Searching Images with MPEG-7 (& MPEG-7-like) Powered Localized dEscriptors: The SIMPLE answer to effective Content Based Image Retrieval

C. Iakovidou, N. Anagnostopoulos, A. Ch. Kapoutsis, Y. Boutalis, S. A. Chatzichristofis

***Abstract*— In this paper we propose and evaluate a new technique that localizes the description ability of the well established MPEG-7 and MPEG-7-like global descriptors. We employ the SURF detector to define salient image patches of blob-like textures and use the MPEG-7 Scalable Color (SC), Color Layout (CL) and Edge Histogram (EH) descriptors and the global MPEG-7-like Color and Edge Directivity Descriptor (CEDD), to produce the final local features' vectors. In order to test the new descriptors in the most straightforward fashion, we use the Bag-Of-Visual-Words framework for indexing and retrieval. The experimental results conducted on two different benchmark databases with varying codebook sizes, revealed an astonishing boost in the retrieval performance of the proposed descriptors compared both to their own performance (in their original form) and to other state-of-the-art methods of local and global descriptors. Open-source implementation of the proposed descriptors is available in c#, Java and MATLAB[1].**

I. Introduction

After many years of research, little is known about the combination of features that best describes an image with respect to its visual properties or its visual content. With image collections growing by the minute in various areas such as medicine, private life, industrial/commercial products, journalism, tourist attractions and art -to name a few- a plethora of Content Based Image Retrieval (CBIR) systems have been introduced in the literature. The main objective of all proposed schemes is to represent images with a feature vector or descriptor that will allow fast access and meaningful retrieval for the user.

Representing images with numerical values in a way that grasps their distinctive visual properties and contents, is a challenging process. The variety of solutions and proposed implementations of description methods in the literature is indicative of the complexity of the problem. A wide collection of early strategies and recent trends on image retrieval can be found in the well-structured studies [1], [2], while feature specific studies evaluating color description [3], [4], [5], texture description [6], [7], [8] and shape description [9], [10] strategies, vividly outline the many directions researchers explored in the quest of effectively representing image content.

In essence, the success of any CBIR system is subject to the user's requirements and the specific characteristics of the image collection. When a query is set with the objective to retrieve visually similar images, for instance natural scenes of mountains or fields and forests, or even images depicting objects with little or no background clutter, a feature vector that treats and describes the images as a whole, is effective [11]. Global Features (GF) such as color, texture and shape are calculated on the entire image to form an informative feature vector representation. The images in a collection are compared with various distance measures to the query. The lower the distance, the higher the rank they achieve in the retrieval process.

On the other side of the spectrum, when searching for images with similar visual and conceptual content, which is the case for verbose images or images where objects appear with partial occlusions [12], global vectors fail to discriminate the constituent parts of an image. For example, objects of the same shape described by a global image vector will not be distinguished even if their local texture information varies significantly. The information concerning localized aspects of an image can be of great importance for representing and classifying images that present high in-class variability.

In order to enrich the representation with a local information component, Local Features (LF) were introduced. In theory, every pixel in an image can be used to define a LF. This would lead to an unmanageable number of features that do not necessarily add to the descriptor's discrimination ability. Thus, LF are vector representations of salient regions of the image. These salient regions, often referred to as points-of-interest (POI), are local extrema of some function of the image, like edges, corners and blobs. Some among the most widely employed POI detectors are corner detectors Harris [13], Shi-Tomasi [14], and FAST [15] and blob detectors SIFT [16], SURF [17], to name a few. In [18] the authors provide some basic guidelines concerning the proper selection of detectors, as they investigate their pros and cons for usage in visual odometry. An overview of the detectors can be found in [19].

After locating the salient region, feature vectors that are to some extent translation, scale and rotation invariant are calculated for every POI. The representation of the image is mapped into a high dimensional local feature space. In order to address the dimensionality problem, resent research has focused on the Bag-Of-Visual-Word (BOVW) model [20]. In BOVW each LF is classified into a class (Visual Word-VW), the total number of VW forms the codebook. Every image is represented by a histogram of the VWs that were located

All the authors are with the Department of Electrical and Computer Engineering, Democritus University of Thrace, Xanthi GR 67100, Greece

[1]http://tinyurl.com/SIMPLE-Descriptors

in it. A recent study of image retrieval implementations employing the BOVW model can be found in [21].

With all said above, a clear question arises; are global features' methods outworn? Should we focus mainly on designing and evaluating novel retrieval methods that incorporate local features, since they outperform the generalized and often holistic representations of images obtained from global descriptors?

In this paper we propose a hybrid approach that recycles the strengths of existing well studied and established global feature representation methods, empowering them by focusing their feature extraction mechanism on various salient patches of the image.

More specifically, we employ the MPEG-7 global descriptors SCD, CLD and EHD [22] and the MPEG-7-like CEDD global descriptor [23] to describe image patches derived by locating salient image regions using the SURF detection mechanism. We are *Searching Images with Mpeg-7 Powered Localized dEscriptors* (hereon referred to as **SIMLPE**). Thus, we produce localized descriptions of the SCD (SIMPLE-SC), CLD (SIMPLE-CL) and EHD (SIMPLE-EH) methods and we also propose the localized equivalent of the CEDD method, hereupon referred to as SIMPLE-CEDD or '**LoCATe**' (Local Color And Texture dEscriptor), which is strongly inspired by the original MPEG-7 descriptors since it is global, compact and quantized.

II. Related Work

More and more approaches are presented in the literature utilizing the effective and compact representations that the MPEG-7 family of global descriptors introduces. The proposed methods are in great abundance, but here we will only focus on a small fraction of implementations that attempt to combine them with local information of same kind. In [24] three content-based image classification techniques are introduced, that fuse various low-level MPEG-7 descriptors. A 'merging' fusion combined with a support vector machine (SVM) classifier, a back-propagation fusion with a KNN classifier and a Fuzzy-ART neurofuzzy network strategies are explored, that can be extended in matching the segments of an image with predefined object models. A classification-driven similarity matching framework is presented in [25] for biomedical image retrieval. In order to generate the feature vectors at different levels of abstraction, both the visual concept feature [26] based on the 'bag of concepts' model (that comprises of local color and texture patches) and various low-level global color, edge, and texture related features are extracted (like CLD, EHD, CEDD and FCTH [27]). The utilization of the multi-class SVM and various classifier combination rules in different aspects of the image feature spaces are explored for the categorization, representation and similarity matching of the images. In [28] authors combine local and global features. In order to index a collection of images, the method extracts SURF local features and five MPEG-7 descriptors (CS, CL, SC, HT, EH) as global features and proceeds by associating each image with six text fields, one corresponding to the bag-of-features obtained from the SURF descriptor and five surrogate text representations, one for each MPEG-7 descriptor. These segments, form the basic units on which search is performed. An approach that evaluates the fusion (baseline fusion and score fusion) of MPEG-7, SIFT, and SURF content-based retrieval techniques to address the event search issue is presented in [29]. The detailed results illustrate that the MPEG-7, SIFT, and SURF are broadly comparable, and also highly complementary. Finally, authors in [30] propose a new method to combine MPEG-7 descriptors with spatial information, by the use of cluster correlograms for image categorization. They employ fixed partitioning and salient points schemes to extract image patches and use 4 MPEG-7 descriptors to represent them. A clustering technique is applied to group similar patterns into a cluster codebook. A correlogram is constructed from the spatial relations between visual keyword indices in an image, in order to obtain high-level information about the relational context. For similarity and matching, the feature vector of each signature is represented by a 2D $m \times m$ matrix where m is the number of clusters. Every image has four different signatures (one for each MPEG-7 descriptor). The m value varies and depends on the number of clusters used in the clustering algorithm. For m clusters the feature dimension size for each image would be $4m^2$.

To sum up, most attempts to combine local information and global descriptions rely on some late fusion method. Fixed partitioning of images and region based image patches obtained to be described with global descriptors, are also presented but suffer in domain specific tasks, where background information and foreground are not easily dissociated. Our proposed implementation of localized MPEG-7 descriptors is very similar in concept with the salient-approach of the method in [30]. However in [30] authors propose a more sophisticated scheme, while we are interested in exploring the fundamentals of a marriage between local detector-global descriptors and actually manage to obtain a descriptor that outperforms some of the most established approaches from the literature.

III. The proposed method. The SIMPL dEscriptor

The method proposed and described in this section is a straightforward combination of the SURF local-points detector and three of the global MPEG-7 descriptors along with the global CEDD descriptor in a Bag-of-Visual-Words scheme for CBIR. In their essence all four implementations share the same architectural principles. The method utilizes the SURF detector to spot points of interest in images. The SURF detector was preferred for the localization of the images' key points, mainly because, as reported several times in the literature, it is faster than the SIFT detector and more robust against different image transformations.

- ***a. Speeded Up Robust Feature (SURF) detector:*** SURF detects points-of-interest in an image and describes them extracting orientation information. The SURF detector uses the determinant of the Hessian to detect both the location and the scale of blob-like structures. The Hessian matrix is approximated, using

a set of box-type filters. The scale space is analyzed by up-scaling the filter size rather than iteratively reducing the image size. Independently of their size, these approximate second-order Gaussian derivatives are evaluated using integral images, significantly speeding up the whole process. The responses are stored in a blob response map, and local maxima are detected and refined using quadratic interpolation.

According to the scale (s) that the points were detected in, we mark a squared area around them with varying sizes ($s \times s$). These image patches are the vocal areas of the image and will be used to identify and describe the image content.

After locating and obtaining the salient image patches we proceed by describing their content with three different descriptors from the MPEG-7 family of global image descriptors and the MPEG-7-like CEDD global descriptor.

Initially, we decided to try and replace the SURF descriptor, which is a descriptor focused on the spatial distribution of gradient information, with the light-weighted SC and CL MPEG-7 descriptors in order to incorporate color information in the patches' descriptions. Color is a very important element for image retrieval [31], [3], [12], [23], [30] tasks.

- ***b. MPEG 7- Scalable Color Descriptor:***
  The scalable color descriptor is a color histogram in a fixed HSV color space achieved through a uniform quantization of the space to 256 bins. An encoding step is performed by a Haar transform, for compression. Then, a number of coefficients is used to represent the descriptor. Its representation is scalable in terms of bin numbers and bits used for accuracy. We followed the default proposed setting of 64 coefficients.
- ***c. MPEG 7- Color Layout Descriptor:***
  The color layout descriptor represents the spatial distribution of the color in images in a compact form. The image is divided into 8×8 discrete blocks and their representative colors in the YCbCr space are extracted. The descriptor is obtained by applying the discrete cosine transformation (DCT) on every block and using its coefficients. The produced descriptor is a 3×64 bin (64-Y, 64-Cb, 64-Cr) representation of the image.

We also decided to experiment with the EH MPEG-7 descriptor, which like the SURF descriptor contains no color information, but describes the visual content based on the achromatic information that edges carry.

- ***d. MPEG 7- Edge Histogram:***
  The edge histogram descriptor represents the spatial distribution of five types of edges in the image. A given image is first subdivided into 4×4 sub-images, and the local edge histogram of five broadly grouped edge types (vertical, horizontal, 45 diagonal, 135 diagonal, and isotropic) is computed. Each edge histogram consists of five bins (one for every edge type). An image subdivided in 16 blocks produces an 80-bins edge descriptor.

Finally we employed the CEDD description method, which combines both color and texture information in a compact manner.

- ***e. Color and Edge Directivity Descriptor:***
  CEDD is originally a global descriptor that divides an image into 1600 rectangular image areas, referred to as Image-Blocks. Those Image-Blocks are then handled independently to extract their color information (through a two staged Fuzzy Histogram Linking procedure that produces a 24-bin color histogram of pre-set colors) and texture information (employing the five digital filters proposed by the MPEG-7 EHD and using a heuristic fuzzy pentagon diagram to threshold the normalized maximum responses so as to form a 6-bin texture vector). The obtained vectors are combined in the end to form the CEDD descriptor of the input image.

All three global MPEG-7 descriptors are adjusted to describe the image patches, detected and defined by the SURF detection mechanism, as though they were autonomous images. Likewise, for the CEDD implementation we divide the patches of interest in a dynamically calculated number of Patch-Blocks, depending on the octave they emerged from during the SURF detection stage and proceed with the calculations as though they were autonomous images.

When all local features from a collection of images have been detected and described by the SIMPL dEscriptor we randomly select a sample of the descriptors to be clustered via the k-means classifier into a preset number of clusters (Visual Words) to form the Codebook. Each image is then described by a histogram of the frequencies of Visual Words that it contains. This is the simplest form of the Bag-Of-Visual-Words model. After the indexing has been completed, both for the image collection and the queries, and before we proceed to test the retrieval effectiveness of the method we apply the weighting schemes [32].

- ***f. Weighting Schemes:***
  We incorporate the common textual term weighting schemes in the BOVW model. 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). A 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. Many times, 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. Last, a normalization can be performed to quantify the similarity between to documents in terms of the cosine similarity of their vector representation. The SMART notation is a compact way to describe combination of weighting schemes in the form of (d.d.d). The first letter denotes the tf weighting method, the second letter denotes the df weighting method, and the third letter specifies the normalization used. Table 1 presents the SMART notation for several $tf.idf$ variants. For more details kindly refer to [12].

The scope of this paper is to present a set of localized MPEG-7 and MPEG-7-like descriptors, in order to test their

TABLE I
SMART NOTATION

tf	df	Normalization
n(natural): $tf_{t,d}$	**n**(no): 1	**n**(none): 1
l(log): $1+log(tf_{t,d})$	**t**(idf): $log(N/df_t)$	**c** (cos): $\frac{1}{\sqrt{w_1^2+\cdots+w_M^2}}$

retrieval effectiveness and performance. Since the SURF detector locates salient blod-like patches in the images, we are replacing the SURF descriptor which also focuses on achromatic information and take advantage of the SCD, the CLD and CEDD description mechanisms to incorporate the local color information of those vocal image patches. We further investigate the performance of the EHD descriptor combined with the SURF detector to test and evaluate the possibility of mix and matching different detection and description methods of image textures. We have kept the rest of the architecture (indexing and retrieval) in the simplest form on purpose, so as to evaluate the proposed descriptors in a fundamental way. The following section presents the experimental set-ups and results in detail.

IV. Experimental Set-ups and Results

To test the proposed descriptors we employed two different well known benchmark datasets. First, experiments were conducted on the UKBench database [33]. UKBench consists of 10200 images arranged in 2250 groups of four images per group. Each group includes depictions of a single object. The images are taken from different viewpoints and slightly different lighting conditions. Only images of the same group are considered relevant. The first 250 images of the first 250 groups were used as queries. Next, we experimented using the UCID database [34]. This database consists of 1338 uncompressed Tagged Image File (TIF) format images on a variety of topics, including natural scenes and man-made objects. Manual relevance assessments among all database images are provided. UCID, includes several query images where the ground truth consists of images with similar visual concept to the query image, without necessarily the co-occurrence of the same objects. Global descriptors, due to the nature of the UCID database, are reported to perform better than local feature descriptors using the BOVW model [12].

Our four proposed descriptors (SIMPLE-SC, SIMPLE-CL, SIMPLE-EH and SIMPLE-CEDD or LoCATe) along with five well established local features descriptors from the literature (SURF, SIFT, opponent-SIFT [5] ORB [35] and BRISK [36]) were tested (using the recently proposed GRIRe [37] open source framework and the OpenCV implementation of the descriptors) for four different vocabulary sizes (32, 128, 512, 2048) that emerged from the k-means classifier randomly employing 15% of extracted features. Results were obtained using 8 different weighting schemes (as illustrated in Table 1). We also conducted experiments for 7 global descriptors (using the img(Rummager) [38] application), including of course the original MPEG-7 SCD, CLD, EHD[2] and CEDD. To evaluate the systems' performance, the precision-at-position, the Mean Average Precision (MAP) and the Average Normalized Modified Retrieval Rank (ANMRR) are calculated [40]. The significance of the results was evaluated with a bootstrap test, one-tailed, at significance levels 0.05 (↑), 0.01 (↑') and 0.001(↑") against the 'baseline'. In each experiment, we assumed as baseline the best performance that can be obtained employing a non-SIMPLE descriptor.

Table 2 and Table 3, present the experimental results on the UKBench and the UCID collections, respectively. Please note that for each local descriptor and for each codebook size, the experiment was repeated for all 8 weighting schemes but only the best obtained result is listed in the tables.

TABLE II
Experimental Results on the UKBench Database.

Descriptor	Size	WS	MAP	P@4	ANMRR
SIMPLE-SC	512	l.t.c.	0.9145↑"	0.8960↑"	0.0713↑"
LoCATe	512	l.t.c.	0.8964↑"	0.8670↑"	0.0879↑"
SIMPLE-SC	128	l.t.c.	0.8941↑"	0.8640↑"	0.0858↑"
SIMPLE-SC	2048	l.t.c.	0.8730↑"	0.8180↑"	0.0871↑"
LoCATe	128	l.n.c.	0.8665↑"	0.8260↑"	0.1104↑"
SIMPLE-CL	512	l.t.n.	0.8446↑'	0.7710-	0.1333-
LoCATe	2048	l.t.c.	0.8280-	0.7580-	0.1207↑'
SURF(baseline)	512	l.n.n.	0.8159	0.7730	0.1535
SIMPLE-CL	128	l.n.n.	0.8112	0.7640	0.1576
CEDD	Global		0.8026	0.7630	0.1690
SIMPLE-SC	32	l.n.n.	0.7956	0.7420	0.1672
LoCATe	32	l.n.n.	0.7806	0.7250	0.1771
SIMPLE-CL	2048	l.n.c.	0.7693	0.6890	0.1706
SURF	128	l.n.n.	0.7634	0.7120	0.1983
Oppo. SIFT	128	n.n.c.	0.7475	0.7010	0.2178
Oppo. SIFT	512	n.n.c.	0.7390	0.6980	0.2243
SIFT	512	l.n.n.	0.6984	0.6580	0.2672
SURF	2048	n.c.c.	0.6911	0.6530	0.2691
SIFT	128	l.n.n.	0.6903	0.6330	0.2642
SIMPLE-CL	32	n.n.n.	0.6857	0.6290	0.2725
SIFT	2048	n.n.c.	0.6638	0.6320	0.2895
Oppo. SIFT	32	n.n.n.	0.6613	0.6070	0.2900
BTDH [41]	Global		0.6468	0.6150	0.3196
SURF	32	l.n.n.	0.6377	0.5840	0.3165
MPEG-7 CLD	Global		0.6181	0.5760	0.3366
Oppo. SIFT	2048	n.t.c.	0.5926	0.4610	0.2935
SIFT	32	l.n.c.	0.5683	0.5230	0.3853
ORB	512	n.n.c.	0.5371	0.4990	0.4191
MPEG-7 EHD	Global		0.5271	0.4890	0.4320
ORB	2048	n.t.c.	0.4913	0.4730	0.4128
ORB	128	n.n.c.	0.4830	0.4410	0.4694
MPEG-7 SCD	Global		0.4716	0.4160	0.4813
SIMPLE-EH	512	n.n.c.	0.4276	0.4010	0.5321
Color Hist.	Global		0.4133	0.3850	0.5398
SIMPLE-EH	2048	n.n.n.	0.4093	0.3280	0.5422
SIMPLE-EH	128	n.n.n.	0.3972	0.3760	0.5590
BRISK	128	l.n.n.	0.3904	0.3570	0.5694
ORB	32	n.n.n.	0.3880	0.3560	0.5656
SIMPLE-EH	32	n.n.n.	0.3570	0.3330	0.5987
BRISK	32	n.n.n.	0.3550	0.3190	0.5979
BRISK	512	l.n.n.	0.3463	0.3240	0.6166
Tamura [42]	Global		0.3130	0.2950	0.6582
BRISK	2048	n.n.c.	0.3096	0.2900	0.6524

The experimental results on the UKBench collection confirmed our initial forecast that enriching the salient patches

[2]It is worth noting that the MPEG-7 descriptors, both, in case of SIMPLE as well as in case of image retrieval using the global form of the descriptors are computed using the LIRE [39] open source library

TABLE III
EXPERIMENTAL RESULTS ON THE UCID DATABASE

Descriptor	Size	WS	MAP	P@10	ANMRR
LoCATe	2048	l.t.c.	0.7811↑"	0.2590↑"	0.1892↑"
SIMPLE-SC	2048	l.t.c.	0.7718↑"	0.2550↑"	0.1968↑"
LoCATe	512	l.t.c.	0.7635↑"	0.2531↑"	0.2054↑"
SIMPLE-SC	512	l.t.c.	0.7648↑"	0.2515↑"	0.2010↑"
LoCATe	128	l.n.n.	0.7332↑"	0.2447↑"	0.2260↑"
SIMPLE-SC	128	l.t.c.	0.7275↑"	0.2382↑	0.2355↑"
SIMPLE-CL	2048	l.t.c.	0.7161↑"	0.2393↑	0.2502↑'
SIMPLE-CL	512	l.n.n.	0.6765-	0.2225-	0.2829-
CEDD(baseline)	Global		0.6748	0.2267	0.2823
LoCATe	32	l.n.n.	0.6570	0.2206	0.2954
SURF	512	l.n.n.	0.6513	0.2088	0.3113
SIMPLE-SC	32	l.n.n.	0.6450	0.2095	0.3118
SIMPLE-CL	128	l.n.n.	0.6291	0.2073	0.3288
SIFT	512	l.n.n.	0.6261	0.2034	0.3353
SURF	2048	l.n.c.	0.6259	0.2011	0.3387
Oppo. SIFT	2048	n.t.c.	0.6244	0.2050	0.3383
Oppo. SIFT	512	n.n.c.	0.6072	0.1962	0.3579
SIFT	2048	n.n.c.	0.6046	0.1943	0.3610
SURF	128	n.n.c.	0.5927	0.1889	0.3687
Oppo. SIFT	128	n.n.c.	0.5872	0.1866	0.3746
SIFT	128	n.n.c.	0.5849	0.1874	0.3752
SIMPLE-CL	32	n.n.n.	0.5610	0.1809	0.3994
SURF	32	l.n.n.	0.5492	0.1763	0.4102
SIFT	32	n.n.c.	0.5453	0.1725	0.4168
MPEG-7 CLD	Global		0.5361	0.1702	0.4322
BTDH	Global		0.5353	0.1676	0.4295
MPEG-7 EHD	Global		0.5326	0.1687	0.4331
Oppo. SIFT	32	n.n.n.	0.5240	0.1664	0.4361
SIMPLE-EH	512	n.n.c.	0.5066	0.1576	0.4609
SIMPLE-EH	2048	n.n.c.	0.5030	0.1599	0.4600
MPEG-7 SCD	Global		0.4998	0.1565	0.4667
SIMPLE-EH	128	n.n.c.	0.4973	0.1553	0.4644
ORB	512	l.n.n.	0.4929	0.1504	0.4746
ORB	2048	n.n.c.	0.4913	0.1485	0.4814
SIMPLE-EH	32	n.n.c.	0.4682	0.1450	0.4948
ORB	128	n.n.c.	0.4642	0.1397	0.5052
BRISK	128	l.n.n.	0.4636	0.1385	0.5070
BRISK	32	n.n.n.	0.4532	0.1370	0.5107
Color Hist.	Global		0.4443	0.1328	0.5231
Tamura	Global		0.4411	0.1317	0.5304
BRISK	2048	n.t.c.	0.4360	0.1328	0.5352
ORB	32	n.n.c.	0.4360	0.1298	0.5332
BRISK	512	l.n.n.	0.4345	0.1347	0.5352

detected by SURF with color information, would lead to an increase in performance. SIMPLE-SC, in particular, managed to significantly outperform all other local and global descriptors for 3 out of 4 codebook sizes. Our best performing SIMPLE descriptor improves MAP by 12% (compared to SURF 512 baseline), P@4 by 16% and ANMRR by 53%. LoCATE and SIMPLE-CL also showed consistent high performance. SIMPLE-EH did not produce the desired results in this collection for any codebook size. The baseline descriptor (SURF, line 5 in table 2) compared to the respective SIMPLE-EH (512 codebook) presents almost double the performance according to MAP.

The results on the UCID database are also very interesting. Again, the proposed LoCATe, SIMPLE-SC and SIMPLE-CL descriptors outperform the next best reported descriptor (which in this case is the CEDD global descriptor). In particular LoCATE and SIMPLE-SC 2048 increase MAP by 14%, P@10 by 12% and ANMRR by 30%. Both descriptors even with a tiny codebook of 32 visual words perform comparable to the best global and local features descriptors (CEDD, SURF 512). SIMPLE-EH seems to perform slightly better in this collection than in UKBench but still fails to even improve the original global EHD or SURF descriptors that it emerged from.

V. DISCUSSION AND CONCLUSIONS

Four novel descriptors were presented in this paper. Based on the simple idea that we could adopt a salient region detector that searches for vocal textural information in an image in multiple scales and then proceed with a descriptor traditionally used for global image descriptions, to individually treat and describe the visual information of those image patches, we were able to obtain a set of powerful yet light-weighted local features descriptors. The SIMPLE-SC and SIMPLE-CL demonstrated excellent performance. What made them so successful is that they provide color information with textural attention. The SURF detector points out the salient textural parts and only those are then described with color information. The retrieval is effective, even though there is no textural description in the final vector representations thus, keeping the vectors relatively small. It is worth noticing that since SIMPLE-SC incorporates color information without spatial distribution, it is rotation invariant and probably, this is one of the keys to its success. On the other hand, SIMPLE-EH descriptor, did not manage to present remarkable performance, or surpass the performance of the original descriptors it emerged from. We still decided to include it in this study, merely for comparison reasons and reference. For instance, authors in [30] could exclude it from their four-signature image representation or replace it with the actual SURF descriptor, who performs better. SIMPLE-CEDD (or LoCATe) which has both local color and texture information also performs exceptionally good. We believe this descriptor to actually be the most retrieval friendly of all our proposed descriptors. The elements that are used to classify an image patch as similar to another (i.e. its color/hue combined with texture) are quantized in discreet areas and thus bring similar visual content of patches closer together in terms of vector distances, while at the same time distance descriptions of different contents further away. In contrast, most local descriptors in the literature, try to describe the key point as detailed and uniquely as possible, forcing retrieval systems that employ them to often fail, due to many possible matching candidates whose vector distance is not apparent.

For future work we plan to combine different local feature detectors with the MPEG-7 and MPEG-7-like global descriptors and/or index images with some or all four SIMPLE descriptors employing some late fusion method. This leads us to our final note. In this paper we tried to explore the possibility of mix and match local-feature and global descriptors, starting with the most popular in their respective field of application. Apart from the obvious contribution that this paper makes, i.e. the introduction of a new set of very effective local feature descriptors for CBIR, we also like to think that, by stripping all the rest of the procedures for

indexing and retrieval to the absolute necessary, we provide some insight to the fundamentals and the effectiveness of combining local and global descriptors.

Although simple in concept, the idea of revisiting global features' description methods and localizing their effect by applying them on local neighborhoods of salient image regions located by detectors from the literature, could introduce a whole new generation of hybrid approaches that revive or extend older techniques. In particular, future approaches that employ the MPEG-7 global descriptors with varying POI detectors or fusion techniques can become part of the SIMPLE family proposed here on our first such attempt. Open source c#, java and MATLAB implementations of the proposed descriptors can be found at http://tinyurl.com/SIMPLE-Descriptors

ACKNOWLEDGMENTS

This research has been co-financed by the European Union (European Social Fund-ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF)- Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund.

REFERENCES

[1] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comput. Surv.*, vol. 40, no. 2, 2008.

[2] T. Deselaers, D. Keysers, and H. Ney, "Features for image retrieval: an experimental comparison," *Inf. Retr.*, vol. 11, no. 2, pp. 77–107, 2008.

[3] O. A. B. Penatti and R. da Silva Torres, "Color descriptors for web image retrieval: A comparative study," in *SIBGRAPI*, 2008, pp. 163–170.

[4] J. Annesley, J. Orwell, and J.-P. Renno, "Evaluation of mpeg7 color descriptors for visual surveillance retrieval," in *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*. IEEE, 2005, pp. 105–112.

[5] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1582–1596, 2010.

[6] S. Loncaric, "A survey of shape analysis techniques," *Pattern Recognition*, vol. 31, no. 8, pp. 983–1001, 1998.

[7] D. Zhang and G. Lu, "Review of shape representation and description techniques," *Pattern Recognition*, vol. 37, no. 1, pp. 1 – 19, 2004.

[8] M. Safar, C. Shahabi, and X. Sun, "Image retrieval by shape: A comparative study," in *IEEE International Conference on Multimedia and Expo (I)*, 2000, pp. 141–144.

[9] F. Xu and Y.-J. Zhang, "Evaluation and comparison of texture descriptors proposed in mpeg-7," *Journal of Visual Communication and Image Representation*, vol. 17, no. 4, pp. 701–716, 2006.

[10] P. Howarth and S. M. Rüger, "Evaluation of texture features for content-based image retrieval," in *CIVR*, 2004, pp. 326–334.

[11] M. Aly, P. Welinder, M. E. Munich, and P. Perona, "Automatic discovery of image families: Global vs. local features," in *ICIP*, 2009, pp. 777–780.

[12] S. A. Chatzichristofis, C. Iakovidou, Y. S. Boutalis, and O. Marques, "Co.vi.wo.: Color visual words based on non-predefined size codebooks," *IEEE T. Cybernetics*, vol. 43, no. 1, pp. 192–205, 2013.

[13] C. G. Harris and J. Pike, "3d positional integration from image sequences," *Image and Vision Computing*, vol. 6, no. 2, pp. 87–90, 1988.

[14] J. Shi and C. Tomasi, "Good features to track," in *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*. IEEE, 1994, pp. 593–600.

[15] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *ECCV (1)*, 2006, pp. 430–443.

[16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[17] H. Bay, T. Tuytelaars, and L. J. V. Gool, "Surf: Speeded up robust features," in *ECCV (1)*, 2006, pp. 404–417.

[18] D. Scaramuzza and F. Fraundorfer, "Visual odometry [tutorial]," *IEEE Robot. Automat. Mag.*, vol. 18, no. 4, pp. 80–92, 2011.

[19] R. Siegwart, I. R. Nourbakhsh, and D. Scaramuzza, *Introduction to autonomous mobile robots*. MIT press, 2011.

[20] O. G. Cula and K. J. Dana, "Compact representation of bidirectional texture functions," in *CVPR (1)*, 2001, pp. 1041–1047.

[21] Y. Chen, X. Li, A. Dick, and R. Hill, "Ranking consistency for image matching and object retrieval," *Pattern Recognition*, vol. 47, no. 3, pp. 1349–1360, 2014.

[22] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada, "Color and texture descriptors," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 11, no. 6, pp. 703–715, 2001.

[23] S. A. Chatzichristofis and Y. S. Boutalis, "Cedd: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval," in *ICVS*, 2008, pp. 312–322.

[24] E. Spyrou, H. L. Borgne, T. P. Mailis, E. Cooke, Y. S. Avrithis, and N. E. O'Connor, "Fusing mpeg-7 visual descriptors for image classification," in *ICANN (2)*, 2005, pp. 847–852.

[25] M. M. Rahman, S. Antani, and G. R. Thoma, "A classification-driven similarity matching framework for retrieval of biomedical images," in *Multimedia Information Retrieval*, 2010, pp. 147–154.

[26] ——, "A medical image retrieval framework in correlation enhanced visual concept feature space," in *CBMS*, 2009, pp. 1–4.

[27] S. A. Chatzichristofis and Y. S. Boutalis, "Fcth: Fuzzy color and texture histogram - a low level feature for accurate image retrieval," in *WIAMIS*, 2008, pp. 191–196.

[28] G. Amato, P. Bolettieri, F. Falchi, C. Gennaro, and F. Rabitti, "Combining local and global visual feature similarity using a text search engine," in *CBMI*, 2011, pp. 49–54.

[29] A. R. Doherty, C. O. Conaire, M. Blighe, A. F. Smeaton, and N. E. O'Connor, "Combining image descriptors to effectively retrieve events from visual lifelogs," in *Multimedia Information Retrieval*, 2008, pp. 10–17.

[30] A. Abdullah, R. C. Veltkamp, and M. A. Wiering, "Fixed partitioning and salient points with mpeg-7 cluster correlograms for image categorization," *Pattern Recognition*, vol. 43, no. 3, pp. 650–662, 2010.

[31] K. Bowyer and P. Flynn, "A 20th anniversary survey: Introduction to 'content-based image retrieval at the end of the early years'," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, p. 1348, 2000.

[32] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge University Press, 2008.

[33] D. Nistér and H. Stewénius, "Scalable recognition with a vocabulary tree," in *CVPR (2)*, 2006, pp. 2161–2168.

[34] G. Schaefer and M. Stich, "Ucid: an uncompressed color image database," in *Storage and Retrieval Methods and Applications for Multimedia*, 2004, pp. 472–480.

[35] E. Rublee, V. Rabaud, K. Konolige, and G. R. Bradski, "Orb: An efficient alternative to sift or surf," in *ICCV*, 2011, pp. 2564–2571.

[36] S. Leutenegger, M. Chli, and R. Siegwart, "Brisk: Binary robust invariant scalable keypoints," in *ICCV*, 2011, pp. 2548–2555.

[37] S. A. Chatzichristofis, Y. S. Boutalis, and M. Lux, "Img(rummager): An interactive content based image retrieval system," in *SISAP*, 2009, pp. 151–153.

[38] L. T. Tsochatzidis, C. Iakovidou, S. A. Chatzichristofis, and Y. S. Boutalis, "Golden retriever: a java based open source image retrieval engine," in *ACM Multimedia*, 2013, pp. 847–850.

[39] M. Lux and S. A. Chatzichristofis, "Lire: lucene image retrieval: an extensible java cbir library," in *ACM Multimedia*, 2008, pp. 1085–1088.

[40] S. A. Chatzichristofis, C. Iakovidou, Y. S. Boutalis, and E. Angelopoulou, "Mean normalized retrieval order (mnro): a new content-based image retrieval performance measure," *Multimedia Tools and Applications*, pp. 1–32, 2012.

[41] S. A. Chatzichristofis and Y. S. Boutalis, "Content based radiology image retrieval using a fuzzy rule based scalable composite descriptor," *Multimedia Tools Appl.*, vol. 46, no. 2-3, pp. 493–519, 2010.

[42] H. Tamura, S. Mori, and T. Yamawaki, "Textural features corresponding to visual perception," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 8, no. 6, pp. 460–473, 1978.