

IAPR Workshop on Pattern Recognition in Information Systems.
Setubal, Portugal, July 6-8, 2001, pp. 34-49.

Image Retrieval via Isotropic and Anisotropic Mappings*

Qasim Iqbal and J. K. Aggarwal

Computer and Vision Research Center
Department of Electrical and Computer Engineering
The University of Texas at Austin
Austin, Texas 78712, USA.
{qasim,aggarwaljk}@mail.utexas.edu

Abstract. This paper presents an approach for content-based image retrieval via isotropic and anisotropic mappings. Isotropic mappings are defined to be mappings invariant to the action of the planar Euclidean group – invariant to the translation, rotation and reflection of image data, and hence, invariant to orientation and position. Anisotropic mappings, on the other hand, are defined to be those mappings that are correspondingly variant. Structure extraction (via a perceptual grouping process) and color histogram are shown to be representations of isotropic mappings. Texture analysis using a channel energy model comprised of even-symmetric Gabor filters is considered to be a representation of anisotropic mapping. Results of retrieval of outdoor images by query and by classification using a nearest neighbor classifier are presented.

1 Introduction

The interest in automatic analysis of images based upon their content has increased with recent developments in the World Wide Web (WWW), digital image collections, networking and multimedia. Active research in content-based image retrieval (CBIR) is geared towards the development of methodologies for analyzing, interpreting, cataloging and indexing image databases. In image analysis, the input and output are functions of \mathcal{R}^2 , and an appropriate notion of isotropy of computations is the Euclidean invariance: any rotation, translation or reflection of the input should produce an identical result under these transformations, thus achieving orientation and position invariance. These image transformations are generated by the action of the planar Euclidean group (the semi-direct product of the orthogonal group and the translation group). Using this notion of isotropy, we present an approach for content-based image retrieval via isotropic and anisotropic mappings.

* This work was supported in part by the Army Research Office under contracts DAAD19-00-1-0044, DAAG55-98-1-0230 and DAAD19-99-1-0012 (Johns Hopkins University subcontract agreement 8905-48168).

We define an *isotropic mapping* as a mapping that is *invariant* to the action of the Euclidean group – invariant to translation, rotation, and reflection of image data. Similarly, we define an *anisotropic mapping* as a mapping that is variant to the action of the Euclidean group. The Euclidean group is the group of isometries of \mathbb{R}^2 – mappings that preserve distances – and its action on the space of positions and directions $\mathbb{R}^2 \times \mathcal{S}^1$, where positions are represented using \mathbb{R}^2 and directions using the unit circle \mathcal{S}^1 , generates isometric geometrical objects. It has been argued that visual computations occur on $\mathbb{R}^2 \times \mathcal{S}^1$, rather than on just \mathbb{R}^2 [1]. The generation of isometries is important for developing a framework for isotropic mappings, as seen later. Isotropic mappings acting on perceptually salient image structures are useful in retrieval, as they illustrate the *similarity* of different structures in an image. On the other hand, anisotropic mappings indicate the *uniqueness* of certain attributes of different images.

Most of the previous work in image retrieval has focused on retrieval by image query [2–5]. However, retrieval by image classification has also gained attention [6–8]. In this paper we develop a methodology for retrieval of outdoor images using both image query and image classification by using a nearest neighbor classifier. Retrieval by image query refers to the retrieval of images similar to a given query image from an image database, whereas retrieval by classification refers to the classification of images into certain known classes for retrieval.

As seen in the next sections, perceptual grouping is a natural candidate for isotropic mappings, as are histograms of pixel color values. On the other hand, lower-level texture analysis via a Gabor filter bank (which possesses affinity for certain preferred directions) operating in a channel energy model is an effective candidate for anisotropic mappings.

1.1 Action of the Euclidean group – Action by translation, rotation, and reflection

It is well-known that the group of all isometries of \mathbb{R}^2 is the Euclidean group. To see this, let Γ be an isometry of \mathbb{R}^2 , and let $\mathbf{b} = \Gamma(0)$. Then $\varrho = \tau_{-\mathbf{b}}\Gamma$ is an isometry of \mathbb{R}^2 , satisfying $\varrho(0) = 0$. It can be shown that if $\varrho(0) = 0$, then ϱ is linear [9], and thus, $\Gamma = \tau_{-\mathbf{b}}^{-1}\varrho = \tau_{\mathbf{b}}\varrho$ is a product of a linear isometry and a translation. Further, it can also be shown that the linear isometries are represented by the orthogonal group $O(2, \mathbb{R})$ of 2×2 orthogonal matrices that represent reflections and rotations. Hence, the product of the translation group and the orthogonal group is the group of isometries of \mathbb{R}^2 (called Euclidean group $E(2)$). The normality of the translation group in $E(2)$ can be used to deduce that $E(2) \cong O(2, \mathbb{R}) \bowtie \mathbb{R}^2$, where \bowtie denotes semi direct product.

The rest of the paper is organized as follows: section 2 explains the perceptual grouping process to extract structure and the color histogram as representations of isotropic mapping, section 3 describes the texture analysis via a channel energy model as a representation of anisotropic mapping, section 4 outlines the integration of isotropic and anisotropic mappings, section 5 describes the results obtained, and finally, section 6 provides the conclusions.

2 Isotropic mapping

We have considered feature extraction from structural analysis of an image via the perceptual grouping process and color histogram as representations of isotropic mappings. The extraction of structure from an image is described first, and then its Euclidean isotropy is established, followed by the description of the color histogram process.

2.1 Perceptual grouping

The human visual system can detect many classes of patterns and statistically significant arrangements of image elements. Perceptual grouping refers to the human visual ability to extract significant image relations from lower-level primitive image features without any knowledge of the image content and group them to obtain meaningful higher-level structure. Research in perceptual grouping was started in 1920's by Gestalt psychologists, whose goal was to discover the underlying principle that would unify the various grouping phenomena of human perception. Gestalt psychologists observed the tendency of the human visual system to perceive *configurational wholes*, with rules that govern the uniformity of psychological grouping for perception and recognition, as opposed to recognition by analysis of discrete primitive image features. The hierarchical grouping principles proposed by Gestalt psychologists embodied such concepts as grouping by *proximity, similarity, continuation, closure, and symmetry* [10].

The grouping of low-level features provides a higher-level structure. These higher-level structures may be further combined to yield another level of higher-level structures. The process may be repeated until a meaningful semantic representation is achieved that may be used by a higher-level reasoning process. Certain scene structures will always produce images with discernable features regardless of viewpoint, while other scene structures virtually never do. This correlation between salience and invariance has suggested that the perceptual salience of viewpoint invariance is due to the leverage it provides for inferring geometric properties of objects and scenes. It has been noted that many of the perceptually salient image properties identified by the Gestalt psychologists such as collinearity, parallelism, and good continuation, are viewpoint invariant [11].

To discover and describe structure, the visual system uses a wide array of perceptual grouping mechanisms. These range from the relatively low-level mechanisms that underlie the simplest principles of grouping and segregation, to relatively high-level mechanisms in which complex learned associations guide the discovery of structure. Perceptual grouping generally results in highly compact representations of images, facilitating later processing, storage, and retrieval [12].

Many computer vision systems implicitly use some aspects of processing that can be directly related to the perceptual grouping processes of the human visual system [13]. Frequently, however, no claim is made about the pertinence or adequacy of the digital models as embodied by computer algorithms to the proper model of human visual perception [14]. Edge-linking and region-segmentation, which are used as structuring processes for object recognition, are seldom considered to be a part of an overall attempt to structure the image [13]. This

enigmatic situation arises because research and development in computer vision is often considered quite separate from research into the functioning of human vision. A fact that is generally ignored is that biological vision is currently the only measure of the incompleteness of the current stage of computer vision, and illustrates that the problem is still open to solution [10].

2.2 Structure extraction – Feature selection

We extract the following features hierarchically in an unconstrained environment, i.e., with no constraints on the viewing angle and depth, using the approach detailed in [8]: *line segments*, *longer linear lines*, *coterminations*, “*L*” *junctions*, “*U*” *junctions*, *parallel lines*, *parallel groups* and “*significant*” *parallel groups* (figure 1(a) - (f)). As an enhancement to that approach, we also extract closed figures comprised of *polygons* (figure 1(g)). Perceptual grouping rules of similarity, continuity, parallelism and closure are used to extract these features.

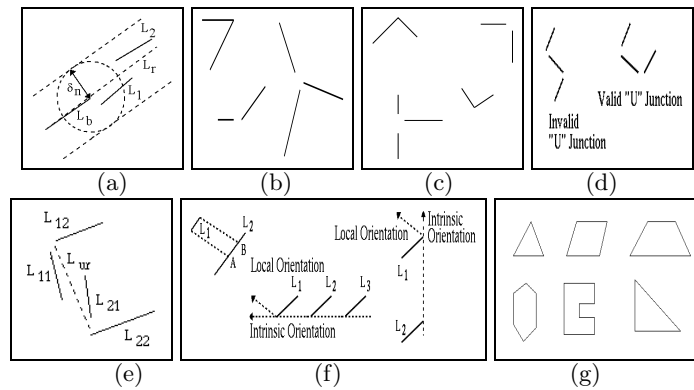


Fig. 1. Visualization of the groupings. (a) Longer linear line (b) Coterminations (c) “L” junctions (d) “U” junction (e) “U” junction (f) Parallel groups (g) Polygons

Burns edge detector [15] is used to detect straight line segments in an image. Longer linear lines are obtained by the extension of approximately collinear fragmented line segments that either overlap or are close to each other. The lines obtained are further pruned to eliminate lines that are very small or have low edge strength. All other features are extracted using the longer linear lines. A set of non-parallel lines terminating at a common point is called a *cotermination*. In practice, a small neighborhood is constructed around a point to allow the lines to terminate in a small common region for cotermination extraction.

The cotermination is an important relation. According to the *proximity* rule of perceptual grouping, the human visual system easily groups coterminous lines. In fact, it has been suggested that the major function of eye movements is to determine coterminous edges [16]. Cotermination is a *non-accidental* relationship and, hence, reflects significant structural information. Coterminations are

grouped to extract “L” junctions, and “L” junctions are grouped to get “U” junctions.

Parallel groups are obtained by constraining the amount of the overlap of the orthogonal projections of parallel lines onto each other and their projections along the x- and y-axis, while incorporating differences in the local and intrinsic orientation of the lines. In other words, we group the parallel lines that significantly overlap each other. “Significant” parallel groups are extracted by further constraining the search to only those parallel groups in which at least one member line is enclosed by an “L” or “U” junction, while accommodating the obliqueness of the viewing angle.

Polygons are closed figures formed by non-parallel lines. A polygon is a significant image relation. According to the *closure* rule of perceptual grouping, human vision tends to complete curves to form enclosed regions [10]. Extracting closed figures corresponds to this feature of human vision. Polygons are non-accidental image relationships, since the coterminations forming them are non-accidental. Hence, polygons represent significant structure in an image.

Elements of graph theory [17] are employed to extract polygons from an image using the cotermination graph. The underlying idea is to take advantage of the one-to-one correspondence between the closed figures comprised of line segments and the circuits in the graph. A set of fundamental circuits is searched and extracted.

Let $G = (V, E)$ be a cotermination graph, where V and E are the set of vertices and the set of edges of G , respectively. Let $\tilde{e}_{ij} \in E$ be an edge connecting vertices $\tilde{v}_i, \tilde{v}_j \in V$. The weight of \tilde{e}_{ij} is defined as $w(\tilde{e}_{ij}) = deg(\tilde{v}_i) + deg(\tilde{v}_j)$, where $deg(\cdot)$ is the *degree* of a vertex, that is, the number of edges incident with the vertex. The edge weights are collected by extracting the adjacency matrix of the graph. The connected components of the graph are found, and each sub-graph corresponding to each component is processed separately. The weight of a spanning tree is the sum of the weights of all the branches in the tree. We search for the maximal spanning tree, which may be found by slightly altering the minimal spanning tree algorithm to incorporate the vertices resulting in maximal-weight spanning tree [17]. The maximal spanning tree is employed to extract the fundamental circuits. Each fundamental circuit represents a closed figure in the image, where edges on this circuit correspond to line segments on the closed figure.

A polygon is defined to be that fundamental circuit extracted that meets the following requirements: (a) *the polygon is simple, i.e., the edges of the polygon do not intersect among themselves*, (b) *the polygon is relatively compact*, (c) *the polygon does not have many cavities*, and (d) *the number of edges on the polygon does not exceed a given threshold*.

The importance of perceptual grouping for typical instances of recognition can not be overemphasized. In the absence of the necessary information for perceptual grouping, it is difficult for humans to make an intelligent decision regarding the structure or recognition of an object. Experiments conducted with line drawings, in which most of the elements of significant collinearity, end point

proximity, parallelism and symmetry were removed, demonstrated the difficulty perceived by humans subjects in recognizing the objects [10]. With the addition of a few elements at key locations, the human subjects were able to perceive the line drawings with remarkable ease. When the elements were added at locations that did not lend themselves to meaningful perceptual groupings, then the response time for the perception of the line drawings was unaltered. The ability to influence recognition times by controlling the formation of perceptual grouping illustrates the search-based nature of this process, and it has been hypothesized that perceptual grouping can be a key element in search space and recognition time reduction.

2.3 Feature extraction

The extracted feature vector $\mathbf{X}_{\mathcal{S}} = (\tilde{\mathbf{x}}_{\mathcal{S}_1}, \dots, \tilde{\mathbf{x}}_{\mathcal{S}_d})^t$, where d is the dimensionality of the feature space, is expressed in the general form as:

$$\tilde{\mathbf{x}}_{\mathcal{S}_i} = \frac{\sum_j \chi_{\omega_{\mathcal{S}_i}}(l_j)}{\sum_k \chi_{\omega_{\Theta_l}}(l_k)} \quad (1)$$

where χ denotes the characteristic (indicator) function, l is a longer linear line, ω_{Θ_l} is the set of all longer linear lines, $\omega_{\mathcal{S}_i}$ is a higher-level structure extracted, and $\tilde{\mathbf{x}}_{\mathcal{S}_i} \in [0, 1]$ ($i \in [1, \dots, d]$), i.e., the feature space is represented by a unit hypercube.

For generating results for retrieval by both image query and image classification, we set $d = 3$, and $\omega_{\mathcal{S}_i}$ represents “L” junctions, “U” junctions, and “significant parallel groups and polygons” for $i \in \{1, 2, 3\}$, respectively, i.e., $\tilde{\mathbf{x}}_{\mathcal{S}_i}$ represents the corresponding normalized number of lines.

2.4 Euclidean isotropy of $\mathbf{X}_{\mathcal{S}}$

Let $\omega = \{\omega_i\}$ represent the collection of objects of interest present in an image, where each object ω_i , is a collection of $\omega_{i_k} = \{\mathbf{r}, \phi\} \in \mathfrak{R}^2 \times \mathcal{S}^1$, where $\mathbf{r} = \{x, y\} \in \mathfrak{R}^2$ is a coordinate pair, \mathcal{S}^1 is the unit circle, $\phi \in \mathcal{S}^1$ represents the orientation of ω_i . At the lowest level of vision ω_{i_k} 's, are represented by points on an edge segment ω_i (where ω_i is obtained by using Burns' edge detector [15]). At the next level of perceptual grouping, certain ω_i will be combined to generate a higher-level structure. Such a structure obtained from the grouping of ω_i 's may be called ω_j for consistency of notation, although it should be understood that ω_j now represents a structure at a higher-level than ω_i . (Refer to figure 2.)

We have defined a mapping $\psi : \omega \rightarrow \mathfrak{R}^d$, (where d is the dimensionality of the feature space), to be isotropic if it is invariant to the action of the Euclidean group:

$$\psi(E \cdot \omega) = \psi(\omega) \quad (2)$$

where E is the Euclidean group $E(2)$ – the semi-direct product of the group of linear isometries and the translation group – such that:

$$E \cdot \omega = \{E_j \cdot \omega_i \mid E_j \in E, \omega_i \in \omega\} \quad (3)$$

The extraction of the feature vector, \mathbf{X}_S , is represented by ψ . The action of the Euclidean group on ω_i transforms each $\omega_{i_k} \in \omega_i$ and is given as (refer to figure 3):

$$\begin{aligned}\tau_{\mathbf{b}} \cdot (\mathbf{r}, \phi) &= (\mathbf{r} + \mathbf{b}, \phi), \quad \mathbf{r}, \mathbf{b} \in \mathfrak{R}^2, \phi \in \mathcal{S}^1 \\ R_\theta \cdot (\mathbf{r}, \phi) &= (R_\theta \mathbf{r}, \phi + \theta), \quad \theta \in \mathcal{S}^1 \\ \kappa \cdot (\mathbf{r}, \phi) &= \acute{\kappa} R_{-2\theta} \cdot (\mathbf{r}, \phi) = (\acute{\kappa} R_{-2\theta} \mathbf{r}, -(\phi - 2\theta))\end{aligned}\tag{4}$$

where $\tau_{\mathbf{b}} \in T(2)$, ($\mathbf{b} \in \mathfrak{R}^2$), represents a member of the translation group of \mathfrak{R}^2 , $T(2)$, such that $\tau_{\mathbf{b}}(\mathbf{r}) = \mathbf{r} + \mathbf{b}$, $\mathbf{r} \in \mathfrak{R}^2$, $R_\theta \in O(2, \mathfrak{R})$ is a rotation by an angle θ , κ is a reflection along an axis in \mathfrak{R}^2 , and $\acute{\kappa}$ is the reflection along the x -axis, $(x, y) \rightarrow (x, -y)$. The action $\kappa \cdot (\mathbf{r}, \phi) = R_\theta \acute{\kappa} R_{-\theta} \cdot (\mathbf{r}, \phi) = \acute{\kappa} R_{-2\theta} \cdot (\mathbf{r}, \phi)$ (by using the identity $R_\theta \acute{\kappa} = \acute{\kappa} R_{-\theta}$), because reflection along an arbitrary axis is equivalent to rotation of \mathfrak{R}^2 by an angle $-\theta$ to align the axis of reflection with the original x -axis, followed by a reflection in the x -axis, and then rotation by an angle θ .

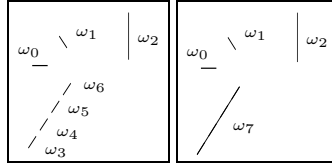


Fig. 2. $\omega_3, \omega_4, \omega_5$ and ω_6 combined to form ω_7 .

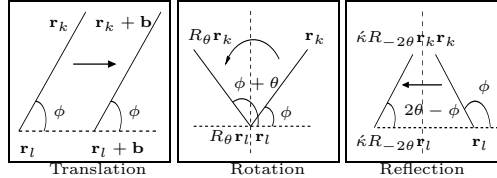


Fig. 3. Action of $E(2)$ on an edge segment, ω_i . \mathbf{r}_k and \mathbf{r}_l represent the end-points of ω_i .

Linear feature modeling: The premise of linear feature modeling is to extract rich descriptions of lower-level local image primitives and use these descriptions for subsequent grouping into higher-level features (linear line segments). The following illustrates the modeling of the perceptual grouping process described in [8] for the collection of edge segments ω_k 's, to form a longer linear line ω_j , (figure 1(a)). Let $\mathbf{r} = \{x, y\}$ denote the x - and y -coordinates of an end-point of an edge segment ω_k , and $\phi \in \mathcal{S}^1$ represents the orientation of the edge segment. We treat \mathbf{r} and ϕ as independent variables, so that all possible orientations for ω_k exist at each corresponding position \mathbf{r} . A certain collection \mathcal{C}_i of ω_k 's is collected,

which will be replaced by ω_j , that maximizes the energy λ_i given as:

$$\lambda_i^{(n)} = \lambda_i^{(n-1)} + \sum_{k \in \mathcal{K}, l \notin \mathcal{K}; \mathcal{K} = \{\bar{k}: \omega_{\bar{k}} \in \mathcal{C}_i\}} \xi_{kl}, \quad \lambda_i^{(0)} = 0 \quad (5)$$

where the superscript n is an iteration index, and (omitting the subscript i), the energy functional $\xi_{kl} : (\omega_b, \omega_k, \omega_l) \rightarrow \mathfrak{R}$ is expressed as:

$$\xi_{kl}(\omega_b | \omega_k, \omega_l) = \Lambda(q) \Lambda(st) \delta(\mathbf{r}_k - \mathbf{r}_l - s\mathbf{e}_{kl}) \delta(\phi_b - \phi_l) \quad (6)$$

where ω_b is a certain *base* edge segment in the collection that is used to determine that all other edge segments are parallel to it, Λ is a weighting function and q is the maximum length of the orthogonal distance of any point of ω_l from ω_b . In the above equation, \mathbf{r}_k and \mathbf{r}_l represent those end-points of two edge segments ω_k and ω_l respectively, (at the lower-level), that are closer to each other, and ϕ_b and ϕ_l are the orientations of ω_b and ω_l , respectively. In addition, δ is the Dirac delta function, \mathbf{e}_{kl} is a unit vector in the direction of $\mathbf{r}_k - \mathbf{r}_l$ and s is a distance parameter along an axis parallel to the direction of $\mathbf{r}_k - \mathbf{r}_l$. The Boolean parameter t is such that $t = 0$ if the length of the orthogonal projection of ω_l on ω_k is greater than zero, otherwise $t = 1$. In our system Λ is represented by a constant function (not equal to zero) with compact support. Specifically, we have selected the constant as 1 and the support is equal to 5 units (pixels). Equation 5 indicates the iterative nature of the grouping. At the start \mathcal{C}_i consists of only one segment ω_b . At the end of each iteration those ω_l 's for which ξ_{kl} is non-zero are put into \mathcal{C}_i . The grouping is started again and continued until there is no increase in λ_i . The higher-level longer linear line ω_j is then obtained by a weighted average of the lengths and orientations of all edge segments in \mathcal{C}_i [8].

The form of the energy functional expressed in equation 6 is similar to the one defined in [18], however, in their model \mathbf{r}_k represents the V1 image of the center of the receptive field of a neuron, and \mathbf{e}_{kl} represents the V1 image of the orientation preference of the neuron. Unlike their model, in our system \mathbf{e}_{kl} points in the direction of $\mathbf{r}_k - \mathbf{r}_l$ and incorporates the non-collinearity of two edge segments to an arbitrary extent (e.g., figure 1). (To further emphasize closer points, unequal weights, as opposed to constant weights in the support of Λ , can be achieved by replacing Λ with an appropriate weighting function, such as a Gaussian function.)

Euclidean invariance of ξ_{kl} : The form of the energy functional expressed in equation 6 has a well-defined symmetry: it is invariant under the action of $E(2)$; it is invariant under translations $\{\mathbf{r}, \phi\} \rightarrow \{\mathbf{r} + \mathbf{b}, \phi\}$, rotations $\{\mathbf{r}, \phi\} \rightarrow \{R_\theta \mathbf{r}, \phi + \theta\}$ and reflections $\{\mathbf{r}, \phi\} \rightarrow \{kR_{-2\theta} \mathbf{r}, -(\phi - 2\theta)\}$.

The argument q , s and t in equation 6 remain unchanged, because as shown in section 1.1 the action of $E(2)$ generates isometric objects, or it can be verified as following. The invariance of $s = \|\mathbf{r}_k - \mathbf{r}_l\|$ can be established as:

$$\begin{aligned} \|\tau_{\mathbf{b}} \rho \mathbf{r}_k - \tau_{\mathbf{b}} \rho \mathbf{r}_l\|^2 &= \langle \tau_{\mathbf{b}} \rho \mathbf{r}_k, \tau_{\mathbf{b}} \rho \mathbf{r}_k \rangle + \langle \tau_{\mathbf{b}} \rho \mathbf{r}_l, \tau_{\mathbf{b}} \rho \mathbf{r}_l \rangle \\ &\quad - 2 \langle \tau_{\mathbf{b}} \rho \mathbf{r}_k, \tau_{\mathbf{b}} \rho \mathbf{r}_l \rangle \\ &= \|\mathbf{r}_k\|^2 + \|\mathbf{r}_l\|^2 - 2 \langle \mathbf{r}_k, \mathbf{r}_l \rangle \\ &= \|\mathbf{r}_k - \mathbf{r}_l\|^2 = s^2 \end{aligned} \quad (7)$$

where \langle, \rangle denotes the dot product and ϱ is either a rotation or a reflection; since, $\langle \tau_{\mathbf{b}} \varrho \mathbf{r}_k, \tau_{\mathbf{b}} \varrho \mathbf{r}_l \rangle = \langle \mathbf{r}_k, \varrho^{-1} \tau_{\mathbf{b}}^{-1} \tau_{\mathbf{b}} \varrho \mathbf{r}_l \rangle = \langle \mathbf{r}_k, \mathbf{r}_l \rangle$, and similarly for the first and second terms in the first line of the above equation. Similarly, the invariance of q and t can also be established.

Translation invariance of equation 6 is evident because:

$$\begin{aligned} & \xi_{kl}(\tau_{\mathbf{b}} \cdot \omega_b \mid \tau_{\mathbf{b}} \cdot \omega_k, \tau_{\mathbf{b}} \cdot \omega_l) = \\ & \quad \Lambda(q) \Lambda(st) \delta((\mathbf{r}_k + \mathbf{b}) - (\mathbf{r}_l + \mathbf{b}) - s\mathbf{e}_{kl}) \delta(\phi_b - \phi_l) \\ & = \Lambda(q) \Lambda(st) \delta(\mathbf{r}_k - \mathbf{r}_l - s\mathbf{e}_{kl}) \delta(\phi_b - \phi_l) \\ & = \xi_{kl}(\omega_b \mid \omega_k, \omega_l) \end{aligned} \quad (8)$$

Invariance with respect to a rotation θ follows from:

$$\begin{aligned} & \xi_{kl}(R_{\theta} \cdot \omega_b \mid R_{\theta} \cdot \omega_k, R_{\theta} \cdot \omega_l) = \\ & = \Lambda(q) \Lambda(st) \delta(R_{\theta} \mathbf{r}_k - R_{\theta} \mathbf{r}_l - sR_{\theta} \mathbf{e}_{kl}) \delta((\phi_b + \theta) - (\phi_l + \theta)) \\ & = \Lambda(q) \Lambda(st) \delta(R_{\theta}(\mathbf{r}_k - \mathbf{r}_l - s\mathbf{e}_{kl})) \delta(\phi_b - \phi_l) \\ & = \Lambda(q) \Lambda(st) \delta(\mathbf{r}_k - \mathbf{r}_l - s\mathbf{e}_{kl}) \delta(\phi_b - \phi_l) \\ & = \xi_{kl}(\omega_b \mid \omega_k, \omega_l) \end{aligned} \quad (9)$$

and invariance under a reflection κ about the an axis holds since:

$$\begin{aligned} & \xi_{kl}(\kappa \cdot \omega_b \mid \kappa \cdot \omega_k, \kappa \cdot \omega_l) = \\ & = \Lambda(q) \Lambda(st) \delta(\acute{\kappa} R_{-2\theta} \mathbf{r}_k - \acute{\kappa} R_{-2\theta} \mathbf{r}_l - s\acute{\kappa} R_{-2\theta} \mathbf{e}_{kl}) \delta(-(\phi_b - 2\theta) + (\phi_l - 2\theta)) \\ & = \Lambda(q) \Lambda(st) \delta(\acute{\kappa} R_{-2\theta}(\mathbf{r}_k - \mathbf{r}_l - s\mathbf{e}_{kl})) \delta(-(\phi_b - \phi_l)) \\ & = \Lambda(q) \Lambda(st) \delta(\mathbf{r}_k - \mathbf{r}_l - s\mathbf{e}_{kl}) \delta(\phi_b - \phi_l) \\ & = \xi_{kl}(\omega_b \mid \omega_k, \omega_l) \end{aligned} \quad (10)$$

It must be noted that the energy functional given in equation 6 incorporates the Gestalt principles of proximity, collinearity, parallelism, and good continuation. Equation 6 is at the heart of the perceptual grouping process. Its Euclidean invariance, as shown above, means that equation 5 remains invariant, and the perceptual grouping process will produce the *same* groupings – longer linear lines. All higher-level structures are extracted using these longer linear lines.

Higher-level structures: The fundamental perceptual grouping proposed in [8] for higher-level structures can be modeled as the following. The proximity of two edge segments ω_k and ω_l can be modeled by the relation $\Lambda(s)\delta(\mathbf{r}_k - \mathbf{r}_l - s\mathbf{e}_{kl})$, whereas the variation in the orientations of ω_k and ω_l can be controlled by the relation $\tilde{A}(p) \delta(\phi_k - \phi_l - p)$, where the variable $p = \phi_k - \phi_l$, $p \in [0, 2\pi]$ and \tilde{A} is a constant function (not equal to zero) with compact support (similar to A). The length of overlap of lines is determined by orthogonal projection, and remains invariant because, as shown in section 1.1, the action of $E(2)$ generates isometric objects. Using an argument similar to the one shown above, it can be verified these relations are invariant under the action of $E(2)$. Hence, equation 1 also remains invariant, i.e., $\mathbf{X}_{\mathcal{S}}$ obtained by the mapping ψ is invariant after the action of $E(2)$ – invariant to orientation and position.

2.5 Color histogram

It can readily be seen that color histogram measures are invariant to both $O(2, \mathfrak{R})$ and $T(2)$, and hence, $E(2)$, because histogram measures are only dependent on summations of identical pixel values and do not incorporate orientation and position. The extraction of the normalized histogram $\mathbf{X}_{\mathcal{H}} \in \mathfrak{R}^{512}$ is used as a representation of an isotropic mapping.

A color space is *perceptually uniform* if a small perturbation to a component value is approximately equally perceptible across the range of that value. The *RGB* color space does not exhibit perceptual uniformity. However, the CIE *LAB* space [19], conceived in 1976, improves the perceptual uniformity of *RGB* space considerably. *LAB* color space is an approximately uniform color space that maps equally distinct color differences into approximately equal Euclidean distances in space. In this space, *L* defines lightness, *A* denotes red/green chrominance and *B* the yellow/blue chrominance. Presently, it is one of the most popular color spaces for color measurement.

Given an image $I_{RGB}(x, y)$ in *RGB* space we generate $I_{LAB}(x, y)$, where the pair (x, y) denotes the coordinates in an image I . A 512-dimensional feature vector $\mathbf{X}_{\mathcal{H}}$, representing the 512-bin normalized histogram, is extracted from the image $I_{LAB}(x, y)$ by uniformly quantizing the *LAB* space, i.e,

$$\mathbf{X}_{\mathcal{H}} = (\tilde{\mathbf{x}}_{\mathcal{H}_0}, \dots, \tilde{\mathbf{x}}_{\mathcal{H}_{511}})^t \quad (11)$$

where $\tilde{\mathbf{x}}_{\mathcal{H}_j}$ (where the index integer $j \in [0, 511]$) represents the normalized value of the j^{th} bin of the histogram such that $\sum_{j=0}^{511} \tilde{\mathbf{x}}_{\mathcal{H}_j} = 1$. This feature space represents a unit hypercube.

3 Anisotropic mapping

In most quantitative channel energy models of texture analysis, an image is processed by channel selective filters along certain fundamental stimulus dimensions such as spatial frequency and orientation. These channels generally contain a non-linearity, such as full-wave rectification, so that they signal the local contrast energy within the bandpass of the channel.

Texture analysis via a channel energy model employing a Gabor filter bank is considered a representation of anisotropic mapping. The representation is accomplished by the extraction of the feature vector $\mathbf{X}_{\mathcal{T}} \in \mathfrak{R}^{48}$, which measures the fractional energy in various spatial channels after treating the input image with the Gabor filter bank. That can readily be verified from the fact that the translation of an image $I(\mathbf{r}) \rightarrow I(\tau_{\mathbf{b}}(\mathbf{r}))$ transforms the Fourier transform of the image $\mathcal{I}(\nu) \rightarrow \mathcal{I}(\nu) e^{j2\pi\langle \mathbf{b}, \nu \rangle}$, where $\mathbf{b} \in \mathfrak{R}^2$ and rotation of $I(\mathbf{r}) \rightarrow I(R_{\theta}(\mathbf{r}))$, where $\mathbf{r} = \{x, y\}$ – the space domain coordinates, transforms the Fourier transform $\mathcal{I}(\nu) \rightarrow \mathcal{I}(R_{\theta}\nu)$, where $\nu = \{u, v\}$ are the Fourier domain co-ordinates. Similar result holds for reflection. Hence, texture analysis is not invariant after the action of $E(2)$ on an image.

The channel energy model employed is based upon multiresolution analysis that is characterized by both orientation and scale. The LAB space is used for multiresolution texture analysis by measuring the fractional energies in the lightness and the two chrominance channels mentioned in the last section. Given an image I , the convoluted sequence $\{I * f_{mn}\}$ defines the multiresolution image texture characteristics, where f_{mn} denotes a base texture extraction function f at scale m and orientation n , and $\|f_{mn}\|^2$ (filter energy) is held constant.

Gabor filters have been used to represent f_{mn} . The impulse response of an even-symmetric 2-dimensional Gabor filter is expressed as:

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right)} \cos(2\pi u_0 x) \quad (12)$$

where $f(x, y)$ represents the response at spatial locations x and y , u_0 is the frequency of a sinusoidal plane wave along the x-axis (i.e., the 0° orientation), and σ_x and σ_y are the spreads of the Gaussian envelope along the x- and y-axis, respectively.

A set of self-similar Gabor filters is obtained by appropriate rotations and scalings of $f(x, y)$ through the generating function:

$$\hat{f}_{mn}(x, y) = k^{-m} f(k^{-m}\hat{x}, k^{-m}\hat{y}), \quad k \geq 1 \quad (13)$$

where m and n are integers, $\hat{f}_{mn}(x, y)$ is the rotated and scaled version of the original filter, k is the scale factor, $n = 0, 1, \dots, N-1$ is the current orientation index, N is the total number of orientations, $m = 0, 1, \dots, M-1$ is the current scale index, M is the total number of scales, and \hat{x} and \hat{y} are the rotated coordinates: $\hat{x} = x \cos \theta + y \sin \theta$, $\hat{y} = -x \sin \theta + y \cos \theta$ where $\theta = \frac{n\pi}{N}$ is the orientation. The scale factor k^{-m} ensures that the filter energy is independent of m . In order to eliminate the sensitivity of the filters to absolute intensity values, we set $F_{mn}(0, 0) = 0$. A total of 16 Gabor filters are selected, with 4 filters in equi-angular orientations at 4 different scales, i.e., $N = 4$, and $M = 4$, starting at 0° orientation. Parameters σ_u , σ_v and k are calculated as described in [20].

Channels L , A and B are treated with the Gabor filter bank described by equation 13. The 48-dimensional feature vector $\mathbf{X}_{\mathcal{T}}$ is constructed using the fractional energies in each of the 16 spatial-frequency channels in the L , A and B channels, i.e.,

$$\mathbf{X}_{\mathcal{T}} = (\tilde{\mathbf{x}}_{\mathcal{T}L_{00}}, \dots, \tilde{\mathbf{x}}_{\mathcal{T}L_{33}}; \tilde{\mathbf{x}}_{\mathcal{T}A_{00}}, \dots, \tilde{\mathbf{x}}_{\mathcal{T}A_{33}}; \tilde{\mathbf{x}}_{\mathcal{T}B_{00}}, \dots, \tilde{\mathbf{x}}_{\mathcal{T}B_{33}})^t \quad (14)$$

where $\tilde{\mathbf{x}}_{\mathcal{T}L_{mn}}$, $\tilde{\mathbf{x}}_{\mathcal{T}A_{mn}}$ and $\tilde{\mathbf{x}}_{\mathcal{T}B_{mn}}$ represent the fractional energy at the output of the filter in the n^{th} orientation and the m^{th} scale, for L , A and B channels, respectively. The fractional energy $\tilde{\mathbf{x}}_{\mathcal{T}L_{mn}}$ is given as:

$$\tilde{\mathbf{x}}_{\mathcal{T}L_{mn}} = \frac{\sum_{y=0}^{W_y-1} \sum_{x=0}^{W_x-1} \hat{L}_{mn}^2(x, y)}{\sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \sum_{y=0}^{W_y-1} \sum_{x=0}^{W_x-1} \hat{L}_{mn}^2(x, y)} \quad (15)$$

where \hat{L}_{mn} is the L channel treated with filter \hat{f}_{mn} , W_x is the width of the image, W_y is the height, and $\sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \tilde{\mathbf{x}}_{\mathcal{T}L_{mn}} = 1$. Due to the fact that

the Fourier transform is a linear isometry (for space and spatial-frequency domains), equation 15 represents energy calculation in the space domain. Similar expressions hold for $\tilde{\mathbf{x}}_{TA_{mn}}$ and $\tilde{\mathbf{x}}_{TB_{mn}}$. This feature space is also represented by a unit hypercube.

4 Integration Framework

A 2-level framework is employed for integrating lower-level and higher-level vision features. Given the isotropic feature vectors \mathbf{X}_S and \mathbf{X}_H and anisotropic feature \mathbf{X}_T extracted from a query image, and \mathbf{X}_{S_j} , \mathbf{X}_{H_j} and \mathbf{X}_{T_j} extracted from the j^{th} image in the database, the first level of the framework maps the feature vectors to a discriminant value within each of the 3 categories, structure, histogram and texture. The respective mappings $\Phi_S: \mathbb{R}^{N_S} \rightarrow \mathbb{R}$, $\Phi_H: \mathbb{R}^{N_H} \rightarrow \mathbb{R}$ and $\Phi_T: \mathbb{R}^{N_T} \rightarrow \mathbb{R}$, where $N_S = 3$, $N_H = 512$ and $N_T = 48$, are selected as ℓ_2 norms: $\Phi_S(\mathbf{X}_{S_j}, \mathbf{X}_S) = \|\mathbf{X}_{S_j} - \mathbf{X}_S\|$, $\Phi_H(\mathbf{X}_{H_j}, \mathbf{X}_H) = \|\mathbf{X}_{H_j} - \mathbf{X}_H\|$ and $\Phi_T(\mathbf{X}_{T_j}, \mathbf{X}_T) = \|\mathbf{X}_{T_j} - \mathbf{X}_T\|$.

At the second level a supra discriminant is generated by utilizing the mapping $\Psi_{SHT}: \mathbb{R}^3 \times \mathbb{R}^3 \rightarrow \mathbb{R}$ that is given as:

$$\Psi_{SHT}(\mathbf{X}_{S_j}, \mathbf{X}_{H_j}, \mathbf{X}_{T_j}, \mathbf{X}_S, \mathbf{X}_H, \mathbf{X}_T) = \mathcal{W}^t \cdot \Phi_{SHT}(\mathbf{X}_{S_j}, \mathbf{X}_{H_j}, \mathbf{X}_{T_j}, \mathbf{X}_S, \mathbf{X}_H, \mathbf{X}_T) \quad (16)$$

where $\mathcal{W}^t = (w_1, w_2, w_3)^t$ is a weight vector such that $\sum_{i=1}^3 w_i = 1$, $\Psi_{SHT} \in [0, 1]$ and $\Phi_{SHT}: \mathbb{R}^3 \times \mathbb{R}^3 \rightarrow \mathbb{R}^3$, such that $\Phi_{SHT} \in [0, 1] \times [0, 1] \times [0, 1]$, is given as:

$$\Phi_{SHT}(\mathbf{X}_{S_j}, \mathbf{X}_{H_j}, \mathbf{X}_{T_j}, \mathbf{X}_S, \mathbf{X}_H, \mathbf{X}_T) = (\hat{\Phi}_S(\mathbf{X}_{S_j}, \mathbf{X}_S), \hat{\Phi}_H(\mathbf{X}_{H_j}, \mathbf{X}_H), \hat{\Phi}_T(\mathbf{X}_{T_j}, \mathbf{X}_T))^t \quad (17)$$

where

$$\begin{aligned} \hat{\Phi}_S(\mathbf{X}_{S_j}, \mathbf{X}_S) &= \frac{\Phi_S(\mathbf{X}_{S_j}, \mathbf{X}_S)}{\max_j \Phi_S(\mathbf{X}_{S_j}, \mathbf{X}_S)} \\ \hat{\Phi}_H(\mathbf{X}_{H_j}, \mathbf{X}_H) &= \frac{\Phi_H(\mathbf{X}_{H_j}, \mathbf{X}_H)}{\max_j \Phi_H(\mathbf{X}_{H_j}, \mathbf{X}_H)} \\ \hat{\Phi}_T(\mathbf{X}_{T_j}, \mathbf{X}_T) &= \frac{\Phi_T(\mathbf{X}_{T_j}, \mathbf{X}_T)}{\max_j \Phi_T(\mathbf{X}_{T_j}, \mathbf{X}_T)} \end{aligned} \quad (18)$$

The above normalizations ensure that $\hat{\Phi}_S \in [0, 1]$, $\hat{\Phi}_H \in [0, 1]$ and $\hat{\Phi}_T \in [0, 1]$ for properly constructing Φ_{SHT} . The index \hat{i} of the image most similar to a given query image is then given as:

$$\hat{i} = \arg \min_i \Psi_{SHT}(\mathbf{X}_{S_i}, \mathbf{X}_{H_i}, \mathbf{X}_{T_i}, \mathbf{X}_S, \mathbf{X}_H, \mathbf{X}_T) \quad (19)$$

The next most similar image is retrieved by removing the i^{th} image from the database and utilizing equation 19 again. The process is repeated for retrieving any number of images most similar to a given query image.

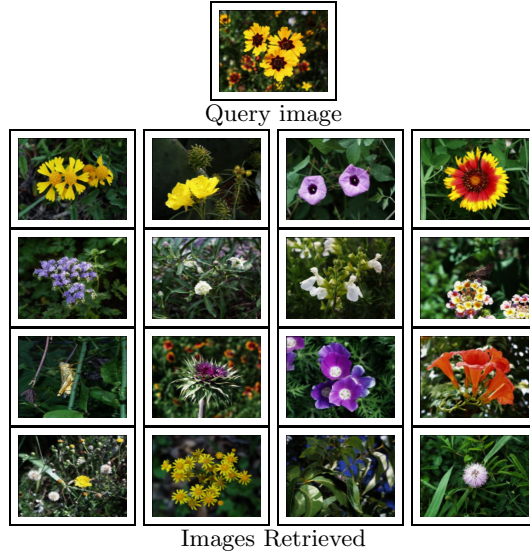


Fig. 4. Retrieval by image query (Databases #1 & #2): Flower, leaves and grass.

The above integration framework has the following advantages over a simple concatenation of vectors $\mathbf{X}_{\mathcal{S}_j}$, $\mathbf{X}_{\mathcal{H}_j}$ and $\mathbf{X}_{\mathcal{T}_j}$. First, the different lengths of these three vectors preclude the proper construction of a concatenated vector that is equally sensitive to all of its components. The 3-dimensional vector output by $\Phi_{\mathcal{SHT}}$ is equally sensitive to all of its three 1-dimensional components. Second, the size of the corresponding weight vector for the concatenated vector will be large, making the selection of proper weights difficult and unfeasible. Third, in our proposed integration, weights are assigned at the *module level*, i.e., structure, histogram and texture, whereas weights in a concatenated vector are assigned at the vector component level without particular regard to the modular structure of the system. The weight vector plays an important role in controlling the content of images retrieved. For a given image query, different weights can be assigned to structure, histogram and texture according to user specification to control the images retrieved.

5 Results obtained

Our image databases consists of 2660 24-bit color images. Database #1 consists of 2139 images of size adjusted to 1024×1024 acquired from two CDs obtained from The Visual Delights Inc. (<http://www.visualdelights.net>). Database #2 consists of 521 images of size adjusted to 512×512 acquired from the ground level using a Sony Digital Mavica camera. The weight vector is chosen as $\mathcal{W} = (1/3, 1/3, 1/3)^t$.



Fig. 5. Retrieval by image query (Databases #1 & #2): A building facade.

Figures 4 - 5 display examples of image retrieval by query from both databases #1 and #2 utilizing equation 19. First 16 images retrieved are shown in both figures. Tables 1 - 4 display results for retrieval by image classification obtained using a nearest neighbor classifier and using Φ_{SHT} 's as patterns (equation 17). The image space is partitioned into three classes, *Structure*, *Non-structure* and *Intermediate*, based upon the measure of structure present in an image. Each class is represented by 10 training samples.

Total	Training	Effective	Correct	RR
T		D	C	(C/D)
521	30	491	363	73.93%

Table 1. Retrieval by image classification (Database #2): Overall retrieval rate. T = Total # of images, D = Effective # of images, C = Correct and RR = Retrieval rate.

Table 1 shows the overall retrieval rate. Table 2 displays class-conditional retrieval performance measured in terms of *recall* and *precision*. Recall is defined as the fraction of the total number of images that are correctly retrieved for a particular class. Precision is defined as the fraction of images retrieved for a particular class that are actually correct. The retrieval statistics are shown fully in the confusion matrix shown in Table 3. Table 4 shows the distribution of images that *actually* belong to a particular class within the “best matches” for that class, in intervals of 100 images, and the corresponding *efficiency* of the system. Efficiency is defined as the ratio of the number of images that actually belong to a particular class in the block of closest best matches, to the size of

the block, where the block size is equal to the number of images corresponding to that class. The best matches were obtained by sorting images in ascending order based upon their distances from the training samples of each class.

Class	T	R	C	Recall (C/T)	Precision (C/R)
Structure	255	222	195	76.47%	87.84%
Non-structure	140	144	114	81.43%	79.17%
Intermediate	96	125	54	56.25%	43.20%

Table 2. Retrieval by image classification (Database #2): Recall and precision. (Database # 2.) T = Total, R = Retrieved, C = Correct.

Class	Structure	Non-structure	Intermediate
Structure	195	14	46
Non-structure	1	114	25
Intermediate	26	16	54

Table 3. Retrieval by image classification (Database #2): Confusion matrix. Entries presented in rows, e.g., 195 Structure class images classified as Structure, 14 as Non-structure, and 46 as Intermediate.

Class	1-100	101-200	201-300	301-400	401-500	501-521	T	Q	Eff.=Q/T
Structure	87	70	57	28	13	-	255	190	74.51%
Non-structure	79	43	11	7	-	-	140	108	77.14%
Intermediate	38	32	13	11	2	-	96	36	37.50%

Table 4. Retrieval by image classification (Database # 2): Distribution of images *actually* belonging to a particular class in the “best matches” for that class, in intervals of 100 images, and the efficiency of the system. T = Total # of images belonging to a certain class, Q = # of images that actually belong to a certain class in the first T best matches for that class, and Eff. = Efficiency.

6 Conclusions

This paper has presented an approach for content-based image retrieval via isotropic and anisotropic mappings. Isotropic mappings were defined to be mappings invariant to the action of the planar Euclidean group – invariant to the translation, rotation and reflection of image data, and hence, invariant to orientation and position. Anisotropic mappings, on the other hand were defined to be those mappings that are correspondingly invariant. Structure extraction (via a perceptual grouping process) and color histogram were shown to be representations of isotropic mappings. Texture analysis using a channel energy model comprised of even-symmetric Gabor filters was considered to be a representation of anisotropic mapping. Results of retrieval of outdoor images by query and by classification using a nearest neighbor classifier were presented. Results obtained show the efficacy of combining structure, histogram and texture for retrieval.

References

1. J. Zweck and L.R. Williams, "Euclidean group invariant computation of stochastic completion fields using shiftable-twistable functions," in *6th European Conference on Computer Vision (ECCV '00), Dublin, Ireland, 2000*, pp. 100–116.
2. Michael J. Swain and Dana H. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.
3. Jonathan Ashley, Ron Barber, Myron Flickner, James Hafner, Denis Lee, Wayne Niblack, and Dragutin Petkovic, "Automatic and semi-automatic methods for image annotation and retrieval in QBIC," in *Proc. SPIE: Storage and Retrieval for Image and Video Databases III*, Feb. 1995, vol. 2420, pp. 24–35.
4. A. P. Pentland, R. Picard, and S. Sclaroff, "Photobook: Content-based manipulation of image databases," *Int. Journal of Computer Vision*, vol. 18, no. 3, pp. 233–254, 1996.
5. J. R. Smith and S.-F. Chang, "VisualSEEK: a fully automated content-based image query system," in *ACM Multimedia*, Nov. 1996, pp. 87–98.
6. Monika M. Gorkani and Rosalind Picard, "Texture orientation for sorting photos "at a glance",", in *Proc. IEEE International Conference on Pattern Recognition*, 1994, vol. 1, pp. 459–464.
7. A. Vailaya, A. K. Jain, and H.-J. Zhang, "On image classification: City images vs. landscapes," *Pattern Recognition*, vol. 31, pp. 1921–1936, December 1998.
8. Qasim Iqbal and J. K. Aggarwal, "Applying perceptual grouping to content-based image retrieval: Building images," in *IEEE International Conference on Computer Vision and Pattern Recognition, Fort Collins, Colorado, June 1999*, vol. 1, pp. 42–48.
9. Frederick M. Goodman, *Algebra, Abstract and Concrete*, Prentice Hall, Inc., 1998.
10. David G. Lowe, *Perceptual organization and visual recognition*, Kluwer Academic publishers, 1985.
11. D.W. Jacobs, "What makes viewpoint invariant properties perceptually salient?: A computational perspective," in *Perceptual Organization for Artificial Vision Systems*, K.L. Boyer, Ed., pp. 121–138. Kluwer, 2000.
12. Wilson S. Geisler and Boaz J. Super, "Perceptual organization of two-dimensional patterns," *Psychological Review*, vol. 107, no. 4, pp. 677–708, 2000.
13. James D. McCafferty, *Human and Machine Vision, Computing Perceptual Organization*, Ellis Horwood Limited, 1990.
14. Martin D. Levine, *Vision in Man and Machine*, McGraw Hill, 1985.
15. J. Brian Burns, Allen R. Hanson, and Edward M. Riseman, "Extracting straight lines," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 4, pp. 425–455, 1986.
16. T. O. Binford, "Inferring surfaces from images," *Artificial Intelligence*, vol. 17, pp. 205–244, 1981.
17. Alan Gibbons, *Algorithmic Graph Theory*, Cambridge University Press, 1985.
18. P.C. Bressloff, J.D. Cowan, M. Golubitsky, P.J. Thomas, and M.C. Wiener, "What geometric visual hallucinations tell us about the visual cortex.," *Neural Computation. To appear.*
19. Gunter Wyszecki, *Color science : Concepts and methods, quantitative data and formulae, 2nd edition*, New York, Wiley, 1982.
20. W. Y. Ma and B. S. Manjunath, "Texture features and learning similarity," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 1996, pp. 425–430.