Invariant Representations for Content Based Image Retrieval

Jan-Mark Geusebroek, Gertjan J. Burghouts, Jan C. van Gemert, and Arnold W. M. Smeulders

Intelligent Sensory Information Systems, Informatics Institute, Faculty of Science, University of Amsterdam, Kruislaan 403, 1098 SJ Amsterdam, The Netherlands;
mark@science.uva.nl

Abstract.

1 Introduction

The human visual system is a signal processing system with extremely high performance. Information about our natural environment is extracted from the enormous quantity of visual information. A general vision system is considered to be able to process visual sensory information to interpret the physical surrounding. The human visual system is a very well adapted example of a general vision system. Computer vision in this respect is disappointing, as there are only a few areas where it can compete with the capabilities of human perception. Two reasons are apparent. First, a major bottleneck is the difference between the physical scene and the observation of the scene. The observation is affected by accidental imaging circumstance, the viewpoint to the scene, the aspects of interaction between light and material, the limited resolution of the sensory system, and several other factors. The semantic interpretation of the scene depends on the intrinsic properties of the viewed object, i.e., the 'observables'. Therefore, observation effects may complicate the interpretation of a scene such as recognition of objects. Only recently computer vision aims at solving this bottleneck by measuring intrinsic properties that are invariant to the irrelevant variation [5, 34, 9, 10, 13].

The second bottleneck comes from knowledge and expectations of what we see. Human perception actively assigns knowledge and anticipation to the observed scene, using semantic information for reasoning on a higher level than from purely visual evidence can be achieved. The difference between the interpretation as can be derived from the data, in respect to the additional semantics added by knowledge, is referred to as the semantic gap [44]. A better knowledge of the assignment of semantic information to visual data is rudimentary for the sustained development in content based image retrieval.

We start with the prerequisite that any general sensory system is adapted to the outside world it is processing, specifically to the statistical structure of the input signals (Barlow 1961). For one, the statistics of the sensory input is

dominated by physical laws of image formation and by the reflection from materials [13]. These physical laws are basically domain independent, as they cover the universally applicable laws of light reflectance from materials. They generate scene specific imaging aspects, which are undesirable for the recognition of objects in the scene and their labelling with general categories. The scene accidental conditions have to be removed first by an invariant description. This is a requirement for general vision systems as is posed in current philosophy as the proper way to describe all conscious perception (Nozick 2002). In addition, statistics of the sensory input may be shaped by the structure of our environment. For example, parts of an image which deviate from the common structure around us are likely to contain perceptually salient details.

Image formation involves physical laws that generate scene specific observation effects. The scene is illuminated by a light spectrum and reflects parts of this spectrum. Reflectance is influenced by the geometry of the scene [5, 13]. The contents of a scene and its physical and geometrical properties cause its observation to be structured [37] and to contain geometrically similar forms [40]. Consequently, the visual stimulus is redundant [7]. Processing of visual information turns out to fit the statistics of the physical image formation process well [7, 3, 4, 41]. Clearly human vision's sensory system is adapted to the more or less stable environment it processes: the visual system and the environment are each other's duals [9].

For instance, color constancy is the ability of the visual system to correct for reflectance deviations caused by a difference in illumination, which is important to determine the reflectance of an object [10]. Likewise, lightness invariance is of interest to human vision, because a rather general class of linear as well as nonlinear transformations of the retinal irradiance distribution has little effect on perception, or at least on recognition [5]. We believe that the visual system counteracts these scene accidental effects by representing the scene in many diverse invariant syntactical representations that simplify the visual interpretation task. Invariant representations require knowledge of the physical variables involved in the stimulus formation.

To counteract the accidental aspects of the scene, a general vision system will represent the image in many diverse invariant representations. We consider the transformation of sensory responses to invariants an important and inevitable information reduction stage in general vision system. The resulting directly observable quantities are believed to be an essential part of human perception (Foster and Nascimento 1994; Koenderink 1984). Computer vision has partly solved the problem of invariant transformations. Koenderink (1984) has made a significant step forward by his work on the structure and scaling behavior of receptive fields. He has been followed by many, among others the subsequent categorization of geometrical invariants by Florack (1991) and Van Gool (1995). The level of invariance may well be the level at which it starts to make sense to abstract from the very local detailing [21]. In our opinion, invariant representations minimize the computational burden arising from the observation task of a visual scene.

In this chapter, we model visual invariants for the purpose of content based image retrieval. The chapter is organized as follows. We identify the observables and associated physical variables important for human vision (Sect. 2). The receptive fields are formalized in Sect. 3. Invariants are extracted from the receptive field measurements, and the relation between invariants and observables is described in Sect. 4.

2 Analysis of Visual Observables and Unwanted Variations

In order to analyze what might be observed from a scene in general without too much a priori knowledge about the scene or the objects in the scene, we follow the light. When we ignore the influence of the medium as well as inter-reflections, the main degrees of freedom are the source, the object, its surroundings and the camera. It is modelled as free parameters in the Kubelka-Munk reflection model [13] as follows.

2.1 Follow the light

The light starts at the source, where there is freedom to have 1 or more *sources*. Each source has a *direction* relative to the scene, a spectral composition and intensity. Likewise, the essential free parameters of the camera are the spectral sensitivity, the gain, its *direction* relative to the scene and its *distance*. As the spectral content of the source and the spectral sensitivity of the camera have practically the same effect, we take them together under the name *spectral content*. The same holds for the source intensity and the gain of the camera under the name of *intensity*.

For the object, the free parameters can be grouped in the *cover*, ranging from glosse to matte objects. Glosse produces specular reflections. The *albedo* describes the true color of the object, and the *texture* describes the spatial layout of the albedo patterns at its surface. The *touch* of an object describes the 3D-nature of the surface as it introduces a large variability in the perception of the object. In this simplification, the final group of object parameters is grouped under *form*.

For a scene, the one group of parameters left is the stage setting where the objects are placed in the scene in a certain depth order with respect to the light, causing *shadows*, and with respect to the view, causing *occlusion* and *clutter*, preventing the object to be delineated amidst similarly appearing objects.

In listing the main groups of accidental, unknown causes of variation in a general scene, we ignore light-emitting, mirroring, fluid, and transparent objects.

2.2 The outer scene

Given all the sources of variations in a scene, Tab. 2.2 gives an overview of what can be observed about a general scene [42].

free parameters	as seen in scene	directly observables	see
source direction	shadows	one source	[15]
	$\sim$ directions	direction source	
	$\sim$ locations	direction source	
	cast $\sim$	depth order	
source extent	specularity		[2]
spectral content	spectral composition	color source	[8]
intensity	contrast composition	contrast	
camera direction	projection	affine distortion	
camera distance	size composition	depth	[30]
stage setting	occlusion	depth order to view	
	clutter	–	
objects	specularities	color source	
	$\sim$ locations	number sources	
	$\sim$ size	shape source	
	self-shadow	one source	
	shading	direction source	
	$\sim$ maxima	number of sources	[2]

Table 1. Inspired by [5], page 122. What is directly observable from the outer scene? Free parameters of the scene and the single features that can be observed in general without a priori knowledge about the specifics of the scene. Methods listed in the references generally assume sufficiently rich scenes.

free parameters	as seen on object	constraint on free parameters	directly observables	see
cover albedo				
gloss	specularity	–	cover type	[14]
	∼ locations	–	facing	
	apparent color	–	object color	[13, 14]
matte	apparent color	white source	object color	[13, 14]
		color source	constancy	[13, 14]
texture				
	albedos	–		[46]
	∼ layout	–		[18]
touch				
	meso-highlights	gloss	roughness	
	meso-shadow	one source	roughness	
	meso-shading	matte	meso-shape	[26]
		one source		
form				
matte	macro-shading	one source	shape	
	discontinuity		direction folds	[39]

Table 2. Directly observables in the inner scene.

From the table, it is clear that many instances of knowledge about the scene parameters are far from complete. And, we treat the causes as independent factors, ignoring any inter-reflections among them. Especially for closely packed, transparent, mirroring or poly-limbed objects this may not be a valid assumption, but we have to start somewhere.

2.3 The inner scene

Table 2.3 provides a list of free parameters of the object and what can be done to find them [42].

3 Visual Observation and Physical Quantities

A general vision system may be considered as a remote sensing device, able to extract directly observables from light measurements. Hence, vision can be viewed as the process of deriving invariants from physical quantities as coded in the energy distribution falling onto the eye. Note that this "coding" is non-trivial; it involves the projection of a three-dimensional world described by infinite physical entities onto a two-dimensional retina only able to capture spatial, spectral, and temporal information. The resulting measurements have a direct or indirect correlate with directly observables. Every measurement is implemented on the

retina by an integration over area, wavelength, and time, the only variation imposed by a task-tuned sensitivity curve to combine these quantities. The shape of these sensitivity curves essentially determine which information is emphasized. Since a limited amount of orhogonal physical quantities can be derived from the visual data, these sensitivities can be catagorized. The categorization yields typical receptive field structures, each related to one physical quantity which is measured when probing the visual stimulus with the receptive field. In what follows, we derive relevant image measurements for obtaining correlates of the directly observables.

3.1 Generic Requirements for Receptive Fields

Before relating the observables to measurements of different receptive fields, we specify their generic requirements.

We have no prior knowledge of what is where within the visual stimulus. Hence, no positions are preferred or prioritized when measuring a physical variable. Consequently, measurements are performed by a linear operator [20, 9] that integrates the energy E of a physical variable x (e.g., space, time, wavelength) over its domain with a receptive field G. Note that a linear integrator tends to be relatively insensitive to corruption of the visual data by noise.

Definition 1 (Receptive Field Measurement). *The measurement* $\hat{E} : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ *of the physical variable* x *with a receptive field* $G : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ *becomes:*

$$\hat{E}(x) \equiv \int E(x')\, G(x' - x)\, dx = E * G(x),$$

where $(*)$ *denotes convolution.*

The receptive fields cannot be infinitesimal. The free scale parameter is also a natural requirement for uncommitted measurements, in this case due to the lack of a priori knowledge of the perceived size [20]. For instance, when we turn to the spatial domain, we have no knowledge of the size of a perceived object: the spatial scale must be a free parameter. A receptive field probing in more than one spatial dimension is tuned to a specific orientation, that is, anisotropic, when scales in the different dimensions are not equal. A lack of knowledge considering orientation therefore imposes the constraint of probing at an abundant range of orientations [9]. Without loss of generality we only consider isotropic receptive fields.

At larger scales no spurious detail should be created by the receptive field measurements, nor should information in the visual stimulus be enhanced without prior knowledge of the information. Respecting such causality in the scale domain singles out the Gaussian family of receptive fields [20].

Definition 2 (Receptive Fields). *The receptive field* $G : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ *probing at position* x *and located at* $x_0 \in \mathbb{R}$ *and scale* σ_x *is defined as:*

$$G^{x_0, \sigma_x}(x) = \frac{1}{\sqrt{2\pi}\, \sigma_x}\, e^{-\frac{(x - x_0)^2}{2\, \sigma_x^2}}.$$

The Gaussian family receptive fields G_{x^i} are then given by:

$$\left\{G_{x^i}^{x_0,\sigma_x}\right\}_{i\geq 0} \equiv H_i(x)\, G^{x_0,\sigma_x},$$

with G_{x^i} the i-th derivative of G with respect to x and H_i the i-th order Hermite polynomial. This family is complete, i.e., fit to represent the physical variable to any accuracy [23].

Including these receptive fields in the front-end is advantageous, because a physical variable can locally be approximated by a Taylor expansion up to order n of derivative receptive field measurements:

$$\hat{E}(x) \approx \sum_{j=0}^{n} \frac{1}{j!} \left(x\, \nabla_x\right)^j \hat{E}(x),$$

where $\nabla_x{}^j\, \hat{E} = \hat{E}_{x^j} = E * G_{x^j}^{x_0,\sigma_x}$.

3.2 Taxonomy of Receptive Fields

The visual stimulus is analyzed in the spectral, spatial, and frequency domain. Probing these domains is necessary to gather information about the observables. Cover reflectance is measured in the spectral domain. Local geometry is apparent in local reflectance changes and determines object shape. Size and distance are spatial properties. Regularity involves a certain frequency tuning. Simultaneous measurements are performed by multi-dimensional receptive fields probing different variables. Due to the separability of the Gaussian [20], we start off by considering the different receptive fields probing a single variable. Simultaneous probing of spatial, spectral, and frequency information is given in the last few paragraphs.

Wavelength Spectrum The wavelength spectrum λ is probed at position λ_0:

$$G_{\lambda^i}^{\sigma_\lambda,\lambda_0}(\lambda) \quad i \in \{0,1,2\}, \tag{1}$$

where the spectral receptive fields measure at scale σ_λ. Knowledge of the photoreceptor's sensitivity curves as imposed by daylight gives $\lambda_0 = 520$nm. Due to smooth reflectance of common surfaces we obtain a fairly large scale: $\sigma_\lambda = 55$nm. Fixing for these parameters gives us the tuned spectral receptive fields. Reflectance information is obtained from the derivatives with respect to λ. The zeroth order derivative G_{λ^0} measures the intensity of the spectral distribution. The first G_{λ^1} and second order derivative G_{λ^2} compare yellow and blue regions of the spectrum, and middle (green) and two outer (magenta) regions of the spectrum, respectively [12]. The spectral receptive fields show resemblance to the Hering basis [16] and are approximately colorimetric with human vision, that is, the sensitivity curves can be approximated by linear transformations of the CIE XYZ sensitivities [12].

Local Geometry Local geometry depends on the differential spatial structure [24]. The local geometry receptive fields probe the spatial structure (x, y) at (x_0, y_0):

$$G_{x^i y^j}^{\sigma_{xy}, x_0, y_0}(x). \tag{2}$$

For instance, first order measurements with the receptive fields $G_{x_i y^j}$ yield the gradient, i.e., a direction and a magnitude (for a framework of directionally polarized measurements we refer to [11]). Second order measurements yield the pair of receptive fields expressed in gauge coordinates (v, w): (G_{vv}, G_w), which enable the visual system to determine the curvature $-\frac{G_{vv}}{G_w}$. For derivations of higher order local geometry we refer to [24].

Since no knowledge of object sizes is involved, we relate the size of perceived objects or surface elements to the appropriate scale of observation of local geometry, $\hat{I}_{\boldsymbol{x}^m}$. For the selection of scale we turn to the scale at which the local geometry receptive fields give maximum responses, that is, the scale over which spatial variation is maximized [31]. However, because at larger scales and higher orders of differentiation m responses inherently decrease, these have to be normalized: $\hat{I}_{\boldsymbol{x}^m}^{norm} = \sigma_{\boldsymbol{x}}^{m\frac{\gamma}{4}} \hat{I}_{\boldsymbol{x}^m}$. The size now relates to the scale $\sigma_{\boldsymbol{x}}$ that maximizes $\hat{I}_{\boldsymbol{x}^m}^{norm}$. For instance, we get the normalized second order local geometry measurement that reflects a blob-like shape: $\hat{I}_{\boldsymbol{x}^2}^{norm} = \sigma_{\boldsymbol{x}}^{\frac{\gamma}{2}} (\hat{I}_{xx} + \hat{I}_{yy})$ [31], for which we take $\gamma = 1$ [35]. We assumed isotropy, that is, $\sigma_x = \sigma_y$, but scales can also be analyzed for the anisotropic cases, in which elongated geometry is regarded.

Spatial Frequency For a receptive field tuned to a frequency we turn to the spatial frequency domain. Probing the spatial frequency (u, v) domain at position (u_0, v_0) and scale σ_{uv} with a Gaussian receptive field $G^{\sigma_{uv}, u_0, v_0}(u, v)$ gives in the spatial domain (x, y) the Gabor spatial frequency receptive field [6]:

$$\tilde{G}^{\sigma_{xy}, x_0, y_0; u_0, v_0}(x, y) = G^{\sigma_{xy}, x_0, y_0}(x, y)\, e^{2\pi i \left(\begin{smallmatrix} u_0 \\ v_0 \end{smallmatrix} \cdot \begin{smallmatrix} x \\ y \end{smallmatrix} \right)}, \quad i^2 = -1, \tag{3}$$

Here, $\sqrt{u_0^2 + v_0^2}$ is the radial center frequency and $\tan^{-1}(\frac{v_0}{u_0})$ the orientation. Note that σ_{xy} is inversely proportional to σ_{uv} (see for a survey [6]).

Spatiospectral Introduction of a spatial extent in the spectral receptive fields yields an expansion at wavelength λ_0 and position (x_0, y_0). The measurements of a spatiospectral energy distribution has a spatial as well as a spectral scale, respectively, σ_{xy} and σ_λ. Probing an energy density volume in a 3-dimensional spatiospectral space at (x, y, λ) requires a spatiospectral receptive field, or color opponent receptive field [12]:

$$G_{\lambda^i x^j y^k}^{\sigma_\lambda, \lambda_0; \sigma_{xy}, x_0, y_0} = G_{\lambda^i}^{\sigma_\lambda, \lambda_0} * G_{x^j y^k}^{\sigma_{xy}, x_0, y_0}. \tag{4}$$

Spatiospectral Frequency Probing the wavelength spectrum in the spatial frequency domain results in the spatial domain in a spatiospectral frequency receptive field [17]:

$$\tilde{G}_{\lambda^i}^{\sigma_{xy},x_0,y_0;u_0,v_0;\sigma_\lambda,\lambda_0} = \tilde{G}^{\sigma_{xy},x_0,y_0;u_0,v_0} * G_{\lambda^i}^{\sigma_\lambda,\lambda_0}. \tag{5}$$

4 Invariance

We have argued that not all of the measured visual stimulus is considered to be relevant, in that measurements *correlate* with directly observables. This is our motivation for identifying invariants that describe the observables, i.e., invariant to accidental observation effects but maintaining discriminative power within the relevant information about the important object properties. Invariants involve knowledge about the physical variables involved and how the irrelevant parameters may be eliminated.

In Sect. 2 we discussed the main degrees of freedom in the variation of receptive field measurements:

- Illumination: direction relative to the scene, spectrum and intensity;
- Scene objects: size, 2- and 3-dimensional shape, surface, cover, distance, motion;
- Scene setting: object order which causes occlusion, inter-reflections, and clutter;
- Inherent measurement properties: spectral sensitivity, intensity gain, viewing direction and distance to the scene.

The illuminant, scene setting and inherent measurement properties may cause general transformations. A property f of the object t of the group of objects T is invariant under the group of transformations W if and only if f_t remains the same regardless the state of condition W: $t_1 \overset{W}{\sim} t_2 \Rightarrow f_{t_1} = f_{t_2}$ [43]. More specifically, object placement causes translational and rotational variation. Scene setting causes accidental effects such as shadow and shading. The illumination spectrum may deviate causing reflectance variation. Viewing distance causes scale variation, and direction, finally, causes affine variation.

4.1 Photometric Invariance

Important for the observation of object reflectance, is the visual system's ability to correct for deviations caused by a difference in illumination spectrum as well as intensity. Furthermore, irrelevant variations may be caused by the observer's viewpoint, object surface orientation, and the direction of the illuminant. The visual system should be enabled to discriminate in a natural scene the shadow, shade and highlight edges from object edges [12].

Modelling the physical process of the formation of the wavelength spectrum stimulus provides insight into the effect of different parameters on object reflectance. Finding invariant properties of the spectral measurements relies on a

reflectance model and the required discriminative power of the measurements performed.

Reflectance Model The formation of the wavelength spectrum stimulus is modelled by means of the Kubelka-Munk theory [27, 19]. We consider the Kubelka-Munk model as a general model for the wavelength spectrum stimulus formation. The reflected spectrum in the viewing direction is given by: $E(\lambda, \boldsymbol{x}) = e(\lambda, \boldsymbol{x})(1 - \rho_f(\boldsymbol{x}))^2 R(\lambda, \boldsymbol{x}) + e(\lambda, \boldsymbol{x})\rho_f(\boldsymbol{x})$, where $\boldsymbol{x}$ denotes the position at the imaging plane and λ the wavelength. Further, $e(\lambda, \boldsymbol{x})$ denotes the illumination spectrum and $\rho_f(\boldsymbol{x})$ the Fresnel reflectance at $\boldsymbol{x}$. The object reflectance is denoted by $R(\lambda, \boldsymbol{x})$.

Required Discrimination and Invariants We consider 'white' or arbitrary illumination wavelength spectrum as an accidental observation circumstance. Note that natural objects mostly have a matte cover type. Different required discrimination abilities result in suited invariant expressions (for derivations we refer to [12]).

The spatiospectral receptive fields are considered the general probes for these invariants. Spatiospectral measurements are denoted by $\hat{E}_{\lambda^i x^j y^k}$.

Edges invariant to shadows With white illumination and no intensity variations, the reflectance model reduces to: $E(\lambda, x) = i\, R(\lambda, x)$, of which the invariant set

$$\mathcal{W}_{\lambda^m x^n} = \left\{ \frac{\hat{E}_{\lambda^m x^n}}{\hat{E}} \right\}_{m \geq 0,\, n \geq 1} \tag{6}$$

can be derived. We denote the intensity measurement $\hat{E} \equiv \hat{E_{\lambda^0 x^0 y_0}}$.

$\mathcal{W}$ determines object reflectance invariant to shadows. $\mathcal{W}_{\lambda w}$ and $\mathcal{W}_{\lambda\lambda w}$ are the color gradient magnitudes of the first and second order spectral derivative, respectively, representing the blue-yellow and green-red color transitions.

Edges invariant to shadows, shading and highlights With white illumination and intensity variations, the reflectance model reduces to: $E(\lambda, x) = i(x) \left\{ (1 - \rho_f(x))^2 R(\lambda, x) + \rho_f(x) \right\}$, of which the invariant set

$$\mathcal{H}_{\lambda^m x^n} = \frac{\partial^{m+n}}{\partial \lambda^m \partial x^n} \left\{ \arctan(\frac{\hat{E}_\lambda}{\hat{E}_{\lambda\lambda}}) \right\}_{m,\, n \geq 0} \tag{7}$$

can be derived (arctan guarantees numerical stability).

$\mathcal{H}$ determines object reflectance invariant to shadows, shading and highlights.

Edges invariant to shading With colored illumination and intensity variations, the reflectance model reduces to: $E(\lambda, x) = e(\lambda)\, i(x)\, R(\lambda, x)$, of which the invariant set

$$\mathcal{N}_{\lambda^m x^n} = \frac{\partial^{m+n-2}}{\partial \lambda^{m-1} \partial x^{n-1}} \left\{ \frac{\hat{E}_{\lambda x} \hat{E} - \hat{E}_{\lambda} \hat{E}_x}{\hat{E}^2} \right\}_{m,\, n \geq 1} \tag{8}$$

can be derived.

$\mathcal{N}$ determines object reflectance invariant to shading and a change of illumination spectrum over time (note that the time parameter not explicitly modelled). Illumination spectrum changes over space are not common in natural scenes. $\mathcal{N}_{\lambda w}$ and $\mathcal{N}_{\lambda\lambda w}$ detect material edges.

The object cover type can be detected by the photometric invariant gradients: high responses of $\mathcal{W}_w$ and $\mathcal{N}_w$ indicate highlights caused by object cover specularity.

The intensity measurement $\hat{E}$ is invariant under the group of general intensity transformations ι [9].

An example of the set of transformations ι is a logarithmic rescaling of the intensity domain. Logarithmic scaling obtains a uniform sampling in the intensity domain and is due to a lack of preference for a scaling in measurement time, and, consequently, in intensity [25]. For an arbitrary intensity unit E_0 the rescaling becomes [25]:

$$\mathcal{I}_0 = \log(\frac{\hat{E}}{E_0}), \tag{9}$$

with $\mathcal{I}_0$ the intensity measurement invariant to shadows and shadings. The logarithmic rescaling allows for a linear operation to correct for 'gamma transformations' (i.e., intensity transformations of the form $i' = \left(\frac{i}{i_0}\right)^{\gamma}$). Gamma transformations are caused by unevenly illuminated scenes or object reflectance gradations [25].

4.2 Geometrical Invariance

Requiring invariance under the intensity transformation group ι amounts to considering equivalence classes of locally defined structures that share a common local iso-intensity, or isophote structure. Due to the remaining isophote structure the relevant ι-invariant local geometrical properties correspond to geometrically invariant local properties of isophotes [9]. For gauge coordinates (v, w), a complete and irreducible set of geometrical invariants is given by [9]:

$$\mathcal{I} = \{\hat{E}_{\lambda^0 v^m w^n}\}_{m,n \geq 0, m \neq 1}. \tag{10}$$

The 2-dimensional Case In the set $\mathcal{I}$ we find the for human vision important geometrical descriptions of intensity distribution. Local geometry describes object shape, and includes edges, corners, curves, etc. Well known instances of the set $\mathcal{I}$ are the gradient $w = \hat{E}_w$ and the curvature $\kappa = -\frac{\hat{E}_{vv}}{\hat{E}_w}$. Corners of or within objects may be detected by high isophote curvature and intensity gradient, whereas junctions (e.g., occlusion of objects) may be derived from those points where a contour ends or emerges, thus where the curvature changes much in the direction of the normal.

Note that we do not consider affine geometrical invariants. Affine invariants may counteract the measurement of corners and junctions under the accidental viewing direction. However, measurement of such constellations of edges is very dependent on the contrast with a background and is therefore dependent on accidental observations circumstances anyway. Consequently, there is no point in invariantly measuring corners and junctions and inherently and unnecessarily losing information.

The 3-dimensional Case General 3-dimensional shape descriptors are the principal curvatures κ_1, κ_2, which, together with their derivatives to the curve arc length a, form a complete and irreducible set of differential geometric invariants of curves [22]. From this set curvedness (i.e., flat, concave, convex) $c = \frac{\kappa_1+\kappa_2}{\kappa_1-\kappa_2}$ and shape index (e.g., spherical, saddle, cylindrical) $s = \frac{1}{2}\sqrt{{\kappa_1}^2 + {\kappa_2}^2}$ can be derived.

The 3-dimensional shape of an object is apparent in the shading of the object cover.

Shading. Shading is informative of 3-dimensional shape. Shading can be obtained by the measured intensity gradients. We relate the gradual changes within the 3-dimensional shape of the object to the changes of tilt and slant of the object surface normal.

4.3 Spatial Frequency Invariance

Regularities within a surface or cover are present under different viewing and illumination directions, which cause unwanted intensity deviations and rotational variation.

Luminance intensity ($\hat{E}$) normalized summed measurements of spatial/spatiospectral receptive fields ($\hat{S}$) yields:

$$s = \frac{\hat{S}}{\hat{E}}. \tag{11}$$

The normalized measurement $\int$ is invariant to shadows and shading [17].

Measuring with spatial/spatiotemporal frequency receptive fields at an abundant range of locations (u_0, v_0) and scales σ_{uv} in the frequency domain implies measuring at an abundant range of orientations $(\tan^{-1}(\frac{v_0}{u_0}))$ in the spatial domain. We consider simultaneous measurements at such a range of orientations to be rotational invariant.

Surface or cover regularities may become apparent within a repeated arrangement of motifs (different from a stochastic process that organizes motifs [33, 28]). Meaningful motifs are organized by translation, rotation, and reflection [32]. Any motif arrangement that is constructed from translation, rotation, and reflection of the motif conforms to one of 17 groups ($\mathcal{S}_i, 1 \leq i \leq 17$) that can be generated by the Cartesian transformations [32].

4.4 Observables and Invariants

Table 3 summarizes observables obtained from the invariants that in turn are associated with the receptive fields measurements. An asterisk ($\star$) denotes invariance that is inherently implied by the receptive field measurement.

5 An Image Retrieval System Based on Local Invariants

A Taylor series yields a point operator to describe local shape. Hence, the Taylor coefficients as measured by the Gaussian derivative receptive fields provide a basis for local image features, characterizing the neighborhood around a pixel. Every pixel may be characterized, creating a redundant representation of the image. The overlap at the boundaries between the patches introduces spatial correlation between local structures. When the spatial ordering is lost, as in a jigsaw puzzle, this correlation makes it possible to reconstruct the original image. Alternatively, the coherence in the Taylor representation allows histogram matching of coefficients while maintaining global similarity.

For image retrieval, discretization of the receptive field measurements is advantageous, such that histograms can be constructed. Labeling each element in a discrete partition of the color N-jet creates prototypical shapes representing the local color structure of an image. Each label then says something about the behavior of the image at a certain pixel. Hence, similar image patches are described with the same label. Local structure with similar color and similar curvature are then grouped. Discretizing the local color N-jet yields a vocabulary that can be used to describe local spatial color structure.

For our example retrieval system, a multidimensional histogram is used for the discretization process. The dimensions of the histogram are equal to the number of derivatives in the color N-jet. Each spatial-spectral derivative is partitioned in equally sized bins each with their own bin number. The concatenation of bin numbers in which a single pixel is classified make up a typical local shape structure. The partitioning of the color N-jet is a form of weak image segmentation, as similar image structures are grouped. The labeling of the color N-jet segments the image into its primary shape primitives. It groups low level image features to describe the local behavior of the image and operates at a low semantic level. The grouping of similar structures for various invariants is visualized in Fig. 1.

An example of the proposed retrieval scheme is shown in Fig. 2. The proposed method eliminates the problems with image segmentation and blocks of pixels.

Physical variable	*Receptive field*	*Observable*	*Extracted information about observable*	*Irrelevant parameter(s)*	*Invariant(s)*
Wavelength	$G^{\sigma_\lambda,\lambda_0;\sigma_{xy},x_0,y_0}_{\lambda^i x^j y^k}$ Geusebroek [12]	Cover reflectance	Reflectance	Shadow, shading	$\mathcal{W}$
				$\sim$ and highlights	$\mathcal{H}$
				$\sim$ and illumination spectrum	$\mathcal{N}$
		Cover type	Highlights		$\mathcal{W}_{\lambda w}, \mathcal{H}_{\lambda w}, \mathcal{N}_{\lambda w}$ Geusebroek [12]
Local geometry	$G^{\sigma_{xy},x_0,y_0}_{x^i y^j}$ Koenderink [24]	Object shape (2-dimensional)	Edges, curvature,	Intensity transformation, object orientation	w, κ
			corners, junctions, ...		... Koenderink [20, 24, 22], Florack [9]
		,, (3-dimensional)	Principal curvatures	,,	c, s Koenderink [22]
			Shading	,,	w Lee [29]
Spatial frequency	$\tilde{G}^{\sigma_{xy},x_0,y_0;u_0,v_0}$ Bovik [6]	Surface/cover regularity	Frequency	Shadow, shading, intensity transformation, object orientation	s Hoang [17]
	$\tilde{G}^{\sigma_{xy},x_0,y_0;u_0,v_0;\sigma_\lambda,\lambda_0}_{\lambda^i}$ Hoang [17]				
		,,	Motif arrangement	Position and orientation of motifs	$\mathcal{S}_i$ Liu [32]
Size	G^{σ_{xy},x_0,y_0} Lindeberg [31]	Object size	Edges, ...	Intensity transformation, object orientation	$\star$
		Surface/cover coarseness	,,	,,	$\star$
Distance	$\tilde{G}^{\sigma_{xy},x_0,y_0;u_0,v_0}$ Sanger [38]	Object distance	Relative distance (near)	Intensity transformation, object orientation	$\star$
	G^{σ_{xy},x_0,y_0} Pentland [36]		Distance (far)	Object size	z
Time Dynamics	$G^{\sigma_t,t_0;\sigma_\lambda,\lambda_0;\sigma_{xy},x_0,y_0}_{t^i \lambda^0 x^j y^k}$ Adelson [1]	Object motion	Relative motion of edges	Object distance, orientation, intensity transformation	$\boldsymbol{v}_z$
	$G^{\sigma_t,\sigma_{t'},t_0;u_0;\sigma_\lambda,lambda_0;\sigma_{xy},x_0,y_0}_{\lambda^i x^j y^k}$	$\sim$ periodicity	Spatiospectral-temporal frequency	Intensity transformation, object orientation	$\star$
	$\tilde{G}^{\sigma_t,\sigma_{t'},t_0;u0_0;\sigma_{xy},x_0,y_0;u1_0,v1_0;\sigma_\lambda,\lambda_0}_{\lambda^i}$				

Table 3. Summary of the different physical variables, measured by the receptive fields that constitute the selectivities of the visual system. The visual measurements aim at obtaining information about observables that relate to physical object properties. Irrelevant parameters in the measurements of observables are eliminated by invariants, i.e., combinations of particular measurements.

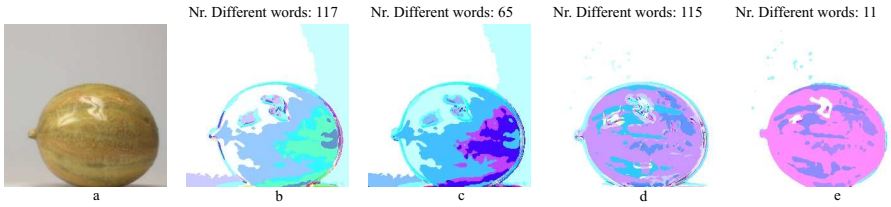

Fig. 1. The effect of using invariance on the vocabulary. Each word is represented with a different color. Identical colors in a single image represent identical words. The size of the vocabulary is displayed above each labeled image. The labeling is done with the second order color jet using 17 bins for each value at a scale of $\sigma = 2$ pixels. (a) The original image. (b) Labeling the image using no invariance. (c) Labeling the image with rotational invariance. Note the contour of the object is labeled everywhere with the same word and the vocabulary size decreases. (d) Labeling the image with rotational and luminance invariance $\mathcal{W}$. Note the disappearance of intensity changes. The vocabulary size increases, as shadows are labeled incorrectly. (e) Labeling the image with rotational and shadow invariance $\mathcal{C}$. Note the significant reduction of the words in the vocabulary and the disappearance of the shadow of the object. For details see [45].

The discrete color jet operates at a low semantic level, yet incorporates image content by the use of invariant properties.

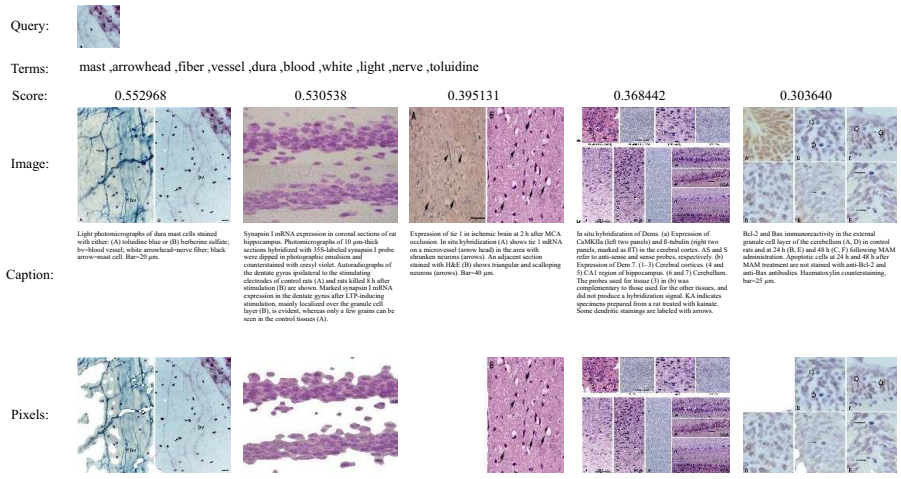

Fig. 2. Example of image searching in the combined images and text space. The query image is part of the first retrieved image (upper right corner). The 5 most relevant images are shown (upper row), together with the most relevant visual structures per image (lower row). Non-relevant pixels are left white. For details see [45].

6 Conclusion

To counteract the accidental aspects of the scene, any content based image retrieval system has to represent the image in many diverse invariant representations. We consider the transformation of sensory responses to invariants an important and inevitable information reduction stage in general vision system. The resulting directly observable quantities are believed to be an essential part of human perception (Foster and Nascimento 1994; Koenderink 1984).

We modeled visual invariants for the purpose of content based image retrieval. Evidently, we considered the observables present in the visual stimulus. A way to conceive the receptive fields as meaningful visual observation units adding up to the front-end is to characterize them syntactically according to their measurement properties, i.e., a constellation of selectivities along the observable dimensions. Invariants represent the observables and consequently provide important cues for the capture, extraction and interpretation of visual information.

In our view, the physical and statistical constrains on the sensory input determines the construction of content based image retrieval systems. The simplification of the sensory input by invariant representation advances towards better retrieval performance. Local features provide robustness to object occlusion and background changes. Invariance includes a low-level of semantic knowlegde, hence achieves a rudimentary level of visual cognition. Rather than aiming for one complete geometrical representation of the visual field, cognition based image retrieval may be based on weak description of the important features in the scene, as long as mutual correspondence between observation and objects in the world is maintained.

References

1. E. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America*, 2(2):284–299, 1985.
2. H.J. Andersen and E. Granum. Classifying the illumination condition from two light sources by color histogram assessment. *J. Opt. Soc. Am. A*, 17(4):667– 676, 2000.
3. H. B. Barlow. The knowledge used in vision and where it comes from. *Philosophical Transactions of the Royal Society of London B*, 352:1141–1147, 1997.
4. H. B. Barlow. Redundancy reduction revisited. *Network: computation in neural systems*, 12:241–253, 2001.
5. M. Baxandall. *Shadows and Enlightment*. Yale University Press, London, 1995.
6. A. C. Bovik, M. Clark, and W. S. Geisler. Multichannel texture analysis using localized spatial filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:55–73, January 1990.
7. D. J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of Optical Society of America*, 4(12):2379–2394, 1987.
8. G. D. Finlayson. Color in perspective. *IEEE Trans. Pattern Anal. Machine Intell.*, 18(10):1034–1038, 1996.

9. L. M. J. Florack. *Image Structure*, volume 10 of *Computational Imaging and Vision Series*. Kluwer Academic Publishers, Dordrecht, 1997.
10. D. H. Foster and S. M. C. Nascimento. Relational colour constancy from invariant cone-excitation ratios. *Proceedings of the Royal Society of London*, Series B(257):115–121, 1994.
11. W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(9):891–906, 1991.
12. J. M. Geusebroek, R. van den Boomgaard, A. W. M. Smeulders, and H. Geerts. Color invariance. *IEEE Pattern Analysis and Machine Intelligence*, 23(12):1338–1350, 2001.
13. J. M. Geusebroek, R. van den Boomgaard, A. W. M. Smeulders, and H. Geerts. Color invariance. *IEEE Trans. Pattern Anal. Machine Intell.*, 23(12):1338–1350, 2001.
14. Th. Gevers and A. W. M. Smeulders. Pictoseek: Combining color and shape invariant features for image retrieval. *IEEE Trans. Image Processing*, 9(1):102–119, 2000.
15. Th. Gevers and H. M. G. Stokman. Classifying color transitions into shadow-geometry, illumination highlight or material edges. In *ICIP*, volume 1, pages 521–525. IEEE Computer Society, 2000.
16. E. Hering. *Outlines of a Theory of the Light Sense*. Harvard University Press, Cambridge, 1964.
17. M. A. Hoang and J. M. Geusebroek. Measurement of color texture. In M. Chantler, editor, *Proceedings of the 2nd International Workshop on Texture Analysis and Synthesis (Texture 2002)*, pages 73–76. Heriot-Watt University, 2002.
18. P. Hong, R. Wang, and T. Huang. Learning patterns from images by combining soft decisions and hard decisions. In *Proc. CVPR*, volume 2, pages 78–83. IEEE Computer Society, 2000.
19. D. B. Judd and G. Wyszecki. *Color in Business, Science, and Industry*. Wiley, New York, NY, 1975.
20. J. J. Koenderink. The structure of images. *Biological Cybernetics*, 50:363–370, 1984.
21. J. J. Koenderink. Operational significance of receptive field assemblies. *Biological Cybernetics*, 58:163–171, 1988.
22. J. J. Koenderink. *Solid Shape*. MIT Press, Cambridge, 1990.
23. J. J. Koenderink and A. J. van Doorn. Receptive field families. *Biological Cybernetics*, 63:291–298, 1990.
24. J. J. Koenderink and A. J. van Doorn. Representation of local geometry in the visual system. *Biological Cybernetics*, 63:291–297, 1987.
25. J. J. Koenderink and A. J. van Doorn. Image processing done right. *Lecture Notes in Computer Science*, 2350:158–172, 2002.
26. J. J. Koenderink, A. J. van Doorn, K. J. Dana, and S. Nayar. Bidirectional reflection distribution function of thoroughly pitted surfaces. *Int. J. Comput. Vision*, 31:129–144, 1999.
27. P. Kubelka. New contributions to the optics of intensely light-scattering materials. *Journal of the Optical Society of America*, 38:448–457, 1948.
28. A. B. Lee, D. Mumford, and J. Huang. Occlusion models for natural images: A statistical study of a scale-invariant dead leaves model. *Int. J. Comput. Vision*, 41:35–59, 2001.
29. C. H. Lee and A. Rosenfeld. Improved methods of estimating shape from shading using the light source coordinate system. *Artificial Intelligence*, 26:125–143, 1985.

30. T. Lindeberg. *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers, Boston, 1994.
31. T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):117–154, 1998.
32. Y. Liu and R. Collins. A computational model for repeated pattern perception using frieze and wallpaper groups. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 537–544, Los Alamitos, 2000. IEEE.
33. G. Matheron. *Random Sets and Integral Geometry*. Wiley, New York, 1975.
34. J. Mundy and A. Zisserman. Geometric invariance in computer vision, 1992.
35. E. J. Pauwels, L. J. van Gool, P. Fiddelaers, and T. Moons. An extended class of scale-invariant and recursive scale space filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(7):691–701, 1995.
36. A. P. Pentland. A new sense of depth of field. *IEEE Pattern Analysis and Machine Intelligence*, 9:523–531, 1987.
37. D. L. Ruderman and W. Bialek. Statistics of natural images: scaling in the woods. *Physical Review Letters*, 73:814–817, 1994.
38. T. D. Sanger. Stereo disparity computation using gabor filters. *Biological Cybernetics*, 59:405–418, 1988.
39. C. Schmid and R. Mohr. Local greyvalue invariants for image retrieval. *IEEE Trans. PAMI*, 19(5):530–535, 1997.
40. M. Sigman, G. Cecchi, C. Gilbert, and M. Magnasco. On a common circle: natural scenes and gestalt rules. *Proceedings of the National Academy of Sciences USA*, 98:1935, 2001.
41. E. P. Simoncelli and B. A. Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24:1193–1216, 2001.
42. A. W. M. Smeulders, J. M. Geusebroek, and T. Gevers. Invariant representation in image processing. In *IEEE International Conference on Image Processing*, volume III, pages 18–21. IEEE Computer Society, 2001.
43. A. W. M. Smeulders, J. M. Geusebroek, and T. Gevers. Invariant representation in image processing. In *IEEE International Conference on Image Processing*, volume 3, pages 18–21. IEEE Computer Society, 2001.
44. A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content based image retrieval at the end of the early years. *IEEE Trans. PAMI*, 22(12):1349–1379, 2000.
45. J. C. van Gemert, J. M. Geusebroek, and A. W. M. Smeulders. Retrieving images as text: Relating local invariant color shapes and words. *Int. J. Comput. Vision*, (submitted), 2003.
46. B. van Ginneken, J. J. Koenderink, and K. J. Dana. Texture histograms as a function of irradiation and viewing direction. *Int. J. Comput. Vision*, 31:169–184, 1999.