Digital Signal Processing 44 (2015) 1–13

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Local detectors and compact descriptors for visual search: A quantitative comparison

S. Bianco [a,*], D. Mazzini [a], D.P. Pau [b], R. Schettini [a]

[a] Department of Informatics, Systems and Communication, University of Milano, Bicocca, Italy
[b] STMicroelectronics, Agrate, Italy

ARTICLE INFO

Article history:
Available online 12 June 2015

Keywords:
Visual search
MPEG CDVS
Compact descriptors
Local descriptors
Color descriptors

ABSTRACT

Local Visual detectors and descriptors have been studied for many years, but their applications (e.g. mobile visual search) in large volume, low-cost, low-power embedded systems have been limited or negligible to date. One reason is the lack of a worldwide industry standard. MPEG Compact Descriptors for Visual Search (CDVS) working group filled this gap by defining a high-performance extraction stage and the bitstream syntax at its output in order to achieve interoperability between different implementations of clients and servers. In a previous work, we presented an analysis of various gray-level interest point detection and description algorithms, which was also contributed to CDVS. This work extends the previous analysis using the MPEG CDVS Test Model framework to consider additional detectors and the use of color descriptors.

1. Introduction

Searching content among billions of images is a very complex task. Text-based, low-level and semantic approaches are widely used but have proven to be less than satisfactory when dealing with massive amount of data stored on the server side, that may be dynamically and frequently updated, which led to the processing of queries by other means, such as the visual ones at the core of Content Based Image Recognition (CBIR). CBIR covers the concept of object search which analyzes the visual content in the image, rather than relying on metadata. Many algorithms and techniques from fields such as statistics, pattern recognition and computer vision were incorporated into CBIR. CBIR has attracted a lot of attention after many years of research and is expanding into the marketplace. CBIR's adaptation to smartphones and tablet context, named Mobile Visual Search (MVS) [1] presents a much more intuitive, seamless, direct way of retrieving information, thus acting as required technology to enable Augmented Reality (AR) with a completely new perspective: pixels are representing "a kind of natural marker". The idea is to interact with the instance of an object itself without intermediation of explicit markers [2].

As a simple scenario, the user takes a photo of a rigid, man-made object and retrieves content and information about it from a remote server, in the form of audio, video, and 3D graphics augmentations [3,4].

Many state-of-the-art algorithms are available to achieve the goal of local visual feature extraction and compression. However, in order to avoid confusing adopters and implementers of Mobile Visual Search client and server systems, and to enable interoperability between them and visual search applications, given such a variety of methods, the Moving Picture Experts Group (MPEG) started in 2010 a standardization initiative called Compact Descriptors for Visual Search (CDVS) [5]. CDVS specifies two procedures for descriptor comparison in visual search systems, pairwise matching and retrieval, that can be implemented as two pipelines in real systems. The pairwise matching pipeline automatically verifies whether two images depict the same objects or scene. The retrieval pipeline accomplishes the search and match of images contained in a large collection that depict the same objects or scenes as those depicted by a query image.

This paper investigates the use of different detectors and color descriptors in the CDVS framework. We demonstrate the advantages of using color descriptors on both the five standard CDVS datasets and on a new dataset introduced here. The rest of the paper is organized as follows: Section 3 presents the detectors and descriptors used for comparison, Section 4 describes the experimental setup, while Section 5 introduces the six datasets used. Experimental results are reported in Section 6, and finally in Section 7 conclusions are drawn.

* Corresponding author.
E-mail addresses: bianco@disco.unimib.it (S. Bianco), mazzini@disco.unimib.it (D. Mazzini), danilo.pau@st.com (D.P. Pau), schettini@disco.unimib.it (R. Schettini).

http://dx.doi.org/10.1016/j.dsp.2015.06.001
1051-2004/

2. CDVS Test Model

CDVS [5] is a technology in the last phase of the ISO standardization process that will enable the design of efficient and interoperable visual search applications and in particular the development of technologies for visual content matching from still images. Visual content matching includes matching of views of objects, landmarks, and printed documents that is robust to partial occlusions as well as changes in vantage point, camera parameters, and lighting conditions. It has the goal of defining a standard bitstream, which encodes in compressed form the information required to perform a search on the server's side. The information encoded consists of a global descriptor, which is a digest computed from compact descriptors features extracted from the image, a compressed descriptor and the associated coordinates.

The CDVS Test Model (TM) [5] implements the required functionality for the extraction and comparison of compact descriptors constrained to a set of predetermined descriptor lengths.

In particular, two procedures for descriptor comparison are implemented in TM, aiming at reproducing two fundamental tasks for real visual search systems: pairwise matching and retrieval. The former regards automated verification of whether two images depict the same objects or scene; in this case, descriptors extracted from a query image are matched against the descriptors of a reference image, in order to determine whether they match or not. The latter regards the search and discovery of images contained within a large collection that depict the same objects or scenes as those depicted by a query image; this requires the database images to be processed for the creation of a database which may be searched using the descriptors extracted from the query.

The Pairwise Matching Stage compares the query and reference image descriptors to determine if the images depict the same object or scene. It uses first local descriptor matching and if the score is below a threshold, it performs global descriptor matching. If the final score is greater than a threshold, the images likely depict the same objects (a match), otherwise the object are different (a non-match).

The Retrieval Stage searches and retrieves relevant images, belonging to a large collection, that depict the same object or scene represented in the query image. At first, an off-line step processes the collection to create a database of local and global visual descriptors which can be matched against the descriptors extracted on the fly from the query. The retrieval stage performs a search in two steps, first using global descriptor to select a shortlist of matching images and then using the shortlist in the next step to compare encoded local descriptor using the Hamming distance. The final ranking score and inlier selection is computed by a geometric consistency check performed to determine the inliers among the interest point matches for the two images. The TM uses the histogram of logarithmic distance ratios (LDR) [6].

3. Algorithms considered for comparison

3.1. Gray-level algorithms

In this section we describe the set of gray-level algorithms evaluated in our previous work [7]. LoG (Laplacian of Gaussian) and SIFT (Scale-Invariant Feature Transform) [8,9] were the algorithms adopted, respectively, for keypoint detection and local visual description in the CDVS Test Model ver. 10 [10]. The following implementations have been considered for testing: Original Lowe's binary code [8,9]; VLFeat [11] and the OpenCV library implementation [12]. Different affine-invariant keypoint detectors [13] were investigated together with the VLFeat SIFT descriptor: Hessian, Hessian Laplace, Multiscale Hessian, Harris–Laplace, Multiscale Harris and Difference of Gaussians (DoG). These detectors normalize the image patch around each detected interest point according to the estimated affine transformation. They also differ in the strategy used to achieve scale-invariance: Laplace automatically selects a single characteristic scale of a keypoint, whereas Multiscale may associate multiple scales to the same keypoint.

Table 1
Gray-levels algorithms investigated.

Name	Patent	Reference
SIFT detector and descriptor	Patent US 6711293 B	[9]
SURF detector and descriptor	Patent US 8165401 B	[14]
Harris detector	Free	[13]
Hessian detector	Free	[13]
FREAK descriptor	Free	[15]
KAZE detector and descriptor	Free	[16]
A-KAZE detector and descriptor	Free	[17]

Table 2
Evaluated color descriptors.

Name	Dimension	Fusion	Reference
RGB SIFT	384	Early	[18]
Opponent SIFT	384	Early	[18]
Transformed Color SIFT	384	Early	[18]
HSV SIFT	384	Early	[18]
C-SIFT	384	Early	[18,19]
rg SIFT	256	Early	[18]
oRGB SIFT	384	Early	[20]
Hue SIFT	164	Late	[18]
Color Names	139	Late	[21,22]
Fuzzy Sets Color Names	139	Late	[23]
Discriminative Color	139, 153, 178	Late	[24]

The other gray-levels algorithms that were considered are: the OpenCV SURF (Speeded-Up Robust Features) [14] implementation; the OpenCV FREAK (Fast Retina Keypoint) [15] implementation which uses SURF as the keypoint detector, and the KAZE [16] and A-KAZE [17] original code.

All tests were made with the default parameters for each algorithm with the exception of the response filter threshold. This parameter was set so each detector produced on average about 1000 keypoints per VGA image, therefore controlling the number of interest points generated. Indeed the distinctiveness of the descriptors produced was analyzed instead of their quantity; therefore their quantity was limited to make fair comparisons. Moreover, in the CDVS processing pipeline the detected keypoints and their descriptors must be sent over the network to a server and, since the network is bandwidth constrained, it was reasonable to limit the maximum number of keypoints produced as an attempt to fit their binary representation into the network bitrate available. Table 1 lists the investigated gray-levels algorithms with their licenses and links to sources or binaries.

3.2. Color descriptors

As an extension to the quantitative comparison presented in the previous work [7], and to evaluate the impact on the accuracy of the visual descriptors when used on the color components, further experiments using color descriptors have been carried out. Table 2 lists all the algorithms considered for comparison.

Color descriptors can be classified in two classes depending on the approach used to combine the shape (luminance) and color information [18,25,26]. Some algorithms make use of an Early Fusion approach and others adopt a Late Fusion approach. As defined by Khan et al. [26]: "Early fusion combines shape and color at the pixel level, which are then processed together throughout the rest of the description pipeline. In late fusion, shape and color are described separately from the beginning and the exact binding between the two features is lost."

Basically, every Early Fusion description pipeline considered consists of the following steps:

1. Transformation of the image channels into a specific color space.
2. Computation of SIFT descriptor on each color space channel.
3. Concatenation of descriptions computed over each channel.

Specifically RGB SIFT [18] computes SIFT descriptors on the original red, green, and blue channels of the image and then concatenates them, thus keeping the image in its original color space. Opponent SIFT [18] instead applies the following transformation from RGB to opponent color space $O_1O_2O_3$:

$$\begin{pmatrix} O_1 \\ O_2 \\ O_3 \end{pmatrix} = \begin{pmatrix} \frac{R-G}{\sqrt{2}} \\ \frac{R+G-2B}{\sqrt{6}} \\ \frac{R+G+B}{\sqrt{3}} \end{pmatrix} \tag{1}$$

Transformed Color SIFT [18] normalizes the RGB channels independently into zero-mean and unity-variance $R'G'B'$ channels:

$$\begin{pmatrix} R' \\ G' \\ B' \end{pmatrix} = \begin{pmatrix} \frac{R-\mu_R}{\sigma_R} \\ \frac{G-\mu_G}{\sigma_G} \\ \frac{B-\mu_B}{\sigma_B} \end{pmatrix} \tag{2}$$

where μ_C is the mean and σ_C the standard deviation of the distribution in channel $C = \{R, G, B\}$. HSV SIFT [18] computes SIFT descriptors over all three channels of the HSV color model:

$$\begin{pmatrix} H \\ S \\ V \end{pmatrix} = \begin{pmatrix} \text{atan2}(\beta, \alpha) \\ 0 \text{ if } V = 0, \frac{\sqrt{\alpha^2+\beta^2}}{V} \text{ otherwise} \\ \max(R, G, B) \end{pmatrix} \tag{3}$$

where $\alpha = (2R - G - B)/2$ and $\beta = \sqrt{3}(G - B)/2$. C-SIFT [18,19] applies the C-invariant [27] to the O_1 and O_2 channels of the opponent color space to eliminate the remaining intensity information from these channels. This can be intuitively seen as the normalized opponent color space O_1/O_3 and O_2/O_3. The rg SIFT transforms the image in the normalized RGB color model, where the chromaticity components r and g describe the color information (b is omitted since it is redundant as $r + g + b = 1$):

$$\begin{pmatrix} r \\ g \\ b \end{pmatrix} = \begin{pmatrix} \frac{R}{R+G+B} \\ \frac{G}{R+G+B} \\ \frac{B}{R+G+B} \end{pmatrix} \tag{4}$$

The last Early Fusion color descriptor considered, i.e. oRGB SIFT [20], maps the image into oRGB color space, which is an opponent color space that is ideal for RGB computation [28]. The mapping consists in two steps: the first one is a linear transformation from RGB to LC_1C_2:

$$\begin{pmatrix} L \\ C_1 \\ C_2 \end{pmatrix} = \begin{pmatrix} 0.299 & 0.587 & 0.114 \\ 0.500 & 0.500 & -1.000 \\ 0.866 & -0.866 & 0.000 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} \tag{5}$$

The second one is the transformation from LC_1C_2 to oRGB, which consists in a compression or decompression of angles depending on which quadrant the linearly transformed point ends up in [28].

Conversely, every Late Fusion description pipeline extracts the shape information from the gray-level image. Then, the color descriptor is computed directly from the original image. Finally, the shape and color descriptions are merged as follows:

1. Normalization of the two parts separately (color and shape descriptions).
2. Multiplication by a fusion factor depending on the specific descriptor.
3. Concatenation (of the color and shape descriptors).
4. Normalization of the overall description.

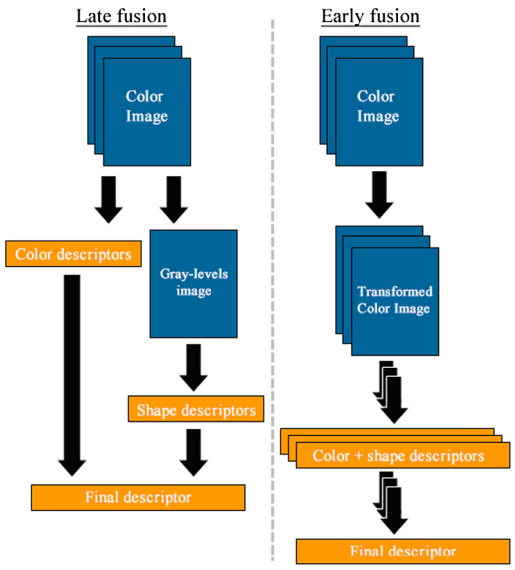

Fig. 1. Late Fusion vs Early Fusion. The Late Fusion approach computes the color descriptors from the original image and the shape descriptors from a gray-levels image; while the Early Fusion approach computes all the descriptors from every color channel (shape and color information are correlated) and then merges them.

Fig. 1 depicts the pipelines of the Early and Late Fusion approaches.

All the Late Fusion descriptors, with the exception of the Hue SIFT algorithm [18], require a prior training phase which determines a quantization function to map each RGB pixel value into a probability vector defining the likelihood of an RGB pixel value to represent a certain color. Every algorithm is characterized by a unique and specific methodology to build the map function. Hue SIFT introduces a concatenation of the robustified hue histogram with the SIFT descriptor. The hue histogram is made more robust by weighing each sample of the hue by its saturation, since the certainty of the hue is inversely proportional to the saturation. Color Names descriptor [22] is trained from weakly labeled images returned by Google Image search. Fuzzy Sets Color Names function [23] trains by using parametric membership functions defined on the basis of psychophysical data obtained from a color-naming experiment. Both the methods based on Color Names are inspired from [26]. Each keypoint is described with a gray-scale SIFT concatenated with the Color Names descriptor computed as follows: the image patch centered on the keypoint (whose width depends on the keypoint scale) is scaled to a fixed size, and for each pixel the Color Name descriptor is computed. The descriptors are weighted by a Gaussian peaked on the center of the resized patch, and then normalized. The Discriminative Color Descriptor [24] performs its training phase by using a statistical method based on information theory: it learns color descriptors which have optimal discriminative power for a specific classification problem. The problem of learning a color descriptor is solved by finding a partition of the color space using the Divisive Information-Theoretic Clustering (DITC) [29].

The computational complexity of color descriptors strongly depends on the fusion approach adopted. Every Early Fusion algorithm reveals a complexity about three times higher than that of the SIFT descriptor computed on the same gray-levels image (except for the rg-SIFT which is about two times higher). The com-

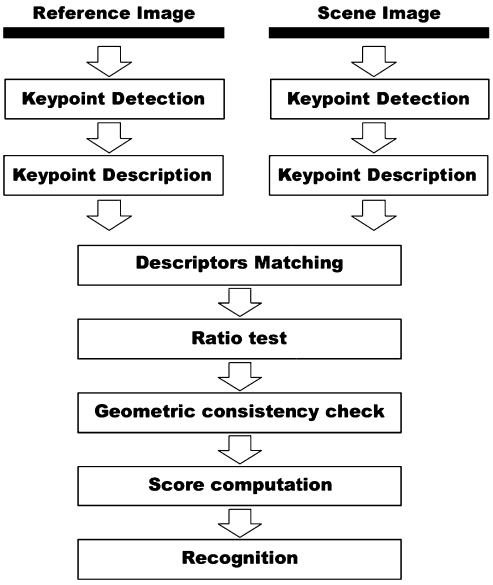

Fig. 2. Evaluation pipeline.

putational complexity of every Late Fusion algorithm is consists of the sum of the complexities of the following steps:

1. SIFT algorithm (computed on the gray-levels image).
2. Pyramid construction.
3. Color quantization by means of a lookup-table (with the complexity depending on the number of keypoints and their size).

Because the MPEG CDVS local visual descriptor is based on SIFT, and because added complexity and memory must be minimized to avoid increasing the cost of embedded systems, the Late Fusion algorithms offer a much better opportunity to limit the extra complexity than the Early Fusion methods, since lookup tables are much cheaper to implement using Read Only Memories (ROMs).

In Table 2, for every color descriptor evaluated, the type of fusion approach and the dimension (in bytes) of the color description produced is reported. To perform fair comparisons, every color descriptor was associated with the same keypoints detector (i.e. the VLFeat implementation of SIFT). Detection phase was performed on the gray-levels image.

4. Experimental setup

A Pairwise Matching Experiment was performed on different datasets in the framework of the CDVS Test Model [10]. In this experiment, a reference image was compared with an image describing the same object (or scene) taken under different illumination conditions, with different acquisition devices, and from different points of view. Fig. 2 shows the complete evaluation pipeline.

Descriptors matching was accomplished by Euclidean distance for integer descriptors (i.e. SIFT, SURF, KAZE, A-KAZE) and by Hamming distance for binary descriptors (i.e. FREAK). The ratio test on candidate matches was the same proposed by Lowe [9]. Let a_1 be a descriptor from image A, b_1 and b_2 the most and second most similar descriptors from image B, and $dist(\cdot,\cdot)$ the distance function between two descriptors. If:

$$\frac{dist(a_1,b_1)}{dist(a_1,b_2)} < threshold \tag{6}$$

then the candidate match (a_1,b_1) was accepted as the best match candidate, otherwise it was rejected. The value of threshold was empirically set. All the accepted matches were further evaluated by a geometric consistency check. This was done using the Logarithmic Distance Ratio [6] algorithm.

Table 3
Datasets adopted in the MPEG CDVS test model.

Dataset	MP	NMP	Origin
1. CDs, DVDs, books, business cards (Mixed text + graphics)	3000	29,903	Stanford Mobile Visual Search
2. Museum paintings	363	3639	Stanford MVS
3. Video frames	399	3999	Stanford MVS
4. Landmarks and buildings (Zurich, Turin)	1789	17,948	ZuBud, Telecom Italia
5. Common object or scenes	2549	21,307	University of Kentucky

The image pair scores were computed using all the geometric consistent matches. First, an image correspondence score based on all matches was computed as follows:

$$w = \sum_{\text{all matches}} \left(\cos\left(\frac{\pi}{2} \sqrt{\frac{dist_1}{dist_2}} \right) \right) \tag{7}$$

where $dist_1$ and $dist_2$ are the best distance and the second best distance for each match. Then, the image correspondence score was transformed by mean of the following equation to obtain the final Image Pair Score (IPS):

$$IPS = \frac{w}{w + wm_{Thresh}} \tag{8}$$

where wm_{Thresh} is a threshold that was empirically found for each detector/descriptor couple. Image pairs with $IPS > 0.5$ were considered a match.

All images were resized to VGA resolution (i.e. minor and major axis length equal to 480 and 640 pixels respectively). Results were reported in terms of True Positives Rate (TPR) at a given level of False Positives Rate (FPR) (also called False Accept Rate or FAR).

5. Datasets

For the gray-level descriptors comparison, the experiments were performed on five datasets of different objects, whereas for the color descriptors, the experiments were performed on the five datasets plus a recent additional dataset that we entitled "SuperMarket Milan".

The CDVS standard datasets are listed in Table 3, together with the number of Matching Pairs (MP) and Non-Matching Pairs (NMP) provided as ground-truth.

The "Mixed-text + Graphics", "Museum paintings", and "Video frames" datasets were collected by the Stanford Mobile Visual Search research group. "Mixed-text + Graphics" consists of 2500 images of five different categories of objects: CDs, DVDs, books, text documents and business cards. The "Video frames" dataset contains 500 images captured by a mobile phone camera shooting a TV screen. Pictures in these three datasets have been shot with different cameras and various different lighting conditions, rotations, scales, and viewpoints. Fig. 3 depicts an object from the "Mixed-text + Graphics" dataset.

The "Landmarks and Buildings" dataset includes images from two different origins: "Zurich Buildings" dataset and "Turin" dataset. "Zurich Buildings" contains pictures of 200 buildings in Zurich, shot by two cameras under different viewing conditions whereas dataset "Turin" contains images of 180 landmarks/buildings in Turin, Italy. 1440 images are still images and 540 images were extracted from videos. Pictures in this datasets exhibit dif-

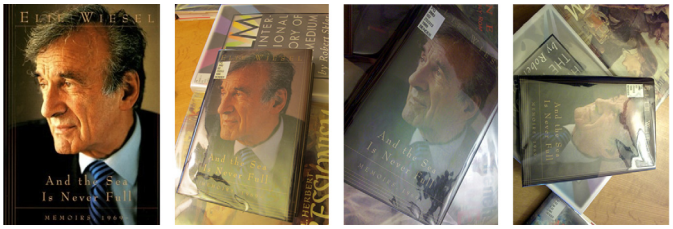

Fig. 3. Images of a typical object (DVD) from the "Mixed-text + Graphics" dataset.

Fig. 4. Images of the same building from different views from the "Turin" dataset.

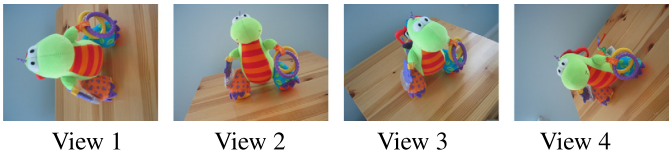

Fig. 5. Images of the same object from different views from the "Common objects and scenes" dataset by University of Kentucky.

ferent point of views of the same landmark/building, occlusions and strong light changes. Example images taken from the "Turin" dataset are depicted in Fig. 4.

The fifth dataset depicts "Common objects or Scenes" collected by University of Kentucky. It contains 10,200 images of 2550 different objects/scenes, each shot from 4 different views. Typical image transformations include rotation, different scales, different points of view and slight changes in illumination. Fig. 5 shows different shots of a typical object from this dataset.

5.1. SuperMarket Milan dataset

"SuperMarket Milan" is a recent ad-hoc dataset composed of 1686 photos of various supermarket products. This dataset was specifically created to test the CDVS Test Model [10] on a new use case and to report the ability of the model to recognize goods commonly available at supermarket premises.[1] Photos were taken under different illumination conditions in different types of physical locations (e.g. home and supermarket shelves). Different camera devices acquired the photos, i.e.: iPhone 4, iPhone 3, Samsung Next Turbo, Samsung S Advance, Samsung Galaxy S3, Lg Optimus L5, and LG Nexus 4. Image resolution ranges from 0.1 to 5.0 MegaPixels. An annotation file containing the coordinates of vertices of the bounding quadrilateral enclosing the object is associated to each image.

The dataset is composed of reference and query images. The reference images are photos of objects taken with the iPhone 4 in the best environmental conditions, which means, for example: no objects in the background, no portions of the object are occluded, etc. The query images are photos of the object in its context, either on the supermarket shelf or at home. Fig. 6 shows an example of

[1] The "SuperMarket Milan" dataset was presented at the 105th MPEG Meeting (Vienna) and tested on the CDVS Test Model. The dataset is available by contacting the authors.

Fig. 6. Images of the same object from different points of view from "SuperMarket Milan" dataset.

Fig. 7. Examples of color significant products from the "SuperMarket Milan" dataset.

an object. The left photo represents the reference image and the others are examples of queries.

There are 1697 query images and 430 reference images (total 2127). About four query images for each reference. Two different pairwise matching experiments were conducted with the "SuperMarket Milan" dataset. In the first experiment (noted as 6a in the rest of this paper) all the matching and non-matching pairs have been chosen randomly. In the second experiment (6b), the matching pairs were also chosen randomly, but the non-matching pairs were chosen randomly among the photos of so-called color significant products. In particular, this experiment measured the discriminative power of color descriptors. The phrase "color significant products" refers to products that differ from each other's by color and not by shape. Fig. 7 shows examples of color significant products.

For the pairwise matching experiment 6a a total of 3350 MP and 33,506 NMP are provided; 250 MP and 2522 NMP are instead provided for the experiment 6b. For the retrieval experiment 1686 queries are provided.

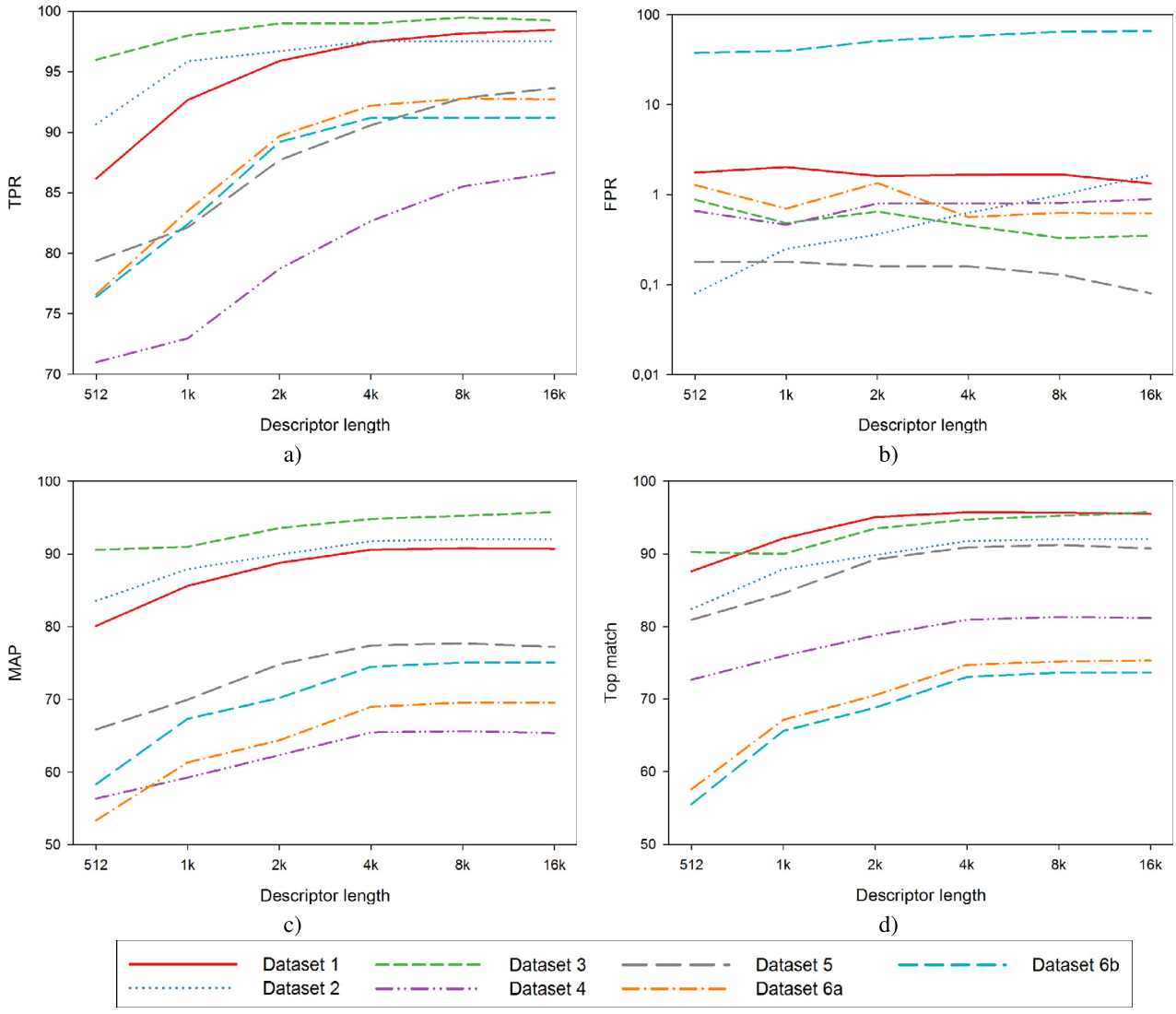

Fig. 8. Plots of the TPR (a), FPR (b), MAP (c) and Top Match (d) with respect to descriptor length for all the datasets considered.

6. Experimental results

6.1. CDVS Test Model—SuperMarket Milan dataset

According to the MPEG CDVS Evaluation procedure protocol [30], a full characterization of the Test Model was performed on the "SuperMarket Milan" dataset. The results were evaluated using two types of experiments: retrieval and pairwise matching. Both experiments were done on the six different datasets described in the previous section. Results were reported for the following operating points (upper bounds on average descriptor lengths in each experiment): 512, 1K, 2K, 4K, 8K, 16K bytes as query length.

The results for the retrieval and pairwise matching were evaluated using different sets of measures. Retrieval results were measured in terms of Mean Average Precision (MAP) and success rate for Top Match.

Pairwise matching results were measured in terms of Success Rate (i.e. TPR) at the average FPR of 1%. In fact, the CDVS Test Model is optimized to give an average FPR of 1% on the five standard CDVS datasets (i.e. dataset 1 to 5). The performances relative to the new datasets (i.e. 6a and 6b) have been obtained applying the Test Model as is. Results of the pairwise matching experiment 6a were in line with those of the other datasets. The success rate varied between 76.60% for lower bitrate descriptors and 92.78% for higher bitrate descriptors (i.e. 8k) and the range of FPR was from 0.62% for the highest bitrate descriptor to 1.35% for the lower bitrate descriptors (i.e. 512–2k). The plot of the success rate and FPR with respect to descriptor length for all the datasets considered are respectively reported in Fig. 8(a) and Fig. 8(b).

For the pairwise matching experiment 6b, Success Rates values were similar to those of the experiment 6a (see Fig. 8(a)), but FPR showed important differences. The range was from 37.83% to 65.94%. The highest level of FPR for this experiment was found in the highest bitrate descriptor, as can be seen in Fig. 8(b).

These levels showed a clear inability of the Test Model to discriminate "color significant products" (as previously defined). In particular, high levels of Success Rate, demonstrated the ability of the Test Model to recognize the class of object of dataset 6 (i.e. "SuperMarket Milan" dataset) but high levels of FPR also revealed that the current Test Model was not able to differentiate between two similar objects. As a consequence, the average FPR levels were extremely high.

On the Retrieval experiment, MAP values of datasets 6a and 6b were similar to those of the datasets 4 and 5. While datasets 1, 2 and 3 had high levels of MAP (i.e. from 76.56% to 95.78%), datasets 4 and 5 showed lower MAP levels (from 56.34% to 77.42%). The retrieval experiments 6a and 6b, concerning the "SuperMarket Milan" dataset showed levels slightly lower than those of

datasets 4 and 5: from 53.34% to 75.06%. The Top Match success ratios were very low with respect to those of the others datasets. The normal range for this value was between 72.78% and 96.27% while the range of Top Match values of the retrieval experiments 6a and 6b were from 55.52% to 75.33%. As expected, the Top Match values for the experiment on dataset 6b were slightly lower than those on dataset 6a. The plot of MAP and Top Match with respect to descriptor length for all the datasets considered are respectively reported in Fig. 8(c) and Fig. 8(d).

The retrieval data confirmed what could be deduced from the pairwise experiment results. Retrieval of SuperMarket Dataset objects was slightly more difficult compared to other kinds of objects of the CDVS Test Model dataset because of the shape of the supermarket objects themselves. Moreover the Top Match statistics showed that the current algorithms of the Test Model were not highly discriminative for this category of objects and in particular they were troubled by very similar "color significant products".

Given the low performance of the CDVS Test Model on the datasets 4, 5, and 6, in the next subsections a set of gray-level and color descriptors will be tested. The experiment is aimed to see if there are alternative descriptors that can consistently improve the performance on all the datasets with respect to those obtained by the descriptor actually used in the CDVS Test Model.

6.2. Gray-level descriptors

To test detectors and descriptors, two different types of experiments were conducted and reported for each dataset. In the first experiment, all the detected keypoints were used in the matching phase. This experiment aims to assess the upper bound performance of the descriptors outside the CDVS Test Model framework. In the latter experiment, a selection of the most discriminant 1024 detected keypoints was made on the basis of the strength of the filter response before the matching phase was performed. This experiment follows the CDVS guidelines, and uses the Test Model disabling the descriptor compression step, which has to be designed ad-hoc for each different descriptor.

True Positives Rates (TPR) were measured at two different levels of FPR: 10% and 1% (as MPEG CDVS requires). The obvious expectation is that the measured TPR should not be lower for any image at the higher FPR level. This expectation was confirmed by the experiments reported in Figs. 9 and 10. Concerning gray-level detectors and descriptors, KAZE performed equal or better than SIFT on the third, fourth and fifth datasets. SURF obtained very good results on the first three datasets. FREAK performed worse than SURF on all the datasets. Performance of A-KAZE are always lower than those of KAZE.

Affine invariant detectors, as expected, worked well under viewpoint changes. Their overall results were poorer than the SIFT and KAZE algorithms for all datasets, except for the "Video Frames" dataset. Conversely, on the "Landmarks and Buildings" dataset, affine invariant detectors achieved the best performances among all the algorithms. Limiting the number of keypoints to 1024 particularly affects the performance of some descriptors: DoG, Harris Laplace, Hessian Laplace, Multiscale Harris, Multiscale Hessian, SURF-FREAK, and SURF. This is particularly evident on datasets 1 to 5 where the loss in TPR shows a magnitude up to 0.20 (SURF-FREAK on dataset 2). The loss in TPR is much lower on datasets 6a and 6b, where the magnitude is lower than 0.032.

Table 4 shows the average computational time required by each algorithm to detect the keypoints and extract the features. Indeed a point of focused attention was not only the accuracy aspect of these algorithms but also their complexity as a measure of the computational burden on the implementation. For each detector the filter response threshold value, the average number of keypoints detected, and the average overall time spent for detection and description were reported. Timings have been taken on the images belonging to the "Mixed text + graphics" dataset (scaled at VGA resolution) with an x86 2.4 GHz single processor with 3 MB of Cache L2 and 8 GB of RAM. As expected, A-KAZE confirms to be the algorithm showing the lowest average computational time [17]. Quite remarkably, the SIFT (OpenCV) implementation is the second among the lowest complexity methods, which is contrary to the common belief, and it was not only the most accurate on average, but also the one showing the best trade-off between performance and speed, making it the most likely candidate for embedded system mapping.

Table 4
Filter response threshold value, average number of keypoints detected, and average overall time spent for detection and description. Timings have been taken on the images belonging to the "Mixed text + graphics" dataset (scaled at VGA resolution).

Algorithm	Peak threshold	Number of points	Comput. time (s)
DoG	3.0e2	869.6	0.981
Hessian (VLFeat)	3.0e2	1061.7	1.074
Hessian Laplace (VLFeat)	3.5e2	1138.2	0.940
Harris Laplace (VLFeat)	1.0e4	1068.3	1.797
Multiscale Hessian (VLFeat)	4.0e2	1565.1	0.908
Multiscale Harris (VLFeat)	2.0e4	1150.4	1.681
SIFT (VLFeat)	0.1e-2	1374.0	0.346
SIFT (OpenCV)	0.1e-2	1374.0	0.164
SURF (OpenCV)	3.0e2	987.5	0.169
SURF-FREAK (OpenCV)	0.1e-2	1321.0	0.483
Opponent SIFT (OpenCV)	0.1e-2	1374.0	0.480
KAZE (original v.1.3)	0.1e-3	1184.6	1.204
A-KAZE (original v.1)	1.0e-3	1271.3	0.112

6.3. Color descriptors

Comparisons of color descriptors are shown with respect to a baseline algorithm. The chosen baseline was the VLFeat implementation of SIFT (gray-levels). The different color descriptors tested correspond to Early or Late fusion of this baseline with the corresponding color information (see Table 2). On most of datasets color descriptors added a little improvement in the discriminative power over the gray-levels descriptor or they performed even worse compared to the baseline. Results are reported in Figs. 11 and 12, where a dashed line represents the TPR of the baseline descriptor at a FPR of 1%. It is possible to notice that on dataset 3 ("Video Frames") no descriptor achieved better results than the baseline. This is due to a lack of invariance of the color descriptors with respect to noise, blurriness, and other kinds of image transformations heavily present in this dataset. On dataset 1 ("Mixed text + graphics") and 4 ("Buildings") only a few color descriptors show better results with respect to the baseline. This is due to the fact that in some images from these datasets, color descriptors are misled by wrong white-balance correction, different lighting conditions and presence of shadows. Fig. 13 shows some examples of image pairs having large difference in color appearance due to imaging conditions. Very good results were obtained by some color descriptors on datasets 5, 6a and 6b. Dataset 5 ("Common objects and scenes") was composed of very colored objects, thus, the color information become highly discriminant. On this dataset, the algorithms achieving the best performances were those using a Late Fusion approach, which have a lower complexity than the Early fusion ones. On dataset 6b ("SuperMarket Milan" with color significant objects), as expected, the gray-levels SIFT baseline performed very low because different objects with different color but the same shape were recognized to be the same object. All color descriptors algorithms achieved better results for the highest level of FPR. Limiting the number of keypoints to 1024 has almost no effect on all datasets with the exception of dataset 1 on which however the loss in TPR is always lower than 0.03.

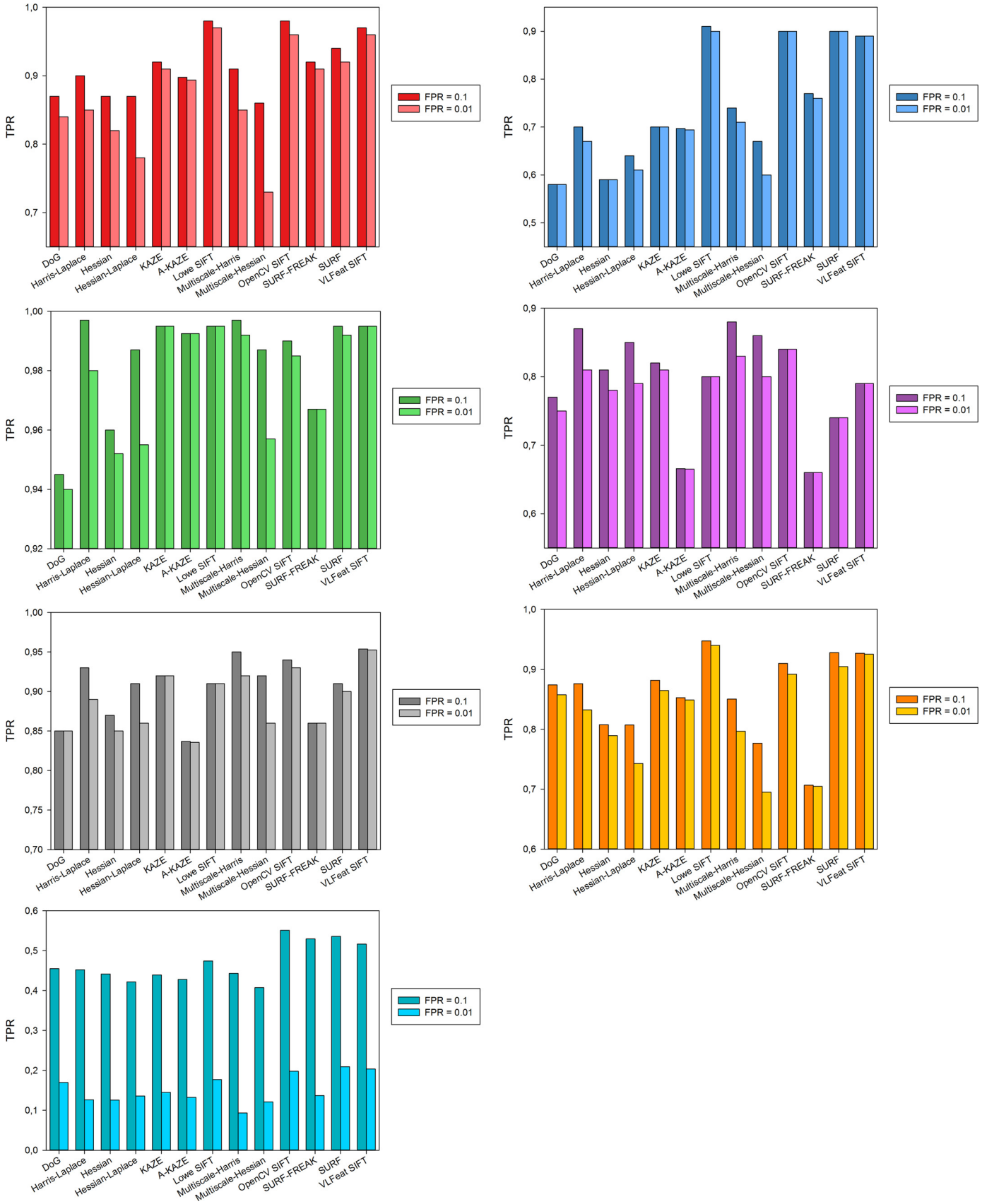

Fig. 9. TPR levels of all the gray-scale algorithms at a FPR of 10% and 1%. No limit on keypoints number. Top to bottom: dataset 1 (mixed text + graphics) and 2 (museum paintings), dataset 3 (video frames) and 4 (buildings and landscapes), dataset 5 (common objects and scenes) and 6a (SuperMarket Milan), dataset 6b (SuperMarket Milan: color significant objects).

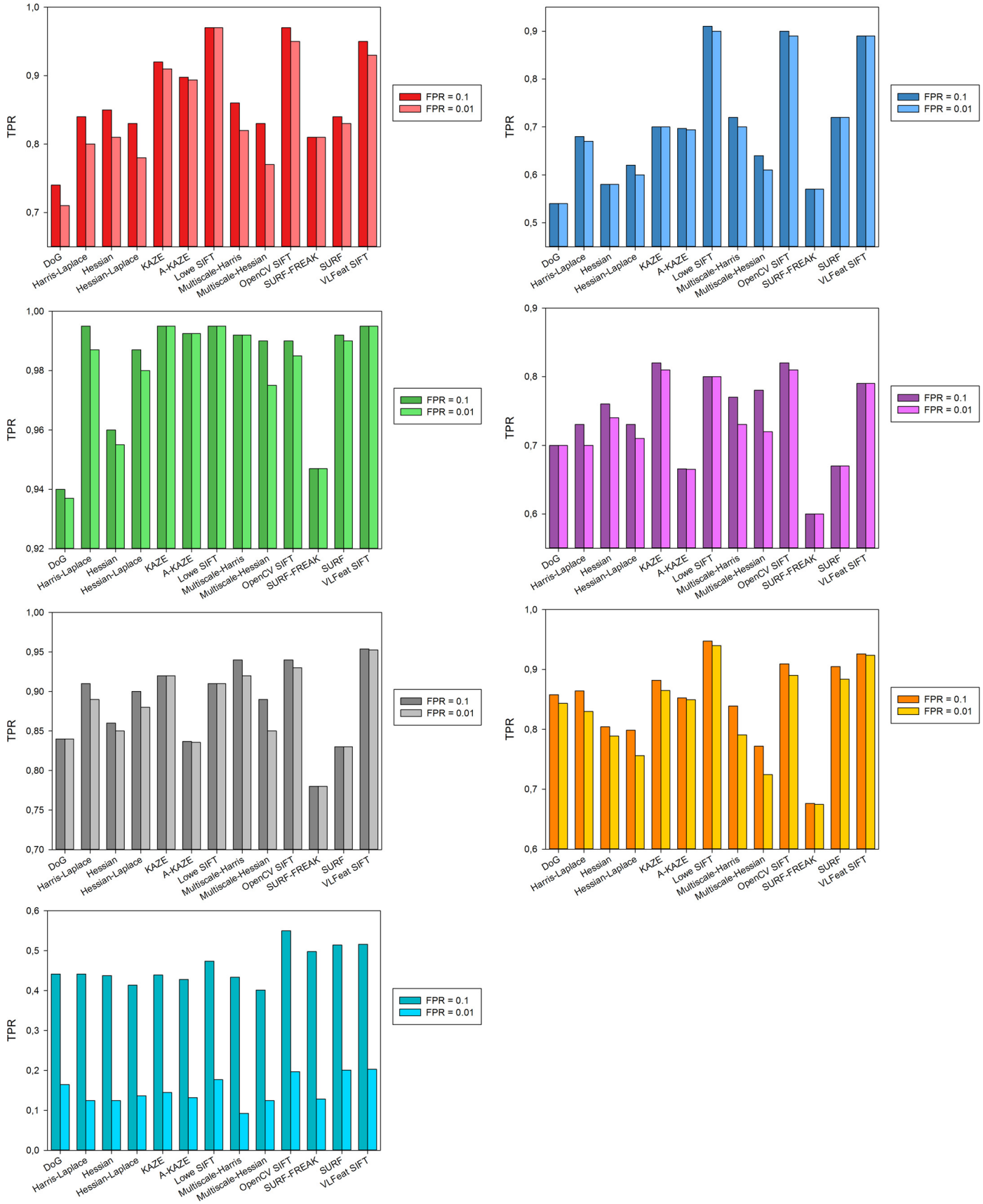

Fig. 10. TPR levels of all the gray-scale algorithms at a FPR of 10% and 1%. Keypoints number limited to 1024. Top to bottom: dataset 1 (mixed text + graphics) and 2 (museum paintings), dataset 3 (video frames) and 4 (buildings and landscapes), dataset 5 (common objects and scenes) and 6a (SuperMarket Milan), dataset 6b (SuperMarket Milan: color significant objects).

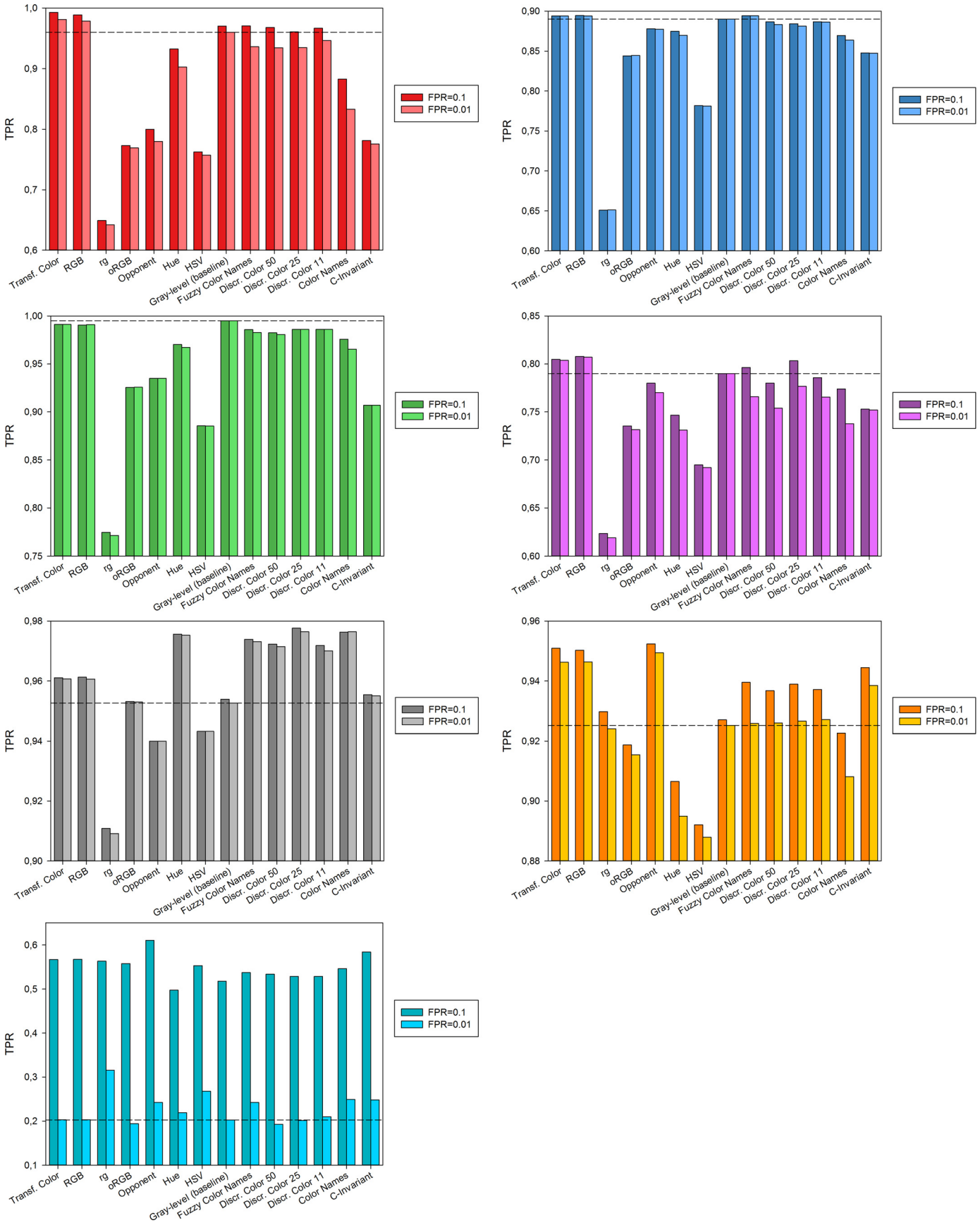

Fig. 11. TPR levels of all the color algorithms at a FPR of 10% and 1%. No limit on keypoints. Top to bottom: dataset 1 (mixed text + graphics) and 2 (museum paintings), dataset 3 (video frames) and 4 (buildings and landscapes), dataset 5 (common objects and scenes) and 6a (SuperMarket Milan), dataset 6b (SuperMarket Milan: color significant objects).

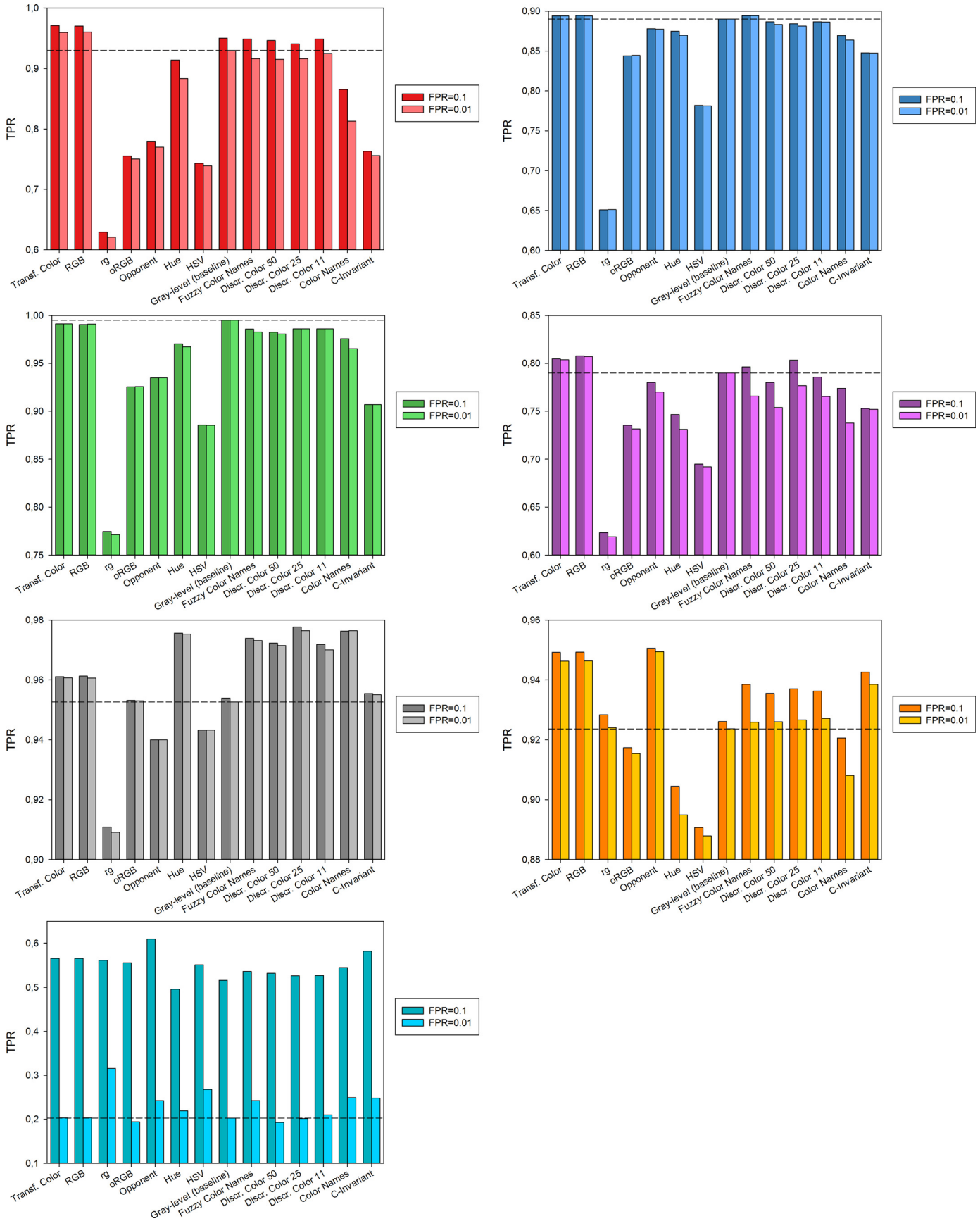

Fig. 12. TPR levels of all the color algorithms at a FPR of 10% and 1%. Keypoints number limited to 1024. Top to bottom: dataset 1 (mixed text + graphics) and 2 (museum paintings), dataset 3 (video frames) and 4 (buildings and landscapes), dataset 5 (common objects and scenes) and 6a (SuperMarket Milan), dataset 6b (SuperMarket Milan: color significant objects).

Fig. 13. Examples of images where color information can be misleading. In each image pair the reference image is shown on the left whereas the query image is shown on the right.

Note that in all datasets the Transformed Color algorithm and the RGB algorithm achieved the same results. That is because all the invariance properties of the Transformed Color space are implicitly included in the SIFT algorithm itself as Van de Sande et al. claimed [18]. Differently from other domains where Opponent SIFT obtained the best results [18,31], on CDVS datasets this is not always true, especially for the dataset 3 ("Video Frames"). We speculate that this difference in performance can be explained by the fact that the acquired images are affected by distortions that cannot be completely modeled by the diagonal-offset model considered in [18].

7. Conclusion

A detailed analysis of thirteen gray-level interest point detectors and descriptors available in the state of art has been performed on six heterogeneous datasets using a pairwise matching procedure similar to the one adopted by the MPEG CDVS Test Model [10]. Lowe's SIFT [11] was confirmed, without a-priori knowledge of the dataset, as the best performing method on average in terms of TPR levels (average TPR: 0.91) among gray-level descriptors. Remarkably, KAZE and SURF performed well on some datasets. For example on "Landmarks and Buildings" dataset KAZE achieved an average TPR of 0.815 whereas Lowe's SIFT attains a TPR of 0.8. Also the affine-invariant detectors achieved good results on the "Landmarks and Buildings" dataset but their results decreased if the number of keypoints detected is limited.

The measured computational times showed that the algorithm with the best trade-off between performance and speed was SIFT, in particular the OpenCV implementation. This was the base complexity adopted by MPEG CDVS against which any color descriptor algorithm has to be measured in terms of extra computational burden added. The use of color information did not achieve interesting performances on all CDVS datasets. This was proven by the fact that the average TPR value of RGB SIFT which was the best performing algorithm showed on average a 1% improvement with respect to the gray-level baseline. This gain was too limited to justify the extra complexity added by the Early Fusion methods. Color descriptors proven to bring some improvements on datasets 5 ("Common objects and scenes") and 6 ("SuperMarket Milan") where color information was certainly more relevant. On dataset 5, Late Fusion algorithms achieved the best TPR values while on dataset 6 the best performing algorithms were some Early Fusion approaches: RGB, Transformed Color, and Opponent. The concern for extra complexity mainly due to look-up handling justified their usage on top of the SIFT complexity baseline.

Results of the current CDVS Test Model on the "SuperMarket Milan" dataset showed that the algorithms were able to recognize the class of objects depicted in the dataset itself, but the algorithms exhibited a difficulty to distinguish what we have called "color significant products". Future works will focus on in-depth studies to understand if the use of color descriptors will be effective to overcome the limited performances on "color significant products" and also investigating more advanced color descriptor algorithms. We also plan to investigate different fusion approaches other than early and late fusion between shape and color information, and the combination of local and global descriptors which might be useful in case of textureless surfaces.

References

[1] B. Girod, V. Chandrasekhar, D.M. Chen, N.-M. Cheung, R. Grzeszczuk, Y. Reznik, G. Takacs, S.S. Tsai, R. Vedantham, Mobile visual search, IEEE Signal Process. Mag. 28 (4) (2011) 61–76.
[2] J. Kim, H. Jun, Vision-based location positioning using augmented reality for indoor navigation, IEEE Trans. Consum. Electron. 54 (3) (2008) 954–962.
[3] S.H. Lee, J. Choi, J.-I. Park, Interactive e-learning system using pattern recognition and augmented reality, IEEE Trans. Consum. Electron. 55 (2) (2009) 883–890.
[4] D. Chen, S. Tsai, C.-H. Hsu, J.P. Singh, B. Girod, Mobile augmented reality for books on a shelf, in: 2011 IEEE International Conference on Multimedia and Expo, ICME, IEEE, 2011, pp. 1–6.
[5] D. Pau, G. Cordara, M. Bober, S. Paschalakis, K. Iwamoto, G. Francini, V. Chandrasekhar, G. Takacs, White paper on compact descriptors for visual search, Incheon, Korea, ISO/IEC JTC1/SC29/WG11 MPEG2013/N13951, April 2013.
[6] S.S. Tsai, D. Chen, G. Takacs, V. Chandrasekhar, R. Vedantham, R. Grzeszczuk, B. Girod, Fast geometric re-ranking for image-based retrieval, in: 2010 17th IEEE International Conference on Image Processing, ICIP, IEEE, 2010, pp. 1029–1032.
[7] S. Bianco, R. Schettini, D. Mazzini, D.P. Pau, Quantitative review of local descriptors for visual search, in: IEEE Third International Conference on Consumer Electronics – Berlin (ICCE-Berlin), 2013, ICCEBerlin 2013, IEEE, 2013, pp. 98–102.
[8] D.G. Lowe, Object recognition from local scale-invariant features, in: The Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 2, IEEE, 1999, pp. 1150–1157.
[9] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[10] MPEG 108 CDVS Test Model 10: compact descriptors for visual search, Valencia, Spain, ISO/IEC JTC1/SC29/WG11 MPEG2014 W14393, April 2014.
[11] A. Vedaldi, B. Fulkerson, Vlfeat: an open and portable library of computer vision algorithms, in: Proceedings of the International Conference on Multimedia, ACM, 2010, pp. 1469–1472.
[12] G. Bradski, The OpenCV library, Dr. Dobb's J. 25 (11) (2000) 120–126.
[13] K. Mikolajczyk, C. Schmid, Scale & affine invariant interest point detectors, Int. J. Comput. Vis. 60 (1) (2004) 63–86.
[14] H. Bay, T. Tuytelaars, L. Van Gool, Surf: speeded up robust features, in: Computer Vision – ECCV 2006, Springer, 2006, pp. 404–417.
[15] A. Alahi, R. Ortiz, P. Vandergheynst, Freak: fast retina keypoint, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2012, pp. 510–517.
[16] P.F. Alcantarilla, A. Bartoli, A.J. Davison, Kaze features, in: Computer Vision – ECCV 2012, Springer, 2012, pp. 214–227.
[17] P.F. Alcantarilla, J. Nuevo, A. Bartoli, Fast explicit diffusion for accelerated features in nonlinear scale spaces, in: Proceedings of the British Machine Vision Conference, BMVA Press, 2013.
[18] K.E. Van De Sande, T. Gevers, C.G. Snoek, Evaluating color descriptors for object and scene recognition, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1582–1596.
[19] A.E. Abdel-Hakim, A.A. Farag, Csift: a sift descriptor with color invariant characteristics, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, IEEE, 2006, pp. 1978–1983.
[20] A. Verma, S. Banerji, C. Liu, A new color sift descriptor and methods for image category classification, in: International Congress on Computer Applications and Computational Science, 2010, pp. 4–6.
[21] J. Van De Weijer, C. Schmid, Applying color names to image description, in: IEEE International Conference on Image Processing, 2007, ICIP 2007, vol. 3, IEEE, 2007, p. III-493.
[22] J. Van De Weijer, C. Schmid, J. Verbeek, D. Larlus, Learning color names for real-world applications, IEEE Trans. Image Process. 18 (7) (2009) 1512–1523.
[23] R. Benavente, M. Vanrell, R. Baldrich, Parametric fuzzy sets for automatic color naming, J. Opt. Soc. Am. A 25 (10) (2008) 2582–2593.
[24] R. Khan, J. Van de Weijer, F.S. Khan, D. Muselet, C. Ducottet, C. Barat, Discriminative color descriptors, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2013, pp. 2866–2873.

[25] F.S. Khan, J. van de Weijer, M. Vanrell, Modulating shape features by color attention for object recognition, Int. J. Comput. Vis. 98 (1) (2012) 49–64.
[26] F. Shahbaz Khan, R.M. Anwer, J. van de Weijer, A.D. Bagdanov, M. Vanrell, A.M. Lopez, Color attributes for object detection, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2012, pp. 3306–3313.
[27] J.-M. Geusebroek, R. Van den Boomgaard, A.W.M. Smeulders, H. Geerts, Color invariance, IEEE Trans. Pattern Anal. Mach. Intell. 23 (12) (2001) 1338–1350.
[28] M. Bratkova, S. Boulos, P. Shirley, oRGB: a practical opponent color space for computer graphics, IEEE Comput. Graph. Appl. 29 (1) (2009) 42–55.
[29] I.S. Dhillon, S. Mallela, R. Kumar, A divisive information theoretic feature clustering algorithm for text classification, J. Mach. Learn. Res. 3 (2003) 1265–1287.
[30] Evaluation framework for compact descriptors for visual search, Torino, IT, ISO/IEC JTC1/SC29/WG11/N12202, July 2011.
[31] S. Bianco, C. Cusano, Color target localization under varying illumination conditions, in: Computational Color Imaging, Springer, 2011, pp. 245–255.

S. Bianco obtained the BSc and the MSc degree in Mathematics from the University of Milano-Bicocca, Italy, respectively in 2003 and 2006. He received the PhD in Computer Science at Department of Informatics, Systems and Communication of the University of Milano-Bicocca, Italy, in 2010, where he currently a post-doc. His research interests include computer vision, optimization algorithms, machine learning, and color imaging.

D. Mazzini obtained his BSc and MSc degree in Computer Science from University of Milano-Bicocca, Italy, respectively in 2010 and 2013. He worked on the Compact Descriptors for Visual Search MPEG standard during his internship in STMicroelectronics. Currently he is a PhD candidate at University of Milano-Bicocca. His research interests are in computer vision, visual search and machine learning.

D.P. Pau is Senior Principal Engineer, Senior Member of Technical Staff at Advanced System Technology, STMicroelectronics, in Agrate Brianza Italy. His research interests are in Computer Vision, Visual Search, Optical Flow and Structure from Motion as well as 3D Graphics for Embedded System and its compression. He is serving as co-chair MPEG Compact Descriptors for Visual Search and Compact Descriptors for Video Analysis groups.

R. Schettini is a professor at the University of Milano Bicocca (Italy). He is head of Imaging and Vision Lab and Vice-Director of the Department of Informatics, Systems and Communication. He has been associated with Italian National Research Council (CNR) since 1987 where he has leaded the Color Imaging lab from 1990 to 2002. He has been team leader in several research projects and published more than 200 refereed papers and six patents about color reproduction, and image processing, analysis and classification. He has been recently elected Fellow of the International Association of Pattern Recognition (IAPR) for his contributions to pattern recognition research and color image analysis.