16

Bridging the Semantic Gap using Human Vision System Inspired Features

Gaëtan Martens, Peter Lambert and Rik Van de Walle
Multimedia Lab - Ghent University - IBBT
Belgium

1. Introduction

In the last decade, digital imaging has experienced a worldwide revolution of growth in both the number of users and the range of applications. The amount of digital image content produced on a daily basis is still increasing drastically. As from the very beginning of photography, those who took pictures tried to capture as much information as possible about the photograph and in today's digital age, the need for appending metadata is even bigger. However, it is obvious that manually annotating images is a cumbersome, time consuming and expensive task for large image databases, and it is often subjective, context-sensitive and incomplete. Furthermore, it is difficult for the traditional text-based methods to support a variety of task-dependent queries solely relying on textual metadata since visual information is a more capable medium of conveying ideas and is more closely related to human perception of the real world. The dynamic image characteristics require sophisticated methodologies for data visualization, indexing and similarity management and, as a result, have attracted significant research efforts in providing tools for content-based retrieval of visual data. Content-based image retrieval uses the visual contents of an image such as color, shape, texture, and spatial layout to represent and index the image.
Early content-based image retrieval systems were based on the search for the best match to a user-provided query image or sketch (Flickner et al., 1995; Mehrotra et al., 1997; Laaksonen et al., 2002). Such systems decompose each image into a number of low-level visual features (e.g., color histograms, edge information) and the retrieval process is formulated as the search for the best match to the feature vector(s) extracted from a query image. However, it was quickly realized that the design of a fully functional retrieval system would require support for semantic queries (Picard, 1995). The basic idea is to automatically associate semantic keywords with each image by building models of visual appearance of the semantic concepts of interest. However, the critical point in the advancement of content-based image retrieval is the semantic gap. The semantic gap is the major discrepancy in computer vision: the user wants to retrieve images on a semantic level, but the image characterizations can only provide a low-level similarity. As a result, describing high-level semantic concepts with low-level visual features is a challenging task. The first efforts targeted the extraction of specific semantics under the framework of binary classification, such as indoor versus outdoor (Szummer & Picard, 1998), and city versus landscape

classification (Vailaya et al., 1998). More recently, efforts have emerged to solve the problem in greater generality through the design of techniques capable of learning semantic vocabularies from annotated training image collections by applying (both unsupervised and semi-supervised) machine learning techniques, e.g. (Duygulu et al., 2002; Feng et al., 2004).
In computer vision one of the traditional goals is the automatic segmentation and interpretation of general digital images of arbitrary scenes. In the literature, certain methods have been proposed to extract the semantics of scenery images using low-level features. Most of these approaches use image partitioning as intermediate step. Wang et al. use a codebook to segment an image based on the statistics of the regions' color and texture features (Wang et al., 2002). At pixel level, color-texture classification is used to form the codebook. This codebook is used in the next stage to segment an image into regions. The context and content of these regions are defined at image level. Zhu et al. partition the image into equally sized blocks and indexes the regions using a codebook whose entries are obtained from the features extracted from a block (Zhu et al., 2000). In (Li & Wang, 2003) a method is described to use 2-dimensional hidden Markov models to associate the image and a textual description. However, most approaches introduced above can not integrate the semantic descriptions into the regions, and therefore cannot support the high-level querying of images. Depalov et al. use a color-texture segmentation algorithm to segment images depicting natural scenes (Depalov et al., 2006). The features of the obtained regions are used as medium level descriptors to extract semantic labels at region level and later at scene level. However, the use of quantized features may result in weaker segmentations. Turtinen and Pietikäinen applied a Self-Organizing Map trained with local binary patterns to classify outdoor scene images (Turtinen & Pietikäinen, 2003). As a means of supervision, the user selects the map nodes with similar appearance and the corresponding samples are retrained using a smaller map in order to reveal if some classes are mixed up in the same node.
Despite all efforts, humans still outperform the best machine vision systems in many aspects. Humans are very good at getting the conceptual category and layout of a scene within a single fixation. So, building a system that emulates the recognition tasks of the cortex is a challenging and attractive idea. However, in computer vision the use of visual neuroscience has often been limited to a tuning of Gabor filter banks (Jain & Farrokhnia, 1991; Clausi & Jernigan, 2000; Zhang et al., 2000). No real attention has been given to biological features of higher complexity so far. Given the fact that the human vision system is best trained to color and texture perception, these low-level features could play an important role in image understanding. Indeed, Renninger and Malik already have concluded that a texture analysis provides useful information for rapid scene identification (Renninger & Malik, 2004). Therefore, the application of the appropriate features is of utter importance.
This chapter deals with the combination of biologically inspired features and Self-Organizing Maps (Kohonen, 2001) for the classification and recognition of real-world textures and the segmentation of textured images. Analogously to the processing principles of the visual cortex, the unsupervised learning capabilities and visualization techniques of a Self-Organizing Map are utilized with the highly efficient color and texture features. The Self-Organizing Maps are particularly well-suited for the combined task of mapping the high-dimensional and non-linear data distribution to a low dimensional plane while conserving the local neighborhood relations for fast and easy-to-use visualization.

The remaining part of this chapter is organized as follows. Section 2 describes the computational model to calculate the biologically inspired texture features and in Section 3 we briefly introduce the basic model of color perception. Section 4 outlines the calculation of the features and explains the data preprocessing with regard to classification. Unsupervised image partitioning experiments on gray-scale and real-life color textures are presented in Section 5. Section 6 describes the automatic interpretation of the obtained image regions and discusses the possible improvements. Final conclusions and future work appear in Section 7.

2. Texture model

Our system is inspired by the standard model of the human visual system (HVS) (Riesenhuber & Poggio, 2003). The standard model summarizes what most visual neuroscientists generally agree on:

- the first few hundred milliseconds of visual processing in primate cortex follows a mostly feed-forward hierarchy for immediate recognition tasks;
- hierarchical build-up of invariances first to position and scale and to the viewpoint, and more complex transformations requiring the interpolation between several different object views;
- in parallel, an increasing size of receptive fields;
- plasticity and learning probably at all stages and certainly at the level of the cortex;
- learning specific to an individual object is not required for scale and position invariance.

In its simplest form, the view based module of the standard model consists of 4 layers of computational units. The first layer of simple cells S1 in the primary visual cortex (also called the striate cortex or V1) represents linear oriented filters followed by an input normalization. Each unit in the next layer (C1) pools the outputs of simple cells of the same orientation but at slightly different positions by using a maximum operation. Each of these units is still orientation selective but more invariant to the scale, similarly to some complex cells. In the next stage signals from complex cells with different orientations but similar positions are combined (in a weighted sum) into simple cells S2 to create neurons tuned to a dictionary of more complex features. The final layer of C2 units is similar to the C1 cells: by pooling together signals from S2 cells of the same type but at slightly different scales, the C2 units become more invariant to the scale but preserve feature selectivity.

2.1 Simple cells

In a first stage, responses are obtained by applying a Gabor filter bank to an input image. As proposed by (Daugman, 1985) the following family of 2-dimensional isotropic Gabor filters are used to model the receptive cells of the HVS:

$$g_{\lambda,\theta,\varphi}(x,y) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{x'^2+\gamma^2 y'^2}{2\sigma^2}\right)\cos\left(2\pi\frac{x'}{\lambda}+\varphi\right) \quad (1)$$

where $x' = x\cos\theta - y\sin\theta$ and $y' = x\sin\theta + y\cos\theta$.

The orientation of the filter is represented by θ and σ denotes the standard deviation of the Gaussian which determines the size of the receptive field of the HVS. The phase offset φ

affects the symmetry of the function. The parameter λ determines the spatial wavelength of the receptive field function (1). Since σ and λ are not independent, the standard deviation is selected to satisfy $\frac{\sigma}{\lambda} = 0.56$ in order to obtain a one-octave spatial frequency bandwidth (Kruizinga & Petkov, 1999). Finally, the parameter γ, also called the spatial aspect ratio, affects the receptive field ellipse. It has been found that γ ranges between 0.23 and 0.92 (Jones & Palmer, 1987) and is set to 0.5 according to (Kruizinga & Petkov, 1998).
The response of the receptive field function to an input image its luminance channel $I(x, y)$ is defined by:

$$r_{\lambda,\theta,\varphi}(x, y) = \iint I(s, t) g_{\lambda,\theta,\varphi}(x - s, y - t) ds dt \tag{2}$$

The response $s_{\lambda,\theta,\varphi}(x, y)$ of a simple cell of the visual cortex, modeled by a receptive field function $g_{\lambda,\theta,\varphi}(x, y)$ to $I(x, y)$, is given by:

$$s_{\lambda,\theta,\varphi}(x, y) = \begin{cases} 0 & \text{if } a_\lambda(x, y) = 0 \\ \chi\left(\frac{\frac{r_{\lambda,\theta,\varphi}(x,y)}{a_\lambda(x,y)} R}{\frac{r_{\lambda,\theta,\varphi}(x,y)}{a_\lambda(x,y)} + C} \right) & \text{otherwise} \end{cases} \tag{3}$$

where R denotes the maximum response level, the average gray value $a_\lambda(x, y) = \iint I(s, t) \exp \frac{(x-s)^2 + \gamma^2 (y-t)^2}{2\sigma^2} ds dt$, C is the semi-saturation constant, and $\chi(t) = t$ for $t \geq 0$ and $\chi(t) = 0$ for $t < 0$ (Kruizinga & Petkov, 1998).

2.2 Grating cells

The next layer corresponds to complex cells which provide some tolerance to shift and size. This tolerance is obtained by taking a maximum across neighboring scales and nearby pixels. Grating cells are orientation selective cells which respond strongly to gratings of appropriate periodicity and orientation, but in contrast to the simple cells or some other complex cells, they do not respond to a single bar (Kruizinga & Petkov, 1999). The computational model of a grating cell consists of two stages:

(i) the calculation of the activity of a grating subunit $q_{\lambda,\theta}(x, y)$ with a preferred orientation θ and frequency $1/\lambda$, see (4)

(ii) the summation of the responses for a given θ, see (6).

A grating subunit $q_{\lambda,\theta}(x, y)$ takes as input the simple cell outputs defined in (3) and will be activated if there are at least 3 parallel bars with orientation θ and frequency $1/\lambda$:

$$q_{\lambda,\theta}(x, y) = \begin{cases} 1 & \text{if } \forall n, M_{\lambda,\theta,n}(x, y) \geq \rho N_{\lambda,\theta}(x, y) \\ 0 & \text{if } \exists n, M_{\lambda,\theta,n}(x, y) < \rho N_{\lambda,\theta}(x, y) \end{cases} \tag{4}$$

where $0 < \rho < 1$ is a threshold in the proximity of 1. As suggested by (Kruizinga & Petkov, 1998), it is assigned 0.9. The quantities $M_{\lambda,\theta,n}$ and $N_{\lambda,\theta}$ are computed as follows:

$$\begin{cases} M_{\lambda,\theta,n}(x, y) = \max\{s_{\lambda,\theta,\varphi_n}(u, v)\} \\ N_{\lambda,\theta}(x, y) = \max\{M_{\lambda,\theta,n}(x, y) | n = -3, -2, -1, 0, 1, 2\} \end{cases} \tag{5}$$

where $\varphi_n = 0$ for $n = \{-3, -1, 1\}$ and $\varphi_n = \pi$ for $n = \{-2, 0, 2\}$. Finally, u and v satisfy the condition:

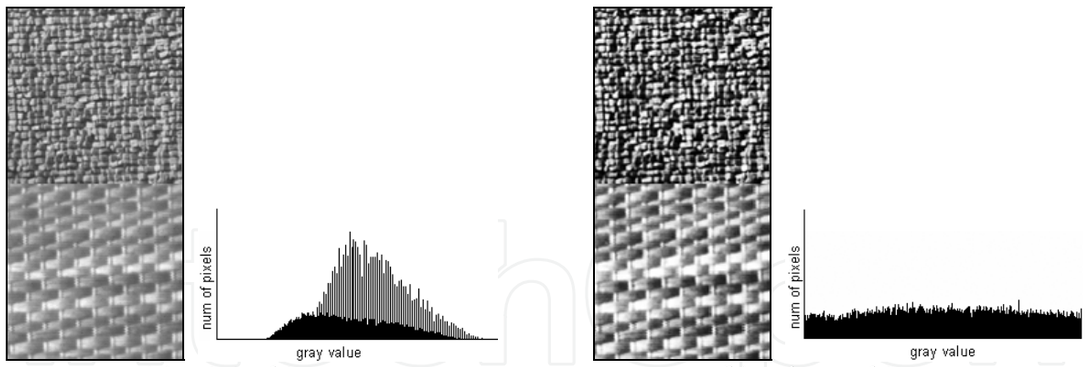
$$\begin{cases} n\frac{\lambda}{2}\cos\theta \leq u - x < (n+1)\frac{\lambda}{2}\cos\theta \\ n\frac{\lambda}{2}\sin\theta \leq v - y < (n+1)\frac{\lambda}{2}\sin\theta \end{cases} \tag{6}$$

A grating subunit will be activated $(q_{\lambda,\theta}(x, y) = 1)$ if for the preferred orientation θ and spatial frequency $1/\lambda$, the receptive field function (2) is alternately activated in intervals of length $\lambda/2$ for $n = -3, -2, .., 2$ and this along a line segment of length 3λ centered on point (x, y). In other words, the condition is fulfilled in case there are at least 3 parallel bars with spacing λ and orientation θ of the normal encountered. In the final stage, the output of the grating cell operator $w_{\lambda,\theta}$ is computed as:

$$w_{\lambda,\theta}(x, y) = \frac{1}{\sqrt{2\pi\sigma}} \iint (q_{\lambda,\theta}(s, t) + q_{\lambda,\theta+\pi}(s, t)) \exp\left(-\frac{(x-s)^2 + (y-t)^2}{2(5\sigma)^2}\right) dsdt \tag{7}$$

2.3 Enhanced grating cell operator

This operator first applies a histogram equalization on the original input image $I(x, y)$ to obtain the enhanced image $\overline{I}(x, y)$. Histogram equalization employs a monotonic, non-linear mapping which re-assigns the intensity values of pixels in the input image such that the intensity values of the output image are more uniformly distributed (*i.e.* a flat histogram) and the image has a higher contrast (see Fig. 1).



(a) original image (b) enhanced image

Fig. 1. Luminance histograms of the original and enhanced image

For each gray level j in the input image $I(x, y)$, the new value $\overline{j}$ is calculated as follows:

$$\overline{j} = l \sum_{i=0}^{j} \frac{n_i}{n} \tag{8}$$

where l is the maximum gray level, n the total number of pixels and n_j the number of occurrences of gray level j in $I(x,y)$. The result of this operation is that $\overline{I}(x,y)$ has not only a higher contrast, but also its details are enhanced such that the salient texture specific periodicities are more distinguishable. The enhanced grating cell features $\overline{w}_{\lambda,\theta}(x,y)$ are obtained by substituting the enhanced image $\overline{I}(x,y)$ in equations (2) and (3). As shown in our previous work (Martens et al., 2007), the application of the enhanced grating cell operator has a positive influence on texture classification results.

3. Color Perception

Since color is the primary visual stimulus, the choice of a color system is of great importance for the purpose of proper image retrieval. Color can be modeled and interpreted in many different ways and color systems have been developed for various purposes, such as *RGB* & *CMYK* for displaying and printing, *YIQ* & *YUV* for television and video transmission efficiency, *XYZ* for color standardization, etc.

The first geometrical model of color perception was created in the 17th century by Isaac Newton. He epitomized his experiments in light and pigment mixing by ingeniously overlapping the red and violet ends of the spectrum to create a hue circle. This circle shows the spectrum as a continuous gradation of color from red to violet, and from violet to red via the mixed colors carmine, magenta and purple. This circular representation of color is also used in the HSI space, where HSI stands for *hue, saturation,* and *intensity* (Gevers, 2001).

In the HVS color vision is mediated by specialized nerve cells in the retina, called cones. The ability to discern different wavelengths of light (*i.e.* colors) gives us more information for detecting and identifying objects than would be provided solely by black and white vision. The human retina has three types of cones which makes color detection possible: red, green and blue cones. By appropriately mixing these three primary colors it is possible to match all of the colors in the visible spectrum. The latter observation is known as the trichromatic theory (von Helmholtz, 1867).

However, the fact that some colors cannot be perceived in combination, e.g. “reddish green” or “bluish yellow”, cannot be explained by the trichromatic theory. This proved to Edward Hering that the visual substances were organized as opponent processes (Herring, 1874). In summary, Hering proposed there are six fundamental color processes arranged as three visual contrasts including two opponent processes:

(i) black versus white,
(ii) red - green opponent process,
(iii) blue - yellow opponent process.

By the middle of the 20th century it was proven that both theories are necessary to explain the physiological processes of color perception. So, color vision is a dual process: the trichromatic theory is correct at photo-pigment level (by conical photoreceptors in the retina) and the opponent theory is correct at the neural level (by opponent cells found in the lateral geniculate nucleus).

4. Features

Scaling of the feature vectors is of special importance since the Self-Organizing Map classifier uses the Euclidean metric to measure the distances between feature vectors, otherwise bigger variables tend to dominate the others. The sigmoidal transformation (also called the softmax transformation) has been applied since it reduces the influence of outliers in the data. We also empirically observed that this normalization gives the best results, i.e.: $x_i' = 1/(1+e^{\widehat{x_i}})$ where $\widehat{x_i} = (x_i - \overline{x})/\sigma_x$ for a vector x with mean $\overline{x}$ and standard deviation σ_x.

4.1 Texture

The texture features are obtained by combining enhanced grating cell features (see Section 2.3) with spatially smoothed Gabor responses. The latter are obtained by convolving the simple cell responses, see (2) where $\varphi = 0$, with a Gaussian with standard deviation 2σ. Smoothed Gabor responses are known to improve the performance for texture analysis (Bovik et al., 1990). The frequencies for the filters are $\sqrt{2}, 2\sqrt{2}, 4\sqrt{2}, 8\sqrt{2}$, and $16\sqrt{2}$ cycles per image and we use eight orientations (θ=0, $\frac{\pi}{8}$,..,$\frac{7\pi}{8}$). This results in an 80-dimensional vector (smoothed Gabor responses + enhanced grating cell responses) to represent a texture feature.

4.2 Color

Digital images are mainly stored in RGB and can thus easily be transformed into color opponent values (COV) using the following transformation:

$$\begin{bmatrix} RG \\ BY \\ KW \end{bmatrix} = \begin{bmatrix} 1/2 & -1/2 & 0 \\ -1/4 & -1/4 & 1/2 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{9}$$

where RG, BY and KW represent the red-green, blue-yellow and black-white channel, respectively. For R, G, and B values between 0 and 255, the values for RG and BY range between -127.5 and 127.5, while KW ranges between 0 and 255. Remark that the transformation from RGB to HSI is computationally more expensive than the transformation into COV. Both the HSI and COV color space are considered in our experiments.

5. Unsupervised segmentation

Texture segmentation experiments are applied on multiple composite images of 256×256 and 512×512 pixels containing gray-scale textures from the Brodatz album (Brodatz, 1966). No pixel adjacency information has been used in this clustering process. Introducing pixel adjacency information generally boosts the classification correctness due to the fact that pixels belonging to the same texture are close to each other, and consequently, they should be clustered together. However, the latter requires a priori information about the image (which is not always available, e.g., in digital photos). Consequently, this method will not perform well if some texture regions are not adjacent in the image.

To segment an image, the extracted texture features (see Section 4.1) are employed to train a 2-dimensional Self-Organizing Map (SOM). In a first stage the map is linearly initialized

along the 2 greatest eigenvectors of the data. Next, the SOM is trained using the well-known batch-training algorithm which is, in contrast to the sequential training algorithm, much faster to calculate and the results are as just as good. A result of the training process is that pixels belonging to the same texture are assigned to the same or adjacent nodes. The number of nodes has evidently an influence on the classification result. A general rule is that a higher number of nodes results in better classification results but a side effect is that overclassification may occur. On the other hand, small-sized maps are more attractive because of their lower computational cost during the training phase. Nevertheless, we are able to use small sized maps to receive decent segmentation results because of the distinguishing characteristics of the proposed texture features. For the classification of images containing 2, 4, 5, and 9 textures, maps of dimension 4×2, 4×4, 8×7, and 8×9 nodes are trained, respectively. Figure 2 exemplifies the segmentation of images containing 4, 5 and 9 Brodatz textures using enhanced grating cell and smoothed Gabor features. Table 1 depicts the precision of the segmentation using real Gabor filter responses, smoothed real Gabor filter responses, enhanced grating cell responses, and the combination of the latter and former features. We notice that the real Gabor filter responses have low discriminating capabilities compared to the other features. The SOM-based classifier produces clearly the most precise segmentation using the combination of the enhanced grating cell features with smoothed Gabor filter responses.

# *textures*	*Real Gabor*	*Smoothed Gabor*	*Enhanced grating cell*	*Enhanced grating cell + smoothed Gabor*
2	0.69	0.97	0.89	0.97
4	0.34	0.85	0.72	0.90
5	0.42	0.85	0.67	0.89
9	0.25	0.81	0.59	0.86

Table 1. Segmentation precision of Brodatz textures.

To investigate the application of the proposed color and texture features for segmenting scenery images, experiments are conducted on 100 composite images of size 512 × 512 pixels containing randomly selected natural color textures. Each collected texture belongs to one of these five classes: (i) *bricks*, (ii) *grass*, (iii) *tree*, (iv) *sky*, and (v) *water*, as exemplified in Fig. 3. An example of the partitioning of natural textures is shown in Fig. 4. It is important to remark that in contrast to the gray-scale textures from the Brodatz album, the intra-variation in terms of orientation and scale of a natural texture class in scenery images is much higher. The latter is exemplified in Fig. 5 which consists of four different grass textures. As can be seen, it is even for the human eye hard to distinct the upper two textures, but the difference with the grass textures at the bottom of the image is much larger. Nevertheless, our segmentation algorithm is, to a certain extent, still capable to distinguish them, even using such a small-sized SOM. The segmentation results depicted in Table 2 are obtained by applying a 4×4 SOM for unsupervised segmentation. As can be seen, the combination of color and texture features gives a relatively stable segmentation result. We also notice that the application of color features gives a slight boost of about 5% while the precision using solely color rapidly decreases. The HSI and COV color space induces almost identical segmentation results. Since the transformation of RGB into COV is computational less expensive than HSI, the COV color space will be used in the next stage of our experiments.

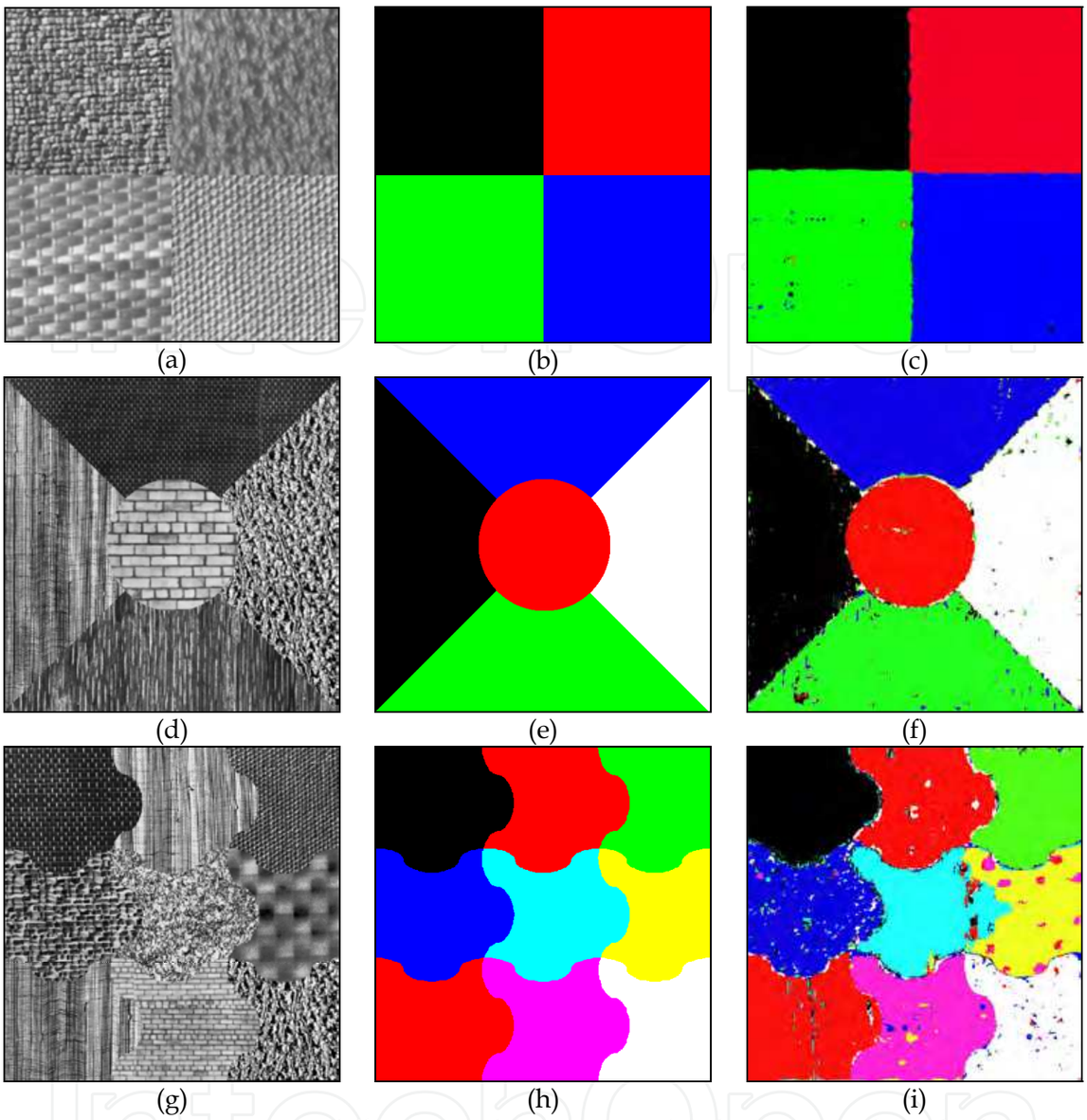
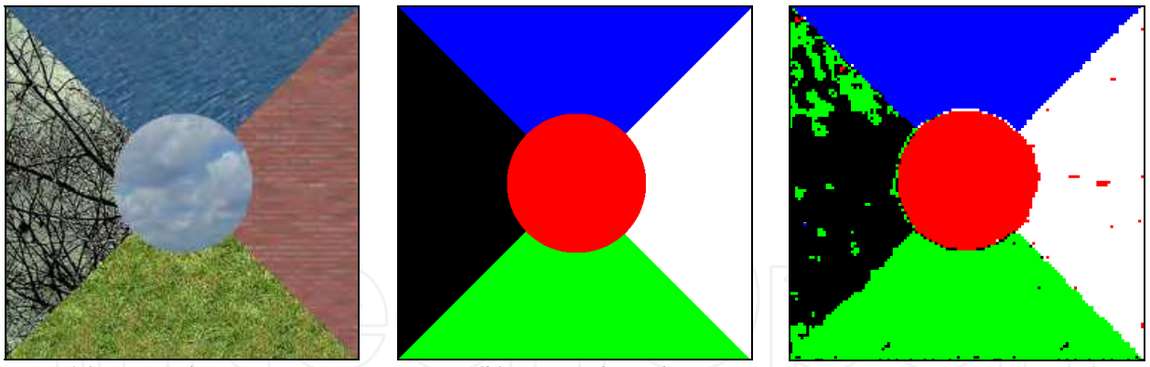


Fig. 2. Image containing different Brodatz textures (a, d, g); ground truth (b, e, h); segmentation using enhanced grating cell and smoothed Gabor features (c, f, i)

Fig. 3 Examples of real life textures: *branches* (a, b); *bricks* (c, d), *grass* (e), *sky* (f, g) and *water* (h, i).

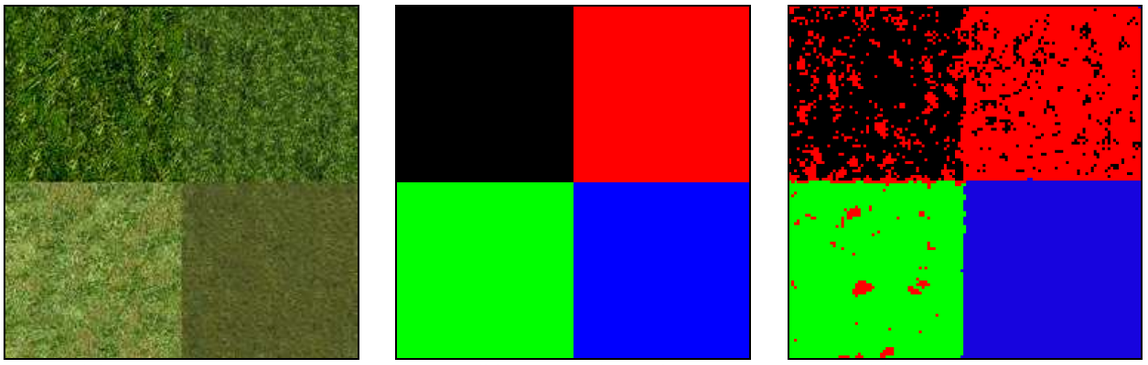
# *textures*	*COV*	*HSI*	*Texture*	*COV + Texture*	*HSI + Texture*
4	0.75	0.76	0.89	0.92	0.89
5	0.74	0.69	0.89	0.94	0.93
9	0.53	0.49	0.83	0.91	0.91

Table 2. Segmentation precision of real-life color textures



(a) natural textures (b) ground truth (c) segmentation

Fig. 4. Example of the segmentation of five natural textures.



(a) grass textures (b) ground truth (c) segmentation

Fig. 5. Segmentation of similar grass textures using COV + texture features.

6. Labeling of image regions

Using the previously computed color-texture samples, a 10×10 SOM is trained such that similar vectors are grouped in the same or in adjacent nodes. After this training phase, each SOM node is assigned a 5-dimensional probability vector v_c by counting the labels of the corresponding training vectors such that $v_c(i)$ is the probability that the label of node c is i. To test the labeling behavior, we employ ten random scenery images from the World Wide Web (containing no other classes than those listed above). To label a segment of an image, the following procedure is applied. At first, the best matching unit c_j for each corresponding feature vector is calculated. Then, the probabilities $v_{c_j}(i)$ are summed together for each i. Figure 6 shows a 2-dimensional mesh visualization of the 10×10 SOM along the 2 highest principle components of the training vectors. The color of each node in Fig. 5 corresponds to the label l with the highest probability, i.e. $v_c(l) > v_c(i), i \neq l$, and illustrates the fact that the different texture classes are nicely clustered.

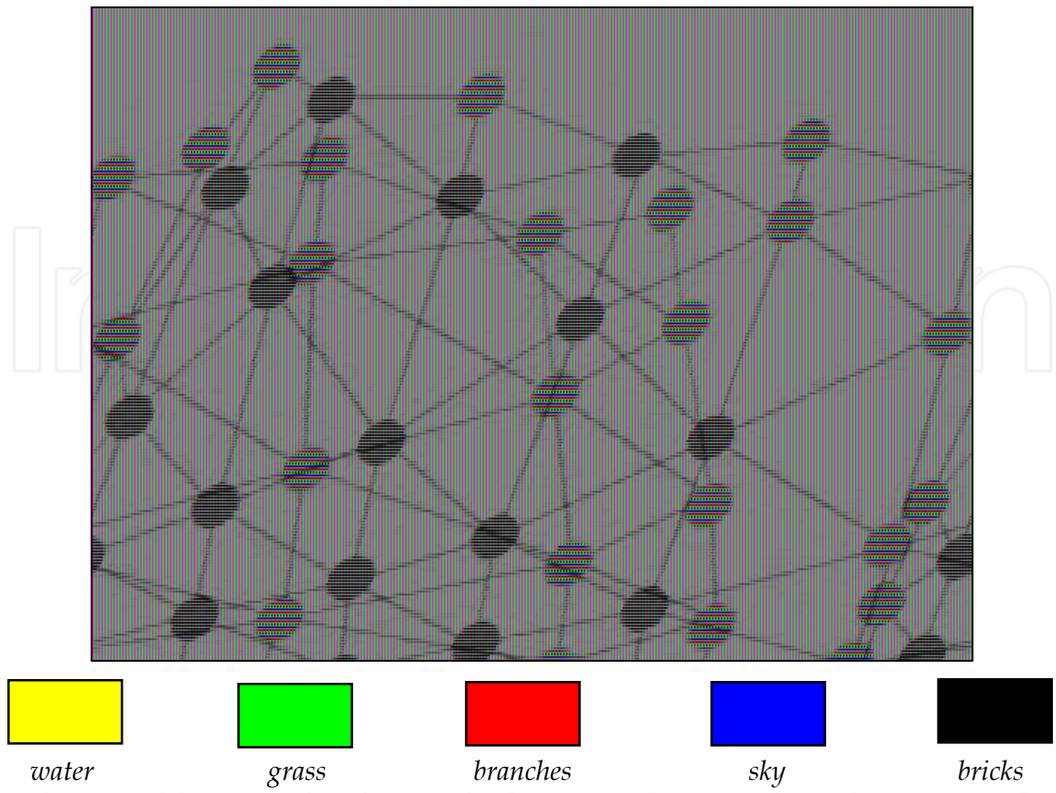


Fig. 6 A 10×10 SOM trained with natural color texture features. The nodes are given the label with the highest probability

Finally, the label of an image partition is obtained by selecting the label k with the highest probability, i.e. $k = max_i \sum_j v_{c_j}(i)$. Using the in Section 4 proposed texture and COV features, the average precision of the classification is 89% of the pixels while the recall is 86%. The labeling precision without an intermediate segmentation step was 72% (Martens et al., 2008). Remark that the ground-truth images are created manually and therefore they should be interpreted as an approximation rather than a certainty. The interpretation of some scenery images is exemplified in Fig. 7. In this figure, different types of errors occur. At first, isolated pixels and edges are misclassified. The latter occurs because they are assigned to the same cluster of nodes in the segmentation step. Consequently, during the labeling step, they will be given the wrong label. To avoid this, small blobs can be filtered out by assigning thresholds to define the minimum size of a partition. Information of enclosing or adjacent regions can then be used to find the most probable label. Another type of error is caused by misclassification in the labeling step. An example hereof is the wrong interpretation of the roof, see Fig. 7 (i). This is mainly caused by the fact that the training set of the class *bricks* doesn't contain any examples which are similar to the corresponding region in Fig. 7 (g). The *sky* blob in the lake of Fig. 7 (d) is also a fault. However, the latter is due to reflection: the increase of brightness causes the contrast to decrease such that the texture significantly alters and a misclassification results. Such errors can only be corrected by incorporating domain knowledge such that, e.g., no blob of bricks can 'float' in the sky.

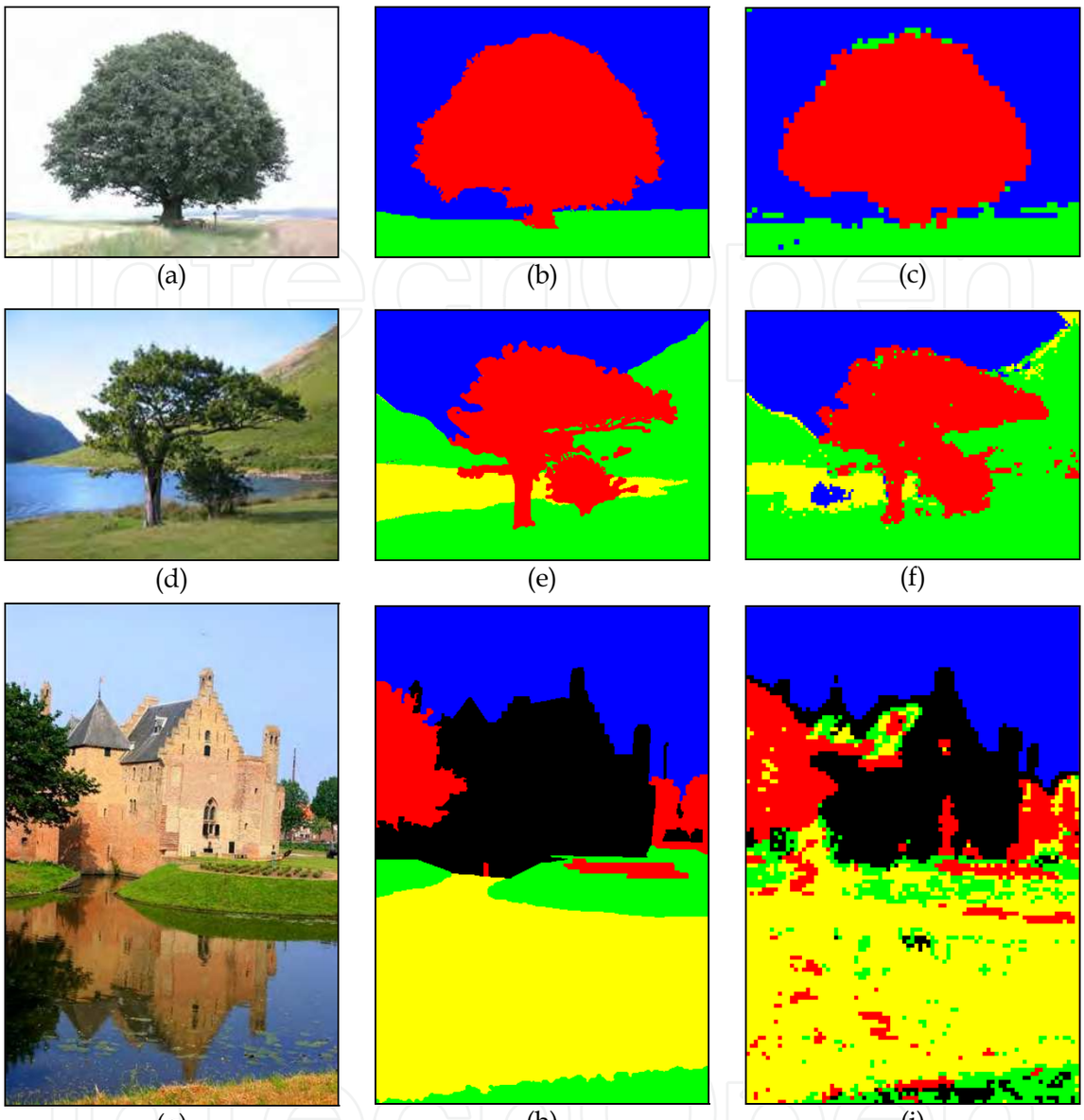


Fig. 7. Interpretation of scenery images: original image (a, d, g), ground truth (b, e, h), and classification (c, f, i).

Other errors, e.g., the *water* detected as *grass* (or vice versa), emerge from the fact that the scaling of certain textures alters due to the perspective. This problem is harder to tackle. Enhancing the segmentation process, e.g. by using a larger SOM, will certainly be helpful but a special approach might be needed for those regions. However, a top-down approach for the detection and correction of misclassifications by embedding domain knowledge is out of scope of this chapter.

7. Conclusion and future work

Due to the semantic gap, the automatic interpretation of images is an intricate task. In this chapter, we have presented a bottom-up approach for the segmentation and interpretation of outdoor scenery images. We established a link between the proposed low-level, biological features and some predefined semantic concepts by applying a SOM for classification. Our method generally consists of two stages. At first, color opponent values and textures are extracted from the image's pixels. Color opponent values induce comparable classification results as colors from the HSI space. Since the transformation of RGB into color opponent values is computationally less expensive than the transformation into HSI, the former are preferred over the latter. The texture features consist of enhanced grating cell features and smoothed Gabor responses and correspond to outputs of cells found in the primary visual cortex of primates and humans. Analogously to the processing principles of the auditory and visual cortex, Self-Organizing Maps are used for the unsupervised segmentation and labeling of textured images. Even using small-sized maps, high precision image segmentations can be obtained (both on gray-scale and on natural color textures). By adding color information, the precision of the segmentation results averagely increased with 5% to a total of 91% of the pixels. In the next stage, the same features are used to train a Self-Organizing Map with textures belonging to one of the 5 predefined classes: (i) *grass*, (ii) *bricks*, (iii) *branches*, (iv) *water*, and (v) *sky*. This map is then used to label the previously obtained image segments. Experiments conducted on randomly collected images from the World Wide Web achieved a precision of 89%. The latter observations indicate that the application of biologically inspired features is very useful for scene interpretation and categorization. We further believe that the classification can be improved by (i) a more accurate segmentation, (ii) a larger, more representative training set, and by (iii) introducing high-level domain knowledge (i.e. a top-down approach). These aspects will be thoroughly investigated. In order to recognize more concepts, we have experienced that when extra texture classes are added to the training set, the number of misclassifications drastically increases. A solution to this problem might be the introduction of hierarchical or tree-structured Self-Organizing Maps (Koikkalainen & Oja, 1990). Nodes containing two (or more) classes are, in a next stage, split up into different clusters what results in the separation of the related concepts. Furthermore, since the visual cortex contains different types of (complex) cells which are tuned to a specific task (e.g., for the detection of edges), we believe that they can also play an important role in image understanding.

8. Acknowledgement

The research activities as described in this paper were funded by Ghent University, the Interdisciplinary Institute for Broadband Technology (IBBT), the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), the Fund for Scientific Research-Flanders (FWO-Flanders), and the European Union.

9. References

Bovik A. C., Clark M. & Geisler W. S., (1990), Multichannel Texture Analysis Using Localized Spatial Filters, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12 (1), pp. 55-73.

Brodatz P. (1966), Textures: A Photographic Album for Artists and Designers, Dover, New York.

Clausi D. A., Jernigan M. E. (2000), Designing Gabor Filters for Optimal Texture Separability, *Pattern Recognition*, 33, pp. 1835-1849.

Daugman J. (1985), Uncertainty Relation for Resolution in Space, Spatial Frequency and Orientation Optimized by Two-Dimensional Visual Cortical Filters, *Journal of the Optical Society of America A: Optics, Image Science, and Vision*, 2, pp. 1160-1169.

Duygulu P., Barnard K., Freitas N. & Forsyth D. (2002), Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary, *Proceedings of the European Conference on Computer Vision*.

Feng S., Manmatha R. & Lavrenko V. (2004). Multiple Bernoulli relevance models for image and video annotation, *Proceedings of the IEEE Computer Vision and Pattern Recognition*.

Flickner M., Sawhney H., Niblack W., Ashley J., Huang Q. , Dom B., Gorkani M., Hafner J., Lee D., Petkovic D., Steele D., & Yanker P.(1995), Query by image and video content: The qbic system. *Computer*, 28 (9), pp. 23–32.

Gever T. (2001), Color-Based Retrieval, In: *Principles of Visual Information Retrieval*, Springer, Lew M. S., (Ed.), pp. 11-49, Springer, 1852333812, London, Great Britain.

Hering E. (1874), Zur Lehre vom Lichtsinne.

Jain A. K., Farrokhnia F. (1991), Unsupervised texture segmentation using Gabor filters, *Pattern Recognition*, 24 (12), pp. 1167-1186.

Jones J., Palmer A., (1987), An Evaluation of the Two Dimensional Gabor Filter Model of Simple Receptive Fields in Cat Striate Cortex, *Journal of Neurophysiology*, 58, pp. 1233-1258.

Kohonen T. (2001), *Self-organizing Maps*, Springer, 3540679219, Berlin, Germany.

Koikkalainen P. & Oja E. (1990), Self-organizing hierarchical feature maps, *Proceedings of International Joint Conference on Neural Networks*, 2, San Diego, CA, USA, pp. 279-285.

Kruizinga P. & Petkov N. (1998) Grating Cell Operator Features for Oriented Texture Segmentation, Proceedings of the 14th International Conference on Pattern Recognition, Brisbane, Australia, pp. 1010-1014.

Kruizinga P. & Petkov N., (1999) Nonlinear Operator for Oriented Texture, *IEEE Transactions on Image Processing*, 8 (10), pp. 1395-1407.

Laaksonen J., Koskela M., & Oja E., (2002), PicSOM - Self-Organizing Image Retrieval With MPEG-7 Content Descriptors, *IEEE Transactions on Neural Networks*, 13 (4), pp. 841-853.

Li J. & Wang J. (2003), Automatic linguistic indexing of pictures by a statistical modeling approach, *IEEE Transactions on. Pattern Analyses and Machine Intelligence*, 25 (9).

Martens G., Poppe C., Lambert P. & Van de Walle R. (2008), Unsupervised texture segmentation using biologically inspired features, *Proceedings of the 2008 IEEE 10th Workshop on Multimedia Signal Processing*, pp. 159-164.

Mehrotra S., Rui Y., Ortega-Binderberger M., & Huang T. S. (1997), Supporting content-based queries over images in mars. *Proceedings of IEEE International Conference on Multimedia Computing and Systems*, pp. 632–633.

Picard R., (1995). Digital Libraries: Meeting Place for High-Level and Low-Level Vision, *Proceedings of the Asian Conference on Computer Vision*.

Renninger, L. W. & Malik, J. (2004), When is scene recognition just texture recognition?, *Vision Research*, vol. 44, pp. 2301-2311.

Riesenhuber M. & Poggio T. (2003). How the visual cortex recognizes objects: The tale of the standard model, *The Visual Neurosciences*, 2, pp. 1640-1653.

Szummer M. and Picard R., (1998). Indoor-outdoor image classification, *Proceedings of the Workshop in Content-based Access to Image and Video Databases*, Bombay, India.

Turtinen M. & Pietikäinen M. (2003). Visual training and classification of textured scene images}, *Proceedings of the 3rd International Workshop on Texture Analysis and Synthesis*, pp. 101-106.

Vailaya A., Jain A., and Zhang H. (1998). On image classification: City versus landscape, *Pattern Recognition*, 31, pp. 1921–1936.

von Helmholtz H. (1867), Handbuch der Physiologischen Optik, Leopold Voss, Leipzig.

Wang W., Song Y., and A. Zhang. (2002), Semantics retrieval by content and context of image regions, *Proceedings of the 15th International Conference on Vision Interface*, pp. 17–24, 2002.

Zhang, J., Tan, T., Ma, L. (2002), Invariant texture segmentation via circular Gabor filter, *Proceedings of the 16th IAPR International Conference on Pattern Recognition*, 2, pp. 901–904.

Zhu L., Zhang A., Rao A., & Srihari R. (2000). Keyblock: An approach for content-based image retrieval}, *ACM Multimedia 2000*, pp. 157–166, Los Angeles, CA, 2000.

Self-Organizing Maps

Edited by George K Matsopoulos

ISBN 978-953-307-074-2
Hard cover, 430 pages
Publisher InTech
Published online 01, April, 2010
Published in print edition April, 2010

The Self-Organizing Map (SOM) is a neural network algorithm, which uses a competitive learning technique to train itself in an unsupervised manner. SOMs are different from other artificial neural networks in the sense that they use a neighborhood function to preserve the topological properties of the input space and they have been used to create an ordered representation of multi-dimensional data which simplifies complexity and reveals meaningful relationships. Prof. T. Kohonen in the early 1980s first established the relevant theory and explored possible applications of SOMs. Since then, a number of theoretical and practical applications of SOMs have been reported including clustering, prediction, data representation, classification, visualization, etc. This book was prompted by the desire to bring together some of the more recent theoretical and practical developments on SOMs and to provide the background for future developments in promising directions. The book comprises of 25 Chapters which can be categorized into three broad areas: methodology, visualization and practical applications.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen