies. This paper presents an effective approach to search and detect illegal copies in large video databases. This automatic detection operates two local descriptors and their paths through the video. This method allows to reduce the temporal redundancy intrinsically linked to the video and to add a context of behavior in these descriptors. Thus, starting from a low-level signal description, our approach allows to achieve a higher level of representation for two reasons; i) it uses a new way to decompose video frames based on a ring decomposition and, ii) it combines a local texture descriptor namely binarized statistical image features (BSIF) extracted using a ring decomposition and local color descriptor. The obtained description is more compact, non-redundant and can be highly robust to rotation and flipping. The evaluation shows a clear improvement in performance against other novel techniques listed in the state of the art while exhibiting better flexibility. It is more real time on a large video base (TRECVID 2009).

Index Terms—Video copy detection, ring decomposition, BSIF descriptor, color descriptor, rotation-flipping attacks.

I. INTRODUCTION

Over the years, the increasing demand on online sharing sites such as YouTube and DailyMotion leads to a waste of storage resources and problems related to copyright infringement. Video copy detection (VCD) is a technology aimed at detecting illegal copies effectively, thereby introducing new vocabularies such as watermark-based technique or content-based technology. Solutions based on watermarking prevent unauthorized copying of multimedia content and Internet distribution [1]. However, if the addition of a trademark does not affect the visual quality of the document, its removal can usually be done also without affecting the quality of the video, after certain types of media processing [2]. On the other hand, content-based technology can be regarded as a passive method since it does not need any extra information, such as a watermark, for detection. By measuring the similarity of some robust features between questionable media and those in the database, we can find the corresponding copies. In this paper, we focus on the content-based technology for video sequence matching.

Content-Based Video Copy Detection (CBVCD) aims to locate the near-duplicates copies of a given query clip from a video database, which means the transformed version after content-preserving operations, such as geometric distortion, blurring, noise contamination, contrast enhancement, and re-encoding. Generally, the robustness of the video feature against the above-mentioned operations is one of the important issues for VCD. The discriminability, as well as the computational and storage complexity, also needs to be considered in the sequence matching algorithms. 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 [3], video authentication [4], video watermarking [5], video forgery detection [6] or video indexing [7].

Rotation or flipping have been frequently exploited to manipulate video data in order to claim copyright. However, most the CBVCD systems are sensitive to these attacks [8]. Furthermore, although color is commonly experienced as an indispensable quality in describing the world around us, many feature-based representations are only based on shape description, and ignore color information. The description of color is hampered by the large amount of variations which causes the measured color values to vary significantly. Indeed, a change in illuminant color, viewpoint, and acquisition material, all influence the color values of the scene.

In this paper, we focus on improving the robustness against rotation and flipping attacks of the work proposed in [13], and then adding a color description for more enhancing the robustness since the color information is an important features in video data. The main contribution of this work is summarized as follows; First, the ring decomposition is used to extract video features. Hence, It has the advantage of being invariant to video rotation/fitpling. second, a photometrically color descriptor which is extracted from local regions is derived by taking into consideration robustness to photometric changes commonly encountered in the real world, varying image quality, from high quality images to snapshot photo quality and compressed internet images. Third, a texture
snapshot description of video sequence is achieved using BSIF. The fusion of both descriptors is then used to represent each video sequence.

The remainder of this paper is arranged as follows. A brief description of some related works is introduced in Section II. The proposed system is presented in Section III. Simulation results against different attacks used, including rotation and flipping are given in Section IV. Conclusion and future works are drawn in Section V.

II. RELATED WORK

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. In this section we briefly described some effective techniques [9], [10], [11], [12], [13], [14] found in literature.

In [9], different efficient histogram-based color descriptors to catch and represent the color properties of a video sequence are presented. One descriptor, called alpha-trimmed average histogram, combine individual frame or image histograms using a specific filtering operation to generate robust color histograms that can eliminate the adverse effects of brightness/color variations, occlusion, and edit effects on the color representation. Through this work, the efficacy of the alpha-trimmed average histograms for video segment retrieval applications is presented, and the way they consistently outperform key frame-based methods is illustrated. Moreover, a color histogram descriptor called the intersection histogram is introduced, it reflects the number of pixels of a given color that is common to all the frames in the video sequence. Finally, the intersection histogram is employed to develop a fast and efficient algorithm for identification of the video segment to which a query frame belongs. In [10], a fuzzy color histogram-based shot-boundary detection algorithm for CBVCD is proposed. After detecting cuts and gradual transitions (fade, dissolve) in a video sequence, a color histogram generated with the fuzzy linking method on $L^*a^*b^*$ color space is computed, the system extracts a mask for still regions and the window of picture-in-picture transformation for each detected shot, which will be useful in a CBVCD system. In [11], the authors accentuate on the content-based technology using color correlation on summarized videos and also incorporating temporal features for matching the video sequences.

In [12], a compact spatio-temporal feature based local binary pattern (STF-LBP) to represent videos and construct an efficient data structure to index the feature to achieve real-time retrieving performance is proposed. This descriptor exploits relative gray-level intensity distribution within a frame and temporal structure of the videos along frame sequence. In [13], an interesting CBVCD system based on the fusion of the BSIF and relative mean intensity is proposed. This approach illustrates good performance against different attacks. However, its performance cannot be appreciated when rotation and flipping attacks are considered. To improve the robustness of this approach against such attacks, a new decomposition that will act more effectively is introduced in [14]. The proposed descriptor is based on binary statistical image features (BSIF) descriptor using a ring decomposition. The ring partition is particularly suitable for rotation/flipping attacks that affect the video frames while keeping good performance against other attacks.

III. PROPOSED SYSTEM

The proposed CBVCD system is based on the fusion of texture descriptor (BSIF) and color descriptor, as shown in Fig. 1. After video decomposition, all the frames are undergone a preprocessing process, where borders are deleted, then each frame is transformed to BSIF domain. The BSIF image is partitioned into different rings, which are then used to construct BSIF histograms. Resulting histograms are concatenated, and video content histogram is finally formed by BSIF coefficients. From another side, a color descriptor is extracted based on Hue transformation and local partition. Finally, the BSIF and color descriptor are concatenated using a weighting parameters to obtained the global descriptor that represents the video sequence.

A. Preprocessing

The performance of the proposed system can be substantially improved by including a simple video preprocessing. In fact, adding borders is a common manipulation used in 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.

B. Binarized Statistical 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 [15]. The ICA is used to represent the data as a linear transformation of some latent independent components. Let
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:

$$p = H \cdot s$$ (1)

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:

$$s = F \cdot p$$ (2)

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. 2 by the matrix performing the preprocessing and obtain:

$$z = H \cdot s$$ (3)

where matrix $H$ 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:

$$r = U \cdot z$$ (4)

where matrix $U$ represents the inverse of matrix $H$. The filter matrix $F$ in Eq. 2 can then be obtained by multiplying the linear transformations $U$ and $V$, i.e.

$$F = U \cdot V$$ (5)

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

$$r = U \cdot V \cdot p$$ (6)

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

$$b_i = \begin{cases} 1 & r_i > 0 \\ 0 & \text{otherwise} \end{cases}$$ (7)

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

### C. Ring decomposition

To make video content resilient to rotation/flipping, in this work we propose to extract the vector descriptors BSIF from rings decomposition instead of blocks decomposition. For this, each BSIF image of the video sequence is divided into different rings and a BSIF histogram is extracted from each ring. 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 obtained histograms from different rings are than concatenated to form a vector which represent the frame content. Finally, all the 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 of proposed ring construction are as follows [16]:

$$A = \pi r_n^2$$ (8)

where $A$ is the area of inscribed circle,

$$\mu_A = \lfloor A/n \rfloor$$ (9)

where $n$ is the total number of rings with a radius $r_n$ 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:

\[
R_1 = \{p(x, y) \mid d_{x,y} \leq r_1\} \quad (13)
\]

\[
R_k = \{p(x, y) \mid r_{k-1} < d_{x,y} \leq r_k\} \quad (k = 2, 3, ..., n). \quad (14)
\]

Later, a sorted vector \(u_k\) is made in ascending order by reorganizing the elements of \(R_k\). 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:

\[
V = [v_1, v_2, v_3, ..., v_n]\quad (15)
\]

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

\section*{D. Color descriptor}

In this section a color representation is introduced by the computation of the hue descriptor. The hue and saturation are computed at each position; these can also be represented as a vector, where the hue and saturation are the angle and the length, respectively. We compute the hue histogram of the patch where the strength of the update is equal to the saturation of the measurement. This guarantees that pixels with low saturation (black-grey-white), where the hue is undefined, have no influence on the final color descriptor. The color descriptor is constructed for each local patch to be robust to photometric changes commonly encountered in the real world, and different video processing attacks [17]. Moreover, the hue descriptor is robust to lighting geometry and specularities when when assuming white illumination. Each video frame is decomposed into a set of patches as it is shown in Fig. 5. Every patch is then represented by a histogram over hue computed from the corresponding RGB values of each pixel according to:

\[
C = \text{hue} = \arctan \left( \frac{\sqrt{3}(R - G)}{R + G - 2B} \right) \quad (16)
\]

\section*{E. Fusion descriptor}

The two above computed descriptors, \((S: \text{BSIF and C: color})\) are combined together using a weighted linear mechanism according to:

\[
\hat{F} = (\gamma_1 \cdot \hat{S}, \gamma_2 \cdot \hat{C}) \quad (17)
\]

where \(\hat{\cdot}\) means that the descriptor is normalized and \(\gamma_1\) and \(\gamma_2\) are the weighting factors.

\section*{IV. Experimental results}

\subsection*{A. Test protocol}

For the experiments, we used the video data from the video database of TRECVID 2009 Copy Detection task [18]. 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 \(352 \times 288\) pixels and 25 fps. To evaluate the robustness of the proposed scheme, several transformations are constructed including rotation, flipping, strong re-encoding (AVC/H.264) compression, logo embedding, gamma correction, picture in picture, etc.

\subsection*{B. Performance Comparison}

To prove the correctness of the proposed ring decompositions, we first compare the global BSIF histogram using blocks decomposition as it was proposed in [13]. The deterioration in performance under different rotation and flipping is depicted in Table I.

\begin{table}[h]
\centering
\caption{Detection rate of [13] for rotation and flipping attacks}
\begin{tabular}{|c|c|}
\hline
\textbf{Attacks} & \textbf{Detection rate (%)} \\
\hline
Rotation (5\textdegree) & 96.29 \\
Rotation (10\textdegree) & 93.72 \\
Rotation (30\textdegree) & 81.52 \\
Rotation (90\textdegree) & 77.11 \\
Flipping & 80.05 \\
\hline
\end{tabular}
\end{table}

To show advantages of the proposed CBVCD system based on the 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 the \(n\) to be 4/8/12/16. Table II illustrates the matching results for rotation and flipping attacks, when the Chi-square distance is chosen as the similarity metric. It is observed that, the whole CBVCD performances will be improved when the number of rings increases and from \(n\) equal to 16, the performance in terms of similarity changes slightly. 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. Is worth noting that the best results
are obtained using $n$ equal to 16 and the weighting parameters $\gamma_1 = 0.01$ and $\gamma_2 = 3$, these parameters are set experimentally.

Afterward, we compare our 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]. The performance of each CBVCD system under rotation and flipping is plotted using its verification rate. This is a plot of the inverse false positive rate $(1 - F_{pr})$ versus the false negative rate $(F_{nr})$. If $N_T$ is considered as the the total number of match tests conducted and $\tau$ is the match threshold. With $F_n$ the number of false negatives (clips that should have matched, but did not) and $F_p$ the number of false positives (clips that matched but are not part of the reference set), we have

$$F_{pr}(\tau) = \frac{F_p}{N_T}, \quad F_{nr}(\tau) = \frac{F_n}{N_T} \quad (18)$$

and

$$\text{Verification rate} = 1 - F_{pr} \quad (19)$$

The verification rate curves are computed by varying $\tau$ from its minimum value to its maximum. Figs. 6 and 7 show the verification rate curves comparison of different descriptor under rotation and flipping attacks. It is observed that the curve of the proposed technique (Color-BSIF-ring) lies very close to the axes, hence the proposed descriptor outperforms the others.

As a final comparison step, other CBVCD systems described in [13], [14] and [22] are implemented and compared to the proposed system. 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 shows the comparison results the TRECVID 2009
TABLE III
 COMPARISON RESULTS FOR TRECVID 2009 DATABASE

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>Picture in picture</td>
<td>93</td>
<td>96.66</td>
<td>95</td>
<td>95.17</td>
</tr>
<tr>
<td>13</td>
<td>Insertion of pattern.</td>
<td>100</td>
<td>100</td>
<td>99</td>
<td>98.11</td>
</tr>
<tr>
<td>14</td>
<td>Strong re-encoding (AVC/H.264).</td>
<td>100</td>
<td>100</td>
<td>99</td>
<td>99.24</td>
</tr>
<tr>
<td>15</td>
<td>Change of gamma</td>
<td>100</td>
<td>100</td>
<td>98</td>
<td>99.15</td>
</tr>
<tr>
<td>16</td>
<td>03 random trans.: blur, gamma change, ...etc</td>
<td>87</td>
<td>93</td>
<td>93</td>
<td>95.05</td>
</tr>
<tr>
<td>17</td>
<td>03 random trans.: crop, shift, contrast, ...etc,</td>
<td>100</td>
<td>100</td>
<td>98</td>
<td>98.65</td>
</tr>
<tr>
<td>18</td>
<td>05 transformations chosen from 12 - 18</td>
<td>100</td>
<td>100</td>
<td>98</td>
<td>98.93</td>
</tr>
<tr>
<td>19</td>
<td>Rotation (90°)</td>
<td>82</td>
<td>77</td>
<td>93</td>
<td>95.88</td>
</tr>
<tr>
<td>10</td>
<td>Flipping</td>
<td>85</td>
<td>80</td>
<td>97</td>
<td>98.73</td>
</tr>
</tbody>
</table>

TABLE II
 MATCHING QUALITIES WITH DIFFERENT n FOR ROTATION AND FLIPPING ATTACKS

<table>
<thead>
<tr>
<th>Ring number (n)/attacks</th>
<th>Rotation (%)</th>
<th>Flipping (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>88.93</td>
<td>92.61</td>
</tr>
<tr>
<td>8</td>
<td>91.72</td>
<td>95.35</td>
</tr>
<tr>
<td>12</td>
<td>95.17</td>
<td>97.09</td>
</tr>
<tr>
<td>16</td>
<td>95.88</td>
<td>98.73</td>
</tr>
</tbody>
</table>

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 more effective to rotation/flipping attacks, although it kept good performance against other attacks.

V. CONCLUSION

In this paper, we have proposed a first version of our proposed CBVCD system (Color-BSIF-ring), in which we have introduced color and texture information and used a ring decomposition, to enhance its robustness and distinctiveness. The positive results achieved on the TRECVID 2009 data base, show its performance and efficiency relative to other popular features descriptors, and demonstrate that this description technique gives more robustness and distinctiveness especially against geometric transformation such as rotation and flipping. In the future, our research will focus on using this descriptor into keyframe based approach to further accelerate the CBVCD framework. Another direction is to fuse the combined spatial-color descriptor to temporal component to improve the performance of proposed CBVCD approach.

REFERENCES