Spatio-chromatic image content descriptors and their analysis using Extreme Value Theory *

Vasileios Zografos and Reiner Lenz

Computer Vision Laboratory, Linköping University, Sweden
zografos@isy.liu.se, reile@itn.liu.se

Abstract. We use the theory of group representations to construct very fast image descriptors that split the vector space of local RGB distributions into small group-invariant subspaces. These descriptors are group theoretical generalizations of the Fourier Transform and can be computed with algorithms similar to the FFT. Because of their computational efficiency they are especially suitable for retrieval, recognition and classification in very large image datasets. We also show that the statistical properties of these descriptors are governed by the principles of the Extreme Value Theory (EVT). This enables us to work directly with parametric probability distribution models, which offer a much lower dimensionality and higher resolution and flexibility than histogram representations. We explore the connection to EVT and analyse the characteristics of these descriptors from a probabilistic viewpoint with the help of large image databases.

1 Introduction

With the considerable increase in online visual content, there has been a great demand for tools to handle efficiently, large and dense collections of image data. Furthermore, online images exhibit a large variation in content, appearance and quality. An automated image search engine must therefore be able to process quickly such large datasets and accurately recover a selection of images that fit a user's query. As a result, many sophisticated feature descriptors [1], are not capable of dealing with image databases comprised of many million samples, in a reasonable time frame.

Motivated by these observations, we suggest a novel spatio-chromatic image descriptor and an associated model selection method that are well suited for very fast search over very large image databases. These descriptors (or filters) are designed to preserve important image information (e.g. colour edges and line features), while being invariant under certain spatio-chromatic changes. Such characteristics can be useful in tasks of object recognition, image retrieval and classification. In this paper, we explore the visual significance of these descriptors and demonstrate that they form effective tools, which may be used to investigate the internal structure of the image databases.

* Funded by the EU FP7/2007-2013 programme, under grant agreement No 247947 GARNICS.

In the rest of this paper, we briefly introduce the theory behind the construction of our descriptors in Sec. 2. In Sec. 3 we review the main properties of EVT and explain how it is connected to the descriptors. In Sec. 4 we propose a simple approach for EVT model estimation and selection. We continue with experiments and their analysis on public image datasets in Sec. 5. Finally, we conclude with a succinct summary discussion in Sec. 6.

2 Spatio-chromatic descriptors

In this work, we propose a number of spatio-chromatic descriptors that have been constructed using the representation theory of finite groups (see [2]). The groups used are the dihedral groups D(3) and D(4). The dihedral group D(n) is defined as the group of all geometry preserving transformations (rotations and reflections) of the regular n-sided polygon, in this case the triangle and the square. The group D(4) exploits the square grid structure of most modern image sensors. The details of the usage of D(4) are described in [3]. The usage of D(3) is based on the observation that in a statistical sense, the three color channels R,G,B are interchangeable. This statistical permutation property suggests the usage of the permutation group S(3) of three elements, which is identical to the group D(3). For an intuitive understanding it might be helpful to identify the three channels R,G and B as corners of the regular triangle. For additional details see [4].

For the descriptor construction, we use only RGB vectors on 4×4 neighborhoods around a pixel. These vectors are all located in a 48-dimensional space. The tools of representation theory are applied to split this space into its smallest subspaces that are invariant under all spatial and RGB transformations in D(4) and D(3). The result is that the RGB space is first transformed into the 1-dimensional R+G+B (intensity) component and the 2-dimensional color opponent space given by the combinations RG=R-G and YB=R+G-2B. This is then followed by a combination with the spatial D(4) filters. The final result is a decomposition of the original 48d space into 24 subspaces of dimensions 1, 2 and 4. The first 12 are spatial filters operating on the intensity component R+G+B whereas the other 12 filters operate on the two-dimensional opponent color space (RG,YB). This decomposition is implemented by an orthonormal transformation and so the norms of the vectors in the subspaces are preserved under the spatial and color operations in D(4) and D(3). To summarize: the original image is first filtered with 48 filters, then the magnitudes of 24 collections of filter results, are computed and the produced images $r_1, ..., r_{24}$ with non-negative pixel/magnitude values provide the spatio-spectral descriptors of the original image. Figure 1 gives an illustration of the relation between the original image and the 24 computed descriptor images. A computer implementation of the filtering process is available from [5].

3 Extreme value Theory

Extreme Value Theory (EVT) deals with the behaviour of the extrema (minima and maxima), of a probability distribution. EVT has been applied to many

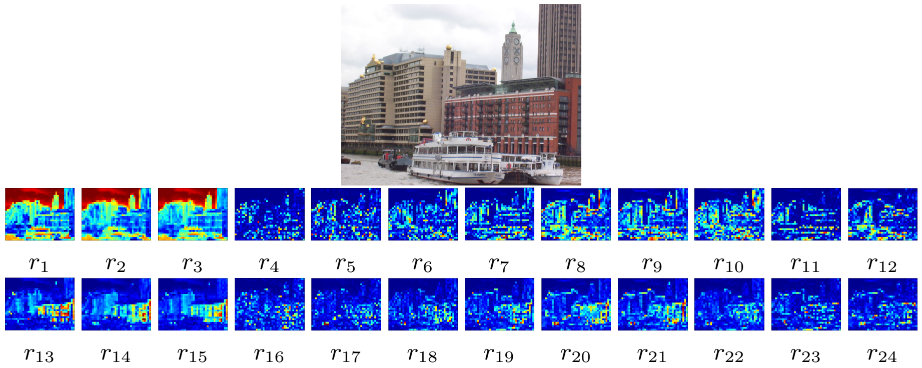


Fig. 1. The intensity (middle row) and colour (lower row) filter results from a typical image. Note that the first three filters represent averaging of pixel values.

natural processes and also in biological and computer vision. In this paper, we suggest a connection between filtered image data and EVT, and we have used the latter to model and analyse the distribution of the former. In the next chapter, we will show experimental results, which demonstrate that the vast majority of examined filtered images follow the EVT model.

3.1 The basics of univariate EVT

EV theory, similarly to the central limit theorem, states that the non-degenerate asymptotic distributions of the sample extremum of a process, must belong to one of just three possible general families regardless of the original distribution function F. Furthermore, it is not necessary to know the detailed nature of F or which limiting form (if any) it gives rise to. As a matter of fact, we just need to know the behaviour of the tails of $F(x)$ for large x, so that a good deal may be said about the asymptotic properties of the extremum.

More formally, suppose that we have an i.i.d. sequence of random variables X_N whose common distribution is $F(x)$=Pr$\{X_i \leq x\}$. Also let s_n=Max$^{(n)}(X_N)$ denote the n$_{th}$ sample maximum of the process. Then Pr$\{s_n \leq x\}$=$F(x)^n$. For non-trivial limit results, and suitable normalising constants a_n>0, b_n, the previous equation converges to Pr$\{a_n(s_n - b_n) \leq x\}$= $F(a_n^{-1}x + b_n) \rightarrow H(x)$. In [6] it is shown that the possible non-degenerate limiting forms of H are:

$$\begin{array}{lll} H(x) = \exp\left(-\exp(\frac{\mu - x}{\sigma})\right) & ,\forall x & \text{Gumbell} \\ H(x) = 1 - \exp\left(-\left(\frac{x-\mu}{\sigma}\right)^k\right) & ,x > \mu & \text{Weibull} \\ H(x) = \exp\left(-\left(\frac{x-\mu}{\sigma}\right)^{-k}\right) & ,x > \mu & \text{Fréchet} \end{array} \tag{1}$$

where μ, σ, k are the location, scale and shape parameters of the distributions respectively.

3.2 A simple stochastic model

The utility of EV theory in the study of low-level vision can be explained with the following simple model: consider a black-box unit U with input X the pixel values from a finite window in a digital image (a similar analogy can be applied to the receptive fields of a biological vision system). The purpose of this black-box is to measure the amount of some non-negative quantity, $X(t)$ that changes over time. We write $u(t)$=$U(X(t))$. We also define an accumulator $s(n)$=$\int_0^n u(t)dt$ that accumulates the measured output from the unit, until it reaches a certain threshold $s(n)$=$\mathrm{Max}^{(n)}(X)$ or a certain period of time, above which the accumulator is reset to zero and the process is restarted. If we consider $u(t)$, $s(n)$ as stochastic processes and select a finite number N of random samples u_1,...u_N, then their joint distribution $J(u_1,...,u_N)$ and the distribution $Y(s_N)$ of s_N, depend on the underlying original distribution $F(X_N)$. At this point we may pose two questions:

1. When $N \to \infty$ is there a limiting form of $Y(s) \to \Phi(s)$?
2. If there exists such a limit distribution what are the properties of the black-box unit U and of $J(u_1,...,u_N)$ that determines the form of $\Phi(s)$?

In [7] the authors have demonstrated that under certain conditions on $Y(s)$ the possible limiting forms of $\Phi(s)$ are the familiar forms in (1) and depend on the tail behaviour of $F(X)$ at large X. In our particular case, we use as units U the black-box that computes the absolute value of the filter result vectors from the irreducible representations of the dihedral groups. The filter vectors not associated with the trivial representation, are of the form s=$\sum(x_i$-$x_j)$ where x_i, x_j are pixel values. We can therefore expect that these filter values are usually very small and that high values will appear very seldom. In addition, these sums are calculated over a small, finite neighbourhood, and for this reason, the random variables are highly correlated. In short, the output for each filter has a form similar to the sums described in [7], and so it should be possible to use the EVT to model their distribution. As we will show experimentally later, the EVT models in (1) provide a good fit to our filtered data, which is a strong indication that the requirements for EVT equivalence from [7] generally hold. We also note, that since we are always dealing with positive quantities (norms of sums) that have a strictly positive support, we do not use the Gumbel model, which is unbounded, but only the Weibull and Fréchet models.

4 Proposed approach

In the previous section, we have discussed the connection between our proposed filters and the EVT models. In this section, we suggest a simple approach for estimating the parameters of these models, using maximum likelihood, and then selecting the model that has the best fit using a residual analysis approach.

Distribution parameter estimation: We begin with a log-likelihood function $\Lambda(\theta)$ that expresses the conditional probability of realising the data sample given the model parameters θ=(μ, σ, k), and then try to determine the choice of parameters (ML estimates) that maximise the likelihood for the available data.

Since the 3-parameter Weibull and Fréchet distributions, do not have closed form expressions of the ML estimates, we need to apply an iterative method, such as the Newton-Rhapson approach. The iteration step, which usually is executed until convergence, is given by $\hat{\theta}_{t+1}$=$\hat{\theta}_t$+p_t, for t=0,1,2..., where p_t=-$\nabla^2 f_t^{-1} \nabla f_t$ is a search (descend) direction on the log-likelihood function. As such, we need expressions for the gradient ∇f_t and Hessian $\nabla^2 f_t$ of the Weibull and Fréchet distributions. For the Weibull, the gradient $\nabla f_t = \left[\frac{\partial \Lambda(\theta)}{\partial \theta}\right]$ is given by:

$$\begin{aligned}
\frac{\partial \Lambda(\theta)}{\partial \mu} &= -(k-1) \sum \tfrac{1}{x_i-\mu} + \tfrac{k}{\sigma} \sum \left(\tfrac{x_i-\mu}{\sigma}\right)^{k-1}, \\
\frac{\partial \Lambda(\theta)}{\partial \sigma} &= \tfrac{k}{\sigma} \left[-n + \sum \left(\tfrac{x_i-\mu}{\sigma}\right)^{k}\right], \\
\frac{\partial \Lambda(\theta)}{\partial k} &= \tfrac{n}{k} - n \log \sigma + \sum \log(x_i-\mu) - \sum \left(\tfrac{x_i-\mu}{\sigma}\right)^{k} \log \left(\tfrac{x_i-\mu}{\sigma}\right),
\end{aligned} \tag{2}$$

and the Hessian $\nabla^2 f_t = \left[\frac{\partial^2 \Lambda(\theta)}{\partial \theta \partial \theta'}\right]$ by:

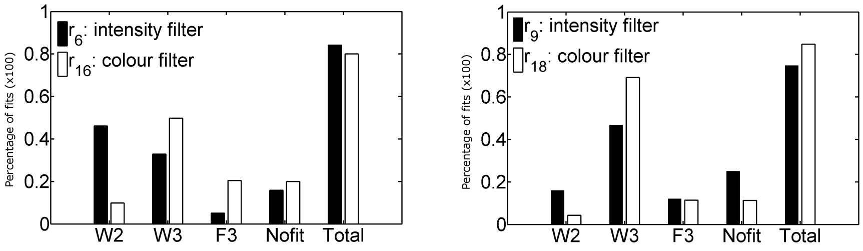
$$\begin{aligned}
\frac{\partial^2 \Lambda(\theta)}{\partial \mu^2} &= -(k-1) \left[\sum \left(\frac{1}{x_i-\mu}\right)^2 + \tfrac{k}{\sigma^2} \sum \left(\tfrac{x_i-\mu}{\sigma}\right)^{k-2}\right], \\
\frac{\partial^2 \Lambda(\theta)}{\partial \mu\, \partial \sigma} &= \frac{\partial^2 \Lambda(\theta)}{\partial \sigma\, \partial \mu} = -\left(\tfrac{k}{\sigma}\right)^2 \sum \left(\tfrac{x_i-\mu}{\sigma}\right)^{k-1}, \\
\frac{\partial^2 \Lambda(\theta)}{\partial \mu\, \partial k} &= \frac{\partial^2 \Lambda(\theta)}{\partial k\, \partial \mu} = -\sum \tfrac{1}{x_i-\mu} + \tfrac{k}{\sigma} \sum \left(\tfrac{x_i-\mu}{\sigma}\right)^{k-1} \log \left(\tfrac{x_i-\mu}{\sigma}\right) + \tfrac{1}{\sigma} \sum \left(\tfrac{x_i-\mu}{\sigma}\right)^{k-1}, \\
\frac{\partial^2 \Lambda(\theta)}{\partial \sigma^2} &= \tfrac{k}{\sigma^2} \left[n - (k-1) \sum \left(\tfrac{x_i-\mu}{\sigma}\right)^{k}\right], \\
\frac{\partial^2 \Lambda(\theta)}{\partial \sigma\, \partial k} &= \frac{\partial^2 \Lambda(\theta)}{\partial k\, \partial \sigma} = -\tfrac{1}{\sigma} \left[n - \sum \left(\tfrac{x_i-\mu}{\sigma}\right)^{k} - k \sum \left(\tfrac{x_i-\mu}{\sigma}\right)^{k} \log \left(\tfrac{x_i-\mu}{\sigma}\right)\right], \\
\frac{\partial^2 \Lambda(\theta)}{\partial k^2} &= -\tfrac{n}{k^2} - \sum \left(\tfrac{x_i-\mu}{\sigma}\right)^{k} \left[\log \left(\tfrac{x_i-\mu}{\sigma}\right)\right]^2.
\end{aligned} \tag{3}$$

Similarly for the Fréchet:

$$\begin{aligned}
\frac{\partial \Lambda(\theta)}{\partial \mu} &= (k+1) \sum \tfrac{1}{x_i-\mu} - \tfrac{k}{\sigma} \sum \left(\tfrac{x_i-\mu}{\sigma}\right)^{-1-k}, \\
\frac{\partial \Lambda(\theta)}{\partial \sigma} &= \tfrac{k}{\sigma} \left[n - \sum \left(\tfrac{x_i-\mu}{\sigma}\right)^{-k}\right], \\
\frac{\partial \Lambda(\theta)}{\partial k} &= \tfrac{n}{k} + n \log \sigma - \sum \log(x_i-\mu) + \sum \left(\tfrac{x_i-\mu}{\sigma}\right)^{-k} \log \left(\tfrac{x_i-\mu}{\sigma}\right),
\end{aligned} \tag{4}$$

$$\begin{aligned}
\frac{\partial^2 \Lambda(\theta)}{\partial \mu^2} &= (k+1) \left[\sum \left(\frac{1}{x_i-\mu}\right)^2 - \tfrac{k}{\sigma^2} \sum \left(\tfrac{x_i-\mu}{\sigma}\right)^{-k-2}\right], \\
\frac{\partial^2 \Lambda(\theta)}{\partial \mu\, \partial \sigma} &= \frac{\partial^2 \Lambda(\theta)}{\partial \sigma\, \partial \mu} = -\left(\tfrac{k}{\sigma}\right)^2 \sum \left(\tfrac{x_i-\mu}{\sigma}\right)^{-1-k}, \\
\frac{\partial^2 \Lambda(\theta)}{\partial \mu\, \partial k} &= \frac{\partial^2 \Lambda(\theta)}{\partial k\, \partial \mu} = \sum \tfrac{1}{x_i-\mu} + \tfrac{k}{\sigma} \sum \left(\tfrac{x_i-\mu}{\sigma}\right)^{-k-1} \log \left(\tfrac{x_i-\mu}{\sigma}\right) - \tfrac{1}{\sigma} \sum \left(\tfrac{x_i-\mu}{\sigma}\right)^{-k-1}, \\
\frac{\partial^2 \Lambda(\theta)}{\partial \sigma^2} &= -\tfrac{k}{\sigma^2} \left[n - (1-k) \sum \left(\tfrac{x_i-\mu}{\sigma}\right)^{-k}\right], \\
\frac{\partial^2 \Lambda(\theta)}{\partial \sigma\, \partial k} &= \frac{\partial^2 \Lambda(\theta)}{\partial k\, \partial \sigma} = \tfrac{1}{\sigma} \left[n - \sum \left(\tfrac{x_i-\mu}{\sigma}\right)^{-k} + k \sum \left(\tfrac{x_i-\mu}{\sigma}\right)^{-k} \log \left(\tfrac{x_i-\mu}{\sigma}\right)\right], \\
\frac{\partial^2 \Lambda(\theta)}{\partial k^2} &= -\tfrac{n}{k^2} - \sum \left(\tfrac{x_i-\mu}{\sigma}\right)^{-k} \left[\log \left(\tfrac{x_i-\mu}{\sigma}\right)\right]^2.
\end{aligned} \tag{5}$$

For a discussion on more advanced iterative ML estimators and appropriate initial estimates for $\hat{\theta}_0$ we refer to the excellent book by [8] on the Weibull distribution. Similar techniques apply for the Fréchet.



UW results ODB results

Fig. 2. Typical EVT model fitting results from the two databases using the R^2 g.o.f. statistic. Note that the numbers are comparable to those in Table 1.

Model selection: Once we have fitted the two models by ML, we can choose the most appropriate of the two, using a goodness-of-fit (g.o.f.) criterion. This criterion is chosen as the deviation between each of the fitted distributions and the data. Given the empirical cumulative distribution function (cdf) $\hat{\varDelta}_n$ of the data sample $(x_1, ..., x_n)$ [9], and cdf F_n (evaluated at the same points as the data sample) from the Weibull and Fréchet distributions separately (equations in (1)), then the g.o.f. measure, called the *coefficient of determination*, is defined as:

$$R^2 = 1 - \frac{(n-1)\sum_{i=1}^{n}(\hat{\varDelta}_n - F_n)^2}{(n-\zeta)\sum_{i=1}^{n}(\hat{\varDelta}_n - \bar{\varDelta}_n)^2}, \text{ with } \zeta = 3 \text{ the model degrees of freedom.} \tag{6}$$

We choose the model with the maximum R^2 value. If in addition we wish to reject a sample ("no-fit"), we can impose a lower threshold on R^2.

5 Experiments

We have used two datasets for our experiments and subsequent analysis. The first is the **UW** database [10], which consists 1109 colour photos of various vacation locations and natural, outdoor scenes e.g. "Barcelona", "Iceland" etc. The images have been obtained by different cameras and resolutions, but most of them are 756×504 pixels. The second dataset, **ODB** [11], contains 30000 thumbnail images (reduced in size so that the maximum size in one direction is 128 pixels), across 15 object categories. These images were automatically crawled from public web pages using a variety of textual keywords.

5.1 Statistical analysis: Goodness of fit

In this section, we show experimentally the following:

I) the R^2 g.o.f. test is more reliable and robust than common statistical g.o.f. tests for model selection.

II) the 3-parameter Weibull-Fréchet models provide a good fit to the distribution of filtered natural images across different datasets.

	F3	W3	W2	no-fit	hard F3	hard W3
Kolmogorov-Smirnov	80.3%	23%	**99.2**%	25%	**93.1**%	1.1%
g-test	0.81%	16%	66.1%	92.4%	19.4%	4%
χ^2	12.4%	31.6%	88%	**98.8**%	0%	0%
R^2	**99.5**%	**88.7**%	89.7%	87.9%	85.5%	**77.3**%

Table 1. Goodness-of-fit comparative results (as percentage of correct classifications).

III) The 3-parameter Weibull-Fréchet models are more flexible and can describe a larger portion of the data, than the 2-parameter Weibull model alone can.

We demonstrate I) on synthetic data, where the ground truth is known, and compare 4 different approaches: the two sample Kolmogorov-Smirnov test, the χ^2 and g-test and the R^2 test from (6). In total, we carried out 6000 tests, with 500 samples drawn from various distributions (2 and 3-param. Weibull "**W2**", "**W3**"; 3-param. Fréchet "**F3**"; and a 2-param. Lognormal, used here as a "**no-fit**" sample), with realistic parameter settings, that is, ones that we are likely to observe in natural images. The results are shown in Table 1. We can see that the R^2 is the only test that performs consistently well along the different samples even for the "hard" W3 and F3 cases (these are samples with parameter choices that lead to problematic ML surfaces). For this reason, we have decided to use the R^2 test in the remainder of our analysis.

II) and III) are demonstrated on the UW and ODB databases. We applied the filters, selected the appropriate model and rejected any fits with a low R^2 value. The results are shown in Fig. 2. Due to space limitations, we have only included 2 filters (one intensity and one colour), but all the other filters exhibit the same typical behaviour. In particular, for the intensity filters, W2 fits a much larger percentage of data than in the colour case (sometimes the W2 model dominates in the intensity filters), with the F3 being the least contributing sub-model. The former is in line with the findings of [12] when intensity gradient filters are used as image patch descriptors (our descriptors are essentially localised gradient filters). Note however, that by combining all the EVT sub-models we can describe well in excess of 80% of the data. This is something that the W2 alone cannot do. This observation becomes more pronounced for the colour filters, where W3 and F3 have a more prominent role, with W3 alone modelling between 50-70% of the data. In this case, W2 is limited to around 10% and thus the approach of [12] cannot be used to model colour edges, unless one applies W2 to each colour channel separately [13].

We note here that around 15-20% of the fits have been rejected. The no-fit portion includes outliers (i.e. non-natural images, trivial filter results etc) and data where the ML estimation did not converge. These numbers are similar to the no-fit results we have observed in the synthetic tests in Table 1, and are therefore related to the characteristics of the algorithm as well as the data.

In conclusion, these experiments indicate that the EVT may be considered as a viable hypothesis for modelling the distribution of our descriptors (or similar types of intensity and colour gradient filters). Moreover, the additional modelling capacity of W3 and F3, relative to W2 alone, has also been demonstrated.

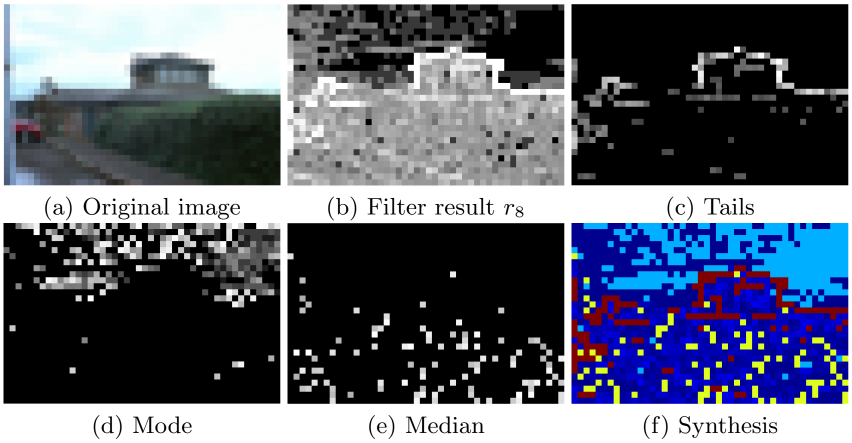

(a) Original image
(b) Filter result r_8
(c) Tails
(d) Mode
(e) Median
(f) Synthesis

Fig. 3. A comparison between the extrema and other regions of a filtered image.

5.2 Further analysis: the σ, k-space

We continue with an analysis of the types of images that are assigned to each submodel (W2, W3 and F3) for a specific filter (r_9) and the image position in the σ, k parameter space. For economy of space, we only demonstrate a single filter on the UW dataset, but the results generalise to all filters and different datasets. We omit the μ parameter since for these datasets it exhibits very little variation and the most important behaviour is observed in the other two parameters. First of all, if we look at Fig. 4 we see a correlated dispersion in the two axes, with the F3 images spanning only a very small region of the space at low σ, k, and well separated from W2 and W3. Also notice how the F3 set typically includes images with near-uniform coloured regions with smooth transitions between them, or alternatively very coarse-textured, homogeneous regions with sharp boundaries. High frequency textures seem to be relatively absent from F3, and on average the image intensities seem to be lower in F3 than in W2 and W3.

On the other hand, the W2 and W3 clusters are intermixed, with W2 mostly restricted to the lower portion of the space. For smaller σ, k values, the W2 images exhibit coarser textures, with the latter becoming more fine-grained as σ, k increase in tandem. Also, there seems to be a shift from low exposure, low contrast images with shadows (small σ, k), to high contrast, more illumination, less shadows when σ, k become large. Furthermore, W2 shows a preference for sharp linear edges associated with urban scenes, whereas W3 mostly captures the "fractal"-type edges, common in nature images.

These observations become more apparent when looking at Fig. 5(a) and (b). In these experiments, we took one (grayscaled) image from the database, and introduced different amounts of noise and smoothing to simulate high and low frequency texture components (Fig. 5(a)) and also linear and nonlinear intensity changes, in order to simulate variations in the amount of illumination (Fig. 5(b)). The image was filtered and the distribution parameters fitted at each instance are shown as trajectories in the σ, k-space. As we have already seen, the images

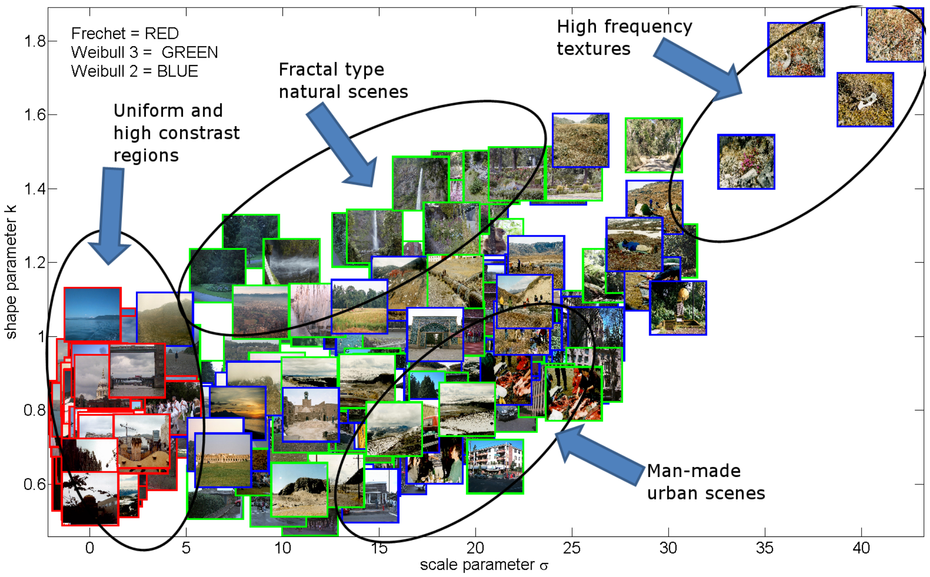
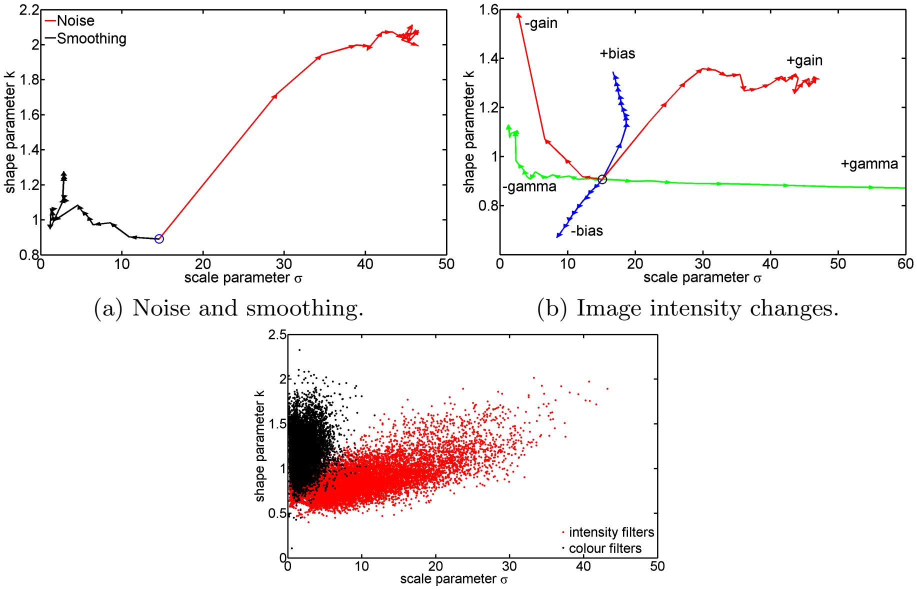


Fig. 4. Image type and model distribution in σ, k-space.

shift to the upper right corner of the space as higher frequency components are added, and for the opposite (smoothing of textures) the images will move towards areas of lower σ and gradually increase in k as the texture homogeneity is increased. For textures that have an approximate constant colour (e.g. sky) the images will cluster on the upper left corner of the space. The UW dataset does not contain such images, and so that space in Fig. 4 remains empty.

If we now look at intensity variations, we see that an increase in gain will move the image toward the upper right corner where all the well-illuminated images lie. When the gain is decreased, we will move towards the upper left corner where the very dark (almost constant) images are. If we now increase the bias, then we see that mostly the k parameter increases (note that the two parameters do not have the same units). Similarly, a decrease in bias will cause a similar decrease in k, while leaving σ relatively intact. Finally, we examine nonlinear changes in intensity (gamma correction). A decrease in gamma value, first reduces the σ parameter only (unlike the bias) and then for additional decreases, the k values start to increase when all the pixels take the same very low (dark) values. Note however, that in this case, the increase in k is much slower and converges to a much lower k, than when the gain was decreased. On the other hand, if we increase the gamma without re-normalising the pixel values between [0,255], then we see a shift towards the lower right corner of the space (increase of σ without increase of k). This region of the σ, k-space is usually empty, but when it is not (depending on the data) it mostly occupied by simple pictorial images such as graphics, designs and logotypes on white background.

In Fig. 5(c) we see a scatter plot for all the images in UW using all the filters (except r_1,...,r_3). We see two very distinct clusters, one for the intensity filters



(a) Noise and smoothing.

(b) Image intensity changes.

(c) Intensity and colour filter scatter plot.

Fig. 5. The behaviour of filtered images in σ, k-space.

that is spread along a σ, k diagonal (as in Fig. 4), and one for the colour filters spread mainly along the k-axis. In conclusion, all the above properties of the σ, k-space are only applicable due to the EV theory and cannot be exploited with histogram representations. The fact that the images exhibit clear clusters and predictable variation in that space, is a good indication of the utility of the EVT framework for retrieval and classification tasks.

Finally, we illustrate the importance of the data at the extrema of a filtered image, as described by the EVT. In Fig. 3(a) we show an image from UW (rescaled for comparison) and its filtered result using r_8 in Fig. 3(b). This is essentially a gradient filter in the x- and y-directions. Next is Fig. 3(c) that shows the response at the tails of the fitted distribution. It it immediately obvious that the tails contain all the important edges and boundary outlines that abstract the main objects in the image (house, roof, horizon, diagonal road). These are the salient features that a human observer will focus on, or that a computer vision system might extract for object recognition or navigation. We also show the regions near the mode in Fig. 3(d). We see that much of it contains small magnitude edges and noise from the almost uniform sky texture. Although this is part of the scene, it has very little significance when one is trying to classify or recognise objects in an image. A similar observation holds for the grass area, which although contains stronger edges than the sky and is distributed near the median (Fig. 3(e)), it is still not as important (magnitude-wise and semantically) as the edges in the tails are. Finally, Fig. 3(f) shows how all the components put together, can describe different regions in the image: the salient object edges in the tails (red); the average response, discounting extreme outliers, (median) in yellow; the most common response in light blue (mode); and the remaining superfluous data in between (dark blue). This is exactly the type of semantic behaviour that the EVT models can isolate with their location, scale and shape parameters, something which is not immediately possible when using histograms.

5.3 Classification and retrieval

We also include a a basic example on how our descriptors may be used, in principle, for classification and retrieval tasks. For this example, we have isolated 4 classes from the ODB dataset, with tags "Andy Warhol", "Claude Monet", "beach" and "garden", each containing 1000 images. After filtering with r_{21} and model selection, we used 75% of the images to train an SVM (with standard settings), and classified the remaining 25%. For the SVM input, we generated 1000 samples from the probability density function of the model chosen for each image.

The overall classification score was 40.5% with the random baseline at 25%. This result is satisfactory considering the many outliers and high variation in the data (due to the automated text-based harvesting) and the lack of specificity in the 4 categories. The 10 top ranked images in each category (one-to-all retrieval) are shown in Fig. 6. The goal here, just like in online image search, is not to retrieve the most representative images for each class (means of the clusters) but the ones that are the furthest away from the SVM decision boundaries (cluster extrema). Therefore, a perfect classification score in CBIR is not as important as fast and accurate retrieval of very few, relevant samples.

Observe in Fig. 6, the differences between the vivid, near-constant colours and sharp edges in the "Warhol" set and the less saturated, softer tones and faint edges of the "Monet" set. In the same way, the "garden" images contain very high frequency natural textures and the "beach" images more homogeneous regions with similarly coloured boundaries. These characteristics are the exact information captured by the filters and the EVT models and which can be used very effectively for image classification and retrieval purposes.

6 Conclusion

In this work, we have presented a set of spatio-chromatic, image content descriptors that are inspired by the theory of group representations. We have demonstrated that by using the EVT to model the output distribution of the descriptors, we can take advantage of specific parametric distribution models that offer a more flexible representation than histograms. Furthermore, additional important characteristics of large image datasets only become visible inside this parametric probability space. These descriptors, combined with the EVT models, offer themselves for very efficient and effective tools for content-based retrieval and classification of image data.

We would like to explain here that the EVT is not the only model one may use to describe similar image properties. In fact [14] have used fragmentation

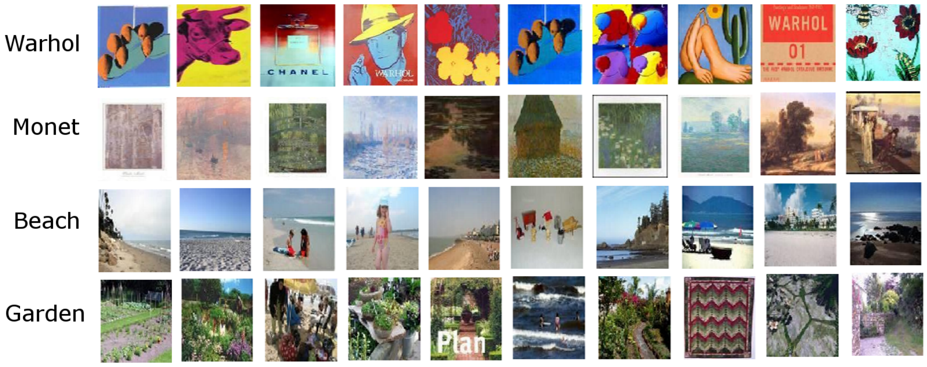


Fig. 6. 4 class image retrieval from the ODB dataset using r_{21} with an SVM.

theory to describe the apparent Weibull distribution of gradient-filtered grayscale images. Despite this, our experiments have shown that EVT is more *flexible*, since [14] advocate a very restrictive fragmentation schedule that might not always apply in practice; more *descriptive*, since EVT has 3 submodels instead of 1 as in [14]; and finally EVT is easily applied to *colour* filters as well.

References

1. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The Pascal visual object classes (voc) challenge. IJCV **88** (2010) 303–338
2. Fässler, A., Stiefel, E.L.: Group theoretical methods and their applications. Birkhäuser, Boston ; (1992)
3. Lenz, R.: Investigation of receptive fields using representations of dihedral groups. Journal of Visual Communication and Image Representation **6** (1995) 209–227
4. Lenz, R., Bui, T.H., Takase, K.: A group theoretical toolbox for color image operators. In: ICIP. Volume 3. (2005) 557–60
5. (`http://people.isy.liu.se/en/cvl/zografos/CBIR`)
6. Gumbel, E.J.: Statistics of Extremes. Columbia University Press, New York (1958)
7. Bertin, E., Clusel, M.: Generalised extreme value statistics and sum of correlated variables. Journal of Physics A: Mathematical and General **39** (2006)
8. Rinne, H.: The Weibull Distribution: A Handbook. CRC Press (2008)
9. Kaplan, E.L., Meier, P.: Nonparametric estimation from incomplete observations. J. Amer. Statist. Assn. **53** (1958) 457–481
10. Li, Y., Shapiro, L., Bilmes, J.: A generative/discriminative learning algorithm for image classification. In: ICCV. Volume 2. (2005) 1605–1612
11. Solli, M., Lenz, R.: Emotion related structures in large image databases. In: ACM CIVR. (2010) 398–405
12. Yanulevskaya, V., Geusebroek, J.M.: Significance of the Weibull distribution and its sub-models in natural image statistics. In: VISAPP. Volume 1. (2009) 355–362
13. Gijsenij, A., Gevers, T.: Color constancy using natural image statistics and scene semantics. IEEE PAMI **99** (2010)
14. Geusebroek, J.M., Smeulders, A.W.M.: Fragmentation in the vision of scenes. In: ICCV. (2003) 130–135