Memory-color segmentation and classification using class-specific eigenregions

Clément Fredembach
Francisco Estrada
Sabine Süsstrunk

Abstract — Memory colors refer to the color of specific image regions that have the essential attribute of being perceived in a consistent manner by human observers. In color correction – or rendering – tasks, this consistency implies that they have to be faithfully reproduced; their importance, in that respect, is greater than that for other regions in an image. There are various schemes and attributes to detect memory colors, but the preferred method remains to segment the images into meaningful regions, a task for which many algorithms exist. Memory-color regions are not, however, similar in their attributes. Significant variations in shape, size, and texture exist. As such, it is unclear whether a single segmentation algorithm is the most adapted for all of these classes. By using a large database of real-world images, class-specific geometrical features, eigenregions, were calculated. They can be used to evaluate how well an algorithm is adapted to segment a given class. A measure of localization of memory colors is given. The performance of class-specific eigenregions were compared to general ones in the task of memory-color-region classification and it was observed that they provide a noticeable improvement in classification rates.

Keywords — *Memory colors, segmentations, evaluation, eigenregions.*

DOI # 10.1889/JSID17.11.1

1 Introduction

By segmenting an image, one effectively decomposes it into a number of disjoint regions. These regions can in turn be analyzed independently and classified according to their content. The various regions and classes that are present in natural images are, however, not of equal importance. Some of the most important ones are the so-called memory colors: blue sky, green vegetation, and skin tones.[10] Human observers locate these classes in very specific areas of the color gamut.[2,20] Thus, many color-rendering and correction algorithms specifically try to map these colors to the correct values. As a result, detecting these regions has been, and still is, a very active area of research.

Detection algorithms generally rely on many different features to classify memory colors: approaches include the use of shape, size, position, color, and texture.[4,21,13,3] Prior to being detected, however, images have to be segmented into meaningful regions. How meaningful a region is depends on the intended application of the segmentation, or image, but most segmentation evaluation methods are predicated on the ideal that all regions are of equal importance. As such, the whole segmentation maps are compared to manually segmented images irrespectively of the image's content.[19]

This work addresses the problem of class-specific segmentation evaluation, where only certain regions are of importance rather than the entire image as well as the localization of memory color regions within natural images. Our framework builds on the eigenregions proposed by Fredembach *et al.*,[11] which are principal component analysis (PCA) based geometrical features that encompass information about the shape, size, and position of regions. The central idea is to calculate class-specific eigenregions, *i.e.*, obtaining different geometrical descriptors for each class. The considered classes have to be reasonably localized across images, *i.e.*, they should usually be found in similar position within images. The classes we consider here (blue sky, green vegetation, and skin tones) generally fulfil, due to physics or photographic composition, this localization criterion.

An objective ground truth for our experiments is obtained by manually segmenting 900 images, 300 per class. These accurate binary segmentation maps are used to calculate class-specific eigenregions that are subsequently compared to the ones resulting from automatic segmentation of the same images. Four segmentation algorithms that exploit very different information are compared: Meanshift (density estimation process),[5] Felzenswalb and Huttenlocher (minimum spanning trees),[8] k-means (Euclidian distances between clusters),[1] and edgeflow (Gabor filter banks).[14]

The comparison is based on the idea that if manual human segmentation is available for a given class, then its N eigenregions provide a reference basis in N-dimension. An algorithm-based segmentation of the same data will, however, provide a *different* basis in the N-dimension. Measuring the distance between these bases effectively quantifies the performance of the algorithm relative to how a set of people would segment it.

The results show a strong class dependency in both the accuracy of segmentation and shape of the eigenregions. The proposed framework can thus be used to quantify, for a given class, the distance between automatic segmentation

Extended revised version of a paper presented at the Sixteenth Color Imaging Conference (CIC-16) held November 10–15, 2008, in Portland, Oregon.
The authors are with the School of Computer and Communications Sciences, Ecole Polytechnique Federic de Lausanne (EPEL), EPEL-IC-LCAV, Station 14, Lausanne, VD 1015, Switzerland; telephone +41-21-693-1273, fax –4312, e-mail: dement.fredembach@epfl.ch.
© Copyright 2009 Society for Information Display 1071-0922/09/1711-01$1.00

and human-generated segmentations, the distance between any two segmentation algorithms, or the influence of input parameters for a given method. In addition, it yields class-specific features that can be used for classification tasks.

2 Segmentation evaluation

When attempting to classify regions, one usually starts by segmenting the image. Because the performance of the region classifier strongly depends on the accuracy of the segmentation, it is often necessary to evaluate the performance of the segmentation algorithm. Such assessment on class-specific data is, however, scarce. In a more global setting, assessing the performance of automatic segmentation is not a new concern and several approaches have been presented that yield a measure of "closeness" or "agreement" with human segmentation. Martin *et al.*[17] first proposed the use of region consistency over a database of human-segmented images[16] to evaluate the performance of automatic segmentation algorithms. These measures of segmentation consistency turned out to be biased toward over- or under-segmentation, so in Ref. 15 the use of precision and recall on region boundaries was suggested instead. A benchmark of several segmentation algorithms based on precision and recall was published in Ref. 6. A different, region-based consistency measure was presented by Ge *et al.* in Ref. 12. Their measure also depends on the overlap between automatic and human segmentations, but it was computed on images that contained only two regions: a salient object and its background. Overlap was measured after deciding (based on the human segmentation) which subset of regions in the automatic segmentation best matched any given human region. More recently, Unnikrishnan *et al.*[19] presented a benchmark based on the Normalized Probabilistic Rand index. This measure compares segmentations through a soft weighting of pixel pairs that depends on the variability of the ground truth data. Other measures of segmentation consistency have been proposed in Refs. 9, 18, and 7. A concise survey of these measures is provided in Ref. 19.

Despite their potential usefulness, each of the above methods for evaluation has its own limitations. First of all, they are global methods that measure the quality of the entire segmentation (all regions are given equal weight, irrespectively of their content); we are here concerned about specific classes. Boundary-based methods will give good scores to under-segmented images, in which two or more distinct (and possibly large) image regions are connected through narrow "leaks." Since most of the boundary is recovered, boundary matching may falsely indicate that the segmentation is accurate. Methods based on overlap such as Ge *et al.* can be biased toward high scores by over-segmenting. In addition, this method assumes some form of expert is available to decide which of the over-segmented regions should be merged together to match human segmentation. The benchmark by Unnikrishnan *et al.*[19] provides interesting insights about the performance of segmentation methods on natural images; however, the question remains of whether particular algorithms are better for specific segmentation tasks, which is one of the fundamental problems addressed in this paper.

3 Eigenregions

Eigenregions were first proposed in Ref. 11 as PCA-based features for image classification. They were obtained by first segmenting a great number of images into regions whose "coverage" was assessed. Working on region coverage allows eigenregions to encompass geometrical attributes, such as shape, size, and position. For the analysis to be tractable, the segmentations are performed on reduced-size images, which is not a concern since downsampling does not alter a region's location or coverage. An illustration of this downsampling procedure is shown in Fig. 1.

Let I be an input image of size $n \times m$, R be a region of I, and p a pixel in the image. For every region R, we have that

$$\forall p \in I : I(p) = 1 \text{ if } p \in R;\ 0 \text{ otherwise.} \tag{1}$$

Let (i, j) be the index of a pixel in the reduced-size image I_d and let $d_1 = n/n_d$ and $d_2 = m/m_d$ be the downsampling factors along the rows and columns of I, respectively. I and I_d are related by

$$I_d(i,j) = \frac{1}{d_1 d_2} \sum_{k=1}^{d_1} \sum_{l=1}^{d_2} I\left[d_1(i-1)+k, d_2(j-1)+l\right]. \tag{2}$$

The pixel (i, j) of I_d is assigned the value of the proportion of white pixels contained within the corresponding $d_1 \times d_2$ sized block in the original binary image.

These downsampled images are the input to the PCA algorithm. In effect, each one is a N dimension feature vector, where $N = n_d m_d$. If we have M regions describing a given class, then X is the PCA input data matrix (of size $N \times M$) and we can write[22]

$$\overline{X} = \mu(X) \text{: the mean of } X, \tag{3}$$

$$Y = X - \overline{X}, \tag{4}$$

$$C = YY^T, \tag{5}$$

FIGURE 1 — Left: an image from our database with a blue-sky region; middle: a binary representation of the sky region's coverage in the original image size; right: the downsampling of the binary image to 6 × 8 pixels, which is used to perform PCA. The gray-scale values represent the relative coverage of the region at a given location: from 0% (black) to 100% (white).

where C can then be expressed, using singular value decomposition, as $C = V \Lambda V^T$, where V is the eigenvector matrix and Λ is the diagonal eigenvalue matrix of C.

The two important elements are the eigenvector and eigenvalues matrices: V and Λ. V defines the new basis vectors, *i.e.,* their orientation, while Λ expresses the relative importance of each basis vector in reconstructing the data. The key insight is that if we are provided with a reference basis, we can calculate its similarity (*i.e.,* distance) to any other basis in a space of identical dimension. In our framework, the reference basis is the eigenregions obtained by human segmentation, while the candidate bases are the eigenregions obtained via automatic segmentation algorithms.

We first note that if two vectors have many common components (that is, two regions' coverage is almost identical), the angle they form is going to be small, *i.e.,* the points they define will be close in space. This property is important since it guarantees that regions that are roughly similar will be located close to one another. Conversely, over-and under-segmented regions will be located much further apart since the number of components they have in common with an exactly segmented region is going to be small.

Let V_i^1 be the i-th eigenvector of a reference segmentation and V_i^2 be the i-th eigenvector of a candidate segmentation. Furthermore, let λ_i^1 and λ_i^2 be the eigenvalues associated with V_i^1 and V_i^2, respectively. We can express the angle between the two vectors as

$$\theta\left(V_i^1, V_i^2\right) = \cos^{-1}\left(\left\langle V_i^1, V_i^2 \right\rangle\right), \tag{6}$$

that is, the inverse cosine of the vectors' inner product. Since we are working with PCA, orientation matters but direction does not, therefore we can further write

$$\theta\left(V_i^1, V_i^2\right) = \min\left[\theta\left(V_i^1, V_i^2\right), 180 - \theta\left(V_i^1, V_i^2\right)\right], \tag{7}$$

where θ is expressed in degrees.

The distance between a reference segmentation method V^1 and a candidate one V^2 can then be defined as the weighted sum of each angle, *i.e.,*

$$\Delta\left(V^1, V^2\right) = \sum_i \lambda_i^1 \left(V_i^1, V_i^2\right). \tag{8}$$

The eigenvalues from the reference method are used as weights because they express the importance of a given orientation in the human segmentation, and thus the importance of committing an error there. This weighting will have the effect of "denoising" the results, only preserving errors that are relevant to the reference segmentation.

Given a reference basis, the proposed distance measure, Eq. (8), is effectively class, algorithm, and parameter independent since it only measures the dissimilarity of two bases in $N - D$ space. It can thus be used to compare the accuracy of different segmentation algorithms, and it can also indicate the relative "difficulty" of segmenting a class compared to others, as shown in the next section. Note that the C matrix of Eq. (5) is a rotation matrix, thus it is a unitary matrix. It follows that for all classes and segmentation methods we have

$$\sum_i \lambda_i = N. \tag{9}$$

Thus, all the distance measures presented in this paper are directly comparable to each other, as N is constant for the entire framework.

In Ref. 11, it was proposed that eigenregions were independent of the segmentation algorithm, and so were the underlying features. While we do not contest this, we point out that this argument was made in light of *general* regions, *i.e.,* all regions were considered equal and were used. We argue, however, that most image classes have a much lower underlying dimensionality than general regions. As a result, their appearance in PCA space will vary significantly and, consequentially, so will the outcomes of different segmentation methods.

4 Experimental setup

The experimental protocol proceeds as follows: first, test images are selected from a database; these images are segmented by hand according to the chosen classes. The images are then segmented using several automatic algorithms and their output is assessed using a simple matching algorithm. Finally, once the data is collected, eigenregions are selected and distances measured.

The database we used consists of 55,000 real-world images. They come in various original formats and quality, and depict a very wide range of scenes. Out of these 55,000 images, 9000 have been manually annotated by photographic experts as containing either one of the memory colors: blue sky, green vegetation, and skin tones. We randomly selected 900 images (300 per class) out of these 9000 for our experiment. Since segmentation is a computationally expensive task, we resized the input images to 64×48 pixels for practical reasons. This downsampling does not, however, alter the location of regions within an image. Examples of images in this database are shown in Fig. 2. These images were segmented by hand. For every image, *only* the relevant class is segmented, which leads to a binary segmentation of the image (see Fig. 3 for an illustration).

The 900 images are also segmented using four different algorithms: k-means (with $k = 8$), edgeflow (with $\sigma = 8$), FH (with $k = 50$), and meanshift (with spatial = 6 and range = 15). For the first two algorithms, the parameters were chosen to match the ones from Ref. 11, while the latter two were chosen so that the number of regions per image was comparable with the first two. Despite parameters being chosen in an informed manner, further optimization for each method was not carried out and would likely depend on the memory color class.

To assess the segmentation results, we look at every region of the segmented image. If a region has a non-null

FIGURE 2 — Example of images present in the database: blue-sky labelled (first row), vegetation labelled (second row) and skin-tones labelled (third row). Images differ greatly with respect to subject, object scale, and capture conditions.

intersection with the human segmentation, *i.e.*, if a segmented region *contains* a given class, this regions is deemed a positive match. A binary map is thus created where the region will appear in white and the rest of the image in black (akin to the ones shown in Figs. 1 and 3). After all the binary segmentations are obtained, they are reduced to a 6×8 image, according to Eqs. (1) and (2). From these output images, 15 sets of eigenregions are calculated: one for each algorithm-class pair (four algorithms + human segmentation).

5 Experimental results

The results are reported in two categories. First, we assess the "localisability" of the memory color classes; that is, whether there is some constancy across images regarding position, shape, and/or size of regions that belong to a specific class. Indeed, if a class is not localized at all within images, a PCA-based framework will be of little help. In a second step, we show the class-specific eigenregions for the three considered classes obtained by manual segmentation and the four segmentation algorithms, and discuss these results in terms of image content and segmentation behavior.

FIGURE 3 — Human binary segmentations examples of the three considered classes. The original images containing sky, vegetation, and skin (top row) and their segmentation (bottom row).

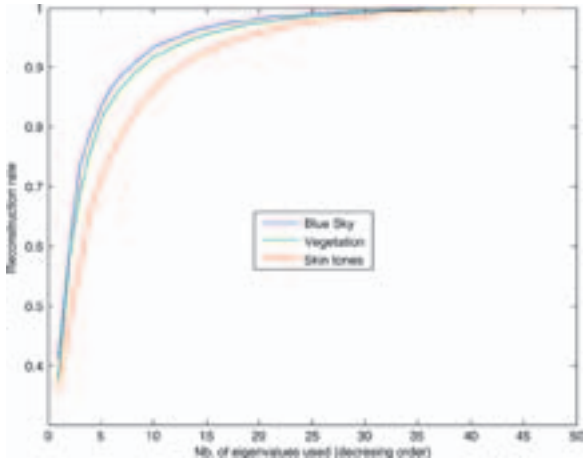

FIGURE 4 — Reconstruction rates for human segmentation. Blue sky and green vegetation are fairly well localized, with 85% of the variance explained by 10% of the eigenregions, while the skin tones reconstruction rate is lower.

5.1 Localized classes

Figure 4 shows how localized our three considered classes are. For blue sky and green vegetation, five eigenregions (*i.e.*, 10% of the available eigenvalues) suffice to explain 85% of the variance. Considering the prevalence of these two classes in landscape images, these results are unsurprising. Conversely, skin tones are not as localised. Since skin tones encompass all of face, hands, arms, body, *etc.*, they are expected to be inherently less localized than sky or grass.

Reconstruction rates, given by the normalized cumulative sum of the eigenvalues, are important because they indicate whether it is judicious to use geometrical features for the detection of a given class. On the other hand, they do not provide a measure of accuracy. A segmentation algorithm that would deterministically partition images into two regions (say top and bottom) would have a very high reconstruction rate. It would, however, be a very inaccurate segmenter.

5.2 Class-specific eigenregions

After eigenvalues, we analyze the eigenregions given by the algorithms on our three classes. The first five eigenregions for each class and each algorithm are shown in Figs. 5–7, where their values have been normalized between 1 (white) and –1 (black) for better visualization. These eigenregions provide important clues regarding the performance of a given segmentation algorithm over a class. First, they allow a visual comparison of class-localization and differences across algorithms. Then, as pointed out in Ref. 11, they can be used as features in image classification; the rationale is

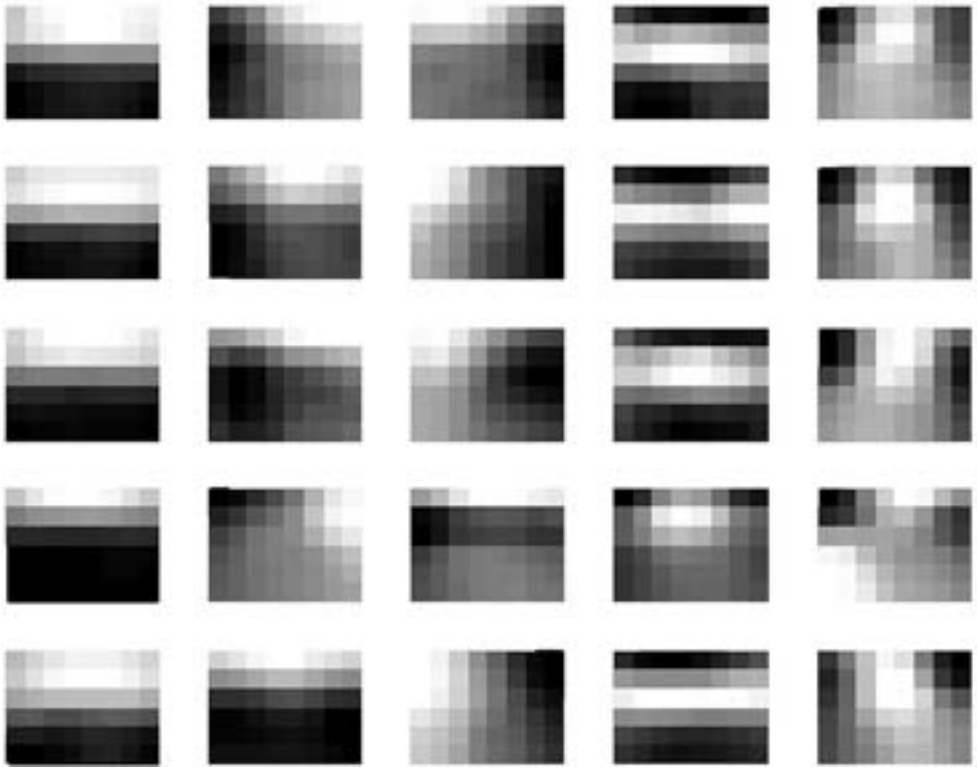

FIGURE 5 — Column-wise: the first five eigenregions for the blue-sky class. Row-wise, from top to bottom: Human segmentation, *k*-means, edgeflow, FH, and meanshift.

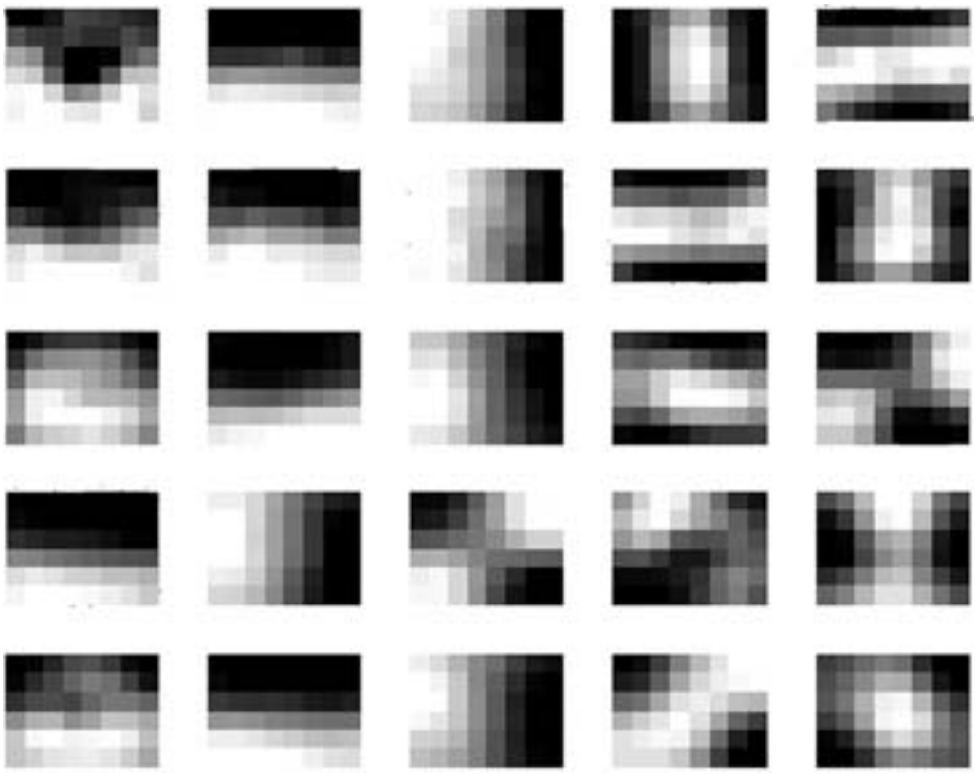

FIGURE 6 — Column-wise: the first five eigenregions for the green-vegetation class. Row-wise, from top to bottom: Human segmentation, *k*-means, edgeflow, FH, and meanshift.

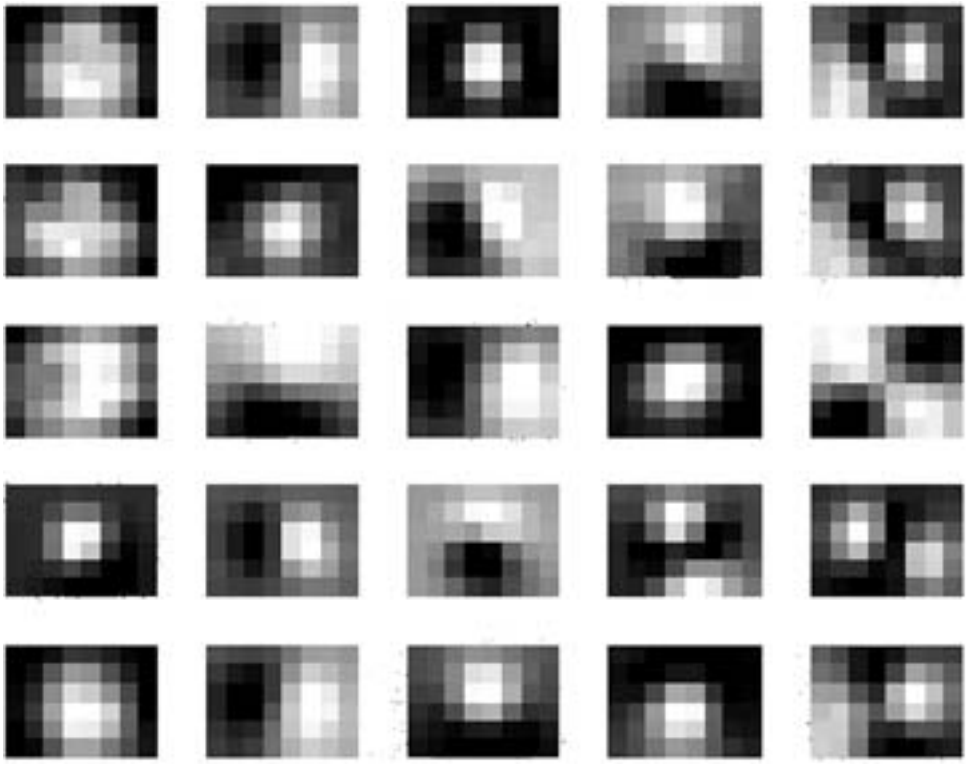

FIGURE 7 — Column-wise: the first five eigenregions for the skin tones class. Row-wise, from top to bottom: Human segmentation, *k*-means, edgeflow, FH, and meanshift.

that our particular eigenregions should actually prove more useful than the general eigenregions since ours are readily tailored to a specific class. Classification using eigenregions is explored in the last section of the paper.

From the results, we observe that sky and vegetation eigenregions appear more coherent than the skin ones. This correlates well with the reconstruction curves shown in Fig. 4 and is easily explained by the fact that sky and vegetation are mostly found in landscape-type images that have a top/down decomposition (or left/right for pictures taken in a portrait orientation). These regions are therefore located in a smaller part of the 48-D space and thus are easier to cluster via PCA.

The blue sky results, Fig. 5, show that while all algorithms correctly find the first eigenregions, *k*-means and edgeflow results appear, in general, much closer to the human segmentation than either FH or meanshift when looking at eigenregions 2–5. In general, however, the eigenregions correspond to our expectations, the first one being a clear clear top/down decomposition, with some variation in the subsequent ones that are likely to originate from images where the sky is partly occluded (trees, buildings, people).

Vegetation eigenregions, Fig. 6, start similarly with a landscape-type decomposition (top/down or left/right, depending on the camera's orientation), but this behavior changes after the first three to indicate the presence of centred objects, *e.g.,* trees or plants in indoor scenes. These latter positions are harder to accurately segment and few algorithms are able to correctly distinguish them. Both meanshift and edgeflow appear to be closer to the ground truth, but their results are still somewhat skewed. *k*-Means performs well on the landscape-type images but is confounded by more complex scenes, while FH misses out one of the first eigenregion.

Finally, skin tones eigenregions, Fig. 7, exhibit various type of centre-surround interactions, *i.e.,* the object of interest is small and located centrally within an image. Since we look for skin tones in general, as opposed to faces only, we expect the results to be somewhat noisy because of the greater location possibilities. The eigenregions express two aspects well: the position and the scale ambiguity. Indeed, while most of them are of center-surround types, the size and the location of the "interest region" vary across eigenregions. Looking at the four algorithms, we see that meanshift is probably the closest to human segmentation while *k*-means is not too far behind. Edgeflow and FH appear to perform worse but for different reasons. In fact, their behavior is complementary with FH not detecting the larger regions (eigenregion 1) and edgeflow wrongly detecting the smaller ones.

Looking at these results, we can draw the following conclusions: class-specific eigenregions have very distinct shapes that express the content of the images well; they are algorithm-dependent, and the closest algorithm to human segmentation does not appear to always be the same. Finally, the shape of the eigenregions correlates well with the reconstruction rates observed, *i.e.,* the simpler the shape of the eigenregion, the better localized the underlying class is.

5.3 Comparison with general eigenregions

The original eigenregions proposed in Ref. 11 were calculated over *general* regions, *i.e.,* all types of regions regardless of their class. A key feature of these general eigenregions, shown in Fig. 8, is that their shape was independent of the segmentation algorithm, a property not verified with the class-specific ones (see Figs. 5–7).

Comparing the first row of Fig. 8 with the first rows of Figs. 5–7, note that general eigenregions are quite different from the class-specific ones calculated over manual segmentation, *i.e.,* the ground truth, suggesting that accurate segmentation is key to achieve relevant features and that class-specific eigenregions are better suited geometric descriptors.

6 Segmentation evaluation and database self-sufficiency

Our proposed distance measure, Eq. (8), allows to numerically compare the performance of automatic segmentation to a manual one, determining whether the algorithms' segmentation accuracy is class dependent. This measure can furthermore be employed to determine how many images are needed to be representative of a given class, *i.e.,* the number of images needed to have a similar region structure than in the entire database. We note that, while all eigenvalues have been used to obtain the results reported in this section, in practice eigenvalues beyond the tenth have little impact on the final result.

6.1 Which algorithm for which class?

The eigenregions themselves give useful information, still, assessing distance in a 48-D space, even when provided with visual cues, is difficult. Using our proposed measure [Eq. (8)], we evaluate the distance between the four algorithms and human segmentation for each class. The results, reported in Table 1, confirm what was visually inferred in the previous section. Looking at the distances as a whole, blue sky is the best segmented region (smallest distance), followed by vegetation and skin tones. This is expected given the much greater variety of position, size, shape, and color of skin tones when compared to blue sky or vegetation, thus making them harder to segment. Also, we see that while *on average* meanshift performs better than the other algorithms, it is not necessarily the best performing one for *every* class.

TABLE 1 — Distance between a given algorithm and human segmen- tation (smaller is better). The results are highly class dependent and there is not a best algorithm overall.

	Blue sky	Green vegetation	Skin tones
k-Means	24.3	34.65	44.15
Edgeflow	**23.6**	28.53	56
FH	40.62	73.2	70.8
Meanshift	36.12	**27.5**	**33.6**

Analyzing the results separately, one observes that *k*-means and edgeflow are equivalent in their sky segmentation, meanshift, and edgeflow are better for vegetation, and meanshift is best for skin tones; these results correlate well with the visual assessment done in the previous section. For all classes, FH is rated as the worst performing algorithm. This comparison brings several questions that have to be answered: why does *k*-means keep up, why is meanshift worse in the simplest class, and why does FH perform so badly?

k-Means' performance can be explained by the choice of classes. Indeed, memory colors are classes that are well located in color space,[2] so a cluster including them will usually be found. As a result, *k*-means can be expected to be accurate. Its performance for vegetation and skin is, however, lower since these classes' luminance and color can be altered by lighting effects (such as shading), thus creating errors.

Edgeflow includes both color and texture information and is therefore expected to yield a good segmentation of our three classes. However, its accuracy for skin tones is not always high. Looking at both the distance and the eigenregions themselves, one observes that its regions are larger than they should be. This is, most likely, the consequence of the

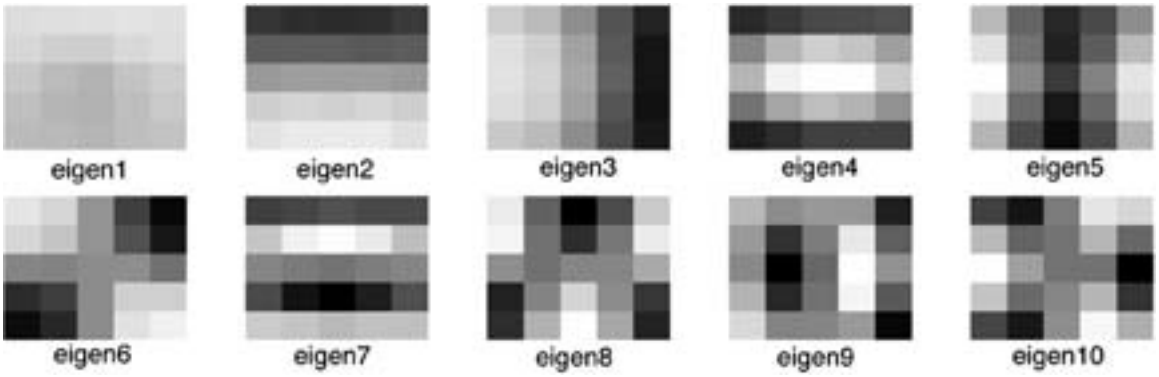

FIGURE 8 — The first ten general eigenregions.

choice of σ that influences the scale at which variations are sought. Additionally, artefacts such as glasses, hats, occlusions or sometimes hair can induce a wrong segmentation.

Perhaps surprisingly, FH is the worst rated algorithm in our test, and this for all classes. The reason here is that the choice of parameters has given rise to chronic over-segmentation. While the number of regions in the image is not overly high (between14 and 28 regions per image), it was very sensitive to noise, vignetting, and small level texture alterations. This is confirmed by looking at the number of regions found for each class. For sky, vegetation, and skin, FH has, on average, 3.2, 5.2, and 2.6 regions per image, respectively, compared to meanshift's 1.2, 2.3, and 1.5, indicating a strong over-segmentation issue. We have found no significant correlation between number of regions and performance for other algorithms than FH, whose over-segmentation was significant.

Finally, meanshift, the best overall algorithm, exhibits a rather unique behavior: its worst class is sky, which is the opposite of every other algorithm. Again, this can be explained by the parameters used. While they were well-suited to vegetation and skin, they tend to under-segment sky, especially in the presence of softer gradients, such as clouds or haze.

Note that we do not advocate here that one algorithm is better than the others. Rather, the results show that a given algorithm (or a given choice of parameters) appears to be measurably better for segmenting a specific class, not all classes in general. It is therefore well possible that meanshift, with other parameters, would have a more accurate sky segmentation. However, this could be detrimental to its segmenting of skin or vegetation. A direct consequence of our results is that the proposed distance measure can not only be used to select one algorithm, but could also be employed to optimise a given algorithm's parameters in order to segment a specific class.

6.2 How many images to form a representative set?

Class-specific eigenregions were calculated for over 300 images (per class). We propose that the suitability of this set's size can be assessed using the proposed distance measure. The insight is that if regions are reasonably well localized in images, then selecting 200 images instead of 300 should not significantly alter the eigenregions and, consequently, the induced distance between the two bases will be small.

For each class, we have selected at random 10, 20, 50, 100, and 200 images out of our set of 300. We repeat each test 200 times and calculate the average distance (over the 200 sets) to the basis obtained using the 300 images (the human segmentation eigenregions reported earlier). If the regions are well-behaved, then we should observe that the error decreases fast as the number of images used increases. The results, Figs. 9–11, show that the error curves are

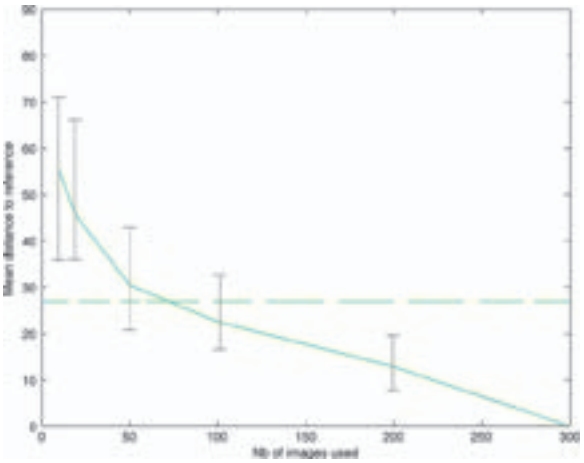

FIGURE 9 — Influence of the number of images on the distance for the vegetation class. When the error becomes low, the implication is that the subset approximates the complete set of images sufficiently well. The vertical lines at each point indicate the range of performance across subsets, which shows that a carefully chosen small subset can be as representative of the entire set as a larger, randomly, selected one. The horizontal line represents the distance of meanshift to human segmentation for the 300 images, to allow visual comparison.

monotonically decreasing. To put the behavior of the distance measures in perspective, we also plot, on the same graphs, the distance between the basis of the meanshift eigenregions and the 300 human segmented ones. The considered classes in this experiment are the three memory colors plus, for comparison purposes, the "normal" class, *i.e.*, all the regions that belong to neither of the memory colors.

We see that the difference between using 200 or 300 images is small compared to the error incurred by automatic segmentation, but using anything less than 200 will necessarily introduce approximations that are not negligible. Knowing, *a priori*, how many images are needed to form a representative subset is valuable in terms of time and resources saved.

In addition to being monotonically decreasing, the curves obtained using our angular error measure are also well correlated with the results obtained previously; that is,

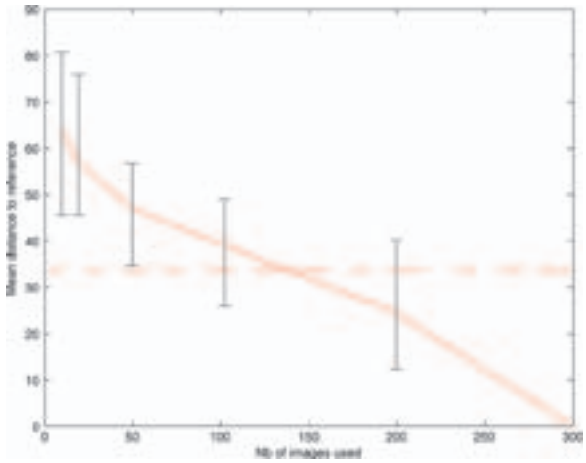

FIGURE 10 — Influence of the number of images on the distance for the skin tone class.

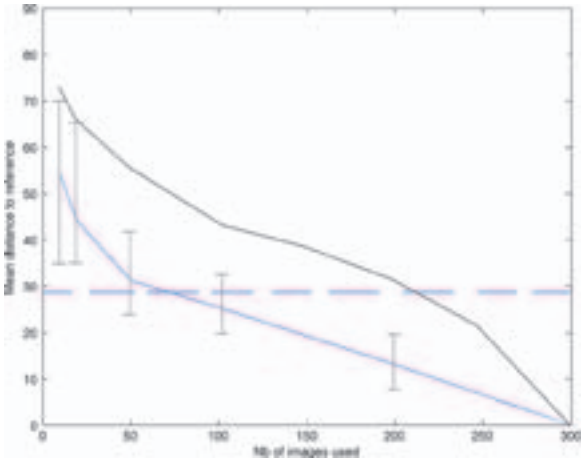

FIGURE 11 — Influence of the number of images on the distance for the sky class (blue with error bars). Note that the behavior of the sky class is very similar to the vegetation one. Also represented on this graph, the performance of meanshift for the sky class (blue, dashed line) and the relationship between number of images and distance for the "normal" class (black, without error bars). One observes that since the normal class is much less localized than the others, its distance is much greater, regardless of the number of images used.

for a given number of images, the relationship between sky, vegetation, and skin and general eigenregions does not vary from the earlier reported reconstruction rates, *e.g.,* regardless of the number of images used, sky is always more localised than vegetation and skin. Furthermore, the slopes of the curves exhibit the same behavior. Given that these reconstruction rates model the variability of these classes well (they corresponds to what one sees), these experiments effectively validate the angular error measure.

Of course, the number of images needed to form a good subset will depend on the inherent localisation of a given class as well as on the variety of images present in a given database. Additionally, we point out that our distance measure can be used so select a specific subset that matches the data structure better, since the results show that there is a large variance across subsets.

7 Classifying memory colors

Eigenregions are geometrical region features, as such they can be used for classification tasks. In Ref. 11, experiments including general eigenregions showed their usefulness in memory color region classification. Here, we assess the improvement that can be obtained by utilizing class-specific eigenregions instead.

This experiment uses the same settings as the ones of Ref. 11. In addition to the first 10 eigenregions, 12 color (mean and standard deviation of R, G, B, L^*, a^*, b^*) and the seven Haralick features[23] are employed.

The classification algorithm is a multivariate Gaussian based on the maximum a posteriori rule, which, being a supervised classification algorithm, requires a ground truth. The ground truth and prior probabilities are calculated over 9874 manually annotated regions, randomly selected from a database of 77,000 regions. The chosen training/testing scheme is a 90/10 recursive decomposition: 90% of the regions are randomly selected to train the classifier and the remaining 10% are classified. This procedure is repeated until all the regions are classified, and the rates are then averaged.

TABLE 2 — Classification rates (average of correct positive and correct negative rates) for the memory-color regions. The rates increase with the use of class-specific eigenregions. Improvements are also noticeable when the best performing segmentation algorithms (according to our measure) are employed to obtain the regions.

Classification rates	Vegetation	Skin	Sky
W/out eigenregions	0.87	0.82	0.85
General eigenregions	0.87	0.85	0.89
Specific eigenregions	0.90	0.89	0.95
Best performing algo.	0.92	0.93	0.95
Manual segmentation	0.94	0.94	0.96

Note that we are not advocating this scheme as the best possible classification algorithm for this task; our aim is to evaluate the performance of class-specific eigenregions and the numerical results should be observed for their relative performance to each other rather than in absolute terms.

Table 2 compares the classification rates without, with the general, and with the class-specific eigenregions calculated over the *k*-mean segmentation (the segmenter used in Ref. 11). Additionally, we provide classification results for regions segmented with each class' best performing algorithm (according to our distance measure) as well as with manual segmentation.

The results show a number of interesting points. First, class-specific eigenregions are better suited to classifications than the general ones. Indeed, an observation was made in Ref. 11 that eigenregions did not help vegetation classification. Class-specific ones, however, also increase vegetation classification successfully.

Importantly, the quality of segmentation also plays a role in classification results, the increase in classification rates is coherent with the segmentation evaluation results given by our distance measure. The combination of class-specific features withadequate segmentation algorithms does provide the best classification.

8 Usage and performance of the framework

This paper has presented the various steps undertaken to obtain class-specific eigenregions, and utilize them in segmentation assessment and region classification. We briefly discuss here the complexity of these steps in terms of user involvement and computation time.

The main required steps to achieve the results presented herein are image segmentation, ground truth acqui-

FIGURE 12 — The graphical user interface used in the ground truth acquisition.

sition, feature extraction, classifier training, and region classification. Among these, feature extraction and region classification can be performed in real time (all numerical times are obtained with an Apple MacBook Pro 1.8-GHz dual core and 4 Gb of RAM), as the calculations are performed on low-resolution images.

Image segmentation can be time consuming depending on the considered segmentation algorithm. Indeed, there is a significant performance difference between *k*-means (<1 sec) and meanshift (a few seconds). Other algorithms, such as normalized cuts[24] take even more time (almost 1 minute) and were thus excluded from our comparison. A trade-off between segmentation accuracy and speed can therefore exist and has to be factored in choosing an algorithm.

The last two steps, ground truth acquisition and classifier training are the most intensive but both can be done offline; training the classifier typically takes a couple of days.

Ground truth, meaning both manual segmentation and region labelling is obtained through a graphical user interface. In a first step, the image is strongly over-segmented (using *k*-means) and a user is asked to click on all regions that correspond to a given class, thus enabling region merging in order to obtain a binary map. Manual and automatic segmentations are then shown to the user (on a per-region basis), where a selection between sky, vegetation, skin, and "normal" is made. An example of the GUI is provided in Fig. 12.

9 Conclusions/future work

We have presented an eigenregion-based framework that evaluates class-specific image information. Using human segmentation and assessment of automatic segmentation algorithms, we were able to show, numerically, that naturally occurring classes in images were neither evenly distributed nor similarly localized. Class-specific eigenregions were shown to outperform general ones in a standard classification framework for all the considered memory-color classes: sky, vegetation, and skin tones, all of them being of critical importance for color rendering or correction tasks.

Moreover, we have proposed a distance measure in $N - D$ space that takes into account the relative weight of a given eigenregion. Using that distance, we showed that different algorithms segment different image classes with varying accuracy compared to human segmentation. Importantly, the algorithms' performance is strongly class dependent, there is no single best algorithm. Finally, if time is a critical factor, the segmentation algorithm cannot be chosen on intrinsic performance alone. The proposed distance measure can, however, be used to optimise the algorithm's settings.

10 Reproducible research

At LCAV, we aim to make our research reproducible by everyone. The matlab code used to obtain the eigenregions and distances measures reported in this paper is therefore available online at http://rr.epfl.ch

References

1 C. Bishop, *Pattern Recognition and Machine Learning* (Springer, 2007).

2 P. Bodrogi and T. Tarczali, "Color memory for various sky skin and plant colors: Effect of the image context," *COLOR Research and Application* **25**(4), 278–289 (2000).

3 C. Carson *et al.*, "Blobworld: Image segmentation using expectation-maximization and its application to image querying," *IEEE PAMI* **24**(8), 1026–1038 (August 2002).

4 J. Chen *et al.*, "Adaptive image segmentation based on color and texture," *IEEE International Conference on Image Processing (ICIP2002)*, 122–126 (2002).

5 D. Comaniciu and P. Meer, "A robust approach toward feature space analysis," *IEEE Trans. PAMI* **24**(5), 603–619 (2002).

6 F. Estrada and A. Jepson, "Quantitative evaluation of a novel image segmentation algorithm," *CVPR*, 1132–1139 (2005).

7 M. R. Everingham *et al.*, "Evaluating image segmentation algorithms using the pareto front," *ECCV*, 34–48 (2002).

8 P. Felzenszwalb and D. Huttenlocher, "Efficient graph-based image segmentation," *Intl. J. Computer Vision* **59**(2), 167–181 (2004).

9 C. Fowlkes *et al.*, "Learning affinty functions for image segmentation," *CVPR* **2**, 54–61 (2003).

10 C. Fredembach *et al.*, "Region-based image classification for automatic color correction," *Proc. 11th IS&T/SID Color Imaging Conference*, 59–65 (2003).

11 C. Fredembach *et al.*, "Eigenregions for image classification," *IEEE Trans. PAMI* **26**(12), 1645–1649 (2004).

12 F. Ge *et al.*, "Image segmentation evaluation from the perspective of salient object extraction," *CVPR* **1**, 1146–1153 (2006).

13 A. Jain, *Fundamentals of Digital Image Processing* (Prentice-Hall International, 1989).

14 W. Ma and B. Manjunath, "Edgeflow: A technique for boundary detection and image segmentation," *IEEE Trans. IP* **9**(10), 1375–1388 (2000).

15 D. Martin, "An empirical approach to grouping and segmentation," Ph.D. Thesis, U.C. Berkeley (2002).

16 D. Martin and C. Fowlkes, The Berkeley Segmentation Database and Benchmark, 2001, http://www.cs.berkeley.edu/projects/vision/grouping/seg.

17 D. Martin *et al.*, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," *ICCV*, 416–425 (2001).

18 M. Meila, "Comparing clusterings by the variation of information," *Proc. Intl. Conf. on Learning Theory*, 173–187 (2003).

19 R. Unnikrishnan *et al.*, "Toward objective evaluation of image segmentation algorithms," *IEEE Trans. PAMI* **29**(6), 929–944 (2007).

20 J. Perez-Carpinell *et al.*, "Familiar objects and memory color," *COLOR Research and Application* **23**(6), 416–427 (1998).

21 J. Da Rugna and H. Konik, "Color coarse segmentation and regions selection for similar image retrieval," *CGIV 2002: IS&T First European Conference on Color in Graphics, Image and Vision,* 241–244 (2002).
22 A. Webb, *Statistical Pattern Recognition* (Arnold, 1999).
23 R. M. Haralick *et al.*, "Textural features for image classification," *IEEE Trans. on Systems, Man and Cybernetics,* 610–621 (1973).
24 J Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* 888–905 (2000).