IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 41, No. 4, August, 2024, pp. 1885-1897
Journal homepage: http://iieta.org/journals/ts

A Novel Correlated Microstructure Elements Descriptor for Image Retrieval

Kevin Salvador Aguilar-Domínguez*, Raúl Pinto-Elías, Gabriel González-Serna, Andrea Magadán-Salazar

Department of Computer Science, National Technological of Mexico/CENIDET, Cuernavaca 62490, Mexico

Corresponding Author Email: kevin.aguilar17ca@cenidet.edu.mx

https://doi.org/10.18280/ts.410419

Received: 19 August 2023
Revised: 21 March 2024
Accepted: 17 May 2024
Available online: 31 August 2024

Keywords:
image retrieval, microstructure descriptor, correlated visual features, structure descriptor, image descriptor, image representation, texture descriptor, color descriptor

ABSTRACT

In recent years, substantial progress has been made in developing new descriptors to enhance content-based image retrieval (CBIR) systems. These advancements often focus on leveraging the relationship between low-level features such as color and texture. This study introduces the Correlated Microstructures Elements Descriptor (CMED), a novel descriptor that integrates three low-level features to improve image retrieval performance. Our experiments on three distinct natural image datasets reveal that CMED significantly outperforms both classical and state-of-the-art descriptors. The proposed algorithm demonstrates superior indexing and retrieval capabilities, achieving up to 26.41% improvement compared to the MPEG-7 standard and 10.75% compared to contemporary state-of-the-art descriptors. The findings underscore CMED's potential to advance the field of CBIR, offering robust solutions for accurately retrieving images based on semantic content.

1. INTRODUCTION

With the high increase in Internet technology and digital devices, it has become easier to obtain images of any object, animals, place, or any content that interests us. Consequently, we can generate a vast number of images in our daily lives. These images are widely utilized across various fields of knowledge. Similarly, these images require filtering, searching, or identification to solve or enhance different processes. Content-based image retrieval (CBIR) systems help address this issue and offer advantages over text-based systems by reducing human workload and the complexities of natural language processing. CBIR systems are currently utilized in a variety of fields such as face retrieval [1], shopping [2], building retrieval [3], clothe retrieval [4], medical image retrieval [5], and others [6-14].

Over the years, ideas for standardizing multimedia content have emerged to facilitate the processes of search, filtering, and retrieval. One such standard is the MPEG-7 standard proposed by the Moving Picture Experts Group (MPEG) [15], which comprises eight parts oriented towards describing multimedia content, including audio, images, and video. However, the problems present in retrieval systems have not yet been fully solved. Recent research has focused on various issues such as user interaction, segmentation, dimensionality reduction, feature indexing, geotag-based image retrieval, high-level image features, content-based image retrieval for privacy preservation, and content-based video retrieval [15]. Specifically, the challenges related to high-level image features revolve around reducing the semantic gap, which refers to the disparity between what a user seeks in an image and what the retrieval system provides [16, 17].

The semantic gap can be associated with the representation of descriptors. Achieving a robust descriptor representation entails it being invariant to certain transformations, including scale, rotation, and translation. Additionally, CBIR systems predominantly utilize low-level features to represent images, such as color, shape, texture, and spatial position. Since low-level descriptors are unlikely to be directly linked to high-level features [18], such as activities, places, objects, emotions, among others [15].

To establish a direct relationship or achieve a better representation of high-level features, some research has introduced improvements based on visual theories or the utilization of various techniques to establish a connection between two or more low-level features. Many proposals primarily focus on texture and color, owing to their effective discrimination and close correlation. Recently introduced descriptors mainly rely on "Textons," which are founded on Julesz's Textons theory [19]. Although different types of descriptors use more than one low-level feature, some descriptors are based on feature integration theory [20], or related descriptors in the similarity measure [21].

This paper presents a novel descriptor to improve the representation of the descriptors present in the state-of-the-art that use Microstructures and elements structures, the proposal descriptors aim to reduce the semantic gap by developing a descriptor that represents high-level image features. Specifically, it utilizes the relationship between low-level features of the image to enhance evaluation on sets of images necessitating level three retrieval, as per the levels outlined by Eakins. As well as, to get better geometry transformation tolerance, present a new methodology to classify the type of structures based on elements instead of shape of structures. According to the metrics, the descriptor gets results above the state-of-the-art even when transformations are applied to

images.

The rest of the paper is organized as follows. Section 2 shows related works in the state-of-the-art, Section 3 presents the proposal model, Section 4 the experiments and results performed by the descriptor and finally, Section 4 reports the conclusions.

2. RELATED WORKS

Various algorithms have been designed to extract color and texture features for image retrieval [22]. Color is widely used due to its invariance to rotation and scale, which makes it powerful in image classification. On the other hand, texture features provide important information about the smoothness, coarseness, and regularity of many real-world objects [23]. Some classical low-level color descriptors used in CBIR systems are proposed by the MPEG-7 standard, such as the dominant color descriptor (DCD), color layout descriptor (CLD), and scalable color descriptor (SCD) [24]. Similarly, in texture-based algorithms, the MPEG-7 standard adopts three texture descriptors: homogeneous texture descriptor (HTD), texture browsing descriptor (TBD), and the edge histogram descriptor (EHD) [24]. Although low-level features yield good results, single low-level visual features are not sufficient to represent high-level semantics due to highly complex visual content [25] To reduce the semantic gap and improve the results of CBIR systems, many methodologies have been proposed in the state-of-the-art.

Considering the powerful discrimination and the close relation between texture and color, CBIR systems such as those presented in the previous studies [26-28] propose a methodology using both color and texture to represent image content and present techniques to combine them for better retrieval results. On the other hand, to represent high-level features, recent descriptor proposals use the Texton theory, as exemplified by the Microstructure descriptor (MSD) by Liu et al. [22], which leverages the relationship between color and texture, utilizing what they term "microstructures". The authors consider microstructures an evolution of Textons as they incorporate both color and texture. In subsequent years, Dawood et al. [25] proposed improvements to the MSD descriptor, such as the Correlated MicroStructure Descriptor (CMSD), which, unlike MSD, identifies microstructures by establishing correlations between texture orientation, color, and intensity features. CMSD also incorporates edge direction differently from MSD by adding 45° and 135° diagonal edges. Similarly, the Structure Elements' Descriptor (SED) proposed by Wang and Wang [29], is another example of descriptors utilizing both color and texture. SED, a scaled invariant descriptor, is based on structures detected in a quantized image, using five different structure elements: 0°, 45°, 90°, 135°, and no direction. Additionally, Chu and Lei [30] present the Multi Integration Features Histogram descriptor (MIFH). Inspired by the feature integration theory, present a model that uses the color and edge features for image representation.

3. CORRELATED MICROSTRUCTURES ELEMENTS DESCRIPTOR

Although the CMSD descriptor provides good results in image retrieval, it has been observed that it only covers small structures because it uses a 3×3 window size, which could affect its retrieval performance. In addition, the CMSD descriptor only considers the correlation of microstructures and not the types of structures present in the images. On the other hand, the descriptor could be limited by the techniques used to generate the feature maps.

Considering microstructure detection based on the Texton theory, the correlated microstructures processed by CMSD, and the methodology of structure element detection as in SED, the proposed descriptor named CMED can be considered an improvement over the CMSD descriptor and SED. The key difference between both descriptors and CMED lies in the utilization of two methodologies and the process employed to generate the elements' histogram. Rather than relying on classical types of structures based on shape like SED, the proposed descriptor adopts a novel methodology based on the number of elements, which remains invariant to geometric transformations.

The feature extraction methodology follows the CMSD and MSD processes and is described in the following six steps: (1) Extraction of low-level features; (2) Generation of microstructure maps using the extracted features; (3) Detection of structure using the microstructure maps obtained in the previous step; (4) Calculation of correlations between the microstructures using the three feature maps, resulting in three correlation maps; (5) Generation of histograms for the detected structure elements in each feature; (6) Concatenation of the histograms to obtain a single histogram of structures. Similarly, the histogram of each microstructure correlation map is obtained and concatenated into a single correlation histogram. Finally, both histograms are used to produce the CMED descriptor vector. Figure 1 illustrates the descriptor process graphically. Detailed descriptions of these steps are provided in the following sections.

3.1 Low level feature extraction

The initial step involves extracting the feature maps and quantizing them to generalize and reduce the required resources. The image is then transformed from the RGB to the HSV color space. This transformation is chosen because the HSV color space is widely reported in the literature as being more akin to human perception [25].

3.1.1 Color map

The quantization of the image in HSV involves uniformly quantizing the H, S, and V values [25]. With $Bh = 8, Bs = 3$ and $Bv = 3$, as depicted in Eqs. (1)-(3), this results in the color map $CM(x, y)$ having the same number of rows and columns as the original image with $8 * 3 * 3 = 72$ distinct values. Consequently, this yields a $matrix(x, y)$ with values ranging from 0 to 71, as shown in the Eq. (4).

$$Q_h(x, y) = H(x, y) \times (B_h / max_h) \tag{1}$$

$$Q_s(x, y) = H(x, y) \times (B_s / max_s) \tag{2}$$

$$Q_v(x, y) = H(x, y) \times (B_v / max_v) \tag{3}$$

$$CM(x, y) = Qh(x, y) \times (B_s \times B_v) + Q_s(x, y) \times B_v + Q_v(x, y) \tag{4}$$

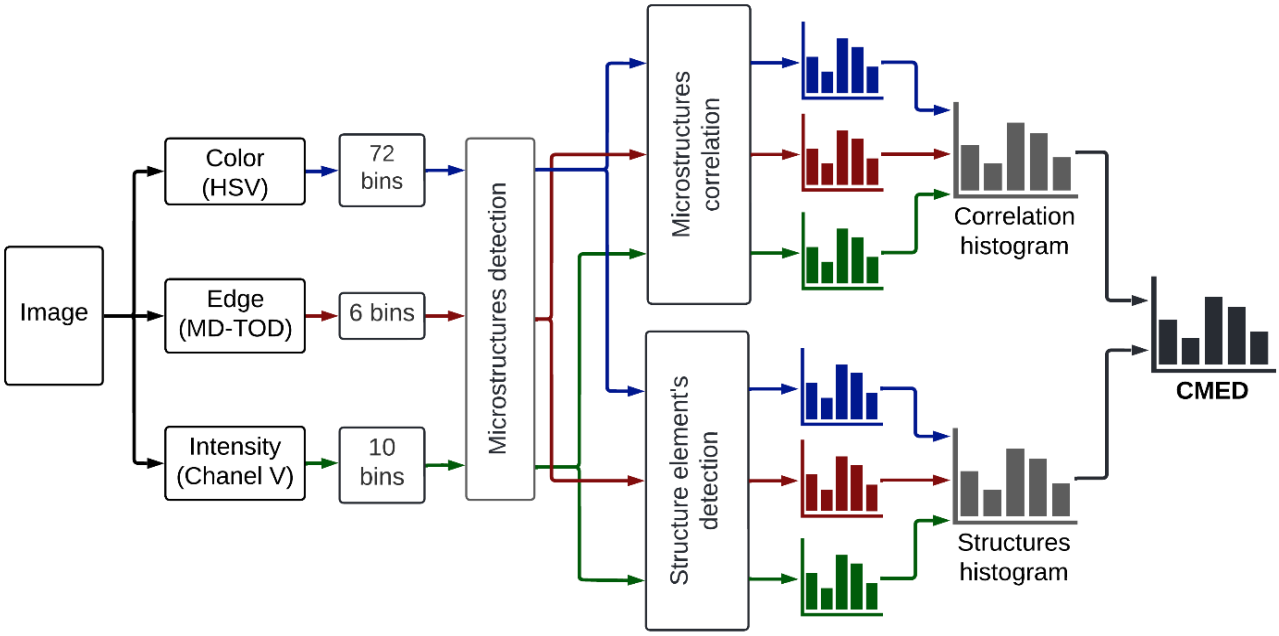


Figure 1. Proposed Correlated Microstructures Elements Descriptor (CMED)

3.1.2 Intensity map

The intensity map $IM(x,y)$ is obtained by quantizing the intensity channel V of the image in HSV into 10 bins. Since the value in channel V range from zero to one, the quantized intensity map is expressed as shown in Eq. (5), where B_I is set to ten, and $V(x,y)$ represents the value in channel V at the coordinates (x,y), in the HSV image. This results in a matrix with values ranging from 0 to 9.

$$IM(x,y) = V(x,y) \times B_I \tag{5}$$

3.1.3 Edge map

Finally, the edge map is obtained with the methodology proposed in the previous study [25], known as Multi-Dimensional Texture Orientation Detection (MD-TOD) method. The edge detection employs Sobel filters in four edge directions, $0°, 45°, 90°$, and $135°$. Since CMSD utilizes the HSV color space, the descriptor transforms from cylindrical to Cartesian coordinates after applying the Sobel filters, as demonstrated in Eqs. (6)-(8). Here, H_c, S_c and V_c represent the new values of the HSV space in Cartesian coordinates.

$$H_c(x,y) = S(x,y) \times \cos\big(H(x,y)\big) \tag{6}$$

$$S_c(x,y) = S(x,y) \times \sin\big(H(x,y)\big) \tag{7}$$

$$V_c(x,y) = V(x,y) \tag{8}$$

With Cartesian coordinates in HSV, the MD-TOD methodology detects diagonal edge vectors denoted by $\widehat{d_{45°}}$ and $\widehat{d_{135°}}$. The diagonal edge orientation map $Orimap_{Diag}$ is obtained using Eqs. (9-13). Here $H_{45°}, S_{45°}$ and $V_{45°}$ represent edges extracted with the Sobel 45° operator for each channel, while $H_{135°}$, $S_{135°}$ and $V_{135°}$ represent edges detected with the Sobel 135° operator.

$$\cos\big(\widehat{d_{45°}}, \widehat{d_{135°}}\big) = \frac{\widehat{d_{45°}} \cdot \widehat{d_{135°}}}{|\widehat{d_{45°}}| \cdot |\widehat{d_{135°}}|} \tag{9}$$

$$\widehat{d_{45°}} \cdot \widehat{d_{135°}} = H_{45°} \cdot H_{135°} + S_{45°} \cdot S_{135°} + V_{45°} \cdot V_{135°} \tag{10}$$

$$|\widehat{d_{45°}}| = \big(H_{45°}^2 + S_{45°}^2 + V_{45°}^2\big)^{\frac{1}{2}} \tag{11}$$

$$|\widehat{d_{135°}}| = \big(H_{135°}^2 + S_{135°}^2 + V_{135°}^2\big)^{\frac{1}{2}} \tag{12}$$

$$Orimap_{diag} = \arccos\big(\cos\big(\widehat{d_{45°}}, \widehat{d_{135°}}\big)\big) \tag{13}$$

The edge orientation map $Orimap_{hv}$ is obtained using Eqs. (14)-(18). Here, the horizontal and vertical edge orientation are denoted by $\hat{h}$ and $\hat{v}$, respectively. H_h, S_h and V_h represents the edges detected for each channel with Horizontal operator, while H_v, S_v and V_v represent the edges with the Vertical operator.

$$\cos\big(\hat{h}, \hat{v}\big) = \frac{\hat{h} \cdot \hat{v}}{|\hat{h}| \cdot |\hat{v}|} \tag{14}$$

$$\hat{h} \cdot \hat{v} = H_h \cdot H_v + S_h \cdot S_v + V_h \cdot V_v \tag{15}$$

$$|\hat{h}| = (H_h^2 + S_h^2 + V_h^2)^{\frac{1}{2}} \tag{16}$$

$$|\hat{v}| = (H_v^2 + S_v^2 + V_v^2)^{\frac{1}{2}} \tag{17}$$

$$Orimap_{hv} = \arccos\big(\cos\big(\hat{h}, \hat{v}\big)\big) \tag{18}$$

The edge orientation map, $OM(x,y)$, is obtained using Eq. (19), which computes the average of the maps previously calculated and quantified into B_O levels, set to 6.

$$OM(x,y) = \frac{\big(Orimap_{hv} + Orimap_{diag}\big)}{2} \times \frac{B_O}{180} \tag{19}$$

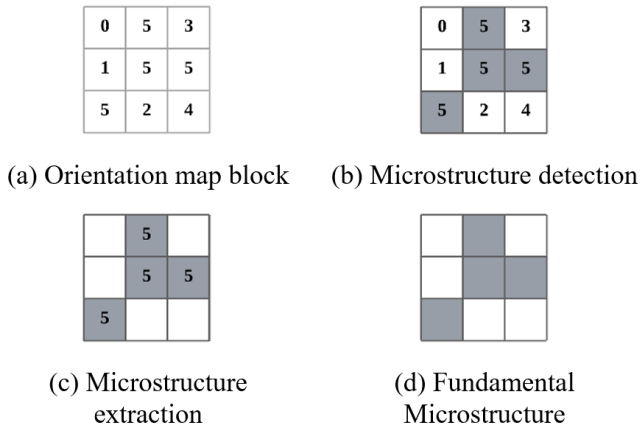
3.2 Microstructure detection

Microstructure detection follows the methodology outlined

in CMSD [25]. Figure 2 illustrates an example of microstructure detection. The method is applied to each featured map using a 3×3 block. Detection occurs on all the featured map from right to left and from top to bottom with four different origins (0,0), (0,1), (1,0) y (1,1), with a stride of 3. The process returns four microstructure maps M_1, M_2, M_3 and M_4, respectively for each different origin.

These four maps are then combined following Eq. (20), to generate a microstructure image, $M_T(x, y)$, as show in Figure 3. Finally, the process of microstructure detection in each feature map returns three microstructures' images, one for each feature.

$$M_T(x,y) = \max\left(M_1(x,y), M_2(x,y), M_3(x,y), M_4(x,y)\right) \quad (20)$$



(a) Orientation map block (b) Microstructure detection

(c) Microstructure extraction (d) Fundamental Microstructure

Figure 2. Fundamental Microstructure detection

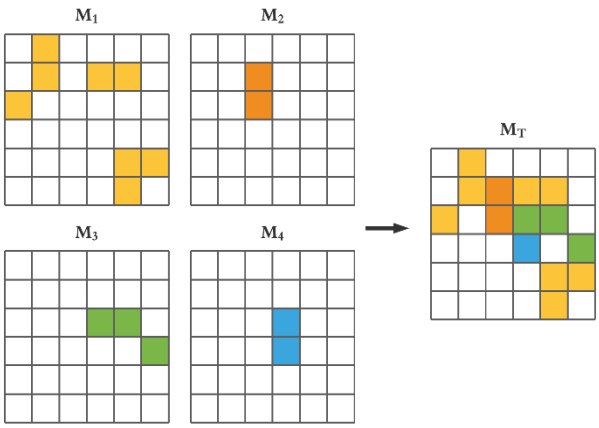


Figure 3. Microstructure image

3.3 Structure's detection

Given the usefulness of structure elements in discriminating between images of different categories, we have decided to integrate them into our descriptor. Since the methodology to obtain the microstructures already generates a map of structures during the process, incorporating this process can enhance our descriptor by leveraging information we already possess from earlier stages. Additionally, we have opted not only to focus on structures present in edges but also to consider structures found in color and luminosity features. This expansion ensures a more comprehensive representation of the image content.

While various types of structures based on shape or direction, as demonstrated in Figure 4, and utilized in previous works such as [29], exist in the literature, we have proposed two new categorizations of structures for our CMED. These new categorizations aim to address the limitations associated with current structures, particularly in terms of their susceptibility to transformations.

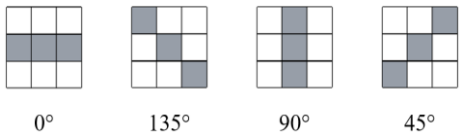


Figure 4. Classical structures

We have proposed two different ways to categorize the structures. One approach is based on the number of elements, while the other is based on identifying structures that remain invariant under rotation and reflection transformations, which we refer to as fundamental structures.

3.3.1 Number of elements

Based on the number of elements in the structures, which provides tolerance to rotation and a reduced number of types, we have established a categorization method. As illustrated in Figure 5, this categorization consists of eight different types, unlike classical structures, which can have more than 16 types. Consequently, the previously detected microstructures are labeled into a type according to the number of elements.

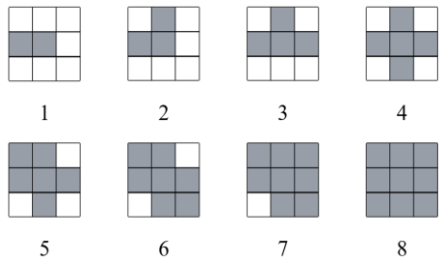


Figure 5. Structures based on number of elements

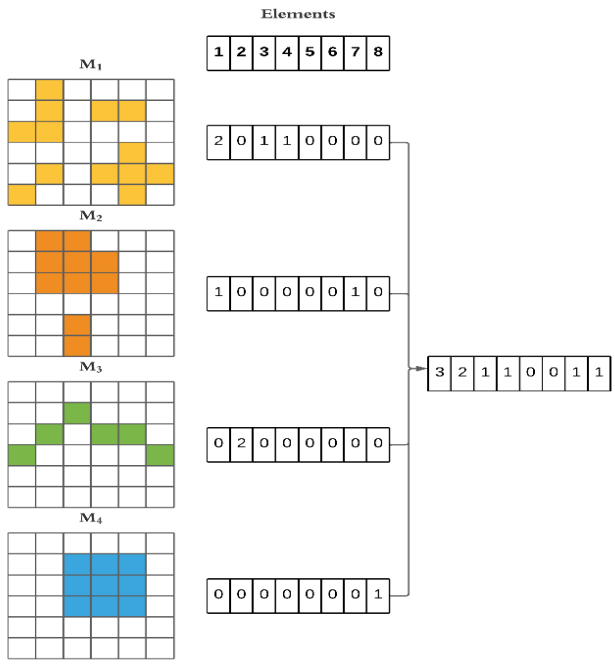


Figure 6. Structure histogram vector

To obtain the structure elements, the detection process begins at the same origin as each feature map and moves a 3×3 block from left to right and top to bottom with a stride of 3. This means that the previously detected fundamental microstructure is assigned a type based on the number of elements, and the structures detected in the microstructure map are saved in a histogram of length ns, representing the number of structure types to consider, in this case, eight or nine if we consider the structure without an element.

The descriptor obtains the structure histogram for each map and combines the four vectors, as shown in Figure 6. Finally, the result vector is normalized. This process is repeated for each feature, resulting in three histograms: one for color, one for edges, and one for luminosity, respectively.

3.3.2 Fundamental structures

Considering that the proposed methodology based on the number of elements in the structures could be overly general, we acknowledge that there is a great variety of different shapes with the same number of elements, as shown in Figure 7. These variations in shape may be important for representing the content in the image.

To obtain a vector with more detailed information in structure detection, we propose considering not only the number of elements but also the shape of the structures detected, while ensuring invariance to reflection and rotation transformations. We classify structures derived from rotation and reflection transformation as the same type. This means that structures obtained from a rotation, as illustrated in Figure 8, and those obtained from a reflection transformation, as shown in Figure 9, are categorized as the same type of structure.

Detecting the type of structure considering rotation and reflection transformations requires more than counting the number of elements. For this reason, we propose a methodology to assign the structure type, considering that structures across transformations are the same structure.

Figure 7. Different shapes with two-element structure

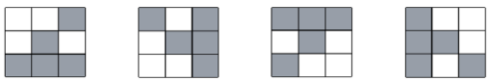


Figure 8. Four-element structure in different orientations

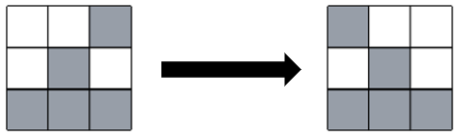


Figure 9. Reflection transformation

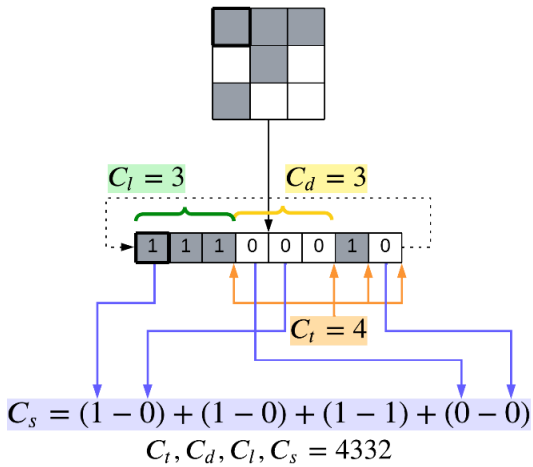


Figure 10. Structure detection

To detect the type of structure, the descriptor extracts the eight elements of the structure, meaning the 3×3 block, and arranges them in a vector. In this vector, we assign a value of one to positions where an element exists and zero to positions without an element, as illustrated in Figure 10.

Once the vector is created, we propose the use of four coefficients to classify the structures:

1) C_t: following the same idea as LBP-U [31], this coefficient counts the number of transitions, indicating the changes from one to zero and vice versa.
2) C_d: This coefficient counts the maximum distance between two elements in the structure.
3) C_l: Represents the maximum length of continuous elements in the structure.
4) C_s : This coefficient represents the number of different symmetric elements, calculated using Eq. (21), where V_e is the vector obtained from the eight neighbors of the structure.

$$C_s = \sum_{i=1}^{8} |V(i) - V(i+4)| \tag{21}$$

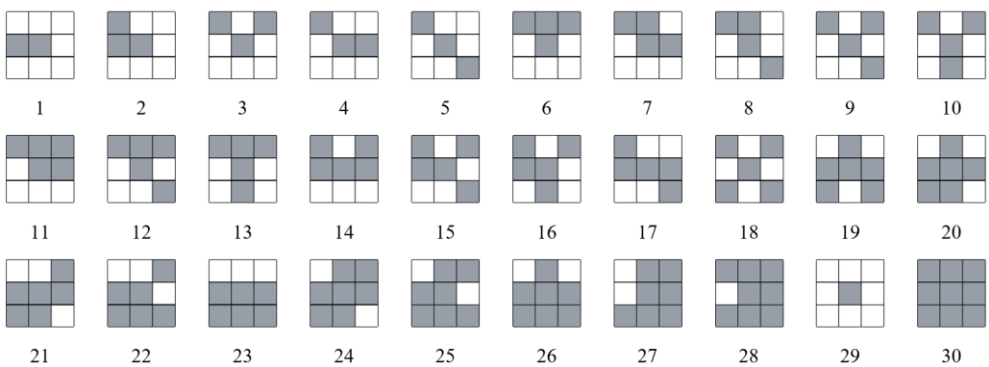


Figure 11. Fundamental structures

To calculate all coefficients, we consider position eight in the vector as preceding position one in a cyclic manner. These four values together form a unique identification code for the structure. Figure 10 visually illustrates the representation of each coefficient.

In clarifying the process and demonstrating how these coefficients classify structures consistently across transformations, we can consider the examples in Figure 9, which is the reflection of the example in Figure 10. In this example, we obtain $V_e = [1,0,0,0,1,1,1,0]$, resulting in the following coefficients: $C_t = 4; C_d = 3; C_l = 3$; and $C_s = 2$. Thus, the identification code for the structure is 4332, which is identical to the code for the structure in Figure 10. This indicates that all structures detected with the same number will be classified as structures of the same type.

Following this process and considering both empty and full structures, we detected a total of thirty different types, as shown in Figure 11. We call these types "fundamental structures".

3.4 Microstructure correlation

To obtain a description that contemplates the relationship between low-level features, our proposed descriptor follows the microstructure correlation process presented by the previous work [25]. Initially, the microstructure map is obtained by utilizing two microstructure images to extract information from a feature map. Subsequently, these microstructure images are combined to yield a correlated microstructure map, denoted as $M_C(x, y)$ as shown in Eq. (22). Finally, M_C is employed to extract information from the feature map, considering the values located within the microstructures. This process yields a micro-feature map. The entire procedure is repeated for each feature, resulting in the Micro-color map, the Micro-orientation map, and the Micro-intensity map. An illustrative example of this process to obtain the Micro-Orientation map is presented in Figure 12, where the intensity and orientation microstructure images are utilized to extract information from the color map.

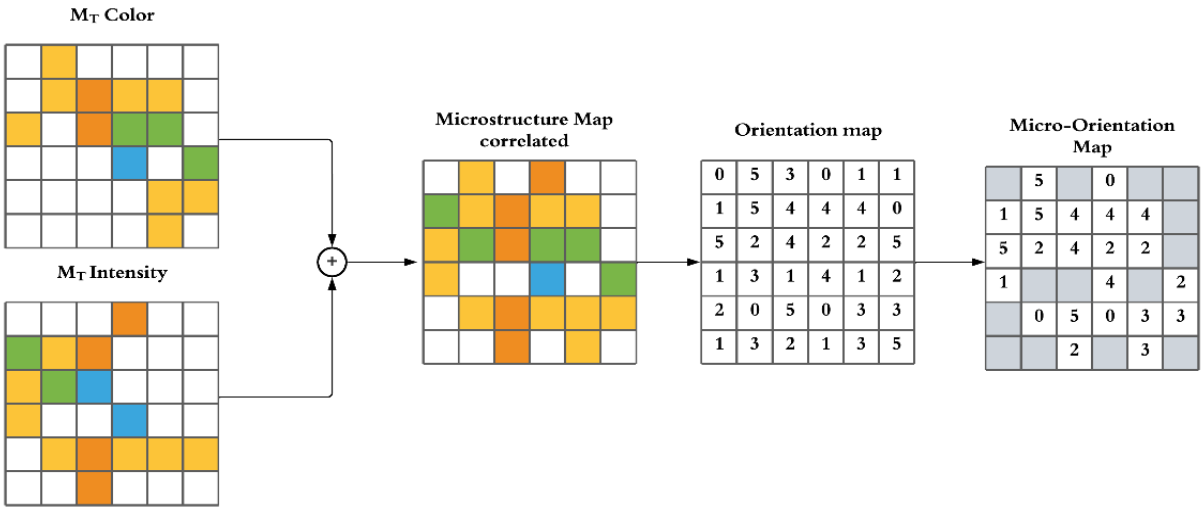


Figure 12. Construction of micro-orientation map

$$M_C(x, y) = \max\left(M_T^1(x, y), M_T^2(x, y)\right) \quad (22)$$

3.5 Feature representation

The descriptor is represented by histograms. Concatenating the histogram of the types of structures in each feature results in a total of $ns * nf$ values, where ns is the number of structures and nf the number of features maps. Additionally, the histogram of correlation of microstructures of each feature is concatenated, resulting in a vector with $72 + 6 + 10 = 88$ values.

Finally, if we consider the fundamental structures in color, orientation, and luminosity (30 fundamental structures, $ns = 30$ and $sf = 3$), the resulting vector length is $120 + 88 = 208$.

On the other hand, when using structures based on the number of elements considering the empty structure ($ns = 9$ and $sf = 3$), we obtain a vector length of $27 + 88 = 115$.

Thus, the CMED descriptor contains information on the occurrence of each type of structure and microstructures. The CMED vector is defined in Eq. (23). Where H_s^C, H_s^O and H_s^I, are the histograms of the structure's elements for color, orientation, and intensity, respectably. Similarly, H_m^C, H_m^O and H_m^I the histograms of the Micro-color, Micro-orientation, and Micro-intensity map, respectably.

$$CMED = [H_s^C, H_s^O, H_s^I, H_m^C, H_m^O, H_m^I] \quad (23)$$

4. EXPERIMENTS AND RESULTS

To evaluate the performance of the proposed CMED descriptor, this section presents experiments conducted to compare different parameter settings, as well as an objective evaluation of its performance compared to the state-of-the-art in content-based image retrieval. We evaluated the performance of our method using the three metrics used in the state-of-the-art, and the metric proposed by the standard MPEG-7.

One of the most used metrics is Precision, which provides a percentage of correctly retrieved images in a query. Its definition is presented in Eq. (24), where $r(x_n)$ can take a value of zero or one following Eq. (25), K represents the total number of retrieved images,x_n denotes the image retrieved at position n, and Ic_q is the set of images in the corresponding category to the query q. Thus, $r(x_n)$ will be one when the image retrieved in position n is relevant.

$$P = |K|^{-1} \sum_{n=1}^{K} r(x_n) \tag{24}$$

$$r(x_n) = \begin{cases} 1, & x_n \in Ic_q \\ 0, & otherwise \end{cases} \tag{25}$$

Recall, a widely used metric, complements Precision by measuring the percentage of relevant images retrieved relative to the total number of images in the query class. Mathematically, it is expressed by Eq. (26). Where R_q represents the total number of images in the class for query q.

$$R = |R_q|^{-1} \sum_{n=1}^{K} r(x_n) \tag{26}$$

Similarly, Mean Average Precision (MAP) could be regarded as a synthesis of the precision and recall processes, utilizing the precision at k, as defined by Eq. (27), for each retrieved image. MAP is calculated using Eqs. (28)-(29). Where k is the position at which the image was retrieved, and Q denotes the total number of queries performed.

$$P_k = |k|^{-1} \sum_{n=1}^{k} r(x_n) \tag{27}$$

$$AP_q = |R_q|^{-1} \sum_{k=1}^{K} P_k \times r(x_n) \tag{28}$$

$$MAP = |Q|^{-1} \sum_{q=1}^{Q} AP_q \tag{29}$$

Finally, the Average Normalized Modified Retrieval Rank (ANMRR) metric, as proposed by the MPEG-7 standard, serves to evaluate retrieval systems by considering the position of the relevant image retrieved. ANMRR provides a percentage of error, where results closer to zero indicate better performance. The metric can be defined by six equations.

Firstly, $Rank_n$ is assigned to each image x_n belonging to the set of images of the query category. $Rank_n$ is determined based on the position k at which the image is retrieved in the query, as shown in Eq. (30). Additionally, C_q is a coefficient obtained by Eq. (31).

$$Rank_n = \begin{cases} k, & k \le C_q \\ 1.25 \times K_q, & otherwise \end{cases} \tag{30}$$

$$C_q = \min\{4 \times R_q, 2 \times \max[R_q, \forall_q]\} \tag{31}$$

AVR_q represents the Average Rank of all images in the category to which the query belongs, and it can be defined as Eq. (32). MRR_q denotes a Modified Retrieval Rank, which is subsequently normalized to obtain $NMRR_q$. Finally, the average of all queries' results yields the ANMRR, as illustrated in Eqs. (33)-(35).

$$AVR_q = \sum_{n=1}^{R_q} \frac{Rank_n}{R_q} \tag{32}$$

$$MRR_q = AVR_q - 0.5 \times [1 + R_q] \tag{33}$$

$$NMRR_q = \frac{MRR_q}{1.25 \times k_q - 0.5 \times [1 + R_q]} \tag{34}$$

$$ANMRR = |Q|^{-1} \sum_{q=1}^{Q} NMRR_q \tag{35}$$

For the experiments we used some of the most popular dataset obtained from Corel Photo Gallery [32]. The Corel-1k dataset is often utilized as a benchmark dataset in image retrieval research [33], It consists of 1,000 images covering a variety of scenes, objects, and textures, providing a diverse range of visual content for evaluation purposes. Each image is annotated and categorized into one of ten distinct classes, facilitating the assessment of retrieval algorithms across different semantic categories.

Similarly, the Corel-5k dataset offers a larger and more varied collection of images, with 5,000 images distributed among 50 categories. However, it is worth noting that these images may contain artifacts on the edges, which can pose additional challenges for retrieval algorithms. Despite this, the dataset remains valuable for evaluating the robustness and generalization capabilities of image retrieval methods.

In contrast, the Corel Database for Content-based Image Retrieval (Corel-CBIR) dataset [34], provides an extensive collection of 10,800 images divided into 80 unbalanced concept groups. This dataset offers a more comprehensive and diverse set of images, covering a wide range of visual concepts and scenes. The unbalanced nature of the concept groups introduces additional complexity, requiring retrieval algorithms to effectively handle varying levels of representation within each category.

Overall, these datasets serve as crucial resources for evaluating and benchmarking the performance of content-based image retrieval systems. Their diverse content and well-defined categories enable researchers to assess the effectiveness and robustness of different retrieval algorithms under various conditions and settings.

For our experiments, we adopted a systematic approach by selecting 10 random images per category to serve as query images. This strategy ensured a balanced representation across the categories and minimized bias in the evaluation process. Consequently, we generated a total of 100 queries (Q=100) from the Corel-1k dataset, 500 queries (Q=500) from the Corel-5k dataset, and 800 queries (Q=800) from the Corel-CBIR dataset.

For practical purposes and considering that people tend to seek results in the first retrieved images, we initially conducted evaluations at *K*=12, meaning the top 12 most relevant images for each query. However, for a comprehensive assessment, we also conducted comparisons with variations of *K* in the range from 1 to 100.

Furthermore, for evaluations using the ANMRR metric, we utilized all retrieved images to ensure proper utilization of the metric.

4.1 Performance evaluation of proposed CMED

To determine the most effective similarity measure for our CMED descriptor, we conducted experiments using three commonly used measures: Manhattan Distance Eq. (36), Euclidean Distance Eq. (37), and the similarity measure

defined in the previous work [29] for the SED descriptor Eq. (38).

$$L1(T,Q) = \sum_{i=1}^{M} |T_i - Q_i| \tag{36}$$

$$L2(T,Q) = \sqrt{\left(\sum_{i=1}^{M} (T_i - Q_i)^2\right)} \tag{37}$$

$$Ls(T,Q) = \sum_{i=1}^{M} \frac{|T_i - Q_i|}{1 + |T_i + Q_i|} \tag{38}$$

These measures were tested using the Corel-1k image set, with the number of retrieved images (K) varying from 1 to 100. The results, depicted in Figure 13, indicate that while the results between the L1 distance and the distance proposed for the SED (Ls) descriptor are quite similar, the Ls distance consistently outperformed the others. The Ls distance demonstrates superior performance throughout the experiment, suggesting that it is a more suitable option for descriptors based on microstructures.

Furthermore, to enable the detection of microstructures and structures of different sizes, we enhanced the proposed CMED descriptor by incorporating a pyramidal sub-sampling approach like the used in the SIFT descriptor [35]. This experiment considered a maximum depth of eight. The results, as illustrated in Figure 14, exhibit a notable decline, likely attributable to the size of the image set. Notably, the smallest set demonstrates diminished results at depth seven, while the largest set experiences a decline at depth five. The most stable depth observed was four or less. Overall, the descriptor achieved an improvement of 1.75% compared to depth four and one. However, depth four entails a higher computational cost due to the processing of four different image sizes.

Table 1 presents the results obtained with the Corel-1k image set, where we evaluated different structures configurations. Specifically, we also tested using 7 and 28 elements, excluding structures with 8 and zero elements. This approach considers the proposed methodologies based on the number of elements and fundamental structures, respectively, resulting in four possible configurations: 7, 9, 28, and 30 types.

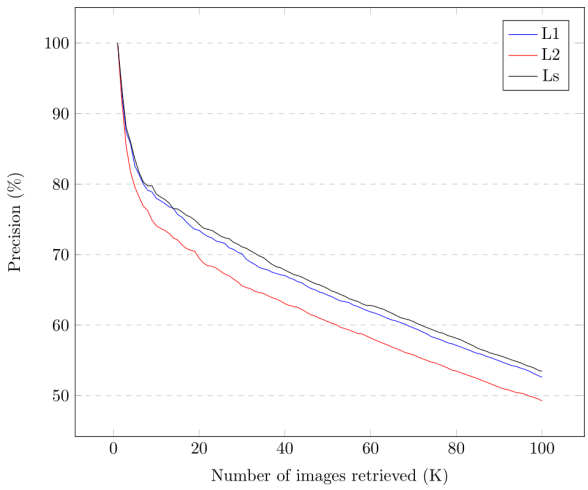


Figure 13. Precision of proposed CMED at different similarity measures with Corel-1k

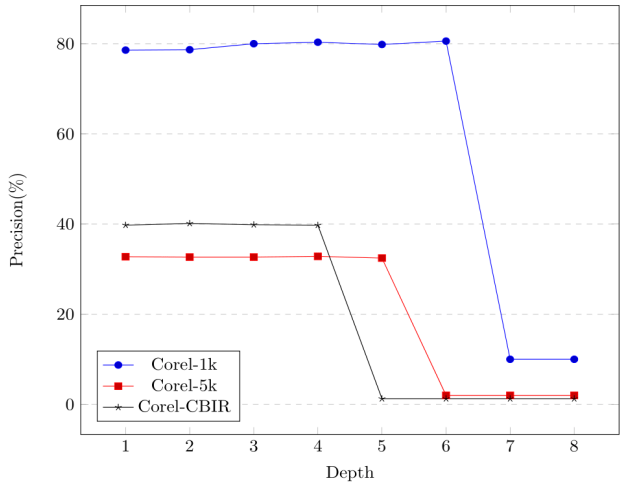


Figure 14. Results of different depths for CMED with precision metric

Table 1. Evaluation of structures and position

Position	Number of Structures			
	7	9	28	30
MM	78.42%	78.58%	78.50%	78.58%
CM	47.08%	44.25%	52.25%	47.83%

Additionally, we explored two methods to obtain the structures: microstructure maps (MM) and correlation maps (CM). This involves considering two microstructure maps derived from different image features.

The results indicate that incorporating more structure types enhances retrieval performance by up to 5%, particularly when the descriptor obtains structures from the correlation maps. However, when structures are detected in the microstructure maps, the descriptor performs better, with no significant difference between considering nine and thirty types of structures. Thus, increasing the number of structures may not necessarily lead to improvement and can even incur additional computational costs.

4.2 Comparison of CMED with other CBIR descriptors

For this experiment, we utilized the image datasets Corel-1k, Corel-5k, and Corel-CBIR. Additionally, we compared our descriptor against several widely used descriptors in content-based image retrieval systems, including those specified in the MPEG-7 standard such as EHD and CLD [24]. We also included state-of-the-art descriptors utilizing microstructures such as MSD [22] and CMSD [25], as well as descriptors based on structures like SED [29] and feature integration such as MIFH [30].

To ensure a fair evaluation, we implemented all descriptors in the same environment and with configurations recommended by their respective authors. For our descriptor, we selected the configuration that exhibited the best performance, which involved using 9 structures based on the number of elements, extracting structures from the microstructure maps, and utilizing the Ls (SED descriptor distance).

Figures 15-17 display the precision results obtained across the three image datasets. The graphs illustrate that our descriptor outperforms others in terms of precision across all image sets. This suggests that CMED can retrieve a greater number of correct images compared to alternative descriptors.

Furthermore, it is evident that CMED exhibits a gentler slope in precision reduction as K increases, consistently maintaining superior performance compared to others. Notably, at K=12, a commonly used threshold for performance evaluation, our descriptor achieved a precision of 78.58% on the Corel-1k dataset.

The results demonstrate a significant improvement of over 19% compared to descriptors proposed by the standard and a 3.5% enhancement compared to the CMSD descriptor, which exhibited the best performance among state-of-the-art descriptors. Moreover, on the Corel-5k dataset at K=12, our descriptor showcases an improvement of more than 22% compared to the MPEG-7 standard and consistently surpasses the CMSD descriptor. Similarly, with the Corel-CBIR dataset, the proposed descriptor achieves an improvement of over 18% compared to standard descriptors and a more than 3.5% enhancement over state-of-the-art descriptors.

The graph in Figure 18 illustrates the results obtained with the Recall metric on the Corel-1k dataset. The Recall metric provides insight into the retrieval capability of the descriptor by indicating how many relevant images it can retrieve in response to a query. Ideally, achieving perfect retrieval would result in a steep and accurate increase in performance as the value of K increases. In the graph, it's evident that the proposed descriptor demonstrates a more linear increase compared to other descriptors, indicating superior retrieval performance.

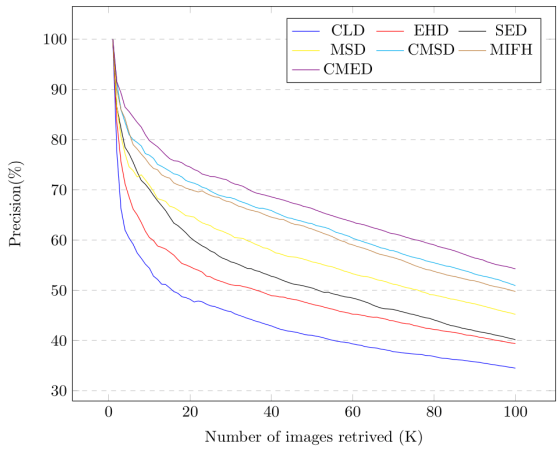


Figure 15. Precision of proposed CMED compared to other state-of-the-art descriptors on Corel-1k

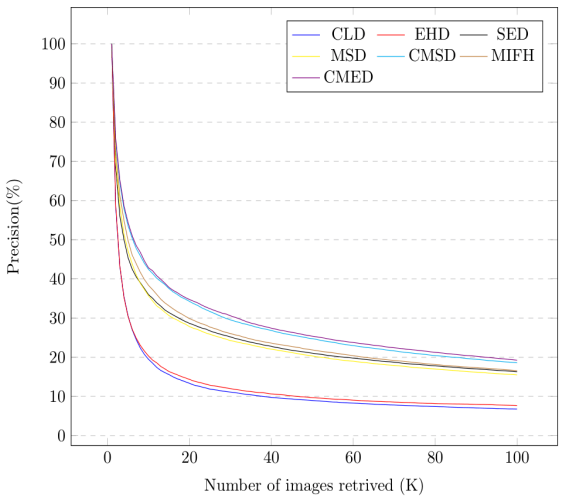


Figure 16. Precision of proposed CMED compared to other state-of-the-art descriptors on Corel-5k

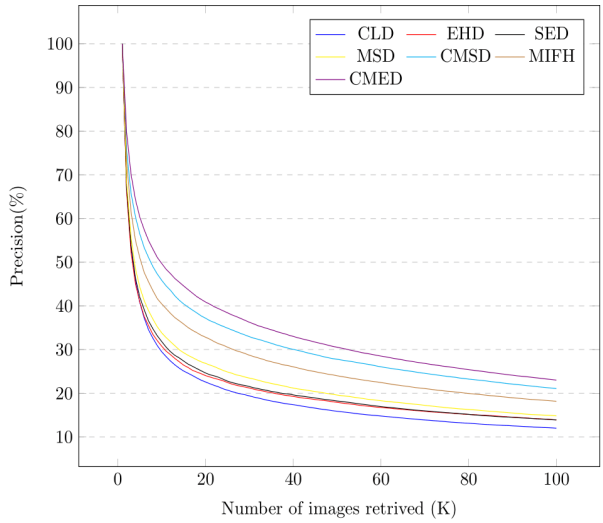


Figure 17. Precision of proposed CMED compared to other state-of-the-art descriptors on Corel-CBIR

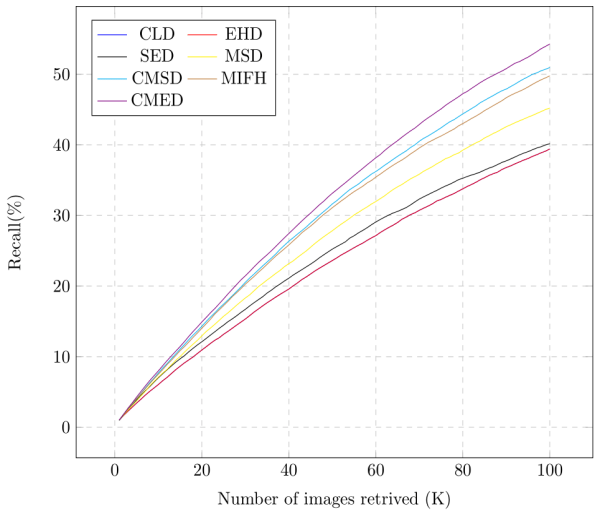


Figure 18. Recall of proposed CMED compared to other state-of-the-art descriptors on Corel-1k

Furthermore, the graph illustrates a growing gap between the proposed descriptor and others, suggesting that it retrieves more images as the value of K increases. Specifically, at K=12, our descriptor achieves a Recall of 9.43% on the Corel-1k dataset, indicating that it retrieves nine to teen images out of twelve relevant ones. This performance surpasses that of other descriptors on the Corel-1k image set.

Similarly, Figures 19 and 20 depict the results for Corel-5k and Corel-CBIR datasets, respectively, showing comparable trends to those observed in Corel-1k. Despite Corel-5k results being closely aligned with CMSD, there is a noticeable divergence favoring our descriptor in the Corel-CBIR dataset, particularly evident with each incremental increase in K.

Figures 21-23 illustrate the graphs depicting the results obtained with the MAP metric. MAP considers both precision and recall, indicating how many relevant images are retrieved relative to the total number of images in each category. The graphs demonstrate that our descriptor outperforms other descriptors consistently. Additionally, similar to the recall metric, our descriptor maintains a considerable lead over others as K increases. Notably, at K=12, the results of our descriptor consistently remain superior to those of other descriptors.

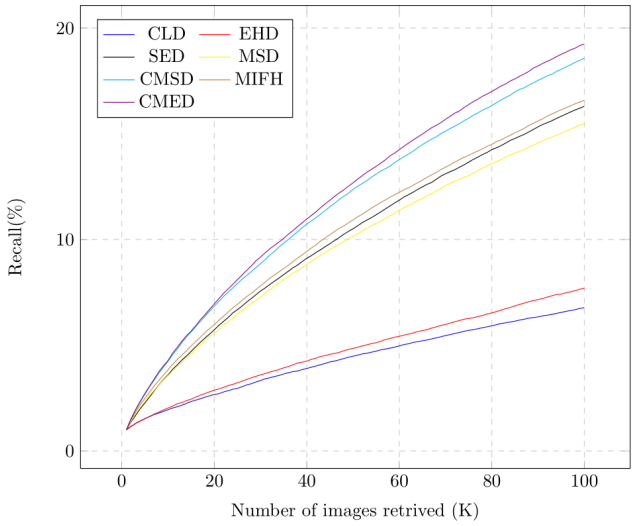


Figure 19. Recall of proposed CMED compared to other state-of-the-art descriptors on Corel-5k

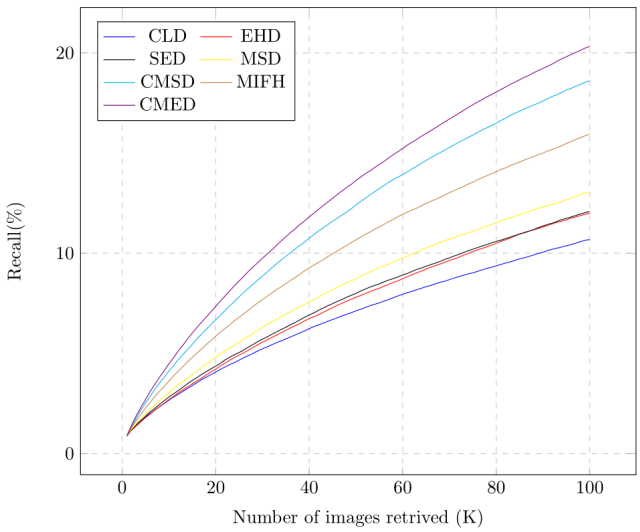


Figure 20. Recall of proposed CMED compared to other state-of-the-art descriptors on Corel-CBIR

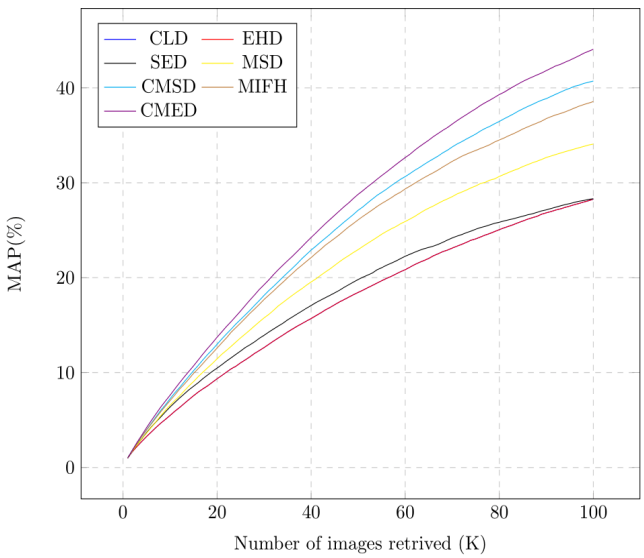


Figure 21. MAP of proposed CMED compared to other state-of-the-art descriptors on Corel-1k

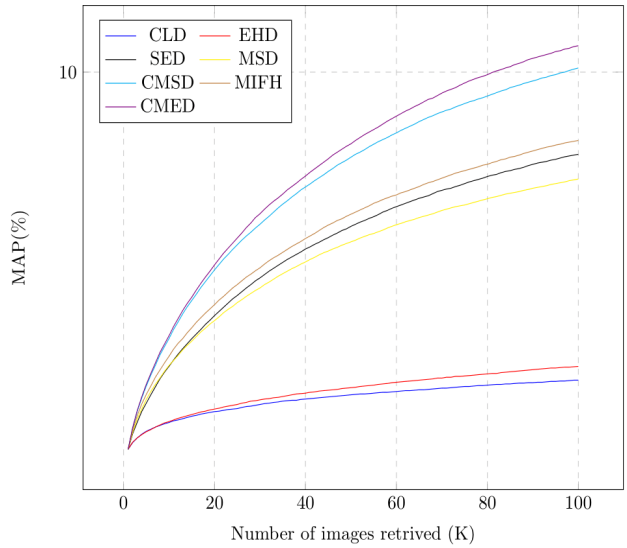


Figure 22. MAP of proposed CMED compared to other state-of-the-art descriptors on Corel-5k

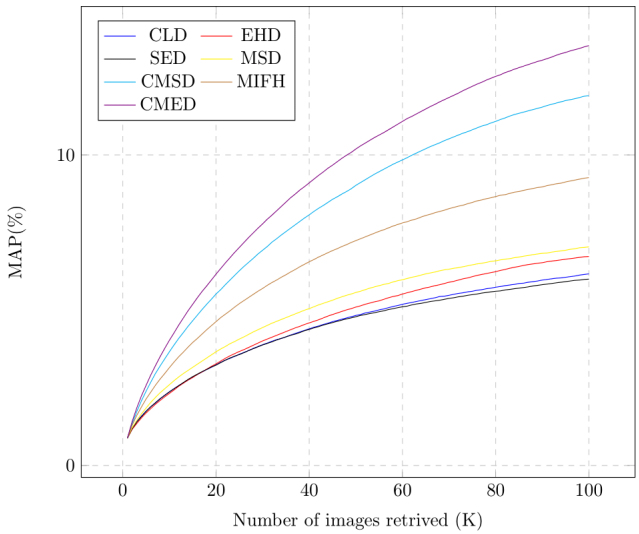


Figure 23. MAP of proposed CMED compared to other state-of-the-art descriptors on Corel-CBIR

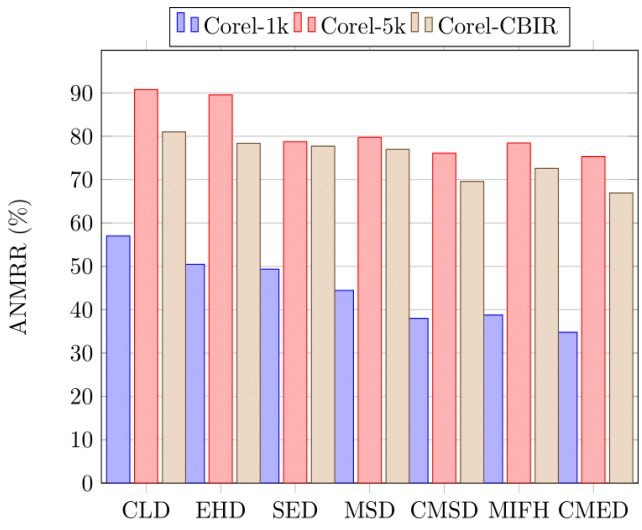


Figure 24. ANMRR metric of proposed CMED compared to classical and state-of-the-art methods on Corel-1k, Corel-5k and Corel-CBIR

Table 2. Category comparison of proposed CMED with other descriptors on Corel-1k

Category	CLD	EHD	SED	MSD	CMSD	MIFH	CMED
Africa	75.72%	65.90%	38.37%	56.85%	46.02%	40.03%	37.49%
Beach	86.08%	73.82%	57.19%	44.31%	38.88%	48.25%	35.14%
Buildings	51.43%	66.97%	74.01%	72.18%	64.07%	62.84%	56.38%
Bus	69.53%	73.96%	52.36%	47.28%	39.10%	39.47%	43.26%
Dinosaur	82.45%	21.96%	41.28%	31.81%	33.49%	27.39%	28.97%
Elephant	0.55%	1.67%	23.61%	4.69%	0.77%	8.01%	0.02%
Flower	59.83%	69.61%	61.85%	54.00%	43.07%	46.45%	40.76%
Horse	47.99%	28.60%	28.10%	35.17%	29.55%	33.64%	27.72%
Mountain	33.15%	32.30%	54.00%	34.99%	27.52%	28.79%	23.56%
Food	63.60%	69.90%	62.67%	63.10%	57.67%	52.96%	54.55%
Average	57.03%	50.47%	49.34%	44.44%	38.01%	38.78%	34.79%

Table 3. Tolerance transformation comparison of proposed CMED with other descriptors on Corel-1k with ANMRR

Transformation	CLD	EHD	SED	MSD	CMSD	MIFH	CMED
Rotation 90°	57.03%	50.63%	49.91%	52.76%	45.60%	53.05%	40.15%
Rotation 180°	57.04%	50.61%	48.35%	45.60%	38.94%	40.41%	35.91%
Scaling 50%	61.52%	56.85%	49.61%	43.75%	38.03%	38.86%	34.81%
Scaling 80%	62.87%	65.14%	52.46%	44.14%	38.40%	39.20%	35.39%
Reflection	57.40%	50.30%	49.82%	44.51%	38.01%	38.86%	34.80%
Average	59.17%	54.70%	50.03%	46.15%	39.80%	42.07%	36.21%

Figure 24 presents the results obtained with the ANMRR metric across the three image datasets. Unlike other metrics, ANMRR considers the position of each retrieved image within the query class, making it unaffected by the number of retrieved images. This metric provides a more comprehensive evaluation as it accounts for both the number and position of retrieved images by each descriptor.

The graph indicates that our proposed CMED descriptor demonstrates superior retrieval performance, achieving an ANMRR of 34.79% on the Corel-1k dataset. This represents a notable improvement of over 15% compared to descriptors from the MPEG-7 standard and over 3% compared to state-of-the-art descriptors. Similarly, on the Corel-5k and Corel-CBIR image datasets, our descriptor achieves ANMRR values of 75.33% and 66.93%, respectively, outperforming other descriptors.

These results suggest that our descriptor not only retrieves a higher number of images on average but also retrieves them in more favorable positions, as indicated by the ANMRR metric.

Table 2 presents the ANMRR results for each category of the Corel-1k dataset with K=12, allowing us to identify categories where the best performance is observed and whether any biases or challenges exist for the descriptor. The table indicates that CLD, EHD, MIFH, CMSD, and our CMED descriptor achieve the best evaluations.

Upon closer inspection, our CMED descriptor consistently outperforms standard descriptors across most categories. Specifically, it demonstrates superior performance in the Africa, Beach, Elephant, Flower, Horse, and Mountain categories. However, our descriptor exhibits lower performance in the building category compared to standard MPEG-7 descriptors.

Overall, while our descriptor may not achieve the highest evaluation in every category, it maintains strong performance on average. Despite not being the top-performing descriptor in each category, it consistently ranks among the best. This suggests that the CMED descriptor offers a more stable performance across diverse image categories compared to other descriptors.

Table 3 presents the results for each descriptor on the Corel-1k dataset under five different transformations: 90° rotation, 180° rotation, 50% rescaling, 80% rescaling, and reflection transformation. The table includes precision with K=12 and the evaluation obtained with the ANMRR metric.

The results demonstrate that the proposed CMED descriptor consistently retrieves the original image in the first position for all queries, indicating robustness to transformations. This suggests that the descriptor performs well even when images are subjected to various transformations. Additionally, the CMED descriptor outperforms other descriptors in terms of both precision and ANMRR under all transformations.

Overall, based on the results obtained with the Corel-1k image set, it can be concluded that the CMED descriptor exhibits higher tolerance to rotation, scaling, and mirror transformations compared to other descriptors.

5. CONCLUSIONS

In addition to the improve in retrieval performance, our CMED descriptor offers several strengths. Firstly, it demonstrates stability across various categories, consistently outperforming other descriptors on average. Moreover, it exhibits robustness to common image transformations such as rotation, scaling, and reflection, ensuring reliable performance in real-world scenarios where images may undergo such alterations.

Furthermore, CMED leverages both structures and microstructures to enhance retrieval of natural images in semantic classes. This comprehensive approach not only improves representation but also increases tolerance to transformations, features that are widely used for content-based image retrieval tasks.

However, it's essential to acknowledge some weaknesses of the CMED descriptor. One limitation is its computational cost, especially when using a depth of four. This may pose challenges in scenarios where computational resources are limited or where real-time processing is required. Additionally, while CMED performs well across most categories, it exhibits lower performance in certain categories, such as buildings and dinosaurs, compared to descriptors from the MPEG-7

Standard. This suggests that further refinement may be needed to optimize performance across all images closely related to low-level features. It's worth considering that these categories had very similar images in terms of color or texture, which could be the reason for the better performance of the low-level descriptors.

Furthermore, the performance of the descriptor is highly dependent on the low-level features used to construct the microstructure and structure maps. Exploring alternative feature representations or combinations may lead to improvements in retrieval performance. For example, incorporating deep learning features extracted from pre-trained convolutional neural networks (CNNs) could enhance the descriptor's ability to capture higher-level semantic information from images.

In real-world scenarios, CMED could benefit applications requiring robust and accurate image retrieval, such as image search engines, content-based image recommendation systems, and multimedia databases. Its ability to handle various image transformations makes it particularly suitable for tasks where images may be captured under different conditions or from different perspectives. Additionally, its effectiveness in retrieving images based on semantic content makes it valuable for applications requiring precise and contextually relevant image retrieval, such as medical image analysis, satellite image processing, and surveillance systems.

ACKNOWLEDGMENT

The authors would like to thank the Tecnológico Nacional de México (TecNM) for their support and the authors who provided their descriptor codes and information. K. Salvador Aguilar-Domínguez would like to express gratitude to the Consejo Nacional de Ciencia y Tecnología (CONACYT) for their support through the doctoral scholarship program.

REFERENCES

[1] Dubey, S.R. (2019). Face retrieval using frequency decoded local descriptor. Multimedia Tools and Applications, 78(12): 16411-16431. https://doi.org/10.1007/s11042-018-7028-8

[2] Archana, N., Menaka, R., Regina, S.B., Prabha, P.L. (2022). Analogous healthcare product identification in online shopping. Healthcare 4.0: Health informatics and precision data management, CRC Press. Chapman and Hall/CRC, pp. 77-96.

[3] Jimenez, A., Alvarez, J.M., Giro-I-Nieto, X. (2017). Class-Weighted convolutional features for visual instance search. British Machine Vision Conference 2017, BMVC. https://doi.org/10.48550/arxiv.1707.02581

[4] Deldjoo, Y., Nazary, F., Ramisa, A., Mcauley, J., Pellegrini, G., Bellogin, A., Noia, T.D. (2023). A review of modern fashion recommender systems. ACM Computing Surveys, 56(4): 1-37. https://doi.org/10.1145/3624733

[5] Shamna, P., Govindan, V.K., Nazeer, K.A. (2022). Content-based medical image retrieval by spatial matching of visual words. Journal of King Saud University-Computer and Information Sciences, 34(2): 58-71. https://doi.org/10.1016/J.JKSUCI.2018.10.002.

[6] Li, X., Yang, J., Ma, J. (2021). Recent developments of content-based image retrieval (CBIR). Neurocomputing, 452: 675-689. https://doi.org/10.1016/J.NEUCOM.2020.07.139

[7] Chen, W., Liu, Y., Wang, W., Bakker, E.M., Georgiou, T., Fieguth, P., Liu, L., Lew, M.S. (2022). Deep learning for instance retrieval: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2022.3218591

[8] Dubey, S.R. (2022). A decade survey of content based image retrieval using deep learning. IEEE Transactions on Circuits and Systems for Video Technology, 32(5): 2687-2704. https://doi.org/10.1109/TCSVT.2021.3080920

[9] Veshkin, A.S., Khvostikov, A.V. (2022). Multiscale content-Based image retrieval for whole-Slide histological images. Computational Mathematics and Modeling, 33(2): 244-254. https://doi.org/10.1007/s10598-023-09569-2

[10] Zhang, X., Bai, C., Kpalma, K. (2023). OMCBIR: Offline mobile content-based image retrieval with lightweight CNN optimization. Displays, 76: 102355. https://doi.org/10.1016/J.DISPLA.2022.102355

[11] Vharkate, M.N., Musande, V.B. (2022). Fusion based feature extraction and optimal feature selection in remote sensing image retrieval. Multimedia Tools and Applications, 81(22): 31787-31814. https://doi.org/10.1007/s11042-022-11997-y

[12] Kishore, D., Rao, C.S. (2020). A multi-class SVM based content based image retrieval system using hybrid optimization techniques. Traitement Du Signal, 37(2): 217-226. https://doi.org/10.18280/ts.370207

[13] Zhang, Y., Wei, Z. (2022). An image classification and retrieval algorithm for product display in E-Commerce Transactions. Traitement Du Signal, 39(5): 1865-1871. https://doi.org/10.18280/ts.390547

[14] Yuan, S. (2022). Classification and retrieval of commodity images oriented to internet marketing. Traitement Du Signal, 39(6): 2087-2093. https://doi.org/10.18280/ts.390621

[15] Tyagi, V. (2017). Content- Based image retrieval (Ideas, Influences, and Current Trends). Springer, Singapore. https://doi.org/10.1007/978-981-10-6759-4

[16] Smeulders, A.W., Worring, M., Santini, S., Gupta, A., Jain, R. (2000). Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(12): 1349-1380. https://doi.org/10.1109/34.895972

[17] Chen, Y., Wang, J.Z., Krovetz, R. (2003). An unsupervised learning approach to content-based image retrieval. In Seventh International Symposium on Signal Processing and Its Applications, Proceedings, Paris, France, 1: 197-200. https://doi.org/10.1109/ISSPA.2003.1224674

[18] Chen, T. (2005). From low-level features to high-level semantics: Are we bridging the gap? In Seventh IEEE International Symposium on Multimedia (ISM'05). IEEE Computer Society, pp. 179-179. https://doi.org/10.1109/ISM.2005.62

[19] Julesz, B. (1981). Textons, the elements of texture perception, and their interactions. Nature, 290(5802): 91-97. https://doi.org/10.1038/290091a0

[20] Treisman, A.M., Gelade, G. (1980). A feature-integration theory of attention. Cognitive Psychology,

12(1): 97-136. https://doi.org/10.1016/0010-0285(80)90005-5

[21] Niu, D., Zhao, X., Lin, X., Zhang, C. (2020). A novel image retrieval method based on multi-features fusion. Signal Processing: Image Communication, 87: 115911. https://doi.org/10.1016/j.image.2020.115911

[22] Liu, G.H., Li, Z.Y., Zhang, L., Xu, Y. (2011). Image retrieval based on micro-structure descriptor. Pattern Recognition, 44(9): 2123-2133. https://doi.org/10.1016/j.patcog.2011.02.003

[23] Gonzalez, R.C., Woods, R.E. (2007). Digital image processing (3rd Edition). Prentice-Hall, Inc. Upper Saddle River, NJ, USA ©2006,: 976.

[24] Manjunath, B.S., Salembier, P., Sikora, T. (Eds.). (2002). Introduction to MPEG-7: Multimedia Content Description Interface. John Wiley & Sons.

[25] Dawood, H., Alkinani, M.H., Raza, A., Dawood, H., Mehboob, R., Shabbir, S. (2019). Correlated microstructure descriptor for image retrieval. IEEE Access, 7: 55206-55228. https://doi.org/10.1109/ACCESS.2019.2911954

[26] Sankara Narayanan, S., Vinod, D., Athisayamani, S., Robert Singh, A. (2022). Combination of local feature extraction for image retrieval. In Proceedings of Third International Conference on Sustainable Computing: SUSCOM 2021. Singapore: Springer Nature Singapore, pp. 319-328. https://doi.org/10.1007/978-981-16-4538-9_32

[27] Kottawar, V., Deshpande, N., Jatti, V.S., Bhoite, S. (2022). Review of feature extraction techniques in content based image retrieval. Journal of Pharmaceutical Negative Results, 13: 7508-7514. https://doi.org/10.47750/PNR.2022.13.S07.905

[28] Zulkurnain, N.F., Azhar, M.A., Mallik, M.A. (2022). Content-Based image retrieval system using fuzzy colour and local binary pattern with Apache Lucene. In Proceedings of Second International Conference on Advances in Computer Engineering and Communication Systems: ICACECS 2021. Singapore: Springer Nature Singapore, pp. 13-20. https://doi.org/10.1007/978-981-16-7389-4_2

[29] Wang, X., Wang, Z. (2013). A novel method for image retrieval based on structure elements' descriptor. Journal of Visual Communication and Image Representation, 24(1): 63-74. https://doi.org/10.1016/j.jvcir.2012.10.003

[30] Chu, K., Liu, G.H. (2020). Image retrieval based on a multi-integration features model. Mathematical Problems in Engineering, 2020(1): 1461459. https://doi.org/10.1155/2020/1461459

[31] Liu, Z.G., Yang, Y., Ji, X.H. (2016). Flame detection algorithm based on a saliency detection technique and the uniform local binary pattern in the YCbCr color space. Signal, Image and Video Processing, 10: 277-284. https://doi.org/10.1007/s11760-014-0738-0

[32] Wang, J.Z., Li, J., Wiederhold, G. (2001). SIMPLIcity: Semantics-sensitive integrated matching for picture libraries. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(9): 947-963. https://doi.org/10.1109/34.955109

[33] Li, J., Wang, J.Z. (2006). Real-time computerized annotation of pictures. In Proceedings of the 14th ACM International Conference on Multimedia, pp. 911-920. https://doi.org/10.1145/1180639.1180841

[34] Bian, W., Tao, D. (2009). Biased discriminant Euclidean embedding for content-based image retrieval. IEEE Transactions on Image Processing, 19(2): 545-554. https://doi.org/10.1109/TIP.2009.2035223

[35] Lowe, D.G. (1999). Object recognition from local scale-invariant features. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, 2: 1150-1157. https://doi.org/10.1109/ICCV.1999.790410