Dihedral Groups and Spatio-Chromatic Filter Systems

Reiner Lenz, Martin Solli

Linköping University,
SE-60174 Norrköping
eMail: reiner.lenz@liu.se; martin.solli@liu.se

Abstract The dihedral groups D(4) and D(3) are the symmetry groups of the image grid and the RGB color channels respectively. The product group D(4) $\otimes$ D(3)is therefore a symmetry group of RGB patches. Using the representation theory of the product group we construct very fast digital filter systems that split the vector space of RGB distributions into minimal group-invariant subspaces. We show that these filter systems can be implemented using additions/subtractions only and that the number of arithmetic operations can be minimized using algebraically defined fast algorithms. The raw filter results are then converted into a polar coordinate system in which the radial component is invariant under all group operations of the original distributions. The angular part encodes the effect of the group operation. We illustrate the performance of these filters by investigating the internal structure of two databases containing approximately 1.5 million images crawled by the Picsearch image search engine. One of the databases contains images indexed with object-related keywords while the second database is organized in emotion-related categories. We characterize images by the statistical distributions of the filter results from the different blocks and use parameters derived from these distributions to characterize the different categories. We also use a standard support vector machine to illustrate the separability of images from different category pairs. The examples illustrate that, despite their simplicity, the features extract relevant visual properties that can be used for indexing, retrieval and classification.

1 Introduction

The introduction of digital cameras, the availability of cheap storage and the increased power of the processing units has led to a situation where huge amounts of visual information is available in digital form. As a consequence there is now a need for tools to handle large collections of image data. Another, less obvious, effect is that search engines

can now automatically link this visual information with related text. This creates large databases with automatically tagged images where the links between visual and text information is much weaker than in databases with manually indexed images but on the other hand the size of these databases makes manual tagging impossible. Such databases constitute thus statistical samplings of a special type of visual environment that is also indexed by text. We can also see image search engines as tools that are used to perform large numbers of psycho-physical experiments: each time a user selects an image from a search page he or she selects actively one of the images from the large number of images in the database. Finally we mention that the interaction between the user and the search engine is almost always in terms of low-resolution thumbnails and typical decisions are made very fast: the typical user does not inspect the thumbnail presented in great detail but makes a fast decision by simply clicking on one of the images shown.

Motivated by these observations we concentrate in this paper on methods that are suitable for usage in very fast search engines for very large image databases. It is also known that human observers can solve complex visual recognition tasks in a very short time. These response times can only be achieved if the process is based on very simple operations. We will therefore always prefer simple operations like addition and thresholding over more complex operations like multiplications. We will also base our approach on thumbnail-sized images since this is the most common image format used in image search. Images contain not only information about the objects contained in them but often they are also generating emotional effects. We produced therefore two databases of thumbnail images, one database contains images tagged by object-related keywords like beach, mountain, flower, cat. The other database contains images indexed by emotion-related tags like: elegant, colorful, spicy, calm. Details about these databases are described in Section 4. These two databases contain approximately 1.5 million images, tagged with keywords and the corresponding usage statistics explained below. We will work with low-resolution thumbnails, which are bigger than the tiny images of Torralba [1] but they are much smaller than the high-quality images from sites like flickr. These images are used in the image search and browsing context where decisions are often made by quickly glancing (for related investigations of human object recognition performance see [2]).

In this paper we describe how methods from the theory of group representations can be used to compute descriptors from images. We explore the visual significance of these features and demonstrate that they are tools that can be used to investigate the internal structure of the image databases at hand.

The theory of group representations is a generalization of the well-known Fourier analysis. The decision to use these tools is motivated by the basic insight in mathematics and theoretical physics of the importance of group theoretical transformation rules and their connections with invariants and laws-of-nature. In this paper we use two types of symmetries: the first uses the fact that the sensors that produce the digital images are located on a regular (usually quadratic) grid. The other, less strict, regularity concerns the color properties, here we assume that the R, G and B channels are, in a statistical sense, interchangeable. We will show that both cases can be treated in the unified framework of the dihedral groups.

Visual processing on the retina is quite complex and powerful. One of the basic organization principles there is the usage of receptive fields where the response of several receptors are combined and processed and only the result of this processing is send to the brain for further evaluation ([3], [4]). We therefore choose to characterize an image by picking a

number of patches in an image, analyze their pixel distributions and combine the result of the analysis into statistical parameters that describe the image content. The strategy used in this paper is therefore more closely related to the GIST descriptors than to the interest-point based strategy (see [5] for a comparison).
The symmetry properties of the sensor is described by the dihedral group D(4) for the square grid. For the three color channels we use the symmetric (permutation) group S(3) as symmetry group which is identical to the dihedral group D(3). Combining them both we will see that for images on the square grid the relevant symmetry group of low-level processing of RGB patches on a regular grid is the tensor product $\mathrm{D}(4) \otimes \mathrm{D}(3)$. Applying the representation theory of finite groups we will then describe the following processing pipeline: Using the projection theorems of the general theory we will construct a filter system that splits the original spatio-color distribution of the pixels into the smallest sub-components by projecting the original RGB-vectors on a patch in the image onto a collection of subspaces. These transformations can be implemented by additions and subtractions only and there is a fast-version of these filters that correspond to the FFT in the case of the discrete Fourier transform. In the next stage of the pipeline we use polar-coordinate systems in these smallest subspaces where the radial part is an invariant under the transformations of the symmetry group and the angular part is related to the transforming group element. Every projector and every image patch results in one such polar descriptor and we characterize the whole image by the probability distribution of the collection of these descriptors. These probability distributions can then be evaluated by clustering methods, Support-Vector-Machine (SVM) classifications or other processing methods.
In the rest of the paper we will first sketch the necessary background from representation theory. Then we will describe how the filter construction and we will briefly illustrate how the filter process can be optimized. We will then derive the descriptors or signatures from the raw filter results. Next we describe the image databases used in the experiments and then we illustrate and evaluate the strategy by characterizing the different image categories in terms of the statistical properties of the descriptors and by showing some classification results that were obtained by running an SVM-classifier on some two-class classification problems.

2 Spatio-Chromatic Filters

Almost all digital images are now captured with sensors located on a square grid. A symmetry group of the grid is a set of transformations that map the points on the grid into itself. For the square grid the symmetry transformations are the rotations around the origin of the square with rotation angles $k{*}90, k = 0, \ldots 3$ degrees and the reflections on the symmetry axes. These transformations form the dihedral group D(4). The construction of filter systems based on the dihedral symmetry group of the sensor grid is described in [6].
For the RGB color channels we can use a similar argument to construct a group theoretical model. We consider the three color channels as points of a equilateral triangle. The symmetry group is the dihedral group D(3) with six elements consisting of three rotations (with rotation angles $k * 120$ degrees) combined with a possible reflection on a symmetry axis of the triangle. This group is identical to the group S(3) of permutations of three elements. The vector space of function defined on the triangle has dimension three (RGB

vectors have length three) and we get again a representation of the symmetry group D(3) on the space of RGB vectors. An application of D(3) in filter design can be found in [7] and in [8],[9] where the permutation group was used to analyze the structures in the space of RGB-histograms.
In the following we will only consider filter kernels defined on a 4×4 square with 16 pixels. We use it since this case is one of the easiest examples to illustrate the general approach and since we found it also useful in our applications which will be described later. The possible distributions of RGB values on this patch span a $4 \times 4 \times 3 = 48$ dimensional vector space V. The representation theory of the dihedral groups provide the tools to split this vector space into a direct sum of subspaces of minimal dimensional that are invariant under all spatial transformations in D(4) and RGB permutations in D(3).
This leads to the following decomposition of the full 48-dimensional space V

$$V = [3V_{ti} \oplus 3V_{ai} \oplus V_{pi} \oplus V_{mi} \oplus 4V_{2i}] \oplus [3V_{tc} \oplus 3V_{ac} \oplus V_{pc} \oplus V_{mc} \oplus 4V_{2c}] \tag{1}$$

where we use the following notational convention: kV_{xy} denotes k copies of the space V_{xy}, the second part of the subscript $y = i$ indicates if the filters operate on the one-dimensional intensity component or $y = c$ on the two-dimensional color opponent vector. The first part x of the subscript denotes to connection to the spatial transformation property of the filters. The dimensions of the representation spaces are collected in Table 1.

Space	Dimension
$V_{ti}, V_{ai}, V_{pi}, V_{mi}$	1
$V_{2i}, V_{tc}, V_{ac}, V_{pc}, V_{mc}$	2
V_{2c}	4

Table 1: Dimensions of subspaces

The filter system on the 4×4 square window is now defined by the projection operators to the subspaces in Equation 1. As an illustration we describe the handling of an RGB vector in this framework. The general results from representation theory show that the three-dimensional RGB space C with D(3) symmetry can be divided into two components C_i, C_c corresponding to the intensity and the chromaticity of the RGB vectors. One basis transformation that splits the RGB space C into these two components ($C = C_i \oplus C_c$) is given by the matrix

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 0 \\ 1 & 1 & -2 \end{pmatrix} \tag{2}$$

The new coordinates correspond to the intensity value (first component) and the Red-Green and Yellow-Blue components respectively. This matrix is orthogonal but not orthonormal. If we normalize the rows to unit length then we see that the first row is a constant vector whereas the other two rows are formed by elements of the form $\sin(2 \cdot k \cdot \pi/3)$ and $\cos(2 \cdot k \cdot \pi/3)$. We can therefore view the new coordinate values as Fourier coefficients of the original RGB measurements.
We see immediately that the first, intensity, component is invariant under all transformations of the group D(3) since the sum (R+G+B) does not change under permutations of the RGB components. If we normalize the last two rows (or scale the resulting raw filtering results) then we can show with a simple calculation that a permutation σ_3 of

the R and G values, keeping the B channel fixed, results in a sign change in the RG filter result, whereas the YB filter result is unchanged. We can also see that a 120 degrees rotation of the RGB triangle leads to a 120 degrees rotation of the two-dimensional (RG, YB) vector and the D(3) transformations in the RGB triangle map to corresponding D(3)transformations in the (RG, YB) plane. From the construction it can be seen that the basic filter operations can be implemented using additions/subtractions only. Furthermore it is possible to compute the partial sum R+G first and save an addition. Using the same technique one obtains fast transforms similar to the Fast Fourier Transform. Scaling of the resulting filter results can be done after the integer-based filtering. For more information on the general background see [10]. The D(4) related theory is described in [6] and the color part in [7].

3 Computation of Signatures

Up to now (see Eq. (1)) we described how a basis transform in the 48-dimensional vector space (containing the 4×4 pattern of RGB vectors) creates a subset of 24 subspaces that are invariant under joint spatio-color changes defined by the actions of the $D(4) \otimes D(3)$group. Simplifying notations we rewrite Eq.(1) as

$$V = V_1 \oplus \ldots \oplus V_{24} \tag{3}$$

were we use the same order of the spaces as in Eq.(1). The first eight spaces have thus dimension one, the next twelve have dimension two and the remaining four have dimension four. For a 48-dimensional vector x from the RGB patch we get a decomposition

$$x = x_1 + \ldots + x_{24} \tag{4}$$

and from that a signature vector

$$R = (r_1, \ldots, r_{24}) = (\|x_1\|, \ldots, \|r_{24}\|) \tag{5}$$

that describes how the energy of the original pixel distribution is distributed over the different subspaces. These values are by definition independent of dihedral transformations of the underlying original vectors. We call them the signatures of the distribution.
From the construction of the subspaces we get the following list of intuitive interpretations of these radial variables

r_1, r_2, r_3**:** Average gray values on the three orbits

$r_9, \ldots, r_{12}$**:** Gray value gradients (edges)

r_{13}, r_{14}, r_{15}**:** Spatially homogeneous color changes across color channels

$r_{21}, \ldots, r_{24}$**:** Spatio-color changes

This analysis can be extended using the group theoretical construction of the filter results. We illustrate this interpretation with the example of the variable x_4. This is a one-dimensional projection originating in the four-point orbit consisting of the four inner pixels and the summation of the RGB channels. This is the intensity distribution on the inner 2×2 patch. The construction of the filter insures that under the reflection,

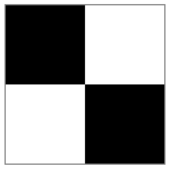


Figure 1: 2×2 Line Filter

or a 90 degree rotation, of the underlying patch the filter value changes its sign. We write $x_4 = |x_4|\, s_4 = r_4 s_4$ where s_4 has value one or minus one. The value of r_4 describes which proportion of the filter result vector is located in subspace V_4 and the sign s_4 divides the patterns into two classes, those that are related to a reflection on the axis or a 90 or 270 degree rotation. A similar interpretation can also be obtained for the other filter functions. The visual appearance of the filter is that of a 2×2 line-filter shown in Figure 1

4 Image databases

We applied the theory to two image databases, both crawled from the Internet by the Picsearch image search engine `http://www.picsearch.com`. The first database, in the following denoted by **objdb**, consists of traditional object- and scene-categories, such as car, animal, beach, garden etc.. It also contains two art-related categories, Andy-Warhol and Claude-Monet. The second database, referred to as **emodb**, contains images that were collected based on emotion related keywords. Examples are calm, elegant, colorful etc. Images in **emodb** were indexed by the Picsearch crawler around March 2009. It is based on 98 main keywords, resulting in 1.2 million images. The database also contains additional tags, meta-data describing how many times each image has been viewed and clicked during March through June 2009. The ratio between number of clicks and number of views is used as a rough estimate of popularity (only for images that have been viewed at least 50 times). The database can be downloaded from our web-site `http://diameter.itn.liu.se/emodb`.
The database **objdb** was created in the same way, but with 29 keywords focusing on objects/scenes, resulting in 320 000 images, all indexed in 2007. Both databases use thumbnail images, with a maximum size of 128 pixels (height or width). Relationships between images and keywords were not checked afterwards and we use the databases as they are. Even if we can find a lot of junk-images in each database, we assume, however, that a majority of the images have some kind of connection to the keyword in use. We also mention that Picsearch uses a family filter for excluding non-appropriate images, otherwise we can expect all kinds of images in the database.

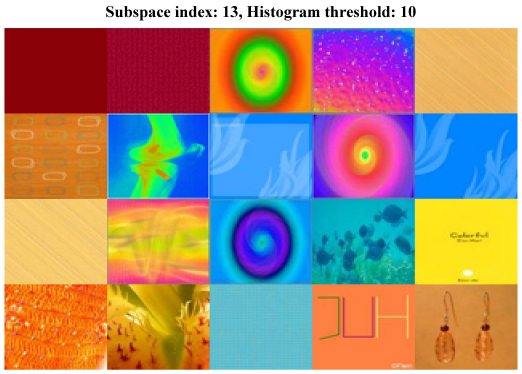
5 Evaluations

In our first experiments we computed the raw-filter images from the original thumbnail images. In the case where the image consisted of only one (intensity) channel we du-

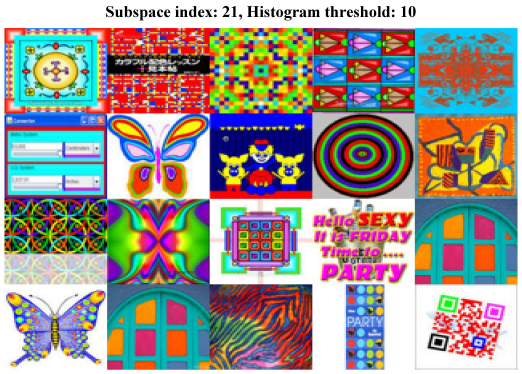
plicated the original image to generate an RGB image with three identical RGB color components. Instead of a convolution based filtering we divided the scaled images into disjoint 4×4 blocks and computed for each block the integer-valued transformation of the block resulting in 48 filter values. The basic descriptor for an image is thus a feature matrix F of size $48 \times B$ where B is the number of selected 4×4 blocks in the image. Next we collected the 48 raw-filter results into the 24 categories described in Eq.(4) above and computed for each category the 24 radial signature values from Eq.(5) that describe which portion of the feature vector falls into the corresponding subspace. The result is a description of a block by 24 non-negative parameters. The image as a whole is characterized by a feature matrix R of size $24 \times B$.

(a) Intensity Line r_4

(b) Intensity Edge r_9



(c) Homogeneous Color r_{13}

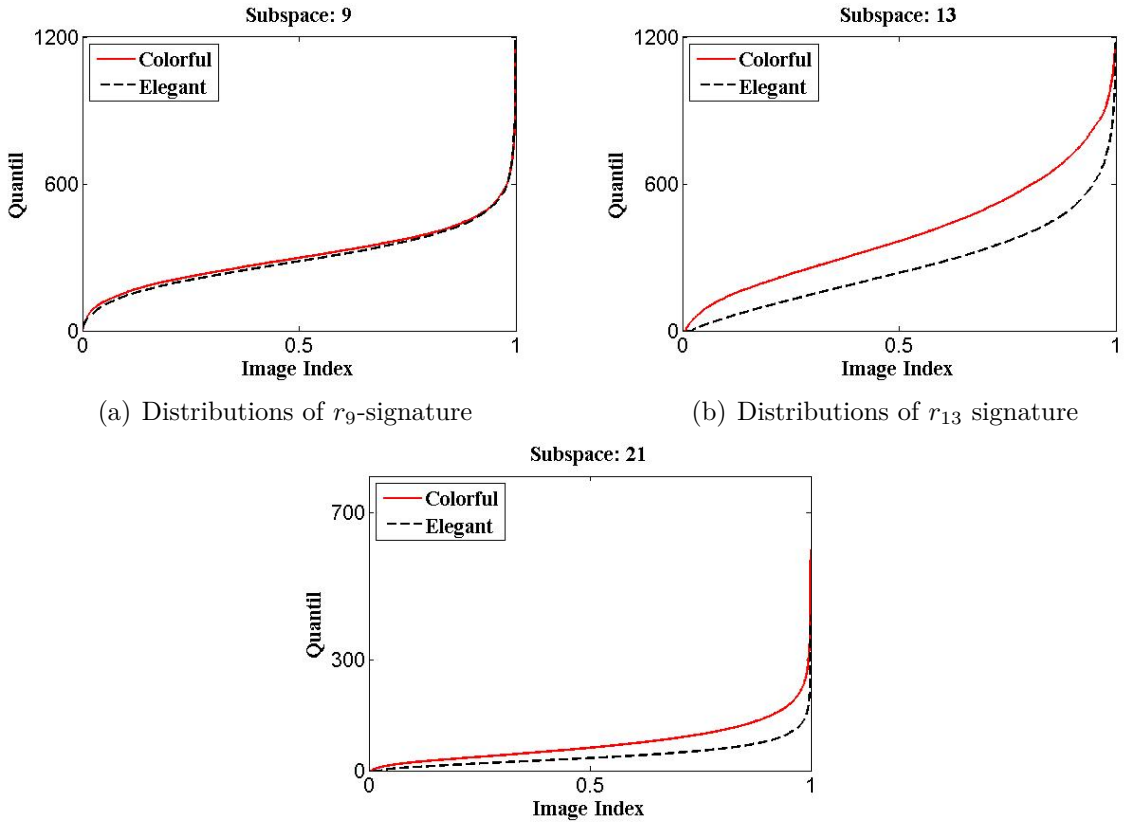


(d) Color Edge r_{21}

Figure 2: Images with highest radial values

Next we illustrate the visual properties of four different subspaces by collecting those 20 images in the colorful category that have the highest signature values in these spaces in a mosaic. This is shown in Fig. 2(a)-2(d). These images were selected as follows: First we computed the histogram of the corresponding signature for every image in the category using 22 bins. We then summed the last twelve histogram values (belonging to the highest signature values) and choose those images in the database that had the highest values for the corresponding sum. We see that images in Fig 2(a) contain many thin line-like structures, those in Fig 2(b) have high intensity contrasts and larger structures. In Fig 2(c) the images have homogenous color variations and Fig 2(d) contains images with

spatio-color variations. This corresponds to the interpretation we gave earlier.
The following figures illustrate the distributions of the 0.9-quantile values for all images in the colorful and the elegant categories and plotting the resulting values as a curve. The Fig. 3(a)-3(c) show the curves for the colorful and elegant categories which both contained roughly 20000 images. The curves are computed from radial signatures r_9, r_{13} and r_{21}. We see that the values obtained for the intensity-edge signature r_9 are almost identically distributed. For the color related signatures r_{13} and r_{21} the figures show that the curves for the images in the colorful category are significantly higher than those computed from the elegant category. We conclude that images automatically tagged with the text descriptor colorful are indeed visually more colorful than the images in the elegant category.



(a) Distributions of r_9-signature

(b) Distributions of r_{13} signature

(c) Distributions of r_{21} signature

Figure 3: Distributions of quantiles

6 Classification examples

A popular approach for large scale image classification is to utilize image histograms together with a supervised learning algorithm, for instance Support Vector Machines. Among histogram descriptors we can mention global histograms, like ordinary RGB-histograms, or histograms derived from the popular bag-of-words (or bag-of-features) approach, where a codebook is used to obtain a histogram from local patch descriptors. For an up-to-date overview of image descriptors, we refer to the paper by van de Sande et al. [11], where various color descriptors, both global and local, are compared and

Table 2: Two-class classification results for three category-pairs

Keywords pair	1:3	9:12	13:15	13:24	21:24	ALL
beach - garden	0.70	0.71	0.57	0.68	0.57	0.76
warhol - monet	0.66	0.81	0.76	0.81	0.80	0.87
colorful - elegant	0.58	0.59	0.60	0.68	0.68	0.69

evaluated. In this work we illustrate a few classification examples using the proposed histogram representation, and a Support Vector Machine implementation, SVMlight, by Thorsten Joachims [12]. For simplicity and reproducibility reasons, all experiments are carried out with default settings.

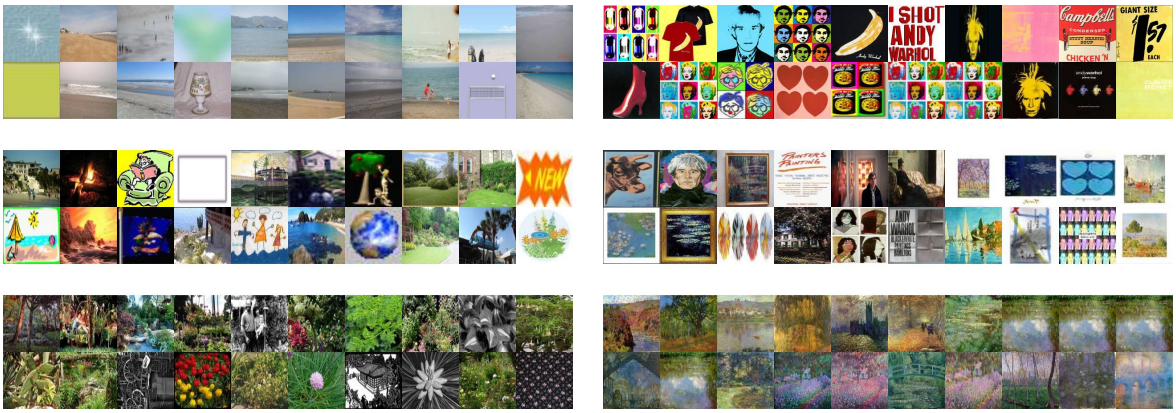
Classification results are illustrated for the following keyword pairs: beach - garden,colorful - elegant, and warhol - monet. We selected these pairs as representative examples of various tasks that can be encountered in image classification. For each keyword we extract a subset of 2000 images from the corresponding database (**objdb** or **emodb**). The subset contains the 1000 most popular images (as described in Section 4), and 1000 images randomly sampled from the remaining ones. For each classification task, we create a training set containing every second image from both keywords in the pair, and remaining images are used for testing.

Two-class classification results are given in Table 2. Different columns correspond to classifications based on different subsets of radial signatures as explained in Sec. 3. The last column displays the accuracy for the case where the entire descriptor is used. The classification accuracy is given by the proportion of correctly labeled images. A value of 0.75 means that 75% of the images were labeled with a correct label. From the table we can see that the homogeneous color signatures 13-15 are slightly less powerful than the color-spatial-filters 21-24 and that all color features 13-24 are significantly better than both of them separately. Also intensity edges are more important than color features and the warhol-monet pair seems to be the easiest task. Note that these tests are very limited and that the processing is based on the very small windows of only 4×4 pixels.

To illustrate the classification result we collected some subsets of classified images. The results are obtained from the classification where all signatures were used (column ALL in Table 2). They are shown in Fig. 4(a), Fig. 4(b) and Fig. 4(c). Each figure shows the 20+20 images, obtained as the most positive and most negative score from the Support Vector Machine. In between we collect 20 images corresponding to scores close to 0 (neutral images). All images are re-scaled to square-size for viewing purposes.

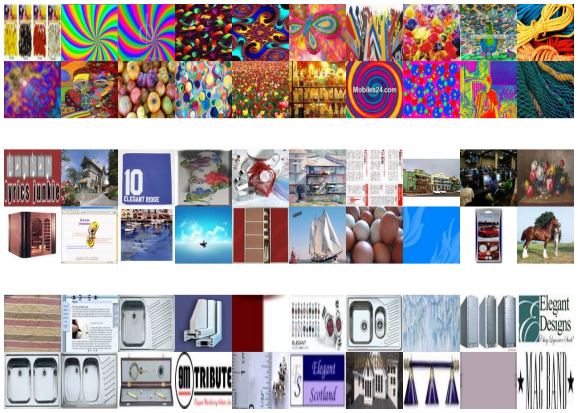
For comparison, we do exactly the same experiments, but with 2000 randomly selected images belonging to each keyword where we did not enforce the popular images to be included in the subset. The classification accuracy can be seen in Table 3. We notice that the accuracy is slightly worse, and conclude that popular images are beneficial, but not crucial, for the classification accuracy.

As a last example we show an example where we used the full decomposition in which both the radial signature, the group elements index and the remaining angle was used. For the three two-dimensional descriptors x_{13}, x_{14} and x_{15} we computed the decomposition and then the histograms for each image. We used six bins for the group elements (one for each element in S(3)), 12 for the remaining angle (5 degrees per bin) and seven for the radial signature. The number of bins was chosen so that the full histogram had approximately 512 bins. We then used the SVM to classify the same 4000 colorful-elegant



(a) Pair beach - garden

(b) Pair Warhol - Monet



(c) Pair colorful - elegant

Figure 4: SVM Classification using all signatures

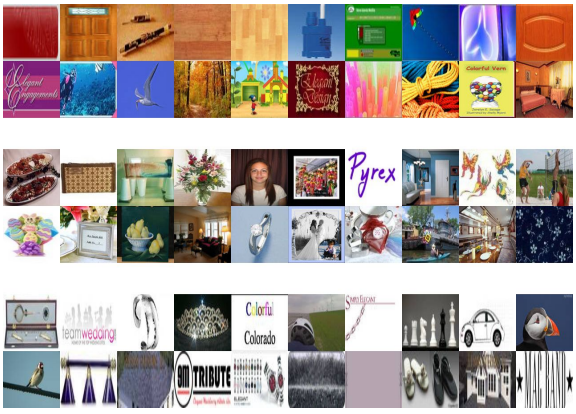
images as in previous experiments. We found that the success rates for using the 504 bin histogram computed from feature x_{13} was 0.6495, if we used all three histograms (with 1512 parameters in total) the success rate decreased to 0.6480. This should be compared to a success rate of 0.5965 when all three radial signatures r_{13}, r_{14}, r_{15} were used. An illustration of the classification examples is found in Figs. 5(a) and 5(b).

7 Conclusions

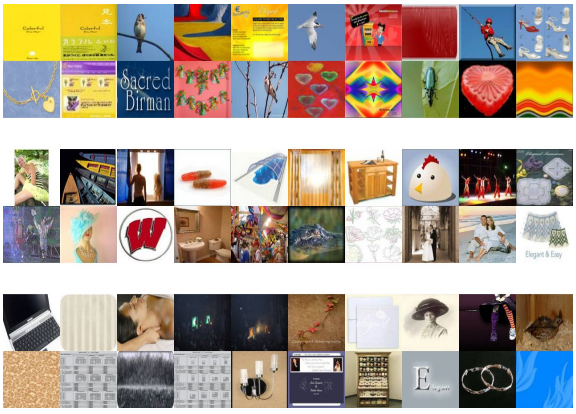
We started from the observation that the sensor grid and the color channels permit symmetry groups D(4) and D(3). Only based on that observation we derived a coordinate transform that transformed patches of RGB vectors to a new coordinate system in which all projected vectors have minimal lengths and transform under the rules determined by the original, underlying symmetry groups. The transform is suitable for fast implementations (also on special hardware like GPU's). We then illustrated the visual interpretation of these filters and used them to analyze the content of two large image databases constructed by the commercial Picsearch engine. We also used them as input to SVM and classified images from two-category pairs. These experiments show that the filter system is not only attractive from an abstract and an computational point of view but that it also allows to investigate and exploit structures in large spaces of visual information.

Table 3: Two-class classification results for 2000 completely randomly selected subsets of images

Keywords pair	1:3	9:12	13:15	13:24	21:24	ALL
beach - garden	0.67	0.69	0.57	0.66	0.58	0.72
warhol - monet	0.66	0.79	0.75	0.79	0.78	0.85
colorful - elegant	0.55	0.57	0.58	0.67	0.65	0.66



(a) Pair colorful - elegant, r_{13}



(b) Pair colorful - elegant, x_{13}

Figure 5: Pair colorful - elegant

We only illustrated the methodology with one of the simplest possible configuration but applications in more complicated combinations are straightforward using the same tools described in this paper.

References

[1] A. Torralba, R. Fergus, and W.T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Trans Pattern Anal Mach Intell*, 30(11):1958–1970, 2008.

[2] S. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *Nature*, 381(6582):520–522, 1996.

[3] D. H. Hubel. *Eye, brain, and vision*. Scientific American Library, New York, 1988.

[4] B.A. Olshausen and D.J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[5] M. Douze, H. Jegou, H. Sandhawalia, L. Amsaleg, and C. Schmid. Evaluation of gist descriptors for web-scale image search. In *ACM International Conference on Image and Video Retrieval, CIVR 2009*, pages 140–147, Santorini Island, 2009.

[6] R. Lenz. Investigation of receptive fields using representations of dihedral groups. *Journal of Visual Communication and Image Representation*, 6(3):209–227, September 1995.

[7] R. Lenz, T. H. Bui, and K. Takase. A group theoretical toolbox for color image operators. In *Proc. ICIP 05*, pages III–557–III–560. IEEE, September 2005.

[8] R. Lenz and P. Latorre Carmona. Transform coding of RGB-histograms. In *International Conference on Computer Vision Theory and Applications (VISAPP)*, pages 117–124. Institute for Systems and Technologies of Information, Control and Communication), 2009. ISBN: 978-989-8111-69-2.

[9] R. Lenz and P. Latorre Carmona. Hierarchical S(3)-coding of RGB histograms. In A. Ranchordas et al., editor, *Selected papers from VISAPP 2009*, volume 68 of *Communications in Computer and Information Science*, pages 188—-200. Springer, 2010.

[10] J.-P. Serre. *Linear representations of finite groups*. Springer-Vlg, New York ;, 1977.

[11] K. van de Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.

[12] T. Joachims. Making large-scale support vector machine learning practical. In *Advances in kernel methods: support vector learning*, pages 169–184. MIT Press, 1999.