2018-03-01

CoMo: a scale and rotation invariant compact composite moment-based descriptor for image retrieval

Vassou, S. A.

Springer International Publishing

Multimed Tools Appl
https://doi.org/10.1007/s11042-018-5854-3

CoMo: a scale and rotation invariant compact composite moment-based descriptor for image retrieval

S. A. Vassou[1] · N. Anagnostopoulos[2] · K. Christodoulou[3] · A. Amanatiadis[4] · S. A. Chatzichristofis[3]

Received: 15 October 2017 / Revised: 20 December 2017 / Accepted: 1 March 2018

Abstract Low level features play a significant role in image retrieval. Image moments can effectively represent global information of image content while being invariant under translation, rotation, and scaling. This paper presents CoMo: a moment based composite and compact low-level descriptor that can be used effectively for image retrieval and robot vision tasks. The proposed descriptor is evaluated by employing the Bag-of-Visual-Words representation over various well-known benchmarking image databases. The findings from the experimental evaluation provide strong evidence of high and competitive retrieval performance against various state-of-the-art local descriptors.

Keywords Content based image retrieval · Low level features · Compact composite descriptors

1 Introduction

Content Based Image Retrieval (CBIR) is a long-standing problem especially in the area of computer and robot vision. In the past decade, visual search has attracted a great deal of attention, even though it has been studied since the early 1990s [33]. Despite the fact that

✉ S. A. Chatzichristofis
s.chatzichristofis@nup.ac.cy

[1] Cyprus University of Technology, Limassol, Cyprus

[2] Microsoft, Prague, Czech Republic

[3] Department of Information Sciences, Neapolis University Pafos, Pafos, Cyprus

[4] Democritus University of Thrace, Xanthi, Greece

for most evaluation benchmarks, the state of the art is currently held by conventional low-level based approaches [27, 34, 55, 57, 58], recent research on CBIR is heavily focused on Deep Learning techniques with Convolutional Neural Networks (CNN) [13]. However, such techniques assume the existence of a significant amount data for training computations, and for generating accurate results. Before a learning machine can perform classification and recognition, it needs to be trained first, and training samples need to be accurately labeled. The labeling process can be both time consuming and error-prone [45]. Notwithstanding the significant improvements introduced by such techniques, in some domains large scale repositories of images are not always accessible. Examples of such domains include: commercial image search engines; where images are not transferred unless payment is done, classification of explicit material; where transmission of the data defeats the purpose of a filter, or large scale investigations of criminal material; for instance, cases involving child abuse. These scenarios indicate examples where training data are not always sufficient or easily accessible, especially in controlled environments; where access to sensitive image data is not permited. In such domains, information from extracted features or image signatures should not allow for images to be reconstucted. Such features, which are typically hand-crafted, are employed for classification and retrieval on these datasets. Deselaers et al. [43]. The goal of any feature-based CBIR method is to vectorize an image so that its unique characteristics and content are captured. Thus, and to make a first rough categorization, existing CBIR approaches can be classified according to the image features considered or deemed meaningful. It has been a long subject of debate the concenrn regarding the most effective way to treat an image for indexing and retrieval [9, 54].

Several approaches for tratitional content based image retrieval have been proposed in the literature, ranging from *global* to *local* features. *Global features*, such as, color, texture, and shape are calculated to form a feature vector representation of an entire image. Such an approach has low computational cost due to the single-feature vector and is very effective for certain retrieval tasks. However, there are cases where global feature vectors fail to discriminate the visual content of an image, providing a rather generalized outline of visual information. The main advantages of extracting global features is the low cost of the single-feature space computations, and the fact that indexing one image is independent of the type or the total number of images in the collection. However, annotating an image solely by a global feature vector often leads to a rather generalized outline of its visual information. Global vectors fail to discriminate between the constituent parts of an image since their scope is built on the assumption that all parts of the image are equally important for the final representation. Therefore, retrieved results to a given query manage to capture visual similarity but lack in extracting semantic similarity. Spatial correlation of color [15, 25, 32], edge or shape occurrences [12, 23] have been explored in an effort to somewhat provide a more focused description.

With retrieval scenarios becoming more demanding, techniques that take into consideration *Local features* were introduced. Such techniques, use local features to represent images obtained from salient regions of the image. Any pixel can define a local feature without, necessarily, introducing any significant support to the retrieval task or improving the descriptor's discrimination ability. For such reasons, salient regions, *a.k.a* Points-Of-Interests (POI), that contain rich local information are used. Examples of popular POI detectors are *corner detectors*, such as, Shi-Tomasi, Harris, and Fast [14, 48, 53], and *blob detectors* e.g., SIFT [29], SURF [2]. Techniques that are based on salient regions face significant computational costs due to the high-dimensional local feature space and therefore

a complexity is introduced. In different domain applications, such as, Visual Simultaneous Localization and Mapping, panorama construction, object tracking etc. these extracted POIs are used directly to recognize one-to-one matches between depictions. In CBIR, however, direct usage of salient regions is impractical even with today's available computational resources.

To overcome the high dimensionality issue, research in the filed of CBMI introduced the Bag-of-Visual-Words (BOVW) model. This model was inspired by the Bag-of-Words model used in text retrieval, were each document is represented by a set of distinct keywords. The difference is that the BOVW model considers Visual Words (VW) i.e., the result of a clustering algorithm over all detected local features in an image collection. The total number of VW, namely the centroid of each cluster, forms a resulting VW dictionary (known as the *codebook*). Having derived the VW dictionary each image is then represented as a histogram of the VW, according to the presence or count of each VW in the dictionary. This approach, solves the high dimensionality challenge, however, from the other hand it introduces the necessity to predict an appropriate codebook size, and a preferred weighting strategy [7].

This type of quantization, however, comes with a loss of the discriminative ability of the features. Therefore, several alternatives have been proposed to solve this issue. The Fisher vector [42] makes use of the Gaussian mixture model by calculating the probability of a feature falling into the Gaussian mixture in order to train the codebook and quantize the features. Alternatively, the soft quantization and soft assignment techniques proposed in [44] and [19], respectively, reduce the quantization error of the original BOVW model, with a cost in terms of memory overload, and higher searching time. Different methods like Hamming embedding [18] provide additional information to filter false positives by generating binary signatures coupling visual words.

Motivation Among the most commonly used features for image retrieval are the *Image Moments*, which can assist in identifying certain key characteristics of images. Their significance, in the fields of image analysis and object representation, is based on the fact that image moments represent global information of image content while being invariant under translation, rotation, and scaling. In the recent years, several methods have been proposed to utilize the advantages of image moment invariants to shape global features [39]. On the other hand, the opportunity of shaping local moment-based descriptors has not been investigated in depth [22]. A Moments' Based local feature would be suitable not only for image retrieval tasks but also for other rigorous applications, such as Simultaneous Localization And Mapping or semantic mapping. This paper describes a novel Moment-Based local and global descriptor, called CoMo, originally introduced in [58]. More specifically, the proposed feature is defined by combining the color information from the color unit of the Color and Edge Directivity Descriptor (CEDD) [5] with the Seven Invariant Moments (SIM), presented by Hu, as the new texture unit. This solution provides a better description that at the same time improves image retrieval due to the independence on rotation and scale variations.

Summary of contributions This paper complements previous work in [58] by contributing the following:

i. a novel low-level feature that utilizes moment invariants along with the CEDD color unit, which is both light-weight and efficient to be used for image retrieval and robot vision tasks;

ii. an experimental evaluation by observing the behavior of the proposed feature over four databases against state-of-the-art local and global features from the literature.

The remainder of this paper is structured as follows. Section 2 presents more details regarding the seven invariant moments as presented by Hu. The process of shaping the descriptor by associating color information with the Hu Moments extraction process is discussed in Section 3 whereas Section 4 describes how CoMo can be extracted either as a local or a global descriptor. Section 5 presents a thorough experimental evaluation of the descriptor complemented by a discussion of results. Finally, Section 6 concludes.

2 HU moments

Moment invariants originated mainly from a well established area of mathematics called algebraic invariants. By using the geometrical, central and normalized image moments, Hu constructed seven moments that are invariant to any translation, scaling and rotation transformation of the image being processed [38]. Hu's approach was based on the work of the 19th century mathematicians Boole, Cayley and Sylvester [10].

For a given image with pixel intensities $f(x, y)$, geometrical image moments M_{pq} are calculated by:

$$M_{pq} = \sum_{X} \sum_{Y} x^p y^q f(x, y) \tag{1}$$

The centroid coordinates are defined as:

$$\overline{X} = \frac{M_{10}}{M_{00}} \quad \text{and} \quad \overline{Y} = \frac{M_{01}}{M_{00}} \tag{2}$$

The central moments μ_{pq} are constructed from geometrical moments:

$$\mu_{pq} = \sum_{x} \sum_{y} (x - \overline{X})^p (y - \overline{Y})^q f(x, y) \tag{3}$$

The nth central moment is translation-invariant, i.e. for any random variable f and any constant e:

$$\mu_n(f + e) = \mu_n(f) \tag{4}$$

Furthermore, invariants η_{pq} with respect to both translation and scale can be constructed from central moments:

$$\eta_{pq} = \frac{\mu_{pq}}{(\mu_{20} + \mu_{02})^{\gamma}} \tag{5}$$

where $\gamma = (p + q + 2)/4$.

The seven invariant moments (SIM) are given as follows:

$$\begin{aligned}
\phi_1 &= \eta_{20} + \eta_{02} \\
\phi_2 &= (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \\
\phi_3 &= (\eta_{30} - 3\eta_{12})^2 + (\eta_{03} - 3\eta_{21})^2 \\
\phi_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{03} + \eta_{21})^2 \\
\phi_5 &= (3\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 \\
&\quad -3(\eta_{21} + \eta_{03})^2] + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03}) \\
&\quad \times[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\
\phi_6 &= (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\
&\quad +4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\
\phi_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 \\
&\quad -3(\eta_{21} + \eta_{03})^2] + (3\eta_{12} - \eta_{30})(\eta_{21} + \eta_{03} \\
&\quad \times[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{30})^2]
\end{aligned} \tag{6}$$

Moment based invariants, in various forms, have been widely used over the years as features for recognition in many areas of image analysis.

3 Shaping the descriptor

The MPEG-7 like global descriptor CEDD utilizes both color and texture information to describe the content of an image and has been widely used in recent literature due to its successful trade-off between effectiveness and efficiency. CEDD is computationally lightweight relative to other feature extraction mechanisms, but has comparable accuracy. Even though CEDD was initially designed so as to globally describe the visual information of an input image, its scalability on characterizing single feature points has already been proven. As shown in [16], the localized equivalent of CEDD outperforms the matching accuracy of many other descriptors, like SIFT [29], SURF [2], ORB [49] and BRISK [26]. The effectiveness of CEDD relies on its ability to combine color and texture information. CEDD is a scale-invariant descriptor and can tolerate small local rotations, but it is not rotation invariant and does not allow for large global rotations and translations.

Similarly to the structure of CEDD, the proposed descriptor is constructed by integrating color information from the *Color Unit* thought a two-staged fuzzy-linking scheme along with texture information associated with the Hu Moments extraction process from the *Texture Unit*. More specifically, texture information is captured by introducing 6 regions, one for each type of texture (see Fig. 8). The number of regions comes as a compromise between the low storage requirements of the application using the proposed descriptor, and the need for more effective retrieval accuracy. Each Texture unit region contains 24 individual regions defined by the Color unit. Overall, the proposed descriptor contains $6 \times 24 = 144$ regions. Figure 1 illustrates the form of the CoMo descriptor. On the completion of the process, CoMo's histogram is normalized within the interval [0,1] and then quantized into three bits/bins.

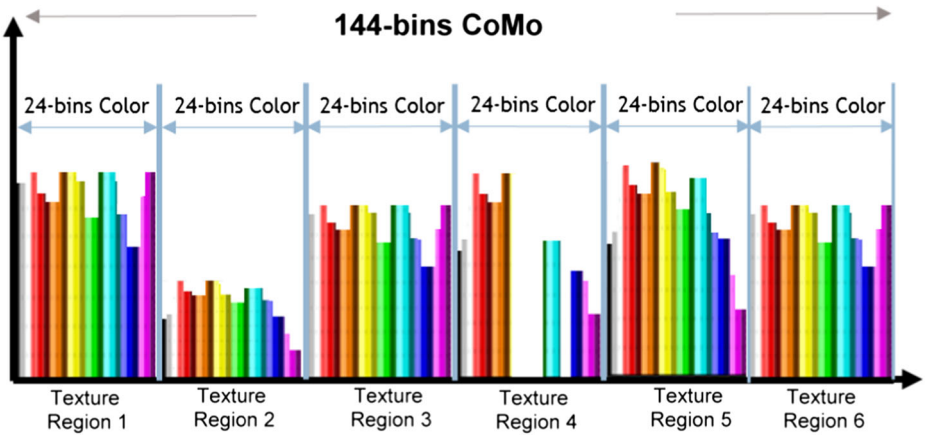

Fig. 1 The CoMo descriptor

3.1 Texture information

To incorporate the texture information, CoMo proposes a novel 6-bin histogram, taking into account the aforementioned set of SIM ($\phi_1, \phi_2, \ldots, \phi_7$). In order to shape the 6 predefined texture regions, we employed 100000 randomly selected images from FlickR. Next, random patches of various sizes from all images were extracted. After calculating the Hu moments from these patches, using a k-means classifier, 6 classes are shaped (see Algorithm 1). It is worth noting that a 7-dimensional vector describes the center of each class. The set of the 6×7 resulted values are hereafter denoted by C.

Algorithm 1 Calculate the First Set of Chromosomes.

```
for i ≤ Number of Random Images do
    Generate Random Number of Patches
    for j ≤ Number of Patches do
        U+= Calculate the 7 Hu Moments from j
    end for
end for
//Array U contains (i × j) 7-dimensional vectors
Using k-Means Classify U into 6 classes (array C)
```

In the sequel, a simple genetic algorithm determines off-line the 6 predefined texture regions to be utilized by the proposed descriptor. The genetic algorithm begins with an initial population of 20 chromosomes, where each chromosome consists of (7 `Hu moments` × 6 `classes`) values. The first chromosome is generated from the set of the 6 classes (array C) as resulted by the k-means classifier, 9 more chromosomes are generated by slightly mutating the first one, and the remaining 10 chromosomes are generated randomly. The chromosomes are in a non-binary form and their initial generation procedure is illustrated in Fig. 2.

For experimental purposes (refer to Section 5 for more details) the `UCID` database is used along with a setup of a simple image retrieval framework. A 6-dimensional vector is

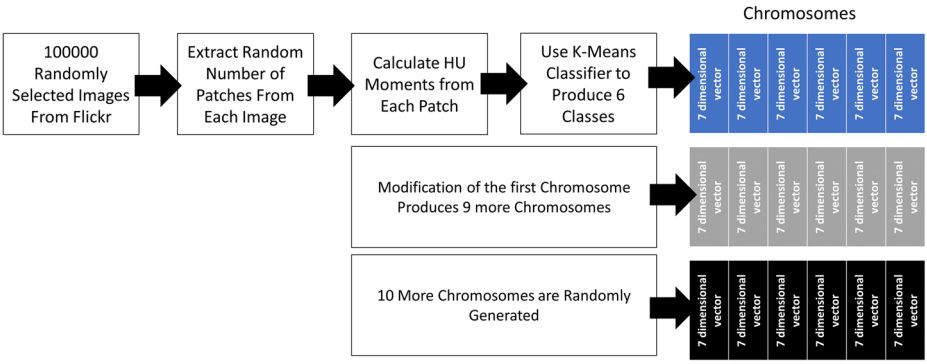

Fig. 2 Shaping the population of the 20 chromosomes

calculated for each image, by considering the values of its Hu moments. This procedure aims to map the texture of a given image into a compact histogram. In order to shape a texture histogram for an input image, the input image is segmented into 64 non-overlapped image blocks. For each image block, the Hu moments are extracted and their distance with the 6 given centers is calculated. Based on the distance with each one of the given centers, the texture histogram is generated. Next, since both, query images and ground truths are known, an image retrieval process is executed and the Mean Average Precision (MAP) is calculated (kindly refer to Section 5 for more details about MAP).

Algorithm 2 Tune the Texture Regions.

```
Ch[20, 6] is a set of (20 × 6) 7-dimensional vectors
for t ≤ 20 do
    for i ≤ Number of Images in UCID do
        Segment the 64 image blocks
        for j ≤ 64 do
            U′= Calculate the 7 Hu Moments from block j
            MinVal=double.MIN
            for k ≤ 6 do
                if MinVal ≤ |Ch[t, k] − U′| then
                    MinVal = |Ch[t, k] − U′|
                    Min= k
                end if
            end for
            Histo[i,Min]++;
        end for
    end for
    Perform Retrieval and Calculate the MAP[t]
end for
Sort Chromosomes based on MAP
```

The procedure is repeated for the set of 20 chromosomes. As a next step the chromosomes are sorted based on the resulted MAP, and the best 10 are kept for the formation of the next generation. A *crossover* function is applied to the next 3 best chromosomes while the next best 3 chromosomes are *mutated* by increasing or decreasing only one contributor value

of the chromosome. Finally, 4 additional chromosomes are randomly inserted. The procedure is repeated until the fitness function is minimized and there is no further improvement. The best chromosome is then used to form the 6 texture areas that the proposed descriptor uses. The entire process is also discussed in Algorithm 2.

3.2 Color information

Moment-based image representation methods extract color information at each color channel independently [28]. In general there exist dependencies caused by linear transform in the color space. In contrary, CoMo shares the same color information extraction unit with CEDD through a two-staged fuzzy-linking scheme.

An effective way to extract color information as feature from an image is by linking the color space channels. Linking is defined as the combination of more than one histograms to a single one. One example is the Scalable Color Descriptor (SCD) demonstrated in MPEG-7 [32]. In the SCD implementation, the HSV color histograms are uniformly quantized into a a single histogram of 256 bins. This new histogram is defined by sixteen levels in H (Hue), four levels in S (Saturation) and four levels in V (Value). The values of H, S and V are computed for every pixel, and then are linearly quantized in the ranges [0, 15], [0, 3] and [0, 3] respectively. At the next step, the function: $H_{Quantized} + 16 \times S_{Quantized} + 64 \times V_{Quantized}$ is used to form a modified version of the histogram. For the proposed image descriptor, color information is incorporated by utilizing the linking approach as proposed in [5]. According to this method, a fuzzy-linking mechanisms fuses the HSV color histograms of a pixel or a region into a 24-bins histogram. Figure 3 illustrates a graphical abstract of the Color Unit used by CoMo.

Fuzzy color histogram At the *first stage*, the employed fuzzy system generates a fuzzy-linking histogram based on the three HSV channels of a given pixel or a region of pixels (hereafter referred as *Tiles*) as inputs, to form a 10-bin histogram as output. Each bin represents a preset color: (0) White, (1) Gray, (2) Black, (3) Red, (4) Orange, (5) Yellow, (6) Green, (7) Cyan, (8) Blue, and (9) Magenta.

Channel H is divided into eight fuzzy areas (Fig. 3 – 1st stage Fuzzy System) defined as follows: (0) Red to Orange, (1) Orange, (2) Yellow, (3) Green, (4) Cyan, (5) Blue, (6) Magenta and (7) Magenta to Red. For more details concerning the boundaries and the shaping of the membership functions we refer the reader to [5].

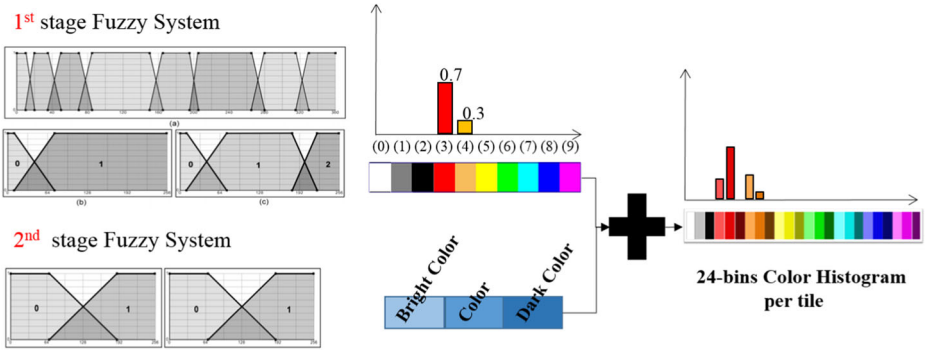

Fig. 3 Graphical abstract of the color unit

Channel S is divided into two fuzzy areas (0, 1) and *Channel V*, is divided into three areas (0, 1, 2). S values that fall in the first area link to a non-color output, black, grey or white, depending on the activation happening for channel V (0, 1, and 2 respectively). S values that fall in the second area link to varying color outputs depending on the activation of H, as long as V is not in its first area. If V falls in the first fuzzy area the output in this case is black, independently from the other input values. For more details regarding the fuzzy rules used to produce the crisp outputs, the reader is refered to [6].

At the *second stage* of the fuzzy-linking system, the method produces a 24-bin histogram as output. Each bin represents a preset color as follows: (0) White, (1) Grey, (2) Black, (3) Light Red, (4) Red, (5) Dark Red, (6) Light Orange, (7) Orange, (8) Dark Orange, (9) Light Yellow, (10) Yellow, (11) Dark Yellow, (12) Light Green, (13) Green, (14) Dark Green, (15) Light Cyan, (16) Cyan, (17) Dark Cyan, (18) Light Blue, (19) Blue, (20) Dark Blue, (21) Light Magenta, (22) Magenta, and (23) Dark Magenta.

The second stage essentially extends the output produced during the first stag; by assigning three different shades to each original color from the 10-bin pallet. To define the different shades (Light Color, Color, and Dark Color) a second fuzzy system is employed that uses the values of S and V as inputs.

Both Channels S and V are divided into two fuzzy regions (Fig. 3 – 2nd stage Fuzzy System). For values of V that fall in the first fuzzy area, and independently of S, the original color (from the 10-bin histogram) is assigned to the respective Dark Color (in the 24-bins histogram). The S values from the other hand suggests whether the assignment of the original color is assigned to Light Color or remains the same in the final 24-bins palette. Note that since the first three bins are already shades of Grey (White, Grey, and Black) their values are transferred directly to the final 24-bins histogram. More details about the color unit are given in [5].

4 Descriptor's implementation

The CoMo descriptor can be extracted either as a local or a global feature. When it is adopted as a global feature, the following procedure is applied on the entire image. Similarly, when the proposed descriptor is used as a local feature, the given methodology applies only on the regions of interest.

As already discussed in Section 3 the CoMo descriptor is compined from considering useful information from both the color and the texture units. More specifically, and to shape the proposed descriptor the input information is decided that is either the entire image of a region of interest on the image. Then the input is separated specifically into 1600 equal size image blocks. Each image block interacts successively with both, the color and the texture information units. When the descriptor is behaving as a local feature it utilizes a random patches' generator to extract the regions of interest from a given image. In other words, rather than attempt to semantically segment the image, e.g. into foreground object and background, the viual content is represented by a set of overlapping (local) regions [8]. An algorothm randomly selects x and y positions in the images, to mark square regions of pixels. As we elaborate in Section 5, employing a random sampling strategy yields results that are directly comparable, and often outperform some of the most sophisticated and complex methods from recent literature [16]. The sizes of the regions were decided as follows: the smallest patch size (defined as Reference Patch, denoted by RP) was set to 80×80 pixels, so as to align with the highest patch size limitation introduced by the CEDD descriptor. Having set the RP we then

employ a scaling factor (sf) to generate larger patches of sizes $RP * sf \times RP * sf$ pixels.

In the Color Unit, the image block is converted to the HSV color space in order to provide the first stage of the fuzzy-linking system with its input. Then, the second sub-unit of the system produces a 24-bin histogram per tile (as shown in Fig. 3).

In the Texture Unit, the Hu Moments of each image block are computed as follows. Briefly, *Shannon* Entropy is used as a statistical measure of randomness,

$$H(X) = -\sum_{i=1}^{n} P(X_i) log_b P(X_i) \tag{7}$$

where, X is a random variable (i.e., image block), n is the number of pixels, $b = 2$ in our case, and $P(X_i)$ is the occurrence probability of each pixel. If the result computed by (7) is less than T_{th}, then the block is not considered in the extraction procedure because the block is assumed to contain insufficient information, and thus is a texture-less block.

Subsequently, the Euclidean distance between the calculated Hu Moments and the 6 predefined texture classes is computed. The distance is normalized within the interval [0, 1], with 0 being the closest to the center of the class. If the resulted value is less than a given threshold, the image block is classified into that texture type. Thus, an image block can participate in more than one texture types. The outcome generated by this unit is a 6-bin histogram. At the final stage, the resulted vectors are combined to form the CoMo histogram of the input patch.

To assist the reader into grasping the fundamentals behinds the extraction procedure the following scenario is described. Defining the bin produced by the texture information fuzzy system as n, [0, 5] and the bin produced by the 24-bin fuzzy-linking system as m, [0, 23], then each image block is placed in the bin position: $n \times 24 + m$. According to the described methodology, an image block interacts with the Texture Unit. For the purposed of our example scenario, let us assume that an image block is classified into the second texture class $n = 1$. Then, in the Color Unit, the same image block is converted to the HSV color space. The mean values of H, S, and V are calculated and become inputs to the fuzzy system that produces the fuzzy 10-bin histogram.

In addition, let us assume that the classification resulted in the $4th$ bin indicates that the color is `red`. If this is the case, then the second fuzzy system (24-bin Fuzzy-linking system), using the mean values of S and V, as well as, the position number of the bin (or bins) resulting from the previous fuzzy ten-bin unit, calculates the hue of the color, and produces the fuzzy 24-bin histogram. For completeness at this stage, let us assume that the Color Unit system classifies this block in the $4th$ bin that indicates the color as `light red` $m = 3$. The combination of the three fuzzy systems will finally classify the image block in the $27th$ bin ($1 \times 24 + 3$). This process is repeated for all image blocks.

To restrict the proposed descriptor's length, the normalized bin values of the descriptor are non-linearly quantized for binary representation in a three bits/ bin quantization. As already discussed, to calculate the CoMo quantization table, 100000 randomly selected images from Flicker were used. First, CoMo vectors are calculated for all images. The combined 100000×144 elements constitute inputs to a k-means classifier which separates the samples volume into eight regions; mapping the bin values from the decimal area [0, 1] into the integer area [0, 7], which can then be represented by 3 bits. It is worth mentioning that the size of the proposed descriptor is equal to the size of CEDD. An implementation flowchart is illustrated in Fig. 4.

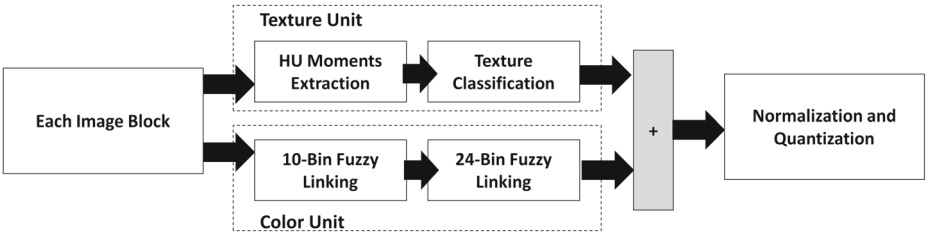

Fig. 4 CoMo: Implementation Flowchart

5 Experimental evaluation

This section describes a comprehensive experimental evaluation of the CoMo descriptor against various state-of-the-art local and global descriptors from the literature. The results highlight that the novel texture unit incorporated enables the descriptor to be invariance to alternations under different viewpoint, rotation, scale and lighting conditions. We demonstrate that our proposed descriptor, at the worst case scenarios, report performances close to the ones reported by CEDD, whereas, in the best case scenarios CoMo outperforms all the comparison descriptors.

The rest of this section proceeds by describing the various datasets and our experimental methodology, followed by a thorough discussion on the experimental results.

Datasets For the evaluation of the retrieval performance of the proposed CoMo descriptor, experiments were conducted on four different benchmarking datasets: UCID [51], UkBench [36], Holidays [18], and ZuBuD [52]. Table 1 summarizes the main attributes of the datasets along with the query mode employed for experimentation purposes.

Our initial experimentation was conducted on the UCID database. This database consists of 1338 images on a variety of topics, including natural scenes, and man-made objects, both indoors and outdoors. Manual relevance assessments among all database images are made available by the benchmark. UCID includes several query images where the ground truth consists of images whose visual concept is similar to the selected image used in the query, even though co-occurrence of the same objects may not exist. An example image is shown in Fig. 5e–g. UKBench database, consists of 10,200 images arranged in 2250 groups. Each group includes 4 images of a single object, captured from different viewpoints and under varying lighting conditions. An example query image together with its ground truth is illustrated in Fig. 5a–d.

INRIA Holidays Database [17] consists of 1491 high-resolution images of natural and man-made scenes. The dataset contains 500 image groups, each representing a distinct scene

Table 1 Databases used for the evaluation with, number of images in the database, number of query images used, average number of relevant images per query, and how the queries are evaluated

Name	# of img	# of queries	Avg no. of relevant img / query	Query mode
UCID	1338	262	3.45	query-in-groundTruth
UkBench	10200	2550	4	query-in-groundTruth
Holidays	1491	500	2.98	query-in-groundTruth
ZuBuD	1005	115	5	queries & dataset are disjoint

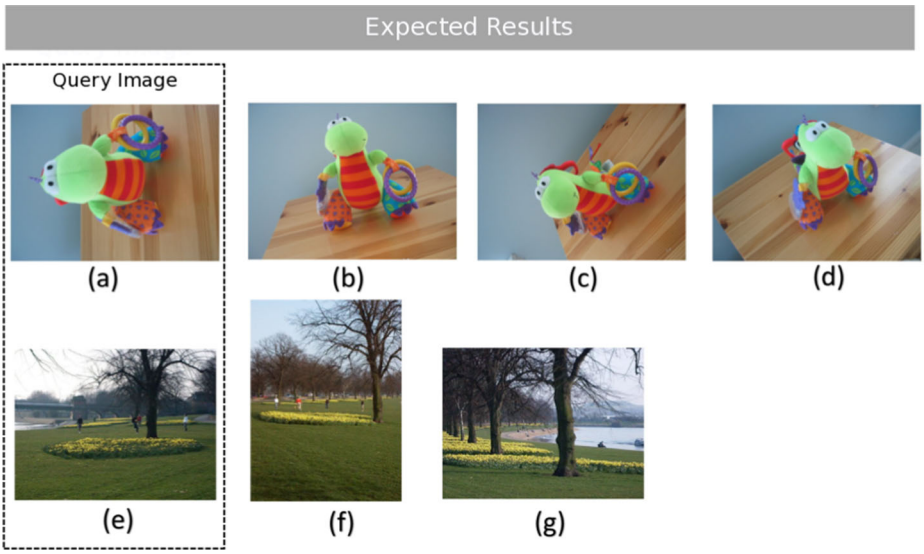

Fig. 5 Examples of visual queries using: **a–d** UKBench, **e–g** UCID database. In both cases, the left-most image is the query image and the images to the right of it are the expected results (ground truth)

or object. The number of images per category ranges from 2 to 13 images. The ground truth of a query per group is also made available. The size of the images varies from 480×640 to 2592×3888 pixels. In contrast to the UKBench and ZuBud databases this database includes several query images where the ground truth contains images with similar visual concept to the query image, without this implying the co-occurrence of the same objects.

The Zurich Building dataset (ZuBud) contains 1005 images with 201 buildings each in five views. All images in the collection have a size of 640×480 pixels. There is also a query dataset containing 115 images for testing the retrieval performance. Query images have a size of 320×240 pixels. The images are taken from random view points, under occlusion and varying lighting conditions from different seasons, weather, and different cameras. In addition this database contains severe viewpoint changes [11]. To give a more precise idea of this database, an example image is shown in Fig. 7.

Experimental setup & metric To evaluate CoMo, all experiments were conducted using the Bag-of-Visual-Word model (BoVW). This model has shown remarkable performance mainly because of its retrieval effectiveness over global feature representations on near duplicate and verbose images, and of course, the significant advantage of the model in terms of efficiency when compared with the local feature representations. The employed codebook used for our experimental purposes consists of 2048 visual words. We note that the codebook size was chosen based on the results of previous investigations as in [16].

To measure the performance of the proposed descriptor, we used Mean Average Precision (MAP) metric,

$$\text{Percision} = P = \frac{\text{Number of relevant images retrieved}}{\text{Total number of images retrieved}} \tag{8}$$

$$\text{Recall} = R = \frac{\text{Number of relevant images retrieved}}{\text{Total number of relevant images}} \tag{9}$$

We compute the Average Precision (AP) using,

$$\mathrm{AP}(q) = \frac{1}{N_R} \sum_{n=1}^{N_R} P_Q(R_n), \tag{10}$$

where, R_n is the *recall* after the nth relevant image retrieved, and N_R the total number of relevant documents for the query. MAP is then computed by,

$$\mathrm{MAP} = \frac{1}{Q} \sum_{q \in Q} \mathrm{AP}(q), \tag{11}$$

where, Q is the set of queries q. An advantage of the MAP is that it contains both *precision* and *recall* oriented aspects that make it sensitive to the entire ranking.

Experimental results & discussion Table 3, presents the experimental results on the UCID, UKBench, Holidays and ZuBud collections. The `WS` field describes the employed weighting scheme using the SMART notation (kindly refer to Table 2). The first weighting factor is the term frequency ($tf_{t,d}$), where a weight is assigned to every term (t) in the codebook according to the number of occurrences in a document (d). The second factor for assigning weights is the document frequency (df_t). This time, df_t is defined as the number of documents that contain the term t.

Often, the inverse document frequency $idf_t = log(N/df_t)$ of a collection is used to determine weights, where N is the total number of documents in the collection. Finally, to quantify the similarity between two documents in terms of the *cosine similarity* of their vector representation a normalization is applied. We note that Table 3 records only the weighing scheme that reported the best result.

Overall, all experimental results confirm that CoMo outperforms various state-of-the-art local and global features from the literature in all databases. The most significant observation is that the global version of CoMo outperforms the complete list of the reported global descriptors in all bases, while its local version exceeds the performance of local CEDD in all databases with UCID as the only exception.

A closer observation on the results reported for the UCID database, one can conclude that the CoMo descriptor, either in its global or on its local form, performs almost identical with CEDD. This observation confirms that the novel texture unit, based on Hu moments, does not have a negative effect on the overall effectiveness of the descriptor. It is worth noting that the UCID database consists only of visually similar images, and ground truths without any rotated images.

On the other hand, experimental results on UKBench database illustrate that CoMo outperforms not only all the other descriptors from the literature and several CNN approaches, but also CEDD. This result confirms that the our proposed texture unit enables the descriptor with invariance to alternations, since this specific dataset consists of groups of objects under different viewpoint, rotation, scale and lighting conditions.

Table 2 SMART notation for tf.idf variants

`tf`		`df`		Normalization	
n	$\mathrm{tf}_{t,d}$	n	1	n	1
l	$1 + log(\mathrm{tf}_{t,d})$	t	$log(N/\mathrm{df}_i)$	c	$\frac{1}{\sqrt{w_1^2+w_2^2+\cdots+w_M^2}}$

Table 3 Experimental evaluation results: show the MAP performance of the CoMo descriptor against other local and global feature on various benchmark databases

	UCID		UKBench		Holidays		ZuBud	
Descriptor	WS	MAP	WS	MAP	WS	MAP	WS	MAP
Local CoMo	ltc	0.779	ltc	**0.929**	ltc	**0.811**	ltc	**0.899**
Local CEDD	ltc	**0.789**	ltc	0.918	ltc	0.808	ltc	0.839
CoMo	Glo.	0.684	Glo.	0.868	Glo.	0.726	Glo.	0.751
CEDD	Glo.	0.674	Glo.	0.806	Glo.	0.726	Glo.	0.723
Neural Codes [3]	–	–	–	–	–	0.749	–	–
LF - AlexNet [40]	–	–	–	–	–	0.793	–	–
LF - PhilippNet [40]	–	–	–	–	–	0.741	–	–
LF (VLAD)- OxfordNet [35]	–	–	–	–	–	0.816	–	–
LF (VLAD)- GoogLeNet [35]	–	–	–	–	–	0.836	–	–
CNNaug-ss [46]	–	–	–	0.911	–	0.840	–	–
S.-A. co-indexing [62]	–	–	–	–	–	0.809	–	–
Contextual Weighting [59]	–	–	–	–	–	0.781	–	–
CNN-ss [46]	–	–	–	0.869	–	0.770	–	–
VLAD [24]	–	–	–	0.847	–	0.558	–	–
MF Re. X^2 [47]	–	0.676	–	0.842	–	0.738	–	0.727
Co-Indexing [61]	–	–	–	–	–	0.809	–	–
IFV [41]	–	–	–	–	–	0.838	–	0.626
C.B. Embed [63]	–	–	–	–	–	0.796	–	–
Sp. CEDD [31]	Glo.	0.732	Glo.	0.883	Glo.	0.797	Glo.	0.772
AHE+Burst [17, 56]	–	–	–	–	–	0.794	–	–
HE+Burst [17, 56]	–	–	–	–	–	0.780	–	–
RWBD [1]	–	–	–	0.618	–	–	–	0.813
Salient colors [50]	–	–	–	–	–	–	–	0.877
SURF	lnc	0.626	ncc	0.691	nnc	0.678	ncc	0.613
Opponent SIFT	ntc	0.624	ntc	0.593	–	–	–	–
Color Moments [22]	ntc	0.617	lnc	0.636	–	–	–	–
SIFT	nnc	0.605	nnc	0.664	ntc	0.691	ntc	0.624
ORB	nnc	0.491	ntc	0.491	–	–	–	–
Fisher [20]	–	–	–	–	–	0.595	–	–
BRISK [26]	ntc	0.436	nnc	0.310	–	–	–	–
JCD	Glo.	0.695	Glo.	0.848	Glo.	0.735	Glo.	0.726
SCD [32]	Glo.	0.496	Glo.	0.468	Glo.	0.537	Glo.	0.351
CLD [32]	Glo.	0.553	Glo.	0.668	Glo.	0.640	Glo.	0.585
RGB Hist	Glo.	0.587	Glo.	0.739	Glo.	0.656	Glo.	0.587
OppHist	Glo.	0.590	Glo.	0.735	Glo.	0.658	Glo.	0.581
ACC [15]	Glo.	0.708	Glo.	0.887	Glo.	0.756	Glo.	0.504
CENTRIST [21]	Glo.	0.560	Glo.	0.564	Glo.	0.605	Glo.	0.029
Sp. CENTRIST [21]	Glo.	0.639	Glo.	0.711	Glo.	0.674	Glo.	0.105
PHOG [4]	Glo.	0.537	Glo.	0.508	Glo.	0.604	Glo.	0.442
EHD [60]	Glo.	0.502	Glo.	0.483	Glo.	0.555	Glo.	0.382

Table 3 (continued)

	UCID		UKBench		Holidays		ZuBud	
Descriptor	WS	MAP	WS	MAP	WS	MAP	WS	MAP
LBP [37]	Glo.	0.533	Glo.	0.530	Glo.	0.558	Glo.	0.132
RILBP [37]	Glo.	0.491	Glo.	0.413	Glo.	0.507	Glo.	0.055

WS corresponds to the selected Weighting Scheme while Glo. corresponds to global descriptors

Bold text indicates best reported performance in each database

Experimental results on Holidays dataset reinforce our conclusions thus far with reagrds to the ability of the proposed descriptor to retrieve both, semantically similar images, as well as, images with a similar visual concept. A study in [7] describes and identifies what family of descriptors is preferable and most suitable for each retrieval scenario. The Holidays dataset is considered a challenging set of images mainly because it contains several images with highly variable poses, and significant amounts of background clutter. Moreover, several classes are characterized by the objects they depict rather than spatial properties of the images. Figure 6 illustrates an indicate example where all images belong to the same ground truth. The reader can easily observe that instances b–f can be considered as relevant since they represent the same visual concept. On the other hand, instance (a) is considered relevant with the rest of the images only because it depicts a semantically similar concept. CoMo performance, especially in its local form, indicates that the proposed descriptor outperforms the vast majority of the listed methods and descriptors.

Further observing the reported results, we noted that some recently proposed approaches reported slightly better retrieval performance against CoMo. Specifically, the CNNaug-ss proposal, introduced in [46], reports an improvement of 3.56% for the Holiday database. At this point and for the specific proposal we highlight that CNNaug-ss is based on a deep learning approach that is highly dependable on a training procedure that requires a

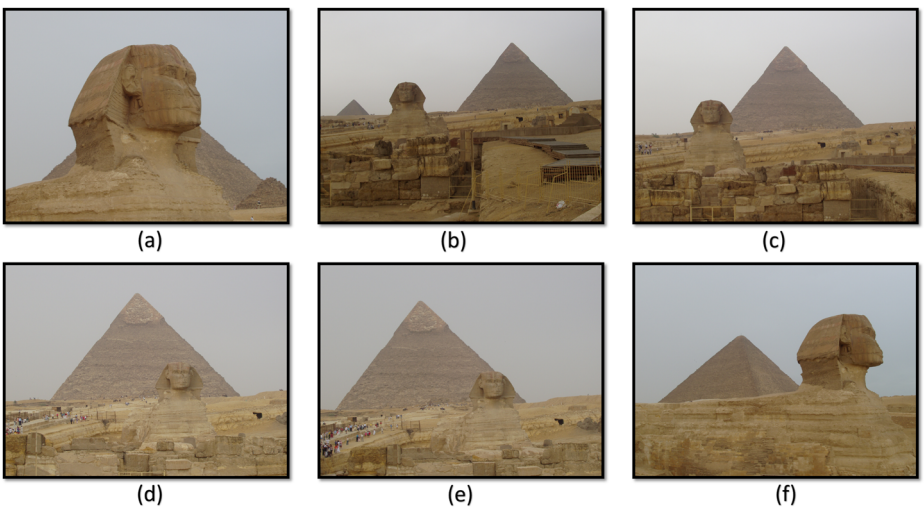

Fig. 6 Holidays: a sample set of relevant images

large number of training data to be made available. In contrast, the CoMo descriptor is a *plug-n-play* method which can be adopted without any prior initialization or training. In addition, compared to other recent CNN-based approaches, CoMo produces better or comparable results. The performance of the proposed descriptor is competitive compare to the deep learning-based method proposed in [35], which involves the pre-trained on imagenet, OxfordNet and GoogLeNet networks. Moreover, the CoMo's retrieval accuracy significantly outperforms the solution propoded in [3].

The foremost advantages of extracting CoMo is the low cost of the single-feature space computations along with the fact that indexing one image is independent of the type or the total number of images in the collection. Moreover, there are domains where training data is not readily available and the image data should either not leave or enter a controlled environment. All domains require hand-crafted image descriptors, from which the image visual content cannot be reconstructed.

Finally, CoMo reported a remarkable performance over the ZuBud dataset. Percentage-wise, local version of CoMo outperformed the local version of CEDD by 18.6%. At the same time, global version of CoMo exceeded the performance of global CEDD by 10%. Moreover, the proposed version of CoMo outperformed significantly all the other reported methods and descriptors. The results highlight the strong ability of CoMo to remain invariant to translation, rotation, and even small modification of the object's aspect (e.g., when an object is partially covered – refer to Fig. 7).

Another significant conclusion worth noting is that in all datasets, the local version of CoMo outperformed its global form. This observation confirms the idea on revisiting global features' description methods and localizing their effect by applying them on local neighborhoods. This idea could introduce a whole new generation of robust local features. Experimental results also indicated that for image retrieval tasks, it is not vital to employ computationally demanding point-of-interest detectors to extract image regions. Accurate image retrieval results may be obtained by adopting a uniform, random, multi-scale image patch generator.

Fig. 7 ZuBud: example of images taken from random viewpoints

Additionally to the above discussion we report that the local version of the CoMo descriptor has the following key advantages:

i. when adopting the BoVW model in real-world databases, there is an urgent need for establishing an efficient method of finding the most appropriate codebook size. Moreover, the BoVW approach adapts textual term weighting schemes for relevance estimation. Identifying the most appropriate weighting scheme for each database is not trivial. For the CoMo case, the best retrieval results were performed using the same codebook size and the same weighting scheme on all databases;
ii. codebook sizes ranging from hundreds to millions of visual words are reported in the literature. For the CoMo case, a compact codebook of 2048 words manages to outperform more than 35 methods and descriptors from the literature.

For completeness, it is worth noting that the proposed descriptor outperforms the only moment-based local descriptor in UCID and UKBench databases. In the case of UCID, the improvement is equal to 21%, while in the case of UKBench, CoMo reports an improvement of 31%.

As reported by our experimental evaluation, both CEDD and CoMo representations demonstrate an impressive performance, thus both methods can be safely recommended as preferable for image retrieval tasks. In all databases, CEDD and CoMo descriptors report stable and robust performances. Therefore, we recommend their usage for the selection of features with high confidence, that will produce accurate retrieval results in a variety of topics and scenarios. Moreover, robustness to several conditions, such as rotation, is often required/desirable. Real-world image retrieval systems, as well as, Simultaneously Localization and Mapping (V-SLAM) mechanisms should be able to successfully handle such kind of images. As mentioned earlier, CEDD can tolerate reasonable rotations, but no prior research has determined its level of invariance.

Discussion on the rotation invariance of CoMo Similarly to CoMo, CEDD's histogram consists of 6 texture areas. The first area describes regions with no texture, the second one corresponds to regions with horizontal activity while the third one to regions with vertical activity. The next 2 areas describe the regions with 45° and 135° respectively. Final, the last region corresponds the the non-uniform texture regions. A graphical representation of the CEDD's texture regions is illustrated in Fig. 8a.

Assume for example that an input image contains only a set of vertical lines. The visual content of this image produces a CEDD descriptor which contains non-zero bin-values only at the third region, as presented in Fig. 8b. In the sequel, lets assume that one rotates the image by 90°. Now, the visual content of the image depicts a set of horizontal lines and the corresponding CEDD descriptor contains non-zero bin-values only at the second region (Fig. 8c). The Euclidean or Tanimoto distance between the original and the rotated image is definitely not equal to zero. This observation reveals that CEDD descriptor is not invariant under the 90° rotation. But what if one rotates the already rotated image by additional 90° - or rotates the original image under 180°. In this case, the rotated image will display a set of vertical lines, and the CEDD descriptor would be identical with the one of the original image. In other words, CEDD descriptor demonstrates zero tolerance under the 90° rotation but in case of 180°, identifies that the images are identical. To conclude, CEDD can be conditionally classified as a rotation invariant feature since it is able to tolerate under specific rotations and retrieve effectively visually similar images on several benchmarking databases. But on the other side of the spectrum, this partial invariance may result to poor and inaccurate retrieval results under real life conditions and V-SLAM scenarios.

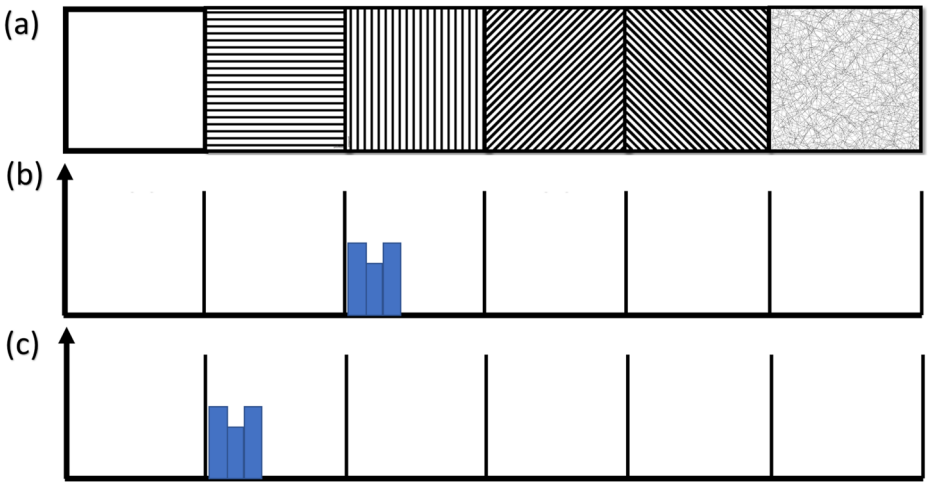

Fig. 8 **a** The 6 texture regions that CEDD uses, CEDD descriptor of an image that contains only vertical lines, CEDD descriptor of an image that contains only horizontal lines

By replacing the CEDD's texture extraction unit with the one that uses the HU moments, we managed to deliver an unconditionally rotation invariant descriptor. Experiments on all databases were repeated, but in this instance, the queries were rotated with rotation angles of 90°, 180° and 270°, respectively.

Table 4 displays the retrieval results and exposes new observations. In all databases, the proposed descriptor, either on its global or local form, accomplished to maintain its original retrieval accuracy. The 90° and 270° rotation experiments revealed the fact that CEDD is

Table 4 Experimental evaluation results: show the MAP performance of the CoMo descriptor against CEDD one on various benchmark databases from the literature under several rotation alternations

Rotation Angle		UCID	UKBench	Holidays	ZuBud
0°	CEDD	0.675	0.803	0.726	0.723
	Local CEDD	0.789	0.918	0.808	0.834
	CoMo	0.684	0.868	0.726	0.751
	Local CoMo	0.779	0.929	0.811	0.899
90°	CEDD	0.347	0.507	0.436	0.385
	Local CEDD	0.695	0.878	0.719	0.582
	CoMo	0.657	0.870	0.720	0.751
	Local CoMo	0.772	0.925	0.806	0.902
180°	CEDD	0.633	0.749	0.717	0.718
	Local CEDD	0.797	0.913	0.811	0.821
	CoMo	0.660	0.869	0.713	0.746
	Local CoMo	0.763	0.926	0.806	0.884
270°	CEDD	0.391	0.508	0.440	0.384
	Local CEDD	0.701	0.877	0.724	0.567
	CoMo	0.683	0.867	0.725	0.742
	Local CoMo	0.779	0.927	0.804	0.889

extremely sensitive to rotations that alternate the texture's orientation. In case of UCID, percentage wise, the CEDD's accuracy was reduced by 48.6%, in case of UKBench by 36.9%, in case of Holidays by 39.9% and in case of ZuBud by 46.8%.On the other hand, on all databases, the 180° rotation experiments demonstrated that CEDD can successfully tolerate these alternations.

6 Conclusion

This paper introduces a new low-level feature for image retrieval. The main novelty of the proposed feature lies in the usage of moment invariants along with the color unit of CEDD as descriptors of local image patches. The findings from the experimental evaluation clearly shown that the proposed descriptor outperforms not only localized CEDD but also other state-of-the-art local descriptors. We plan to extend the experiments by benchmarking the descriptor against other databases used in image retrieval research. The proposed descriptor and its source code is part of the LIRE [30] library[1] and can be used under the GNU GPL license.

References

1. Aslan S, Akgül CB, Sankur B, Tunali ET (2017) Exploring visual dictionaries: a model driven perspective. J Vis Commun Image Represent 49:315–331. https://doi.org/10.1016/j.jvcir.2017.09.009
2. Bay H, Tuytelaars T, Van Gool L (2006) Surf: speeded up robust features. In: Computer vision–ECCV 2006. Springer, pp 404–417
3. Babenko A, Slesarev A, Chigorin A, Lempitsky VS (2014) Neural codes for image retrieval. In: Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I, pp 584–599, https://doi.org/10.1007/978-3-319-10590-1_38. https://dblp.org/rec/bib/conf/eccv/BabenkoSCL14. dblp computer science bibliography, https://dblp.org
4. Bosch A, Zisserman A, Munoz X (2007) Representing shape with a spatial pyramid kernel. In: Proceedings of the 6th ACM International Conference on Image and Video Retrieval, CIVR 2007, Amsterdam, The Netherlands, July 9-11, 2007, pp 401–408, https://doi.org/10.1145/1282280.1282340. https://dblp.org/rec/bib/conf/civr/BoschZM07. dblp computer science bibliography, https://dblp.org
5. Chatzichristofis S, Boutalis Y (2008) Cedd: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval. Comput Vis Syst 5008:312–322
6. Chatzichristofis SA, Zagoris K, Boutalis YS, Papamarkos N (2010) Accurate image retrieval based on compact composite descriptors and relevance feedback information. Int J Pattern Recogn Artif Intell (IJPRAI) 2:207–244
7. Chatzichristofis SA, Iakovidou C, Boutalis YS, Oge M (2013) Co.vi.wo.: color visual words based on non-predefined size codebooks. IEEE Trans Cybern 43(1):192–205
8. Csurka G, Dance C, Fan L, Willamowski J, Bray C (2004) Visual categorization with bags of keypoints. In: Workshop on statistical learning in computer vision, ECCV, vol 1
9. Deselaers T, Keysers D, Ney H (2008) Features for image retrieval: an experimental comparison. Inf Retr 11(2):77–107
10. Eisa M, Eletrebi A, Elhenawy E (2013) Enhancing the retrieval performance by combing the texture and edge features. CoRR, arXiv:abs/1301.2542
11. Fond A, Berger M-O, Simon G (2017) Facade proposals for urban augmented reality. In: 16th IEEE International symposium on mixed and augmented reality (ISMAR)
12. Gholipour F, Ebrahimnezhad H (2014) An efficient content based image retrieval using edge orientation co-occurrence matrix. In: 2014 6th Conference on information and knowledge technology (IKT). IEEE, pp 67–72

[1] http://www.lire-project.net/

13. Gordo A, Almazàn J, Revaud J, Larlus D (2016) Deep image retrieval: learning global representations for image search. In: Computer Vision - ECCV 2016 - 14th European conference. Amsterdam, The Netherlands, October 11-14, 2016, Proceedings Part VI, pp 241–257
14. Harris CG, Pike JM (1988) 3d positional integration from image sequences. Image Vis Comput 6(2):87–90
15. Huang J, Kumar SR, Mitra M, Zhu WJ (2001) Image indexing using color correlograms. US Patent 6,246,790 12:1–16
16. Iakovidou C, Anagnostopoulos N, Kapoutsis A, Boutalis Y, Lux M, Chatzichristofis SA (2015) Localizing global descriptors for content-based image retrieval. EURASIP J Adv Signal Process 2015(1):80
17. Jain M, Jégou H, Gros P (2011) Asymmetric hamming embedding: taking the best of our bits for large scale image search. In: ACM Multimedia. ACM Multimedia, Nov 28 - Dec 1, Scottsdale, Arizona, USA. ACM, Scottsdale, pp 1441–1444
18. Jegou H, Douze M, Schmid C (2008) Hamming embedding and weak geometric consistency for large scale image search. In: Computer Vision - ECCV 2008, 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I, pp 304–317, https://doi.org/10.1007/978-3-540-88682-2_24. https://dblp.org/rec/bib/conf/eccv/JegouDS08, dblp computer science bibliography, https://dblp.org
19. Jégou H, Douze M, Schmid C (2010) Improving bag-of-features for large scale image search. Int J Comput Vis 87(3):316–336
20. Jègou H, Perronnin F, Douze M, Sànchez J, Pèrez P, Schmid C (2012) Aggregating local image descriptors into compact codes. IEEE Trans Pattern Anal Mach Intell 34(9):1704–1716
21. Jianxin Wu, Rehg JM (2011) Centrist: a visual descriptor for scene categorization. IEEE Trans Pattern Anal Mach Intell 33(8):1489–1501
22. Karakasis EG, Amanatiadis A, Gasteratos A, Chatzichristofis SA (2015) Image moment invariants as local features for content based image retrieval using the bag-of-visual-words model. Pattern Recogn Lett 55(0):22–27
23. Kim NW, Kim TY, Choi J-S (2005) Edge-based spatial descriptor for content-based image retrieval. In: Image and Video Retrieval, 4th International Conference, CIVR 2005, Singapore, July 20-22, 2005, Proceedings, pp 454–464, https://doi.org/10.1007/11526346_49. https://dblp.org/rec/bib/conf/civr/KimKC05. dblp computer science bibliography, https://dblp.org
24. Le D, Liang Y, Kong G, Zhang Q, Cao X, Izquierdo E (2016) Holons visual representation for image retrieval. IEEE Trans Multimed 18(4):714–725
25. Lei Z, Fuzong L, Bo Z (1999) A cbir method based on color-spatial feature. In: TENCON 99. Proceedings of the IEEE Region 10 conference, vol 1. IEEE, Cheju Island, South Korea, South Korea. pp 166–169
26. Leutenegger S, Chli M, Siegwart RY (2011) Brisk: binary robust invariant scalable keypoints. In 2011 IEEE International conference on computer vision (ICCV). IEEE, Barcelona, Spain. pp 2548–2555
27. Li X, Larson M, Hanjalic A (2015) Pairwise geometric matching for large-scale object retrieval. In: IEEE Conference on computer vision and pattern recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, pp 5153–5161
28. Li C, Huang Y, Zhu L (2017) Color texture image retrieval based on gaussian copula models of gabor wavelets. Pattern Recogn 64:118–129
29. Lowe DG (1999) Object recognition from local scale-invariant features. In: The proceedings of the seventh IEEE international conference on computer vision, 1999, vol 2. IEEE, Kerkyra, Greece, Greece. pp 1150–1157
30. Lux M, Chatzichristofis SA (2008) Lire: lucene image retrieval: an extensible java CBIR library. In: Proceedings of the 16th International Conference on Multimedia 2008, Vancouver, British Columbia, Canada, October 26-31, 2008, pp 1085–1088, https://doi.org/10.1145/1459359.1459577. https://dblp.org/rec/bib/conf/mm/LuxC08. dblp computer science bibliography, https://dblp.org
31. Lux M, Anagnostopoulos N, Iakovidou C (2016) Spatial pyramids for boosting global features in content based image retrieval. In: 14th International Workshop on Content-Based Multimedia Indexing, CBMI 2016, Bucharest, Romania, June 15-17, 2016. IEEE, pp 1–4, https://doi.org/10.1109/CBMI.2016.7500248. https://dblp.org/rec/bib/conf/cbmi/LuxAI16. dblp computer science bibliography, https://dblp.org
32. Manjunath BS, Ohm JR, Vasudevan VV, Yamada A (2001) Color and texture descriptors. IEEE Trans Circ Syst Video Technol 11(6):703–715
33. Mei T, Rui Y, Li S, Qi T (2014) Multimedia search reranking: a literature survey. ACM Comput Surv (CSUR) 46(3):38

34. Mikulík A, Perdoch M, Chum O, Matas J (2010) Learning a fine vocabulary. In: Computer vision - ECCV 2010, 11th European conference on computer vision, Heraklion, Crete, Greece, September 5-11, 2010. Proceedings, Part III, pp 1–14
35. Ng JY-H, Yang F, Davis LS (2015) Exploiting local features from deep networks for image retrieval. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 53–61
36. Nister D, Stewenius H (2006) Scalable recognition with a vocabulary tree. In: Proc. CVPR, vol 5. Citeseer
37. Ojala T, Pietikäinen M, Mäenpää T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell 24(7):971–987
38. Papakostas GA, Koulouriotis DE, Karakasis EG (2009) A unified methodology for the efficient computation of discrete orthogonal image moments. Inf Sci 179(20):3619–3633
39. Papakostas GA, Koulouriotis DE, Karakasis E, Tourassis VD (2013) Moment-based local binary patterns: a novel descriptor for invariant pattern recognition applications. Neurocomputing 99:358–371
40. Paulin M, Douze M, Harchaoui Z, Mairal J, Perronin F, Schmid C (2015) Local convolutional features with unsupervised training for image retrieval. In: Proceedings of the IEEE international conference on computer vision, pp 91–99
41. Perronnin F, Liu Y, Sánchez J, Poirier H (2010) Large-scale image retrieval with compressed fisher vectors. In: The Twenty-Third IEEE conference on computer vision and pattern recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010, pp 3384–3391
42. Perronnin F, Sánchez J, Mensink T (2010) Improving the fisher kernel for large-scale image classification. In: Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV, pp 143–156, https://doi.org/10.1007/978-3-642-15561-1_11. https://dblp.org/rec/bib/conf/eccv/PerronninSM10, dblp computer science bibliography, https://dblp.org
43. Petscharnig S, Lux M, Chatzichristofis S (2017) Dimensionality reduction for image features using deep learning and autoencoders. In: Proceedings of the 15th international workshop on content-based multimedia indexing CBMI '17, pp 23:1–23:6
44. Philbin J, Chum O, Isard M, Sivic J, Zisserman A (2008) Lost in quantization: improving particular object retrieval in large scale image databases. In: IEEE Conference on computer vision and pattern recognition, 2008. CVPR 2008. IEEE, pp 1–8
45. Qi G-J, Hua X-S, Rui Y, Mei T, Tang J, Zhang H-J (2007) Concurrent multiple instance learning for image categorization. In: 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA, https://doi.org/10.1109/CVPR.2007.383152. https://dblp.org/rec/bib/conf/cvpr/QiHRMTZ07. dblp computer science bibliography, https://dblp.org
46. Razavian AS, Azizpour H, Sullivan J, Stefan C (2014) CNN features off-the-shelf: an astounding baseline for recognition. In: IEEE Conference on computer vision and pattern recognition, CVPR Workshops 2014, Columbus, OH, USA, June 23-28, 2014, pp 512–519
47. Reta C, Solis-Moreno I, Cantoral-Ceballos JA, Alvarez-Vargas R, Townend P (2017) Improving content-based image retrieval for heterogeneous datasets using histogram-based descriptors. Multimedia Tools and Applications, https://doi.org/10.1007/s11042-017-4708-8. ISSN: 1573-7721
48. Rosten E, Drummond T (2006) Machine learning for high-speed corner detection. In: ECCV (1), pp 430–443
49. Rublee E, Rabaud V, Konolige K, Bradski GR (2011) Orb: an efficient alternative to sift or surf. In: ICCV, 6-13 November, Barcelona, Spain. IEEE, Barcelona, pp 2564–2571
50. Sajjad M, Ullah A, Ahmad J, Abbas N, Rho S, Baik SW (2018) Integrating salient colors with rotational invariant texture features for image representation in retrieval systems. Multimed Tools Appl 77(4):4769–4789. https://doi.org/10.1007/s11042-017-5010-5. https://dblp.org/rec/bib/journals/mta/SajjadUAARB18, dblp computer science bibliography, https://dblp.org
51. Schaefer G, Stich M (2004) Ucid: an uncompressed color image database. Storage Retriev Methods Appl Multimed 5307:472–480
52. Shao H, Svoboda T, Van Gool L (2003) Zubud-zurich buildings database for image based recognition. Computer Vision Lab, Swiss Federal Institute of Technology, Switzerland, Tech. Rep. 260:20
53. Shi J, Tomasi C (1994) Good features to track. In: Conference on Computer Vision and Pattern Recognition, CVPR 1994, 21-23 June, 1994, Seattle, WA, USA, pp 593–600, https://doi.org/10.1109/CVPR.1994.323794. https://dblp.org/rec/bib/conf/cvpr/ShiT94. dblp computer science bibliography, https://dblp.org

54. Shyu C-R, Brodley CE, Kak AC, Kosaka A, Aisen A, Broderick L (1998) Local versus global features for content-based image retrieval. In: IEEE Workshop on content-based access of image and video libraries proceedings. IEEE, Santa Barbara, USA. pp 30–34
55. Tolias G, Hervė J (2014) Visual query expansion with or without geometry: refining local descriptors by feature aggregation. Pattern Recogn 47(10):3466–3476
56. Tolias G, Avrithis YS, Jėgou H (2013) To aggregate or not to aggregate: selective match kernels for image search. In: IEEE International conference on computer vision, , ICCV 2013, Sydney, Australia, December 1-8, 2013 pp 1401–1408
57. Tolias G, Avrithis YS, Hervė J (2016) Image search with selective match kernels: aggregation across single and multiple images. Int J Comput Vis 116(3):247–261
58. Vassou SA, Anagnostopoulos N, Amanatiadis A, Christodoulou K, Chatzichristofis SA (2017) Como: a compact composite moment-based descriptor for image retrieval. In: Proceedings of the 15th international workshop on content-based multimedia indexing, CBMI 2017, Florence, Italy, June 19-21, 2017. pp. 30:1–30:5
59. Wang X, Yang M, Cour T, Zhu S, Yu K, Han TX (2011) Contextual weighting for vocabulary tree based image retrieval. In: IEEE International conference on computer vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011. pp 209–216
60. Won CS, Park DK, Park SJ (2002) Efficient use of mpeg-7 edge histogram descriptor. Etri J 24(1):23–30
61. Zhang S, Yang M, Wang X, Lin Y, Qi T (2013) Semantic-aware co-indexing for image retrieval. In: Proceedings of the IEEE international conference on computer vision, pp 1673–1680
62. Zhang S, Yang M, Wang X, Lin Y, Tian Q (2015) Semantic-aware co-indexing for image retrieval. IEEE Trans Pattern Anal Mach Intell 37(12):2573–2587
63. Zheng L, Wang S, Qi T (2014) Coupled binary embedding for large-scale image retrieval. IEEE Trans Image Process 23(8):3368–3380

S. A. Vassou is a Computer Engineering and Informatics student at the Cyprus University of Technology, Cyprus. His research focuses on machine learning, computer vision and multimedia retrieval.

N. Anagnostopoulos after his graduation from the Department of Electrical and Computer Engineering, Democritus University of Thrace, Greece., Nektarios was employed as Project/ Research Assistant at Alpen-Adria-Universität (Klagenfurt University) and worked on research, development and evaluation in visual information retrieval for a year. Moreover, while being there, he also held an External Lecturer position for a semester, teaching Introduction to Multimedia Technology, a course on web development concepts and practices. Finally, a year ago he relocated to Prague, to join the newly established Microsoft Dynamics CRM Development Center as a Software Engineer, however, his interest in the visual information retrieval field remains active and he keeps participating in research projects since then.

Dr. K. Christodoulou is a Lecturer in Informatics. Klitos joined the Department of Informatics at the Neapolis University in Cyprus since the Department's creation in September 2015. He holds a doctorate (PhD) in Computer Science from the University of Manchester, UK. In his doctoral research he worked under the supervision of Prof. Norman W. Paton and Dr. Alvaro A. A. Fernandes undertaking research in the area of Linked Data while exploring automated Data Integration techniques when these are applied on the Semantic Web.

Dr. A. Amanatiadis is an Adjunct Lecturer in the Department of Electrical and Computer Engineering of Democritus University of Thrace, Greece. His research focuses on embedded robotics, machine vision, and real-time FPGA/GPU based systems. He received his Diploma and PhD from the Democritus University of Thrace, Greece, in Electrical and Computer Engineering, developing hardware designs for robotic and computer vision systems. He has received the Stavros Niarchos Award for Promising Young Scientists, the State Scholarships Foundation (IKY) Postdoc Scholarship and awarded as outstanding reviewer in several journals.

S. A. Chatzichristofis pursued both, the Diploma and the Ph.D. degree (with honors) from the Department of Electrical and Computer Engineering, Democritus University of Thrace, Greece.

Currently, he serves as a faculty member at the School of Informatics, Neapolis University Pafos (NUP), Cyprus, as an Associate Professor and director of the Robotics Lab. His research is mainly focused on Cybernetics and Artificial Intelligence together with their applications in the fields of Computer Vision, Multimedia/Multimodal Retrieval, Robotics, Optimization and Pattern Recognition (forensic and industrial applications). He has 10+ years of experience on information technology, with emphasis on topics related to multimedia information retrieval systems and machine vision, reporting more than 60 publications in these fields.