Chapter 1

IMAGE CONTENT ANALYSIS AND DESCRIPTION

Xenophon Zabulis, Stelios C. Orphanoudakis
Institute of Computer Science, Foundation for Research and Technology - Hellas,
Vassilika Vouton, P.O. Box 1385, GR-71110 Heraklion, Crete, Greece
and
Department of Computer Science, University of Crete,
P.O. Box 1470, GR-71409 Heraklion, Crete, Greece
{zabulis,orphanou}@ics.forth.gr

Abstract In this chapter the task of representing, describing, and analyzing visual information is discussed, in the context of image retrieval by content. Initially some basic specifications of the problem are presented and a classification of visual features, in a way compatible with human visual perception, is proposed. Through this discussion, it is realized that scale is an important attribute of visual content and a central issue in its description. Thus, a significant part of this chapter is devoted to the estimation and representation of primitive image content at different scales and a generic framework for this purpose is introduced. Finally, this chapter briefly considers the problem of how to derive and match descriptions of the visual content of an image in a perceptually correct manner.

Keywords: Image retrieval by content, visual information retrieval, scale-space, perceptual grouping, feature extraction, similarity matching.

1.1 Introduction

The large volume and variety of digital images currently acquired and used in different application domains has given rise to the requirement for intelligent image management and retrieval techniques. In particular, there is an increasing need for the development of automated image content analysis and description techniques in order to retrieve images efficiently from large collections, based on their visual content. Large collections of images can be found in many application domains such as journalism, advertising, entertainment,

R.C. Veltkamp et al. (eds.), State-of-the-Art in Content-Based Image and Video Retrieval, 1–19.

weather forecasting, map production, remote sensing, computer aided design, architecture, vision-based robot navigation, medicine, etc. Thus, an important functionality of next generation image database and multimedia information systems will undoubtedly be the search and retrieval of images based on visual content. In the near future, this functionality will also be supported by "intelligent" search engines used to access multimedia documents on the world-wide web.

Before a general solution to the problem of image browsing based on visual content can be found, there are many difficulties to be overcome. These difficulties stem primarily from the following facts or observations: 1) what constitutes image content in general is not well defined, 2) the degree of image similarity or dissimilarity is often context and goal dependent, 3) the types of images used and the requirements for content-based retrieval of such images are different for different application domains, and 4) mechanisms for selecting the image features to be used in content description and matching techniques are not well understood. Specifically, depending on the user's goal associated with a specific similarity search, a query by image content may be constructed based on either abstract or specialized image features. Image features used may also be global or local. The features used affect the precision of the response to a query by image content and the cardinality of the returned set of similar images. Precision and cardinality are also dependent on whether queries, using spatial and visual feature predicates, are exact or approximate. Exact queries require that a specific set of content descriptive criteria are necessarily satisfied, while approximate queries typically retrieve image with respect to their similarity with one or more visual examples.

The image type and context of use often determine those regions of interest and features that are characteristic of image content. The same visual stimuli may have different interpretations when observed in different contexts or by different observers. Furthermore, there is a semantic gap between the pictorial properties of an image and its linguistic interpretation. Thus, given these difficulties, the efficient, objective, and qualitative description of image content for the purpose of image similarity search is a complex task. A fundamental component of image content is structure, which resides at different scales between the scale defined by the sampling interval (pixel size) and the one corresponding to the size of the image itself. Therefore, in order to focus attention at structures of different sizes, it is important to have the ability to select the appropriate scale. If the size of a particular structure is known, the problem of estimating properties of this structure is simpler to solve. In general, scale selection is applicable to almost all image processing, feature detection, and image description tasks. Since scale selection appears to be an important factor in image content analysis and description, a significant part of this chapter

is devoted to the estimation and representation of primitive image content at different scales.

In describing the visual content of images and using such descriptions to retrieve similar images, the use of primitive image features may not be sufficient. One may also need to rely on more complex features obtained through perceptual grouping of primitive ones. In fact, a better understanding of human visual perception will undoubtedly contribute to the development of biologically relevant image content descriptions and more efficient mechanisms of image retrieval based on visual similarity. In this context, it is particularly important to define image similarity metrics, which correspond to known mechanisms of human visual perception. In this chapter, we also examine the role of human visual perception, and perceptual grouping in particular, in deriving descriptions of image content at a higher level than that afforded by primitive structure alone. Finally, this chapter briefly considers the problem of how to derive descriptions of the visual content of an image, which preserve information about its primitive, global, and perceptual features, while permitting salient regions within the image and selected features of such regions to carry an additional weight in image comparisons.

1.2 Problem Definition

The task of image retrieval by content may be subject to a number of requirements with regard to query types supported, retrieval precision, and the number of images retrieved. In all cases, one must first analyze and describe the visual content of the query image and match it to similar descriptions of images in a database. Before the central problem of image content analysis and description is addressed, a number of related problems and constraints are discussed below:

- *Image segmentation.* Segmenting an image into parts that are meaningful with respect to a particular application is critical in image understanding. However, segmenting an image into regions that correspond to distinct physical objects, using solely two-dimensional visual information is difficult or impossible to achieve. This is due primarily to the lack of three-dimensional models for every possible identifiable physical object and missing information regarding image acquisition parameters.
- *Motion and stereo vision* are sources of rich visual information. Visual cues provided by motion and stereo facilitate the extension of object boundaries, as well as the estimation of scene structure. On a semantic level, certain types of motion may constitute intense attractors, dominating an observer's attention. In static images there is no such visual cue. Similarly, stereoptic images can be used to estimate scene structure, thus contributing to the identification of distinct physical objects and scene understanding. This information is not available, in single images.

- *Lighting.* Knowledge of scene lighting plays an important role in the correct estimation of an object's reflectance spectrum. Human perception normalizes perceived spectra with respect to global scene illumination, a phenomenon known as "color constancy". However, in the general case of image acquisition, the scene illumination is neither known nor homogeneous. Specialized cases of color normalization, given certain assumptions about lighting conditions and / or an object's reflectance spectrum, exhibit interesting results, but the full reconstruction of an object's reflectance spectrum from a 3 band RGB image, is not trivial.

- *Object recognition.* The ability to identify specific objects in images would support the retrieval of semantically similar images. Images containing the same or "similar" objects, or even a contextually relevant object, may be considered as semantically related. Furthermore, object semantics may vary depending on the image observation goal and context, as well as the expectation of finding a particular object in a certain visual scene. For example, a tree trunk, which has been cut and is lying on the ground, may be characterized as a chair when taking a walk in the forest, while it could not be matched with any chair, stool or sofa model [1].

- *Context.* As already mentioned, the context of a query by image content and the type of images used have a strong effect on how the content of these images is described and compared. Contextual information and knowledge of the world are essential in deriving an appropriate image representation and may influence the role and significance of specific objects in such interpretations. Furthermore, the target class of images in a search and retrieval by visual content exercise may play an important role in determining which preprocessing methods are to be used for feature extraction.

- *Time.* Biological visual systems employ several physiological adaptation behaviors through time, such as lightness or chromatic adaptation [2], as well as motion adaptation [3]. Furthermore, given enough observation time, certain image features or details may be emphasized in the viewer's perception, depending on his / hers cognitive background and goal of observation. In this study a contextually uncommitted analysis of visual content is attempted, taking only into account only the early stages of visual perception.

- *Feedback.* Image feature extraction in biological vision systems may be adjusted depending on viewpoint, lighting conditions, query target, learning, adaptation and other factors. Feedback connections exist in the visual cortex, however their functionality has not yet been clearly understood. Certain image preprocessing methodologies may use feedback to improve feature extraction, but a generic framework for this is yet to be found.

In this chapter, a phenomenological thesis is adopted concerning the description of primitive image content and the evaluation of generic visual similarity. It is argued that visual content, once objectively represented, could also be appropriately interpreted, with respect to the context of use. However, most prominent of all problems mentioned above seems to be the quantification of qualitative visual attributes, such as image feature impression or the holistic perception of a scene.

1.3 Image Content Representation

Image features, including form and color or intensity distributions, as well as their spatial organization, compose primitive image content. However, some visual features reside in the perceptual domain, often defined by specific types of primitive feature arrangements. Some of these features may be detected by applying perceptual grouping rules [4] and are of strong descriptive power regarding the visual perception of a scene. Also, specific feature distributions may indicate regions of special interest in an image, such as regions attracting our preattentive attention, thus constituting qualitative information about image content. The scale at which features are observed is a central issue in image description. In this section, the importance of size in all cases of feature extraction is considered and some tools for dealing with feature scale are introduced. Methods for the estimation and classification of visual image features are also presented and discussed. In certain cases, an analogy is drawn between the applied methods and corresponding human visual perception mechanisms.

After a brief discussion of global image features and their role in image content description, this section emphasizes the estimation of primitive image features at selected scales and the representation of primitive feature distributions. Additional topics presented in this section are the perceptual grouping of primitive features into more complex ones and the role of regions of interest in images as attractors of attention in image retrieval by content.

1.3.1 Global features

A global statistical description of image features has been widely used in image analysis for image description, indexing, and retrieval. Such global feature descriptors include the image's color histogram, edge statistical infor-

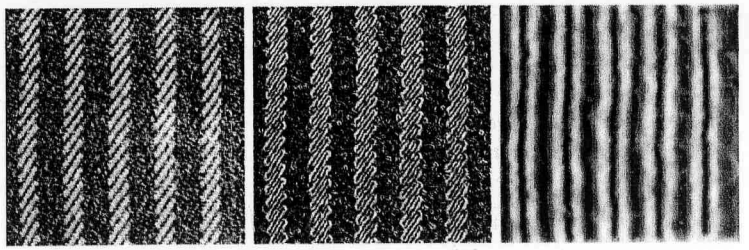

Figure 1.1 An original image (left) analyzed for edges at a fine (middle) and a coarse scale (right).

mation, wavelet or Fourier decomposition statistics etc.. Although these global attributes may be computed efficiently and often do succeed in capturing partial information about image content, they do not capture information about internal structure and cannot make use of any prior knowledge about a user's notion of image similarity based on specific interest in certain aspects of image content.

Despite the small discriminating power of such descriptors and the fact that they do not capture the spatial distribution of image features, their importance cannot be underestimated. Context based heuristics may be also used i.e. the detection of images containing man made structures may be achieved by searching for ones that are rich in straight line segments. In general, global image descriptors may offer important hints about overall visual appearance of an image, the image type, and certain possibly characteristic image properties. Using this information, images may be classified into categories, thus restricting the search space of image queries. Furthermore, knowledge of the image type often permits the selection of more suitable content analysis methods.

1.3.2 Primitive Features

Primitive features, such as edges, corners, blobs etc, model specific types of pixel distributions and constitute the "building blocks" of image content. They are highly correlated with the scale of observation and it is expected that their classification with respect to scale will contribute to the refinement of visual query formulation. As illustrated in Fig. 1.1, different aspects of visual content are observed as scale increases. Therefore, a full description of image content cannot be obtained by considering a single scale.

Observing the image input signal at all scales [5], using gradually coarser sampling, reveals the image content at each scale. Features at each scale can be detected by applying the appropriate operator at these scales. The operator response indicates the intensity (or probability) of feature presence at each pixel

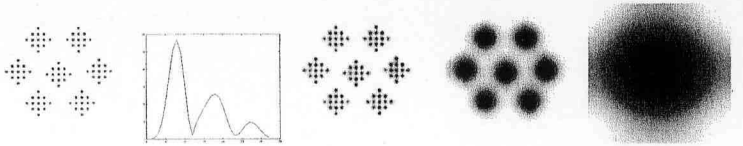

Figure 1.2 An image (left) and the plot of the blob scale selector response for the image point corresponding to the central dot. The three rightmost images present the image scales corresponding to the response function modes.

and with its use, feature points at each scale may be extracted. It may be scale normalized as in Eq. (1.1), where $\tau = \log t$ is the logarithmic scale parameter, $\vec{x}$ the pixel coordinates, $\mathcal{F}(\vec{x}, \tau)$ the feature operator and h is any strictly increasing function, chosen according to the nature of the feature detector.

$$\frac{1}{\int_0^\infty h(t\mathcal{F}(\vec{x}, \tau))\, d\tau} h\left(t\mathcal{F}(\vec{x}, \tau)\right). \tag{1.1}$$

In the example shown in Fig. 1.1, the feature detectors used, were the squared norm of the image gradient for edge detection, and the Harris [6] operator for corner detection. The example shown in Fig. 1.2, illustrates the response of the scale normalized Laplacian blob detector ($\mathcal{F}(\vec{x}, \tau) = |\frac{\partial^2}{\partial x^2} L(\vec{x}, \tau) + \frac{\partial^2}{\partial y^2} L(\vec{x}, \tau)|$, where L is the image scale-space), over all scales for the central point of an image, where the horizontal axis maps the logarithmic scale parameter τ. The bottom row presents the image scales corresponding to each of the detector's response modes. In general, the scale-normalized feature operator response reveals peaks, where the feature presence is most intense. Often more than one peak exists in a pixel's response over scale.

The intrinsic importance of scale in visual perception is observed in primate visual systems, where the sampled signal is passed as input to M and P retinal ganglion cells, that respond to spatial and temporal illumination change [7]. Different spatial "samplings" of ganglion cells are separately projected to the Lateral Geniculate Nucleus (LGN), and terminate in different regions of the visual cortex [8]. Coarse samples provided by the M ganglion cells project to magno cells in LGN, which are color-blind, high-contrast sensitive and with a fast neural response. In contrast, fine samples projected from P cells to parvo cells in LGN are color and low contrast sensitive, but have a slower response. Fig. 1.3 illustrates the described circuitry, representing M and P ganglion cells using black and white circles respectively. It is argued that the multiscale feature representation described in this section, is analogous to the primitive content representation observed in LGN.

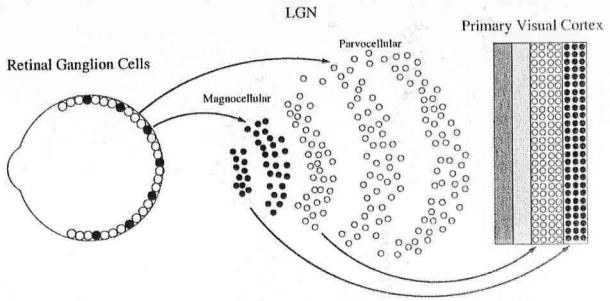

Figure 1.3 Physiology of visual receptors in LGN and the Primary Visual Cortex.

For each image point, the modes of the feature operator response may be represented in scale space, indicating the scales that the feature presence is most intense. Selecting the scale function's mode is a way to select the feature's scale [9], however it is argued that neighboring scales are equally useful, concerning the stability and smoothness of the result. Existing scale selection algorithms employ tracking of features in this three dimensional space in order to select a single feature scale, but in a computationally expensive fashion. Linearly summarizing visual content over neighboring scales [10] yields a more stable result and reduces the dimension of the scale-space to be searched. The Scale Summary Representation (SSR) is defined as a weighted sum,

$$J(\vec{x}) = \sum_{\tau} w(\vec{x},\tau)\mathcal{F}[L(\vec{x},\tau)], \tag{1.2}$$

$$L(\vec{x},\tau) = G(\vec{x},t) * I(\vec{x}), \tag{1.3}$$

$$\sum_{\tau} w(\vec{x},\tau) = 1, \tag{1.4}$$

where I is the original image, L is the image scale-space, and w the probability of feature presence. The summarization of feature content, restricted to a neighborhood of scales, may be achieved with the introduction of Scale Focusing (SF), achieved by multiplying the scale selector function at each pixel with a Gaussian function given by $w'_{m,s}(\vec{x},\tau) = (1/(\sqrt{4\pi s(\vec{x})}))\exp(-((\tau - m(\vec{x}))^2/(4s(\vec{x}))))$ where m is the scale of interest, and s the width of the scale neighborhood. Images in Fig. 1.1 were actually generated by scale focusing visual content over neighborhoods of fine and coarse scale. Fig. 1.4 illustrates the SSR response, for blobs and edges over all scales of an image.

Adjusting the scale of observation and image analysis with respect to local structure is important in morphologic image content description, since attention

Figure 1.4 An image (left) and the SSR for edges (middle) and blobs (right).

may be accurately drawn to structures of different sizes. Locally adaptive image processing methodologies are often employed in order to cope with the continuum of different scales of image structure. For example, adaptive image smoothing may be used as an image description preprocessing tool before image morphologic segmentation, since it simplifies the signal by reducing its variance. The instability of explicit scale selection is demonstrated in Fig. 1.5, where image smoothing is carried out by locally applying a smoothing kernel that fits local image structure. The size of the kernel, for each pixel, is estimated by using the Laplacian as the feature detector $\mathcal{F}$ in Eq. (1.1) and selecting the first mode of the response. In the second case, neighboring scales are combined using SF, reducing the effect of signal discretization and yielding a smoother estimation of scale. The last image demonstrates SF at the dominant mode of the feature response over scale. In this case, the Gaussian is centered at the maximum of the scale selector function.

Summarizing visual content over spatial and scale neighborhoods may very well be closer to human perception, if one considers the analog nature of signal propagation through neurons. Neurophysiological evidence shows that the visual signal is subject to both temporal and spatial smoothing [11]. Spatial averaging of information is performed by horizontal retinal cells, which sum-

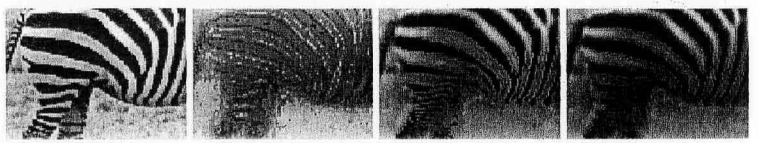

Figure 1.5 Application of SF in image smoothing. From left to right: Original image, explicit scale selection, first mode SF, and dominant mode SF.

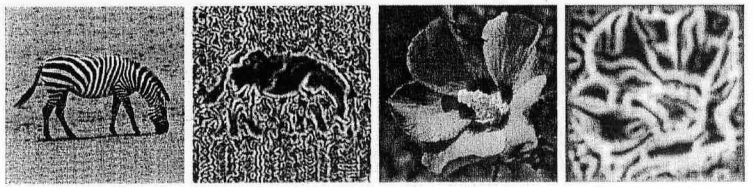

Figure 1.6 The first pair of images shows an image and the magnitude of scale summarized (with respect to image gradient) histogram gradient for the orientation feature. The second pair displays a reddish flower over a green background and the corresponding magnitude of the color histogram gradient.

marize neighboring receptor's outputs, giving emphasis to samples synapsed near their center. The non-exact regularity of the receptor grid implies mixing of neighboring scales information and supports the compatibility of scale neighborhood summarization with human perception. At a higher level, different spatial frequency responses are to be linearly (due to the computational nature of neurons) combined towards a single image perception. Focusing on certain scales for feature extraction is analogous to the attentional activation of winner-take-all neural networks [12].

The classification of primitive features with respect to scale plays an important role in visual perception. Furthermore, abstract scene features encountered at coarse scales may be exploited by image database queries in order to optimize retrieval time. Images that are dissimilar on a coarse level cannot exhibit generic visual similarity. The spatial layout of image features is also part of image content and may be adequately described through topological graphs.

1.3.3 Primitive Feature Distributions

Image content resides in other types of pixel distributions as well, rather than edges, corners, blobs etc. The statistical description of pixel intensities over space, referred to as texture, has been thoroughly studied in literature (see [13] for an overview). Such distributions may not only be characterized by pixel intensities, but also color, local orientation, periodicity etc., and also scale of observation. A generalized representation of such content may be generated by computing the local histogram $h_s(\vec{x})$ of the feature distribution at different scales. By varying the sampling area size, a scale space of such images is defined as $H(\vec{x}, \tau)$ [14].

Through the combination of distribution descriptors and attributes, the dissimilarity of visual feature patterns over a region may be quantified. The histograms' gradient magnitude ($|\vec{\Delta h}| = (\partial\vec{h}/\partial x)^2 + (\partial\vec{h}/\partial y)^2$) or other distri-

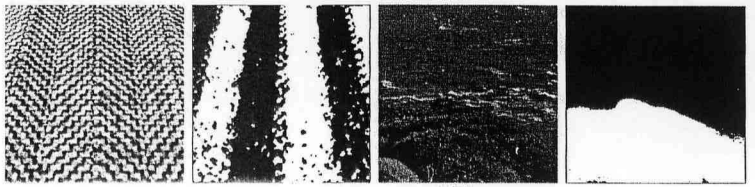

Figure 1.7 Clustering of scale summarized local histograms for orientation (left) and intensity distributions (right).

bution dissimilarity metrics [15], may be used as the distance function in the discrimination and comparison of individual feature distributions (Fig. 1.6 illustrates two such examples). Dimensionality of the described feature analysis may be reduced into a single histogram image, using scale focusing or summarization, and clustering of local histograms can discriminate regions of coherent feature distributions, regardless of scale. In the next example (see Fig. 1.7) the original images have been clustered with respect to intensity and orientation distributions, after computing the histogram SSR for all image scales. As observed in the images, the scale of feature observation varies. Clustering was carried out using K-means clustering algorithm, without taking spatial layout into account. The spatial arrangement of distributions, may be taken into account by associating local distributions to graph nodes and segmenting that graph [16], however using vast computational effort. Other distribution descriptors, such as the mean value, variance or entropy, may be used in combination, for the discrimination of different or salient feature distributions types, as well. Fig. 1.8 shows an image and the entropy of local orientation histograms in contrast with

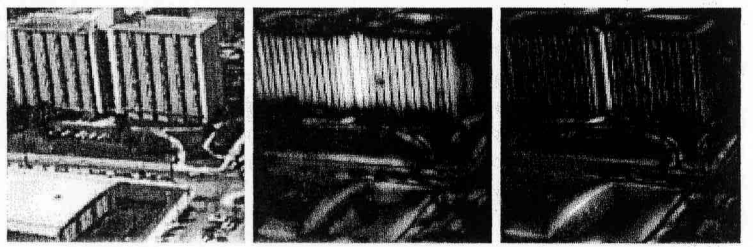

Figure 1.8 An image (left) and the orientation distribution visualizations using entropy (middle) and scale summarized gradient (right).

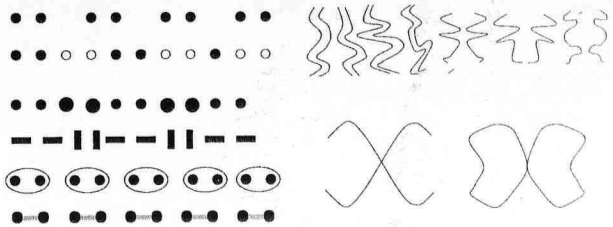

Figure 1.9 Cases of perceptual grouping.

the image gradient, summarized over all scales for both images. Notice that in the entropy image, regions of high informational organization stand out.

1.3.4 Perceptual Features

Observation of an image typically results in the perception of whole connected objects (or visual entities), consisting of simpler parts, which shows that the spatial layout of features is an important component of image content. The human visual system dynamically organizes specific primitive feature arrangements, into perceptual entities of significant contextual importance in the visual impression of an image. Shape, an important visual cue, is often defined by the grouping of individual primitive features. Rules that describe such visual behaviors originate from Gestalt psychology [17], while novel ones have been recently formulated by vision scientists [18][19]. The embodiment of perceptual grouping rules in the techniques of machine vision is expected to yield a more descriptive representation of visual content and, consequently, intuitively more precise responses to image queries by visual content.

Rules of perceptual organization are based on symmetry, parallelism, collinearity, proximity, closure, connectivity, contour continuity, color or size similarity and other. Fig. 1.9 shows the perceptual grouping of different primitive features. Physiological evidence that supports the process of feature aggregation into perceptual entities may be found in cases of brain disease [20]. Ways of measuring the significance of feature groupings are discussed at length in [21], from both the computational and psychophysical points of view. However, a serious drawback of perceptual grouping theory is the lack of a general purpose scheme, for integrating several potential factors into an overall outcome. This fact points to the need for different information source integration.

Most often, perceptual grouping of features composites the form of visual entities. In everyday practice and without effort, features are assembled into significantly important visual entities, despite occlusion or cluttered background. Visual form may originate by perceptual features as well, rather than just edge

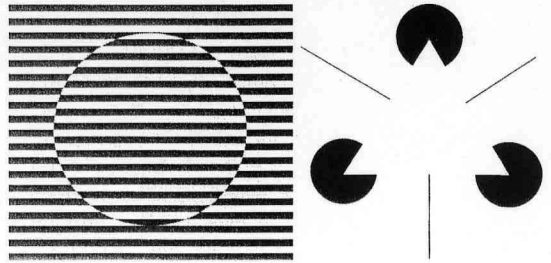

Figure 1.10 Cases of perceptual form definition, through illusionary contours.

segments, as Fig. 1.10 illustrates. The entity's boundary is commonly used to represent visual form and is many times an important visual cue (see [22] for an overview of shape representation and matching). Naturally, scale is an important component of visual form and characterizes all features of an object's boundary (such as curves, corners, or straight segments).

Psychological and physiological observations concerning the nature of the perceptual representation of form indicate that shape regions of high curvature as well as edge endings are of noticeable significance in visual perception [23] [24]. Except from curvature, the scale of such regions also characterizes their impact in visual perception of form. In analogy with the methodology proposed in Sec. 1.3.2 the curvature feature may be studied in a scale space of the contour. The sum of scale-normalized curvature over scale, is introduced as a quantification of a region's contribution in form definition. Apart from psychophysical reasons, regions responding with a high value to this summation are computationally more stable to noise, and thus suitable for form description and matching [25]. Fig. 1.11 illustrates the scale-normalized curvature response over scale for three shape regions: one of very high spatial frequency (observed as noise rather than structure and at coordinates 220, 140 of the left figure), one of a medium scale but sharp peak (at 330,70) and one of a large scale mild curve (at 250,270). The right figure plots the scale normalized curvature for these three image points over scale, where the horizontal axis maps the logarithmic scale parameter and the vertical one the curvature response. The noise dent corresponds to weakest response, the sharp curve to the one that takes the maximum value among the three. The large scale curve corresponds to the longest surviving in scale response.

Finally, a quite subjective image feature, with respect to the goal of image observation, is the existence of face-like structures. Although that in many applications the feature is of no use, it may be an attention-attracting feature, closely related to image content, in others. The significance of the feature is

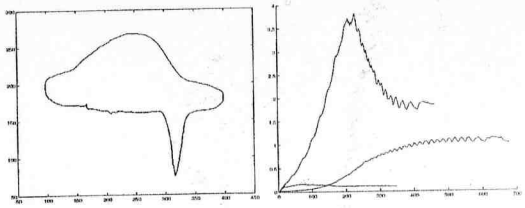

Figure 1.11 A shape and the scale normalized curvature response over scale for three of its points.

supported by neurophysiological observations that indicate the existence of a specialized brain module, used for face recognition, in humans' brains [26].

1.3.5 Salient Regions of Interest

Depending on the type and content of the image, the goal of observation and the observer's cognitive background, certain regions of the image may preattentively attract the viewer's attention. In a generalized phenomenological approach towards image content description, regions that contain attention attracting features are of interest, since they tend to indicate characteristic and discriminative image attributes. The definition of such regions is also time dependent, depending on the duration of observation and the adaptation of perception, and is an open issue in the fields of cognitive and vision sciences.

Rules that define early attentional attracting regions may include context sensitive and individually adapted criteria, however visual attention theories (and everyday experience) indicate that some of these behaviors are commonly shared among all humans. Experiments indicate that the exceptional values of features (like color, intensity, gradient magnitude, orientation, symmetry etc.), in combination with their spatial layout and size are strongly correlated with the existence of attention attractors in an image.

A psychological theory that models competition among different perceptual stimuli is the "Pandemonium Model" [27], which proposes the use of a system of perceptual analyzers, a series of demons who shriek with a loudness consonant with the degree and frequency of stimulation. The one that better stimulates perception dominates over others.

The use of regions of interest may be used to generate a more qualitative image description or to reduce its size by disregarding detail in non salient regions. In Section 1.4 a voting approach is proposed for the technical implementation of the combination of competitive behaviors.

1.4 Similarity Matching of Content Descriptions

This section discusses the use of image visual feature estimation in the context of formalizing an image description and the use of the latter in image similarity estimation. Images may be characterized as similar or not according to more than one criteria and a way of combining their influence, compatible with visual perception, is yet to be found. A voting approach is proposed as an experimental tool for this problem, as well as other information source integration topics, discussed in this study. Other open issues in this field are the dependencies among visual descriptors and the metrics used for the comparison of features and content.

The image description should represent all types of features detected, preserving all necessary information for content comparison. Thus, representations of the image's global, primitive and perceptual features as well as the definition of salient regions within the image, should be present in an image's visual content description. If available the strength of feature observation should be embodied in this description as well. Spatially defined features or distributions are to be attributed with their image location, while a topological graph is found sufficient for a qualitative representation of feature spatial layout [28]. It is proposed that all image features, except from global ones, should be also characterized with their scale of observation. This way not only the refinement of visual queries is made possible but also an abstract description of content will be rapidly accessible.

An open issue regarding image description is the dependencies of the significance of image features, with respect to human visual perception. Certain features, or combinations of them may be of outstanding importance for the detection of image content.

Similarity quantification is widely adopted in the automated retrieval of images as a way to classify retrieval results. Usually the use of some distance function that sums feature similarity in a quantitative way is employed. Following this type of approach, a big percentage in similar features could be enough to characterize two images as similar. However other cases of visual similarity exist, such as the case of the outstanding similarity of a pair of image features that could be responsible for the overall similarity impression. In another case two images that contain different features may be defined as similar due to an overall qualitative similarity, such as the feature arithmetic analogies. Besides the number of matches and the intensity of feature resemblance, spatial layout is also another aspect of visual similarity. It is realized that content similarity depends not only on feature matching and similarity, but also on "global" qualitative content similarity estimators.

Similarity measures that reflect visual similarity perception may be found in psychology literature and with respect to these, the widely adopted model

of Euclidean distance among features seems not suitable enough [29]. Other types of distances are proposed to more adequately represent human perception of visual similarity, while a dissimilarity component, that is not complementary to the similarity one, among features should be also taken into account [30]. In addition, similarity assessment is not always reflexive, meaning that which feature is the prototype and which one is the test subject is important in their comparison [31]. Concerning overall image similarity assessment, many systems adopt the weighted summation of different criteria. Although this is certainly an aspect of similarity perception, cognitive sciences indicate that there are more factors to this function. A review of several studies discussing ways of comparing multidimensional visual data sets, can be found at [32], indicating that similarity impression may not be quantitative, but ordinal.

The interaction and merging of different similarity modules can be studied through autonomous agent modeling of different similarity behaviors, casting their votes regarding the similarity two images. Each agent's vote is based on its specialized knowledge of image content and some similarity assessment method. Depending on image content and comparison task, different behaviors should be activated. An experimental platform featuring voting procedures using a variety of voting systems, was implemented for the study of such behaviors.

The use of a voting infrastructure may be used in the evaluation of different content similarity evaluation strategies, exploiting the dynamic formulation of the set of voters and thus yielding a system where a variety of experiments may take place, without technical effort. Also, the capability of dynamically tuning the similarity assessment method of a visual information system is found useful in applications where context is not specified and a variety of content matching behaviors are required (e.g. image search engines, robotics and other).

1.5 Discussion

Through this chapter ways of describing and analyzing image content were discussed. In this effort various aspects of visual content were taken into account and specific emphasis was given to the scale component of visual features. The understanding of visual perception is speculated to contribute to the appreciation of the results of image retrieval by content applications, by refining queries in a qualitative way. Clearly, many issues related with the design of content-based image browsing applications, remain unspecified, such as the subjectivity of the result due to variations in estimation of similarity among individuals, different retrieval goals, and lack of knowledge concerning the contribution of features in overall visual impression. Current trends in image database research propose the employment of user interaction in query formulation for the specification of such parameters, through visual languages [33].

It is argued that the specification of visual perception's fundamental components, if obtained, can be specialized, in order to serve image retrieval applications. In particular, the set of descriptional and matching processes may be directed towards specific image types or features and also enhanced with specific rules given image context and retrieval goal. In addition, the obtainment of more qualitative results may also be supported by the use of machine learning methods, allowing a system to satisfy individual user preferences, given feedback on similarity criteria and voting systems' effectiveness and suitability.

Concluding, the methodologies discussed in this study could be incorporated in the generic field of multimedia management and browsing. Visual content cues may be fused with other information sources such as text, video, and audio for the retrieval of multimedia documents.

Acknowledgments

This work was supported by EC Contract No. ERBFMRY-CT96-0049 (VIRGO `http://www.ics.forth.gr/virgo`) under the TMR Programme. The authors would like to acknowledge the significant contributions to this research of J. Sporring and E. Tzova, as well as the useful discussions with P. E. Trahanias and A. Argyros, members of the Computer Vision and Robotics Laboratory (CVRL) of ICS - FORTH.

References

[1] James J. Gibson. *The ecological approach to visual perception.* Houghton Mifflin, Boston, 1979.

[2] J. D. Mollon and P. G. Polden. Post-receptoral adaptation. *Vision Research*, 19:35–40, 1979.

[3] A. G. Goldstein. Judgments of visual velocity as a function of length of observation time. *Journal of Experimental Psychology*, 54:457–461, 1957.

[4] S. Sarkar and K.L. Boyer. Perceptual organization in computer vision: A review and a proposal for a classificatory structure. *IEEE Transactions on Systems, Man, and Cybernetics*, 23:382–399, 1993.

[5] T. Lindeberg. *Scale-Space Theory in Computer Vision.* The Kluwer International Series in Engineering and Computer Science. Kluwer Academic Publishers, Boston, USA, 1994.

[6] C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. 4th Alvey Vision Conference*, pages 147–152, 1988.

[7] R. Shapley and V. H. Perry. Cat and monkey retinal ganglion cells and their visual functional roles. *Trends in Neurosciences. Specias Issue: Information processing in the retina.*, 5(9):229–235, 1986.

[8] M. S. Livingstone and D. H. Hubel. Psychophysical evidence for separate channels for the perception of form, color, movement, and depth. *Journal of Neuroscience.*, 11(7):3416–3468, 1987.

[9] T. Lindeberg. Feature detection with automatic scale selection. Technical Report ISRN KTH/NA/P--96/18--SE, Dept. of Numerical Analysis and Computing Science, KTH, May 1996.

[10] X. Zabulis, J. Sporring, and S. C. Orphanoudakis. Scale summarized and focused browsing of primitive image content. In *Visual 2000, Lyon, France*, pages 269–278, 2000.

[11] De Valois R. and De Valois K. *Spatial Vision*. Oxford Science Publications, Oxford, 1988.

[12] C. Koch D.K. Lee, L. Itti and J. Braun. Attention activates winner-take-all competition among visual filters. *Nature Neuroscience*, 2(4):375–381, 1999.

[13] Todd R. Reed and J. M. Hans du Buf. A review of recent texture segmentaiton and feature extraction techniques. *CVGIP: Image Understanding*, 57(3):359–372, May 1993.

[14] Jan J. Koenderink and Andrea J. van Doorn. The structure of locally orderless images. *International Journal of Computer Vision*, 31(2/3):159–168, 1999.

[15] J. Puzicha and et al. Empirical evaluation of dissimilarity measures for color and texture. In *Proceedings of the International Conference on Computer Vision*, 1999.

[16] J. Shi and J. Malik. Normalized cuts and image segmentation. In *Proc. of IEEE CVPR Puerto Rico*, pages 731–737, June 1997.

[17] M. Wertheimer. *A sourcebook of Gestalt phychology*, chapter 1. Gestalt Theory, pages 1–11. The Humanities Press, New York, 1924.

[18] S. E. Palmer. Common region: A new principle of perceptual grouping. *Cognitive Psychology*, 24:436–447, 1992.

[19] S. E. Palmer and I. Rock. Rethinking perceptual organization: The role of uniform connectedness. *Psychonomic Bulletin and Review*, 1(1):29–55, 1994.

[20] A. J. Marcel. Conscious and unconscious perceptions: An approach to relations between phenomenal experience and perceptual processes. *Cognitive Psychology*, 15:197–237, 1983.

[21] D. G. Lowe. *Perceptual Organization and Visual Recognition*. Kluwer, Boston (MA), 1985.

[22] Remco C. Veltkamp and Michiel Hagedoorn. State-of-the-art in shape matching. In Michael Lew, editor, *Principles of Visual Information Retrieval*. Springer, 2001.

[23] F. Attneave. Some informational aspects of visual perception. *Psychological Review*, 61(3):183–193, 1954.

[24] E. Peterhans and R. von der Heydt. Mechanisms of contour perception in monkey visual cortex II, Contours bridging gaps. *Journal of Neuroscience*, 9(5):1749–1763, 1989.

[25] J. Sporring, X. Zabulis, P. E. Trahanias, and S. C. Orphanoudakis. Shape similarity by piecewise linear alignment. In *ACCV, Taipei, Taiwan*, pages 306–311, January 2000.

[26] N. Kanwisher, J. McDermott, and M. M. Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience*, 17(11):4302–4311, 1997.

[27] A. Selfridge. Pandemonium: a paradigm for learning. In *Symposium on the Mechanization of Thought Processes*, London: HM Stationery Office, 1958.

[28] J. R. Smith and S.-F. Chang. Integrated spatial and feature image query. *Multimedia System Journal*, 7(2):129–140, 1999.

[29] F. Attneave. Dimensions of similarity. *Psychological Review*, 63:516–556, 1950.

[30] M. D. Ennis and et al. A multidimensional stochastic theory of similarity. *Journal of Mathematical Psycology*, 32:449–465, 1988.

[31] E. Rosh. Cognitive reference points. *Cognitive Psychology*, 7:532–547, 1975.

[32] S. Santini and R. Jain. Similarity matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(9):871–883, 1999.

[33] S. Santini. Exploratory interfaces for visual information systems. In *Proceedings of Vision Interface '99*, Trois Rivieres, Quebec, CA, 1999.