Class-based Color Bag of Words for Fashion Retrieval

Costantino Grana, Daniele Borghesani, Rita Cucchiara
Dipartimento di Ingegneria dell'Informazione
Università degli Studi di Modena e Reggio Emilia
Modena, Italy
costantino.grana@unimore.it, daniele.borghesani@unimore.it, rita.cucchiara@unimore.it

***Abstract*—Color signatures, histograms and bag of colors are basic and effective strategies for describing the color content of images, for retrieving images by their color appearance or providing color annotation. In some domains, colors assume a specific meaning for users and the color-based classification and retrieval should mirror the initial suggestions given by users in the training set. For instance in fashion world, the names given to the dominant color of a garment or a dress reflect the fashion dictact and not an uniform division of the color space. In this paper we propose a general approach to implement color signature as a trained bag of words, defined on the basis of user defined color classes. The novel Class-based Color Bag of Words is a easy computable bag of words of color, constructed following an approach similar to the Median Cut algorithm, but biased by color distribution in the trained classes. Moreover, to dramatically reduce the computational effort we propose 3D integral histograms, a 3D extension of integral images, easily extensible for many histogram-based signature in 3D color space. Several comparisons in large fashion datasets confirm the discriminant power of this signature.**

***Keywords*-color classification, image retrieval, fashion, optimized histogram**

I. Introduction

Internet shopping has grown incredibly in the last years, and fashion was one of the latest to appear on the net, possibly because clothes-to-body fitting is a very personal and physical experience. Nevertheless due to the high costs of luxury goods, the fashion designers attraction, and the entrance in the consumer classes of younger generations, in latest years also clothing and fashion has began to seriously appear on the web. In Italy, for example, this sector has raised 490 million euro in 2010 with a growth of 43% with respect to 2009. 70% of this revenue is created by operators working only online, such as yoox.com, able to attract 8 million unique visits per month.

Most of these web sites propose hundreds of thousands of clothes of every type, so that metadata and annotations are essential for both producers and customers for achieving an efficient and effective search on the web. As well as the product type annotation, color is one of the more effective cue to filter or sort results (even Google Images introduced this kind of filtering options on generic image search). In the fashion world a correct color annotation if fundamental in order not to suggest a wrong garment to an unwilling buyer, or, on the contrary, not to show an available garment to someone who would like to buy something with that particular color.

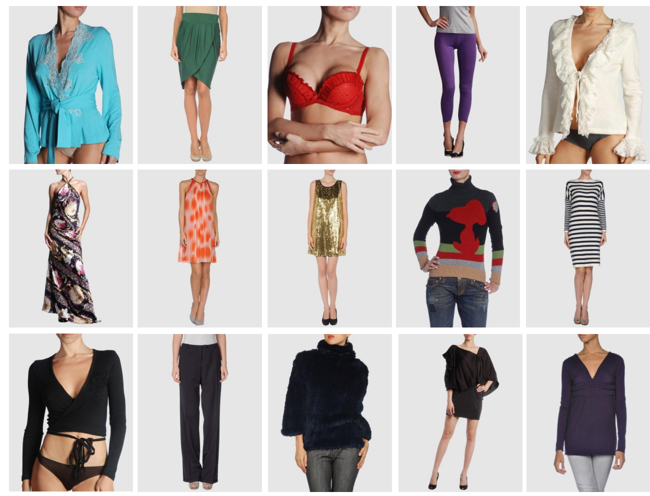

Figure 1. Some samples of the pieces of clothing considered in this work. In the first row, mainly uniform colors are depicted. In the second row, more complex patterns make the classification into a single color category problematic. Finally, in the third row, different colors occur in their lowest luminance version (respectively black, charcoal, dark blue, dark brown and dark violet), making more difficult, even for a human operator, a consistent classification.

However, while in generic object detection and scene categorization contexts the classification, especially for a human operator, turns out to be quite straightforward, in the context of fashion the problem is more complicated. First, color distribution in clothes is not uniform and the color annotation for a dress refers to its dominant color only (Fig. 1). Second, colors are defined with many gradations and nuances, according with the fashion jargon: for instance in the seventh row of Fig. 2, purple clothes are divided in purple, violet, lilac, mauve, aubergine, indigo classes). Then colors are retrieved differently by humans, depending on the material but also on the gamma correction of the screen, the quality of the screen itself, level of view resize and so on. All these points are major sources of annotation errors by human operators.

Therefore in many domain-dependent datasets the color annotation does not reflect an uniform distribution in the color space but is affected by the human experience and the social and cultural background of customers. This must be taken into account in many e-commerce services on the web, in digital catalogs of products (ceramic tiles, furniture, accessories..), in beauty products description etc.

In this paper we define a new specific signature for describing dominant colors of objects in an image. The signature is a color-based bag of word, but is not obtained by working uniformly on the color space, neither is refined according with the data distribution on the space. Instead, the selection of the codebook is biased by the classes of color suggested by users by a training set. We propose an optimization of the color histogram composition according with the learned classes of color, providing a simple ad effective Class-based Color Bag of Words influenced by the color jargon of the context. It provides noticeable improvements both in terms of mAP over similarity retrieval and accuracy in color classification w.r.t. similar signatures. The feature extraction is highly computational intensive, since the optimization works on the 3D color space accounting for all the examples of the very large dataset. Consequently we propose 3D integral histograms, a 3D extension of integral images, which dramatically reduce the computational time.

An application for color classification of fashion clothes and accessories is eventually described to support the quality control step of on-line resellers and to generate additional metadata and suggestions to customers.

II. Related work

In the majority of retrieval system, either dedicated to concept detection, scene categorization, classification or similarity retrieval, a very common approach exploits the bag-of-words paradigm, in which the centroids of a clustered set of training descriptors (*vocabulary*) is used to count the occurrences of descriptors belonging to every class within the objects of interest [1].

Often local descriptors are employed in bag of words used capable to describe perceptively meaningful local invariants such as corner points or interest points. Starting from the older but still very used proposals of SIFT [2] and SURF [3], exploiting gradient information only, a lot of improvement have been proposed to include the color information. A recent comparison on the topic, proposed by van de Sande *et al.* in [4], highlights many of them, including RGB-SIFT, Opponent-SIFT and C-SIFT.

Nevertheless, some retrieval contexts do not require local information. Let's consider for example the case of interest in this paper, which is the automatic annotation and retrieval of pieces of apparel and fashion garments by means of their dominant and more salient color. Here local information is poorly significant, leading to an inaccurate visual summary, mostly driven to boundaries (the shape) and distracting interest points (folds, illumination changes due to the body of the model and the position of lights). Instead, after an adequate shape segmentation and target detection (the main apparel worn by a model) we need global features that provide a compact summary of the visual content, typically by aggregating some chromatic information extracted at every pixel location of the target. One of the earliest and still commonly available in most toolboxes is the Color Histogram [5], which describes the visual content ignoring its spatial arrangement. Many variations of it have been proposed in the last two decades, and also different distance measures (histogram intersection, Bhattacharyya, χ^2).

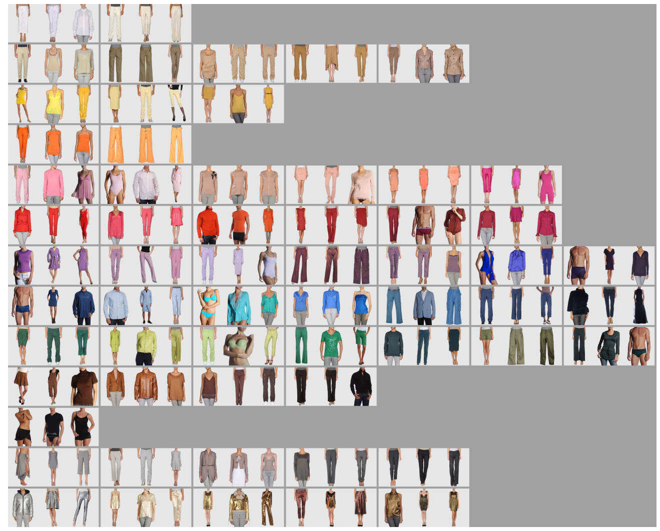

Figure 2. Example images for every color class, defined by standard fashion practices.

A consistent improvement in accuracy is reached with the adoption of adaptive binning on every image [6], [7]. Since the resulting histograms can vary in length, the comparison requires more complex solutions. The Earth Movers Distance (EMD) [8] is a cross-bin distance that addresses this alignment problem. EMD is defined as the minimal cost that must be paid to transform one histogram into the other, where there is a "ground distance" between the basic features that are aggregated into the histogram. The EMD as defined by Rubner is a metric only for normalized histograms. However, recently Pele and Werman [9] suggested $\widehat{EMD}$ and showed that it is a metric for all histograms, providing also a fast version [10].

Recently the bag-of-colors approach has been proposed by Wengert *et al.* [11] for image search: the color signature of the image is produced by a k-means quantization over a training set. This procedures is shown to improve retrieval accuracy in terms of mAP.

III. Optimizing color histograms for color class identification

Working on large datasets for search and classification purposes fixed length signatures are often adopted, being

Algorithm 1 Class Based Color Space Partitioning

```
Compute cumulative histograms of training images p_j for all classes
b ← (0,0,0),(255,255,255)    ▷ Start with the whole color space
FINDBESTSPLIT(b)
INSERT(list, b) ▷ list contains the color space partition
while SIZE(list) < N do
    b ← MAX_DELTA(list)
    b_0, b_1 ← GETSPLITS(b)
    FINDBESTSPLIT(b_0)
    INSERT(list, b_0)
    FINDBESTSPLIT(b_1)
    INSERT(list, b_1)
end while
```

more easily used in different hashing and indexing tasks. Color histograms are a common solution, and using a non uniform division of the color space could better describe the distribution of color aspects in the dataset. Since we aim at exploiting color for visual object classification, we would like to employ a dynamic binning which emphasizes the classes peculiarities. This is different from extracting a set of colors based on the data only (for example by clustering in the color space), but is a task of feature selection by incorporating data classification information in the definition of the color signature.

A color in a color space $\mathcal{C}$ is denoted by c. Given an image I, the color distribution for the image is

$$p(c|I) = \frac{\#\{I(x,y) = c\}}{\#I}. \tag{1}$$

Given a class of images C_j, with $j = 1, \ldots, J$, described by a training set of images I, we can define $p_j(c)$ as the L1-normalized sum of the color distributions of all images in that class. We approach the problem of finding a class optimized binning with a greedy procedure inspired to the median cut algorithm [12].

A box is the set of colors contained within a parallelepiped defined by two extreme colors low (l) and high (h), with $l, h \in \mathcal{C}$:

$$b = \{c \in \mathcal{C} : l_k \leq c_k \leq h_k, k = 0, 1, 2\} \tag{2}$$

To simplify the equations, from now on we will assume a three channels color space. We will equivalently write $b = (l, h)$. We call $m_j(b)$ the mass of box b in class j, such as

$$m_j(b) = \sum_{c \in b} p_j(c). \tag{3}$$

The total mass of b is

$$M(b) = \sum_{j=1}^{J} m_j(b). \tag{4}$$

We also denote $C(b)$ as the class associated to box b, that is the class with maximum mass for the box:

$$C(b) = \arg\max_j m_j(b). \tag{5}$$

The error induced by considering colors in b to be all of class C_j is defined as:

$$E(b) = \sum_{j \neq C(b)} m_j(b) = M(b) - m_{C(b)}(b). \tag{6}$$

We define a split of a box as $s = (v, k)$, meaning that we divide the box along channel k at position v. Splitting a box has the purpose of better describing the colors of that box, thus it is reasonable to assume that this will lower the error. We call $\delta(b, s)$ the difference between the current error caused by the box b and the one obtained after the splitting s. $\delta(b) = \max_s \delta(b, s)$ is the error induced by the *best split*. We will then choose to split the box which maximizes its δ.

The algorithm employs a list of boxes, initially containing a single box enclosing the whole 3D color space, described as b_0. For example in an 8-bit RGB color space $b_0 = ((0,0,0),(255,255,255))$. At each iteration step we extract from the list the box which has the maximum delta value, then it is split such as to minimize the sum of the errors after the split. The resulting boxes are put back in the list. The algorithm proceeds until the required number of boxes/histogram bins is obtained. Pseudo code is given in Algorithm 1.

IV. 3D Integral Color Histograms

The algorithm described in previous section requires to compute the mass of a box, and the search for the best split requires to compute it at all possible positions of a split. The straightforward solution to this problem is to integrate $p_j(c)$ for all c which belong to b in three different directions, that is the three color channels. This is computational demanding and must be performed on all class distributions, making it unfeasible as soon as the distributions are not heavily quantized.

In this work we propose to simplify the selection of the best search by defining a 3D extension of the "integral images" approach introduced in [13]. The integral image contains at every point (x', y') the sum of all pixels with $x < x'$ and $y < y'$. This allows to compute the sum of all values of a rectangle by combining just four values of the integral image. This was successfully employed in the past also for extracting histograms on an arbitrary rectangular region of an image [14].

We propose to apply the same process to 3D color histograms, to compute *in constant time* the mass of a box. To this aim the first step is to extract the 3D integral color histogram from a conventional 3D color histogram. Having $p(c)$, the normalized histogram at color c, the 3D integral histogram $ih(c)$ is defined as

$$ih(c) = \sum_{\substack{x:x_k<c_k \\ k=0,1,2}} p(x) \tag{7}$$

As for integral images, it is possible to compute the 3D integral histogram by a single sweep over the original histograms, taking advance of the recursive nature of the definition. In particular:

$$\begin{aligned} ih(c) = p(c) &+ ih(c_0-1,c_1,c_2) + ih(c_0,c_1-1,c_2) \\ &- ih(c_0-1,c_1-1,c_2) + ih(c_0,c_1,c_2-1) \\ &- ih(c_0-1,c_1,c_2-1) - ih(c_0,c_1-1,c_2-1) \\ &+ ih(c_0-1,c_1-1,c_2-1) \end{aligned} \tag{8}$$

This assumes that the value of the 3D integral histogram is 0 whenever any of the color coordinates is negative. Given a 3D integral histogram, the computation of the mass within a box $b=(l,h)$ is quite similar to the use of integral images for rectangle area calculations:

$$\begin{aligned} m(b) = ih(h_0,h_1,h_2) &- ih(l_0-1,h_1,h_2) \\ &- ih(h_0,l_1-1,h_2) + ih(l_0-1,l_1-1,h_2) \\ &- ih(h_0,h_1,l_2-1) + ih(l_0-1,h_1,l_2-1) \\ &+ ih(h_0,l_1-1,l_2-1) - ih(l_0-1,l_1-1,l_2-1) \end{aligned} \tag{9}$$

This formulation holds, again assuming that $ih(c)=0$ if $c_k=-1$ for any k.

A further observation may help in exhaustively searching for the best split. We define the quantity *column* of b along dimension d at height x as the mass of the box with limits $(\tilde{l},\tilde{h})$, where

$$\tilde{l}_k = \begin{cases} 0 & \text{if } k=d \\ l_k & \text{otherwise} \end{cases} \quad \tilde{h}_k = \begin{cases} x & \text{if } k=d \\ h_k & \text{otherwise} \end{cases} \tag{10}$$

This means that, using Eq. 9, for example when $d=0$ we have

$$\begin{aligned} col(b,0,x) = ih(x,h_1,h_2) &- ih(-1,h_1,h_2) \\ &- ih(x,l_1-1,h_2) + ih(-1,l_1-1,h_2) \\ &- ih(x,h_1,l_2-1) + ih(-1,h_1,l_2-1) \\ &+ ih(x,l_1-1,l_2-1) - ih(-1,l_1-1,l_2-1) \\ = ih(x,h_1,h_2) &- ih(x,l_1-1,h_2) \\ &- ih(x,h_1,l_2-1) + ih(x,l_1-1,l_2-1) \end{aligned} \tag{11}$$

Plugging back Eq. 11 in Eq. 9 we can write:

$$m(b) = col(b,k,h_k) - col(b,k,l_k-1), \tag{12}$$

thus enabling us to compute the best split by keeping fixed the second term (the "column base"), while progressively increasing the first one. The pseudo code is provided in Algorithm 2.

Algorithm 2 Pseudo code of fast best splitting algorithm

procedure FINDBESTSPLIT(b, ih)
  $error \leftarrow +\infty$
  for $k \in [0,2]$ **do** ▷ For all channels
    for $j \leftarrow 1, J$ **do** ▷ Precompute columns limits
      $m\_min_j = col(b,k,l_k-1)$
      $m\_max_j = col(b,k,h_k)$
    end for
    for $x \in [l_k, h_k-1]$ **do** ▷ For all possible splits
      $M^0 \leftarrow 0$ ▷ Total masses for splits
      $M^1 \leftarrow 0$
      for $j \in [1,J]$ **do**
        $c \leftarrow col(b,k,x)$
        $m^0_j \leftarrow c - m\_min_j$
        $M^0 \leftarrow M^0_j + m^0_j$
        $m^1_j \leftarrow m\_max_j - c$
        $M^1 \leftarrow M^1_j + m^1_j$
      end for
      $E^0 \leftarrow M^0 - \max_j m^0_j$ ▷ Errors of splits
      $E^1 \leftarrow M^1 - \max_j m^1_j$
      if $error > E^0 + E^1$ **then**
        $error \leftarrow E^0 + E^1$ ▷ Update best split
        SETSPLIT(b, k, x)
        $\delta \leftarrow E(b) - error$
      end if
    end for
  end for
end procedure

The construction of the color signature requires to identify the box in which every pixel color falls. This has the same complexity of finding the nearest centroid in a k-means based solution, which may be solved with a 16MB look up table. If memory becomes a bottleneck, the fact that boxes are axis-aligned allows to optimize a decision tree with a dynamic programming approach similar to the one proposed in [15]

V. EXPERIMENTS

This signature has been defined and tested for a task of automatic annotation of more datsets of about 10000 images each one, collected from popular fashion companies according with luxury clothes and accessories categories. The datasets are provided with annotation metadata describing the type of clothing, other commercial information like available sizes, price, etc. and the color manually selected from a standard fashion palette by human operators, often prone of errors and mistakes. The images have been photo retouched to provide a more uniform look when exported on the web(background is removed, body hair and tattoos are deleted, images are centered and the face is cropped), so colorimetric information is not available and cannot be restored by conventional hypothesis (such as Grey World

Table I
COMPARISON OF IMAGE RETRIEVAL RESULTS IN TERMS OF MEAN AVERAGE PRECISION.

Descriptor	bins	mAP
RGB Histogram	512	0.457
RGB Histogram	4096	0.496
L*a*b* Histogram	784	0.391
Bag of colors signature	512	0.492
Class-based Color Bag of Words	512	0.558
Class-based Color Bag of Words	1024	0.566

Table II
ACCURACY RESULTS USING SVMS WITH RBF AND HISTOGRAM INTERSECTION KERNELS

Descriptor	RBF kernel	HI kernel
RGB Histogram	71.11	69.96
Bag of colors signature	69.48	72.27
Class-based Color Bag of Words	73.80	74.38

assumptions). We provided an initial automatic preprocessing to remove skin components and mannequin parts, in order to focus only on clothing. GrabCut [16] is employed to select the clothing parts not pertinent with the "main" element depicted (such as boots in skirts images, or trousers under t-shirts, etc.) that is usually the most central piece of clothing. Results in segmentation are satisfactory, while more complex is the classification of the dominant color.

The choice of assigning a piece of apparel to a single color is a complex operation, often ill-posed since the color classes defined by fashion name (e.g. asphalt gray, mouse gray, pearl gray, powder , anthracite,...) are sometimes visually overlapped. There are some very challenging situations: objects with an uncertain color lying between two color classes, objects with complex patterns like flowers or prints, and objects with simple and well defined color patterns (stripes mostly) that however must be associated with a single dominant color. Therefore the automatic classification cannot reach a very high accuracy, but at the same time also the users often disagree in their manual classification. to test the signature described in our test we asked to fashion operators to select only the images with good segmentation and agreement on an unique color classification. Exemplars for all color categories are provided in Fig. 2. The final annotated dataset consists of 5181 images, of which half was used for testing and half for training.

We compared four different color descriptors:

- RGB Color Histogram: the three channels are uniformly quantized to 8 or 16 values, giving a 512 or 4096 bins descriptor.
- L*a*b* Color Histogram: the image is converted to the CIE L*a*b* color space, with the hypothesis that the source images are in the sRGB color space. The three channels are then quantized to 4,14,14 bins as suggested in [17].
- Bag-of-colors signature: this approach consists in clustering the colors extracted from the training set with k-means, followed by Inverse Document Frequency weighting, power law, and L1 normalization [11].
- The proposed Class-based Color Bag of Words.

The first test was performed to assess the ability of the different descriptors to retrieve similar images in terms of colors. Histograms were compared with the histogram intersection metric. Results are shown in Table I. It is clear that the CIE L*a*b* color space fails to correctly represent the color information. This may be due to the assumption of sRGB color space which is not confirmed. While this color space could help in a general image retrieval context, when the collection is uniform, the raw RGB values are probably more reliable. Adapting the histogram binning to the dataset characteristics is indeed useful, even if these results do not show the consistent improvement reported in [11]. The proposed bin selection shows the best performance which is not matched by the RGB color histogram even rising the number of bins to 4096. Some example results are shown in Fig. 3.

The second test tackles the classification task. We trained a multiclass SVM using a 1-vs-1 learning strategy using the three methods with 512 bins. We compared the RBF kernel and the Histogram Intersection kernel optimizing C and σ (only for RBF) with grid search. Table II reports the accuracy results. The bag-of-colors signature has the lowest accuracy with RBF kernel, which is much improved by the HI kernel. The opposite happens with RGB histograms. Our solution is still better, even if the improvement obtained using the HI kernel is not so significant.

The computational time of the proposed solution is much lower than what required by k-means, and most of the complexity is the precomputation of the 3D integral histograms, which requires (without quantization) to sweep $256^3 \cdot N_c$ values, with N_c the number of classes. The main limitation is the memory requirements of this approach which is $256^3 \cdot N_c \cdot 4$ bytes when using 32 bits floats, that is 64 MB for each class. In our experiments we were using 60 classes for a total memory allocation of 3.75 GB.

VI. CONCLUSIONS

We presented a novel solution for generating color signatures optimized in terms of class separability. The approach may be applied to any color space, and a computationally affordable solution for 3D color spaces (the nearly entirety of them) was described. Results confirm a performance improvement both for similarity searches and for classification tasks, both with respect to classical uniform quantization and to other dataset clustering solutions.

ACKNOWLEDGMENT

The authors would like to thank Roberto Vezzani for the invaluable discussions during the algorithm design.

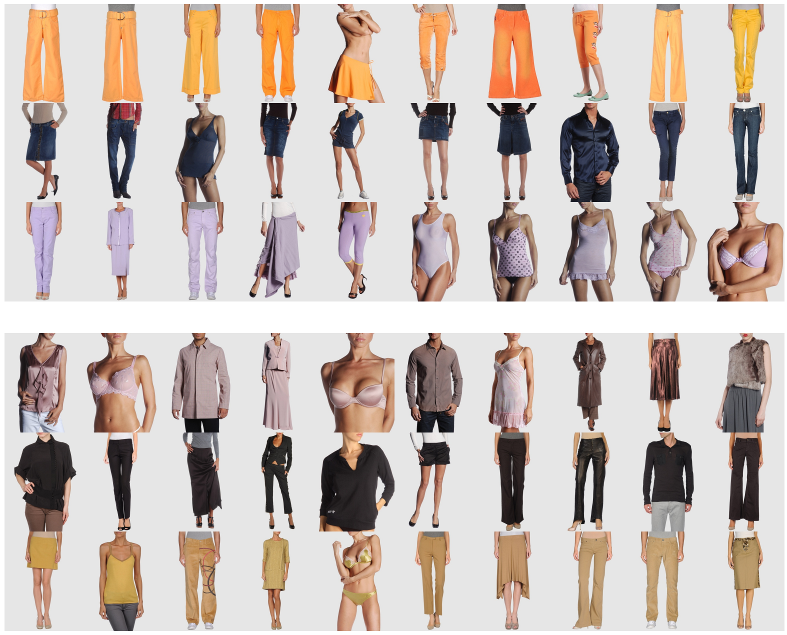

Figure 3. Example of images retrieved by color similarity. The first three examples are correctly classified by the SVM, while the other three examples are assigned to a different class. It is possible to note that it is hard to detect the mistake, basing purely on a quick inspection.

REFERENCES

[1] C. R. Dance, G. Csurka, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *ECCV Workshop on Statistical Learning in Computer Vision*, 2004, pp. 1–22.

[2] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[3] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-Up Robust Features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[4] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582–1596, 2010.

[5] M. J. Swain and D. H. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.

[6] W. K. Leow and R. Li, "Adaptive binning and dissimilarity measure for image retrieval and classification," in *IEEE International Conference on Computer Vision and Pattern Recognition*, vol. 2, 2001, pp. 234–239.

[7] R. Melli, C. Grana, and R. Cucchiara, "Comparison of color clustering algorithms for segmentation of dermatological images,," in *Proceedings of the SPIE Medical Imaging - Image Processing*, San Diego (CA), U.S.A., Feb. 2006, pp. 3S1–9.

[8] Y. Rubner, C. Tomasi, and L. J. Guibas, "A metric for distributions with applications to image databases," in *IEEE International Conference on Computer Vision*, 1998, pp. 59–66.

[9] O. Pele and M. Werman, "A linear time histogram metric for improved sift matching," in *European Conference on Computer Vision*, vol. 3, 2008, pp. 495–508.

[10] ——, "Fast and robust earth mover's distances," in *IEEE International Conference on Computer Vision*, 2009, pp. 460–467.

[11] C. Wengert, M. Douze, and H. Jégou, "Bag-of-colors for improved image search," in *ACM International Conference on Multimedia*, 2011.

[12] P. Heckbert, "Color image quantization for frame buffer display," in *ACM SIGGRAPH*, 1982, pp. 297–307.

[13] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, pp. 137–154, May 2004. [Online]. Available: http://dl.acm.org/citation.cfm?id=966432.966458

[14] F. Porikli, "Integral histogram: a fast way to extract histograms in cartesian spaces," in *IEEE International Conference on Computer Vision and Pattern Recognition*, vol. 1, 2005, pp. 829–836.

[15] C. Grana, D. Borghesani, and R. Cucchiara, "Optimized Block-based Connected Components Labeling with Decision Trees," *IEEE Transactions on Image Processing*, vol. 19, no. 6, Jun. 2010.

[16] C. Rother, V. Kolmogorov, and A. Blake, ""GrabCut": interactive foreground extraction using iterated graph cuts," in *ACM SIGGRAPH*, 2004, pp. 309–314.

[17] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *IEEE International Conference on Computer Vision and Pattern Recognition*, june 2010, pp. 3485–3492.