Int J Comput Vis (2008) 80: 143–165
DOI 10.1007/s11263-008-0138-4

Perceptual Scale-Space and Its Applications

Yizhou Wang · Song-Chun Zhu

Received: 10 April 2006 / Accepted: 3 April 2008 / Published online: 17 May 2008
© Springer Science+Business Media, LLC 2008

Abstract When an image is viewed at varying resolutions, it is known to create discrete perceptual jumps or transitions amid the continuous intensity changes. In this paper, we study a *perceptual scale-space* theory which differs from the traditional *image scale-space* theory in two aspects. (i) In representation, the perceptual scale-space adopts a full generative model. From a Gaussian pyramid it computes a *sketch pyramid* where each layer is a primal sketch representation (Guo et al. in Comput. Vis. Image Underst. 106(1):5–19, 2007)—an attribute graph whose elements are image primitives for the image structures. Each primal sketch graph generates the image in the Gaussian pyramid, and the changes between the primal sketch graphs in adjacent layers are represented by a set of basic and composite *graph operators* to account for the perceptual transitions. (ii) In computation, the sketch pyramid and graph operators are inferred, as hidden variables, from the images through Bayesian inference by stochastic algorithm, in contrast to the deterministic transforms or feature extraction, such as computing zero-crossings, extremal points, and inflection points in the image scale-space. Studying the perceptual transitions under the Bayesian framework makes it convenient to use the statistical modeling and learning tools for (a) modeling the Gestalt properties of the sketch graph, such as continuity and parallelism etc; (b) learning the most frequent graph operators, i.e. perceptual transitions, in image scaling; and (c) learning the prior probabilities of the graph operators conditioning on their local neighboring sketch graph structures. In experiments, we learn the parameters and decision thresholds through human experiments, and we show that the sketch pyramid is a more parsimonious representation than a multi-resolution Gaussian/Wavelet pyramid. We also demonstrate an application on adaptive image display—showing a large image in a small screen (say PDA) through a selective tour of its image pyramid. In this application, the sketch pyramid provides a means for calculating information gain in zooming-in different areas of an image by counting a number of operators expanding the primal sketches, such that the maximum information is displayed in a given number of frames.

Keywords Scale-space · Image pyramid · Primal sketch · Graph grammar · Generative modeling

1 Introduction

1.1 Image Scaling and Perceptual Transitions

It has long been noticed that objects viewed at different distances or scales may create distinct visual appearances. As an example, Fig. 1 shows tree leaves in a long range of distances. In region A at near distance, the shape contours of leaves can be perceived. In region B, we cannot see individual leaves, and instead we perceive a collective foliage texture impression. In region C at an even further distance, the image loses more structures and becomes stochastic texture. Finally, in region D at a very far distance, the image appears to be a flat region whose intensities, if normalized to [0, 255], are independent Gaussian noise.

A short version was published in ICCV05 (Wang et al. 2005).

Y. Wang · S.-C. Zhu (✉)
Computer Science and Statistics, UCLA, 8125 Math Science Bldg., Box 951554, Los Angeles, CA 90095, USA
e-mail: sczhu@stat.ucla.edu

Y. Wang
e-mail: Yizhou.Wang@pku.edu.cn

These perceptual changes become more evident in a series of simulated images shown in Fig. 2. We simulate the leaves by squares of uniform intensities over a finite range of sizes. Then we zoom out the images by a 2×2-pixel averaging and down-sampling process, in a way similar to constructing a Gaussian pyramid. The 8 images in Fig. 2 are the snapshots of the images at 8 consecutive scales. At high resolutions, we see image primitives, such as edges, boundaries, and corners. At middle resolutions, these geometric elements disappear gradually and appear as texture. Finally at scale 8 each pixel is the averaged sum of 128×128 pixels in scale 1 and covers many independent "leaves". Therefore, the pixel intensities are iid Gaussian distributed according to the central limit theorem in statistics.

In the literature, the image scaling properties have been widely studied by two schools of thought.

Fig. 1 A scene with tree leaves at a long range of distances. Leaves at regions A, B, C, D appear as shape, structured texture, stochastic texture, and Gaussian noise respectively

One studies the natural image statistics and invariants over scales, for example, Ruderman (1994), Field (1987), Zhu et al. (1997), Mumford and Gidas (2001). A survey paper is referred to Srivastava et al. (2003). One fundamental observation is that the global image statistics, such as, the power spectrums of the Fourier transform, histograms of gradient images, histograms of LoG filtered images, are nearly invariant over a range of scales. That is, the histograms are the same if we down sample an image for a few times. This is especially true for scenes with big viewing depth or scenes that contain objects of a large range of sizes (infinite in the mathematical model Mumford and Gidas 2001). The global image statistics for the image sequence in Fig. 2 is not scale invariant as the squares have a smaller range of sizes than objects in natural images and this sequence has 8 scales, while natural images usually can be only scaled and down-sampled 4 to 5 times.

The other is the well known *image scale-space* theory, which was pioneered by Witkin (1983) and Koenderink (1984) in the early 1980s, and extended by Lindeberg (1993, 1994) and others (ter Haar Romeny 1997; Ahuja 1993). Although the global statistics summed over the entire image may be invariant in image scaling, our perception of the specific objects and features changes. For example, a pair of parallel edges merge into a single bar when the image is zoomed out. In the image scale-space theory (Witkin 1983; Koenderink 1984; Lindeberg 1994; ter Haar Romeny 1997), two multi-scale representations have been very influential in low level vision—the Gaussian and Laplacian pyramids. A Gaussian pyramid (Simoncelli et al. 1992; Sporring et al. 1996) is a series of low-pass filtered and down-sampled images. A Laplacian pyramid consists of band-passed images which are the difference between every two consecutive images in the Gaussian pyramid. Classical scale-space theory studied discrete and qualitative events, such as appearance

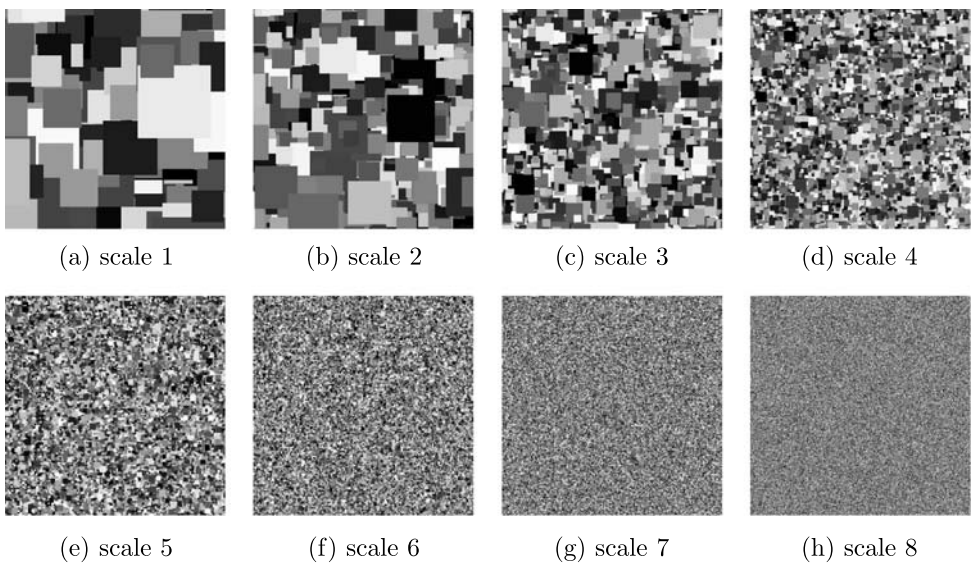

Fig. 2 Snapshot image patches of simulated leaves taken at 8 consecutive scales. The image at scale i is a 2×2-pixel averaged and down-sampled version of the image at scale $i - 1$. The image at scale 1 consists of a huge number of opaque overlapping squares whose lengths fall in a finite range $[8, 64]$ pixels and whose intensities are uniform in the range of $[0, 255]$. To produce an image of size 128×128 pixels at scale 8, we need an image of $128^8 \times 128^8$ at scale 1

of extremal points (Witkin 1983), and tracking inflection points. The image scale-space theory has been widely used in vision tasks, for example, multi-scale feature detection (Lindeberg 1993, 1998a, 1998b; Lowe 2004; Ahuja 1993), multi-scale graph matching (Shokoufandeh et al. 2002) and multi-scale image segmentation (Lifshitz and Pizer 1990; Olsen and Nielsen 1997), super-resolution (Xie et al. 2003), and multi-resolution object recognition (Xu et al. 2005), It has been shown that the performance of these applications is greatly improved by considering scale-space issues.

We study the problem in a third approach by adopting a full generative model, bringing in statistical modeling and learning methods, and making inference in the Bayesian framework.

1.2 Perceptual Scale-Space Theory

In this subsection, we briefly introduce the *perceptual scale-space* representation as an augmentation to the traditional *image scale-space*. As Fig. 3 shows, the representation comprises of two components.

1. A pyramid of primal sketches $\mathbf{S}_0, \mathbf{S}_1, \ldots, \mathbf{S}_n$. $\mathbf{S}_i$ is an attribute graph and it produces the image $\mathbf{I}_k$ in the Gaussian pyramid by a generative model $g()$ with a dictionary Δ_k (Guo et al. 2003a, 2007),

$$\mathbf{I}_k = g(\mathbf{S}_k; \Delta_k) + \text{noise}, \quad k = 1, 2, \ldots, n. \tag{1}$$

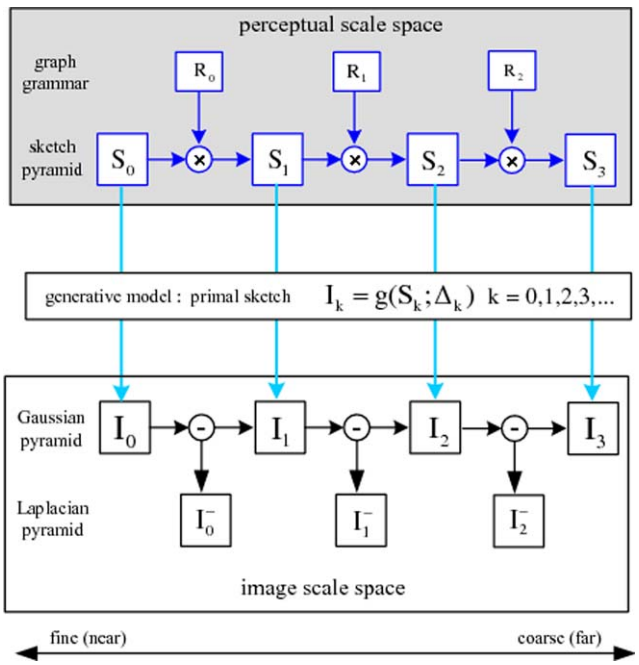

Fig. 3 A perceptual scale-space representation consists of a pyramid of primal sketches represented as attribute graphs and a series of graph operators. These operators are represented as graph grammar rules for perceptual transitions. The primal sketch at each level generates an image in a Gaussian pyramid using a dictionary of image primitives

Δ_k includes image primitives, such as, blobs, edges, bars, L-junctions, T-junctions, crosses etc. $\mathbf{S}_k$ is a layer of hidden graph governed by a Markov model to account for the Gestalt properties, such as continuity and parallelism etc., between the primitives.

2. A series of graph operators $\mathbf{R}_0, \mathbf{R}_1, \ldots, \mathbf{R}_{n-1}$. $\mathbf{R}_k$ is a set graph operators for the structural changes between graphs $\mathbf{S}_k$ and $\mathbf{S}_{k+1}$. Thus it explicitly accounts for the perceptual jumps and transitions caused by the addition/subtraction of the Laplacian pyramid image $\mathbf{I}_k^-$. In later experiments, we identify the top 20 most frequent graph operators which covers nearly all the graph changes.

In the perceptual scale-space, both the sketch pyramid and the graph operators are hidden variables and therefore they are inferred from the Gaussian image pyramid through Bayesian inference by stochastic algorithm. This is in contrast to the deterministic transforms or feature extraction, such as computing zero-crossings, extremal points, and inflection points in traditional image scale-space. We should compare the perceptual scale-space with recent progress in the image scale-space theory in Sect. 1.4.

Studying the perceptual transitions under the Bayesian framework makes it convenient to use the statistical modeling and learning tools. For example,

1. modeling the Gestalt properties of the sketch graph, such as continuity and parallelism etc.
2. learning the most frequent graph operators, i.e. perceptual transitions, in image scaling.
3. learning the prior probabilities of the graph operators conditioning on their local neighboring sketch graph structures.
4. inferring the sketch pyramid by maximizing a joint posterior probability for all levels and thus producing consistent sketch graphs.
5. quantifying the perceptual uncertainty by the entropy of the posterior probability and thus triggering the perceptual transitions through the entropy changes.

1.3 Contributions of the Paper

This paper makes the following main contributions.

1. We identify three categories of perceptual transitions in the perceptual scale-space. (i) Blurring of image primitives without structural changes. (ii) Graph grammar rules for graph structure changes. (iii) Catastrophic changes from structures to texture with massive image primitives disappearing at certain scales. For example, the individual leaves disappear from region A to B.
2. In our experiments, we find the top twenty most frequently occurring graph operators and their compositions. As the exact scale at which a perceptual jump occurs may vary slightly from person to person, we asked

seven human subjects to identify and label the perceptual transitions over the Gaussian pyramids of fifty images. The statistics of this study is used to decide parameters in the Bayesian decision formulation for the perceptual transitions.

3. We infer the sketch pyramid in the Bayesian framework using learned perceptual graph operators, and compute the optimal representation upwards-downwards the pyramid, so that the attribute graphs across scales are optimally matched and have consistent correspondence. Experiments are designed to verify the inferred perceptual scale-space by comparing with Gaussian/Laplacian scale-space and human perception.
4. We demonstrate an application of perceptual scale-space: adaptive image display. The task is to display a large high-resolution digital image in a small screen, such as a PDA, a windows icon, or a digital camera viewing screen. Graph transitions in perceptual scale-space provide a measure in term of description length of perceptual information gained from coarse to fine across a Gaussian pyramid. Based on the perceptual information, the algorithm decides to show different areas at different resolutions so as to convey maximum information in limited space and time.

1.4 Related Work in the Image Scale-Space Theory

In this subsection, we review some of most related work in image scale-space theory and compare them with the proposed perceptual scale-space theory.

One most related work is the *scale-space primal sketch* representation proposed by Lindeberg (1993). It is defined in terms of local image extrema, level curves through saddle points, and bifurcations between critical points. Lindeberg proposed a method for extracting significant blob structures in scale-space, and defined measures for the blob significance and saliency. He also identified some *blob events*, such as, annihilation, merge, split, and creation, and proposed the scale-space lifetime concept.

In Lindeberg (1998a, 1998b), Lindeberg argued that "local extrema over scales of different combinations of γ-normalized derivatives are likely candidates that correspond to interesting structures", and defined *scale-space edge and ridge* as a connected set of points in scale-space at which the gradient magnitude is a local maximum in the gradient direction, and a normalized measure of the edge strength is locally maximal over scales. Based on this representation, Lindeberg proposed an integrated mechanism for automatically selecting scales for feature detection based on *γ-normalized differential entities* and the maximization of certain strength measure of image structures. In this way, it "allows more direct control of the scale levels to be selected".

In Shokoufandeh et al. (2002), Sholoufandeh et al. proposed a framework for representing and matching multi-scale feature hierarchies. The qualitative feature hierarchy is based on the detection of a set of blobs and ridges at their corresponding hierarchical scales, which are determined by an automatic scale selection mechanism based on Lindeberg (1998b).

Lifshitz and Pizer (1990) studied a set of mathematical properties of Gaussian scale-space including the behaviors of image intensity extrema, topology of *extremum path*, and the containment relations of extremal region paths. Then they proposed a hierarchical *extremal region tree* representation in scale-space, so as to facilitate image segmentation task. Similar ideas have also been applied to watersheds in the gradient magnitude map for image segmentation (Olsen and Nielsen 1997).

In comparison to the above work, the perceptual scale-space representation studied in this paper describes scale-space events from a *Bayesian inference* perspective. In the following we highlight a few important difference from the image scale-space theory.

1. In *representation*, previous scale-space representations, including (a) the trajectories of *zero-crossing* (Witkin 1983), (b) the *scale-space primal sketch* (Lindeberg and Eklundh 1992), (c) the *scale-space edge and ridge* (Lindeberg 1998b), and (d) the *extremal region tree* (Lifshitz and Pizer 1990)—are all extracted deterministically by differential operators. In contrast, the perceptual scale-space representation adopts a full generative model where the primal sketch is a parsimonious token representation, as conjectured by Marr (1983), that can realistically reconstructs the original image in the Gaussian pyramid. This is different from the scale-space representations mentioned above. In our representation, the scale-space events are represented by a large set of basic and composite graph operators which extend the blob/extrema events studied in the image scale-space theory (Lindeberg and Eklundh 1992; Lindeberg 1998b; Lifshitz and Pizer 1990). As the Bayesian framework prevails in computer vision, posing the scale-space in the Bayesian framework has numerous benefits as it is mentioned in Sect. 1.2. Most of all, it enable us to introduce prior probability models for the spatial organizations and the transition events. These models and representations are learned through large data set and experiments.
2. In *computation*, the image scale-space representations above are computed deterministically through local feature detection (Lindeberg and Eklundh 1992; Lindeberg 1998b). In our method, we first use human labeled sketches at each scale level to learn the probability distribution of the scaling events conditioned on their local/neighboring graph configurations in scale-space. The primal sketches in the sketch pyramid are inferred by a

stochastic algorithm across scales to achieve global optimality under the *minimum descriptive length* (MDL) principle.

Graph operators are introduced to the scale-space by Shokoufandeh et al. (2002) for encoding the hierarchical feature graphs. Their algorithm intentionally introduces some perturbations to the graph adjacency matrices by adding or deleting graph nodes and edges. The objective of using these graph operators is to analyze the stability of the representation under minor perturbations due to noise and occlusion. Our representation includes a much larger set of basic and composite operators for the perceptual transitions caused by image scaling.

This work is closely related to a number of early papers in the authors' group. It is a direct extension of the primal sketch work (Guo et al. 2003a, 2007) to multi-scale. It is also related to the study of topological changes in texture motion (Wang and Zhu 2004). A specific multi-scale human face model is reported in Xu et al. (2005) where the perceptual transitions are iconic for the facial components. In a companion paper (Wu et al. 2007), Wu et al. studied the statistical models of images in a continuous entropy spectrum, from the sparse coding models (Olshausen and Field 1996) in the low entropy regime, to the stochastic texture model (Markov random fields) in the high entropy regime. Thus they interpret the drastic transitions between structures to textures shown in Figs. 1 and 3 by the perceptual uncertainty.

The paper is organized as follows. In Sect. 2, we set the background by reviewing the primal sketch model and compare it with the image pyramids and sparse coding models. Then, we introduce the perceptual uncertainty and three types of transitions in Sect. 3. The perceptual scale-space theory is studied in three consecutive sections—Sect. 4 presents a generative model representation. Section 5 discusses the learning issues for the set of graph operators and their parameters, and Sect. 6 presents an inference algorithm in the Bayesian framework with some experiment results of computing the sketch pyramids. Then we show the application of the sketch pyramids in adaptive image display in comparison with image pyramid methods. The paper is concluded with a discussion in Sect. 8.

2 Background: Image Pyramids and Primal Sketch

This section briefly reviews the primal sketch and image pyramid representations and shows that the primal sketch model (Guo et al. 2007) is a more parsimonious symbolic representation than the Gaussian/Laplacian pyramids and sparse coding model in image reconstruction.

2.1 Image Pyramids and Sparse Coding

A Gaussian pyramid is a sequence of images $\{\mathbf{I}_0, \mathbf{I}_1, \ldots, \mathbf{I}_n\}$. Each image $\mathbf{I}_k$ is computed from $\mathbf{I}_{k-1}$ by Gaussian smoothing (low-pass filtering) and down-sampling

$$\mathbf{I}_k = \lfloor G_\sigma * \mathbf{I}_{k-1} \rfloor, \quad k = 1, 2, \ldots, n, \qquad \mathbf{I}_0 = \mathbf{I}.$$

A Laplacian pyramid is a sequence of images $\{\mathbf{I}_1^-, \}$. Each image $\mathbf{I}_k^-$ is the difference between an image $\mathbf{I}_k$ and its smoothed version,

$$\mathbf{I}_k^- = \mathbf{I}_k - G_\sigma * \mathbf{I}_k, \quad k = 0, 2, \ldots, n-1.$$

This is equivalent to convolving image $\mathbf{I}_k$ with a Laplacian (band-pass) filter ΔG_σ. The Laplacian pyramid decomposes the original image into a sum of different frequency bands. Thus, one can reconstruct the image by a linear sum of images from the Laplacian pyramid,

$$\mathbf{I} = \mathbf{I}_0, \mathbf{I}_k = \mathbf{I}_k^- + G_\sigma * \lceil \mathbf{I}_{k+1} \rceil, \quad k = 0, 1, \ldots, n-1.$$

The down-sampling and up-sampling rates will be decided by σ in accordance with the well-known Nyquist theorem (see Mallat 1998 for details).

One may further decompose the original image (or its band-pass images) into a linear sum of independent image bases in a sparse coding representation. We denote Ψ as a dictionary of over-complete image bases, such as Gabor cosine, Gabor sine, and Laplacian bases

$$\mathbf{I} = \sum_{i=1}^{N} \alpha_i \psi_i + \epsilon, \quad \psi_i \in \Psi.$$

ϵ is the residue. As Ψ is over-complete, $\alpha_i, i = 1, 2 \ldots, N$ are called coefficients and usually $\alpha_i \neq \langle \mathbf{I}, \psi_i \rangle$. These coefficients can be computed by a matching pursuit algorithm (Mallat and Zhang 1993).

The image pyramids and sparse coding are very successful in representing raw images in low level vision, but they have two major problems.

Firstly, they are not effective for representing low entropy image structures (cartoon components). Figures 4(a) and (b) are respectively the Gaussian and the Laplacian pyramids of a hexagon image. It is clearly that the boundary of the hexagon spreads across all levels of the Laplacian pyramid. Therefore, it consumes a large number of image bases to construct sharp edges. The reconstruction of the Gaussian pyramid by Gabor bases is shown in Figs. 4(c) and (d). The bases are computed by the matching pursuit algorithm (Mallat and Zhang 1993). Even with 500 bases, we still see blurry edges and aliasing effects.

Secondly, they are not effective for representing high entropy patterns, such as textures. A texture region often consumes a large number of image bases. However, in human perception, we are less sensitive to texture variations. A theoretical study on the coding efficiency is referred to in a companion paper (Wu et al. 2007).

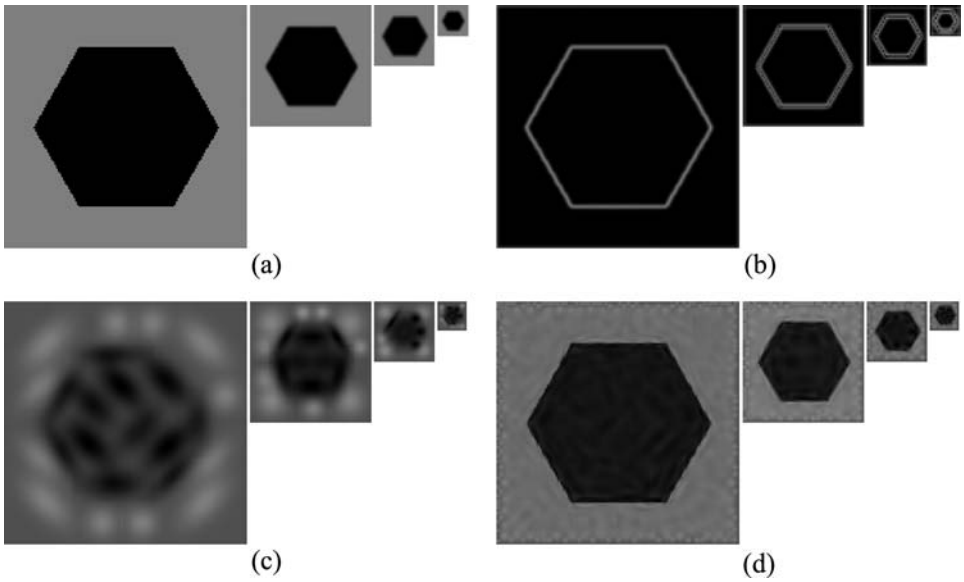

Fig. 4 The image pyramids for an hexagon image. (**a**) The Gaussian pyramid. (**b**) The Laplacian of Gaussian pyramid. (**c**) Reconstruction of (**a**) by 24, 18, 12, 12 Gabor bases respectively. (**d**) Reconstruction of (**a**) by Gabor bases. The number of bases used for a reasonable satisfactory reconstruction quality is 500, 193, 80, 38 from small scale to large scale respectively

These deficiencies suggest that we need to seek for a better model that (i) has a hyper-sparse dictionary to account for sharp edges and structures in the low entropy regime, and (ii) separates texture regions from structures. This observation leads us to a primal sketch model.

2.2 Primal Sketch Representation

A mathematical model of a primal sketch representation was proposed in (Guo et al. 2003a, 2007) to account for the generic and parsimonious token representation conjectured by Marr (1983). This representation overcomes the problems of the image pyramid and sparse coding mentioned above. We use the primal sketch to represent perceptual transitions across scales.

Given an input image $\mathbf{I}$ on a lattice Λ, the primal sketch model divides the image domain into two parts: a "sketchable" part Λ_{sk} for structures (e.g. object boundaries) and a "non-sketchable" part Λ_{nsk} for stochastic textures which has no salient structures

$$\Lambda = \Lambda_{\text{sk}} \cup \Lambda_{\text{nsk}}, \qquad \Lambda_{\text{sk}} \cap \Lambda_{\text{nsk}} = \emptyset.$$

Thus we write the image into two parts accordingly.

$$\mathbf{I} = (\mathbf{I}_{\Lambda_{\text{sk}}}, \mathbf{I}_{\Lambda_{\text{nsk}}}),$$

The structural part Λ_{sk} is further divided into a number of disjoint patches $\Lambda_{\text{sk},k}, k = 1, 2, \ldots, N_{\text{sk}}$ (e.g. 11×5 pixels)

$$\Lambda_{\text{sk}} = \bigcup_{k=1}^{N_{\text{sk}}} \Lambda_{\text{sk},k}, \quad \Lambda_{\text{sk},k} \cap \Lambda_{\text{sk},j} = \emptyset, \ k \neq j.$$

Each image patch represents an image primitive $\mathbf{B}_k$, such as a step edge, a bar, an endpoint, a junction ("T" type or "Y" type), or a cross junction, etc. Figure 7 shows some examples of the primitives which can also be called textons (Julesz 1981). Thus we have a generative model for constructing an image $\mathbf{I}$ below, with the residue following *iid* Gaussian noise

$$\mathbf{I}(u, v) = \mathbf{B}_k(u, v) + \epsilon(u, v), \quad \epsilon(u, v) \sim N(0, \sigma_o),$$
$$\forall (u, v) \in \Lambda_{\text{sk},k}, \ i = 1, \ldots, N_{\text{sk}},$$

where k indexes the primitives in the dictionary Δ for translation x, y, rotation θ, scaling σ, photometric contrast α and geometric warping $\vec{\beta}$,

$$k = (x_i, y_i, \theta_i, \sigma_i, \alpha_i, \vec{\beta}_i).$$

Each primitive has $d + 1$ points as landmarks shown on the top row of Fig. 7. d is the degree of connectivity, for example, a "T"-junction has $d = 3$, a bar has $d = 2$ and an endpoint has $d = 1$. The $(d + 1)$'th point is the center of the primitive. When these primitives are aligned through their landmarks, we obtain a sketch graph $\mathbf{S}$, where each node is a primitive. A sketch graph is also called a Gestalt field (Zhu 1999; Guo et al. 2003b) and it follows a probability $p(\mathbf{S})$, which controls the graph complexity and favors good Gestalt organization, such as smooth connection between the primitives.

The remaining texture area Λ_{nsk} is clustered into $N_{\text{nsk}} = 1 \sim 5$ homogeneous stochastic textures areas,

$$\Lambda_{\text{nsk}} = \bigcup_{j=1}^{N_{\text{nsk}}} \Lambda_{\text{nsk},j}.$$

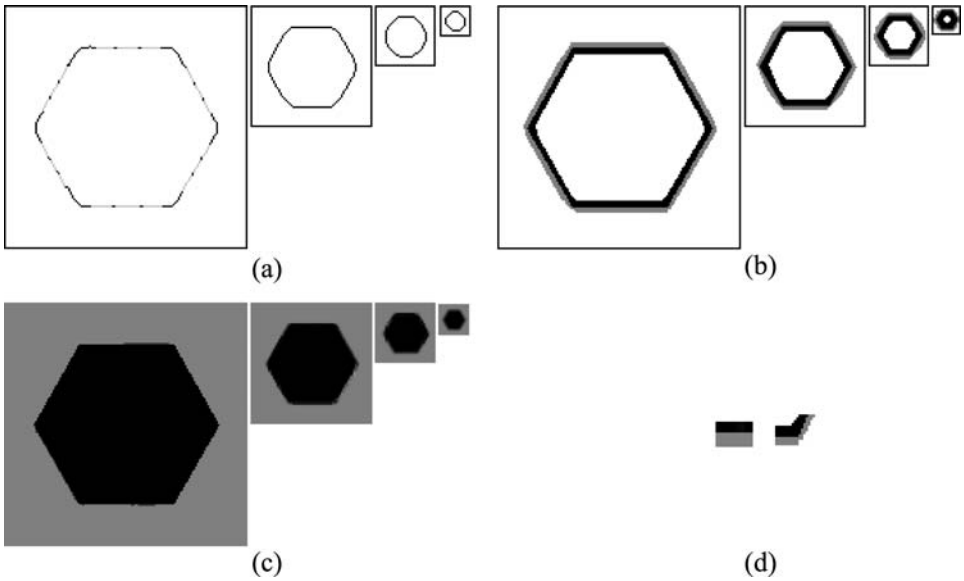

Fig. 5 Representing the hexagon image by primal sketch. (**a**) Sketch graphs at four levels as symbolic representation. They use two types of primitives shown in (**d**). The 6 corner primitives are shown in black, and step edges are in grey. The number of primitives used for the four levels are 24, 18, 12, 12, respectively (same as in Fig. 4c). (**b**) The sketchable part $\mathbf{I}_{\Lambda_{\rm sk}}$ reconstructed by the two primitives with sketch **S**. The remaining area (white) is non-sketchable (structureless): black or gray. (**c**) Reconstruction of the image pyramid after filling in the non-sketchable part. (**d**) Two primitives: a step-edge and an L-corner

Each follows a Markov random field model (FRAME) (Zhu et al. 1997) with parameters η_j. These MRFs use the structural part $\mathbf{I}_{\Lambda_{\rm sk}}$ as boundary condition

$$\mathbf{I}_{\Lambda_{{\rm nsk},j}} \sim p(\mathbf{I}_{\Lambda_{{\rm nsk},j}}|\mathbf{I}_{\Lambda_{\rm sk}};\eta_j), \quad j=1,\ldots,N_{\rm nsk}.$$

For a brief introduction, the FRAME model for any image $\mathbf{I}_A$ in a domain A (here $A=\Lambda_{{\rm nsk},j}$, $j=1,2,\ldots,N_{\rm nsk}$ respectively) given its neighborhood ∂A is a Gibbs distribution learned by a minimax entropy principle (Zhu et al. 1997)

$$p(\mathbf{I}_A|\mathbf{I}_{\partial A};\eta)=\frac{1}{Z}\exp\left\{\sum_{\alpha=1}^{K}\langle\lambda_\alpha, H_\alpha(\mathbf{I}_A|\mathbf{I}_{\partial A})\rangle\right\}. \tag{2}$$

Where $H_\alpha(\mathbf{I}_A|\mathbf{I}_{\partial A})$ is the histogram (vector) of filter responses pooled over domain A given boundary condition in ∂A. The filters are usually Gabor and LoG and are selected through an information theoretical principle. The parameters $\eta=\{\lambda_\alpha,\alpha=1,2,\ldots,K\}$ are learned by MLE and each λ_α is a vector of the same length as the number of bins in the histogram H_α. This model is a generalization to traditional MRF models and is effective in modeling various texture patterns, especially textures without strong structures.

The primal sketch is a two-level Markov model—the lower level is the MRF on pixels (FRAME models) for the textures; and the upper level is the Gestalt field for the spatial arrangement of the primitives in the sketchable part. For detailed description of the primal sketch model, please refer to (Guo et al. 2003a, 2007). We show how the primal sketch represents the hexagon image over scales in Fig. 5. It uses two types of primitives in Fig. 5(d). Only 24 primitives are needed at the highest resolution in Fig. 5(a) and the sketch graph is consistent over scales, although the number of primitives is reduced. As each primitive has sharp intensity contrast, there is no aliasing effects along the hexagon boundary in Fig. 5(c). The flat areas are filled in from the sketchable part in Fig. 5(b) through heat diffusion, which is a variation partial differential equation minimizing the Gibbs energy of a Markov random field. This is very much like image inpainting (Chan and Shen 2001).

Compared to the image pyramids, the primal sketch has the following three characteristics.

1. The primitive dictionary is much sparser than the Gabor or Laplacian image bases, so that each pixel in $\Lambda_{\rm sk}$ is represented by a single primitive. In contrast, it takes a few well-aligned image bases to represent the boundary.
2. In a sketch graph, the primitives are no longer independent but follow the Gestalt field (Zhu 1999; Guo et al. 2003b), so that the position, orientation, and intensity profile between adjacent primitives are regularized.
3. It represents the stochastic texture impression by Markov random fields instead of coding a texture in a pixel-wise fashion. The latter needs large image bases to code flat areas and still has aliasing effects as shown in Fig. 4.

3 Perceptual Uncertainty and Transitions

In this section, we pose the perceptual transition problem in a Bayesian framework and attribute the transitions to the increase in the perceptual uncertainty of the posterior probability from fine to coarse in a Gaussian pyramid. Then, we identify three typical transitions over scales.

3.1 Perceptual Transitions in Image Pyramids

Visual perception is often formulated as Bayesian inference. The objective is to infer the underlying perceptual representation of the world denoted by W from an observed image $\mathbf{I}$. Although it is a common practice in vision to compute the modes of a posterior probability as the most probable interpretations, we shall look at the entire posterior probability as the latter is a more comprehensive characterization of perception including uncertainty

$$W \sim p(W|\mathbf{I};\Theta), \tag{3}$$

where Θ denotes the model parameters including a dictionary used in the generative likelihood. A natural choice for qualifying perceptual uncertainty is the entropy of the posterior distribution,

$$\mathcal{H}(p(W|\mathbf{I})) = -\sum_{W,\mathbf{I}} p(W,\mathbf{I};\Theta)\log p(W|\mathbf{I};\Theta).$$

It is easy to show that when an image $\mathbf{I}$ is down-sampled to $\mathbf{I}_{\text{sm}}$ in a Gaussian pyramid, the uncertainty of W will increase. This is expressed in a proposition below (Wu et al. 2007).

Proposition 1 *Down-scaling increases the perceptual uncertainty,*

$$\mathcal{H}(p(W|\mathbf{I}_{\text{sm}});\Theta) \geq \mathcal{H}(p(W|\mathbf{I};\Theta)). \tag{4}$$

To be self-contained, we provide a brief proof and explanation of the above proposition in Appendix 1.

Consequently, we may have to drop some highly uncertain dimensions in W, and infer a reduced set of representation W_{sm} to keep the uncertainty of the posterior distribution of the pattern $p(W_{\text{sm}}|\mathbf{I}_{\text{sm}};\Theta_{\text{sm}})$ at a reasonable small level. This corresponds to a model transition from Θ to Θ_{sm}, and a perceptual transition from W to W_{sm}. W_{sm} is of lower dimension than W.

$$\Theta \to \Theta_{\text{sm}}, \qquad W \to W_{\text{sm}}.$$

For example, when we zoom out from the leaves or squares in Figs. 1 and 2, some details of the leaves, such as the exact positions of the leaves, lose gradually, and we can no longer see the individual leaves. This corresponds to a reduced description of W_{sm}.

3.2 Three Types of Transitions

We identify three types of perceptual transitions in image scale-space. As they are reversible transitions, we discuss them either in down-scaling or up-scaling in a Gaussian pyramid.

Following Witkin (1983), we start with a 1D signal in Fig. 6. The 1D signal is a horizontal slice from an image of a toaster in Fig. 6(a). Figure 6(b) shows trajectories of

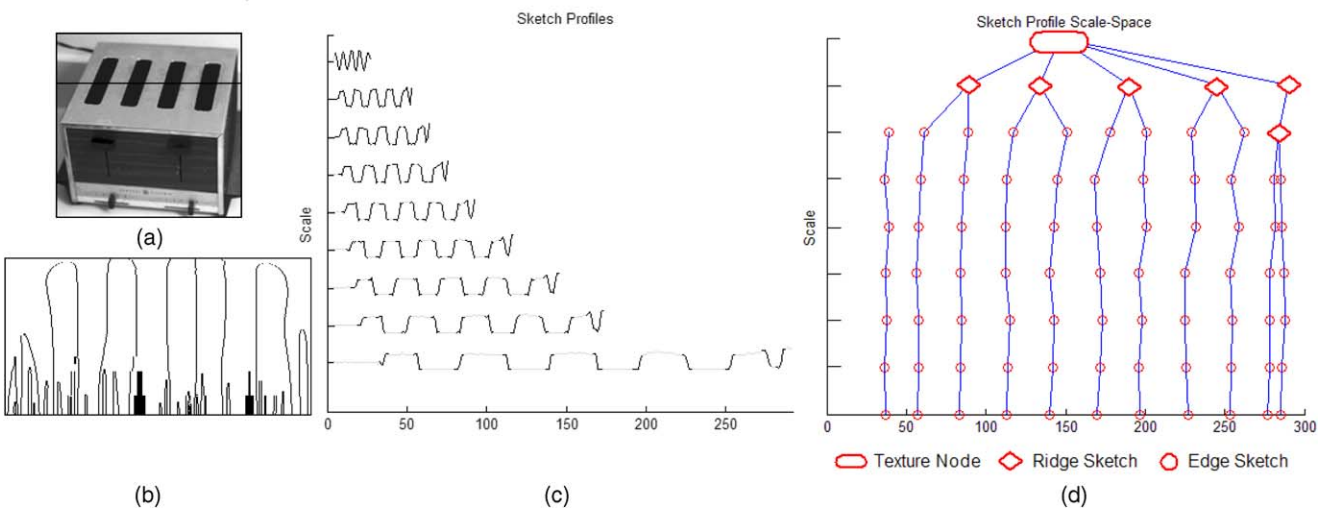

Fig. 6 Scale-space of a 1D signal. (**a**) A toaster image from which a line is taken as the 1D signal. (**b**) Trajectories of zero-crossings of the 2nd derivative of the 1D signal. The finest scale is at the bottom. (**c**) The 1D signal at different scales. The *black segments* on the curves correspond to primal sketch primitives (step edge or bar). (**d**) A symbolic representation of the sketch in scale-space with three types of transitions

zero-crossings of the 2nd derivative of the 1D signal. These zero-crossing trajectories are the signature of the signal in classical scale-space theory (Witkin 1983). We reconstruct the signal using primal sketch with 1D primitives—step edges and ridges in Fig. 6(c) where the gray curves are the 1D signal at different scales and the dark segments correspond to the primitives. Viewing the signal from bottom-to-top, we can see that the "steps" are getting gentler; and at smaller scales, the pairs of steps merge into ridges. Figure 6(d) shows a symbolic representation of the trajectories of the sketches tracked through scale-space and is called a sketch pyramid in 1D signal. Viewing the trajectories from top to bottom, we can see the perceptual transitions of the image from a texture pattern to several ridge type sketches, then split into number of step-edge type sketches when up-scaling. Figure 6(d) is very similar to the zero-crossings in (b), except that this sketch pyramid is computed through probabilistic Bayesian inference while the zero-crossings are computed as deterministic features. The most obvious differences in this example are at the high resolutions where the zero-crossings are quite sensitive to small noise perturbations.

Now we show several examples for three types of transitions in images.

Type 1: Blurring and sharpening of primitives. Figure 7 shows some examples of image primitives in the dictionary Δ_{sk}, such as step edges, ridges, corners, junctions. The top row shows the $d+1$ landmarks on each primitive with d being the degree of connectivity. When an image is smoothed, the image primitives exhibit *continuous blurring* phenomenon shown in each column, or "sharpening" when we zoom-in. The primitives have parameters to specify the scale (blurring).

Type 2: Mild jumps. Figure 8 illustrates a perceptual scale-space for a cross with a four-level sketch pyramid $\mathbf{S}_0$, $\mathbf{S}_1 0$, $\mathbf{S}_2$, $\mathbf{S}_3$ and a series of graph grammar rules $\mathbf{R}_0, \mathbf{R}_1, \mathbf{R}_2$ for graph contraction. Each $\mathbf{R}_k$ includes production rules $\gamma_{k,i}, i = 1, 2, \ldots, m(k)$ and each rule compresses a subgraph g conditional on its neighborhood ∂g.

$$\mathbf{R}_k = \{\gamma_{k,i} : g_{k,i}|\partial g_{k,i} \to g'_{k,i}|\partial g_{k,i},\ i = 1, 2, \ldots, m(k)\}.$$

If a primitive disappear, then we have $g'_{k,i} = \emptyset$. Shortly, we shall show the 20 most frequent graph operators (rules) in Fig. 10 in natural image scaling.

Graph contraction in a pyramid is realized by a series of rules,

$$\mathbf{S}_k \xrightarrow{\gamma_{k,1} \cdots \gamma_{k,m(k)}} \mathbf{S}_{k+1}.$$

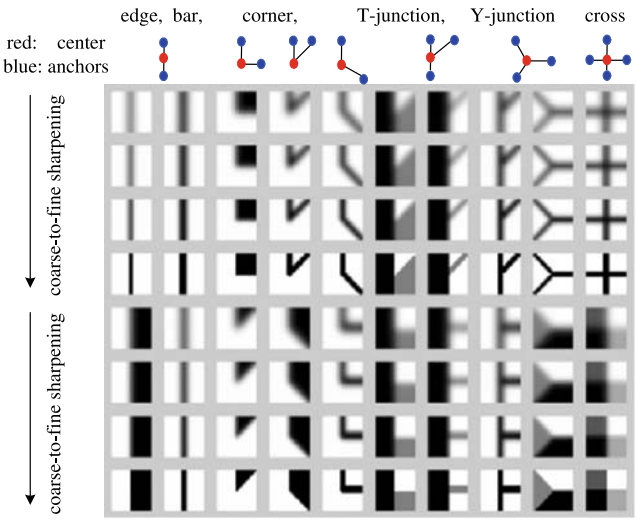

Fig. 7 Image primitives in the dictionary of primal sketch model are sharpened with four increasing resolutions from top to bottom. The blurring and sharpening effects are represented by scale parameters of the primitives

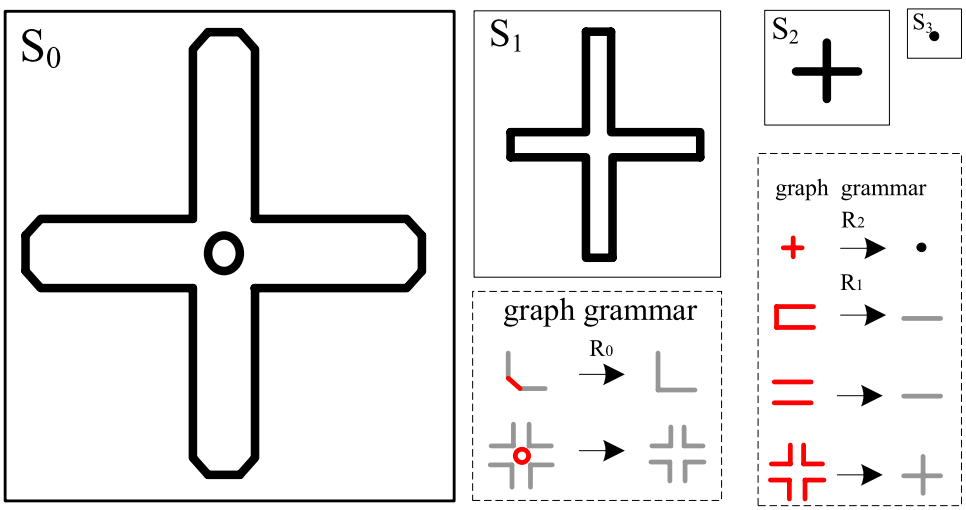

Fig. 8 An example of a 4-level sketch pyramid and corresponding graph operators for perceptual transitions

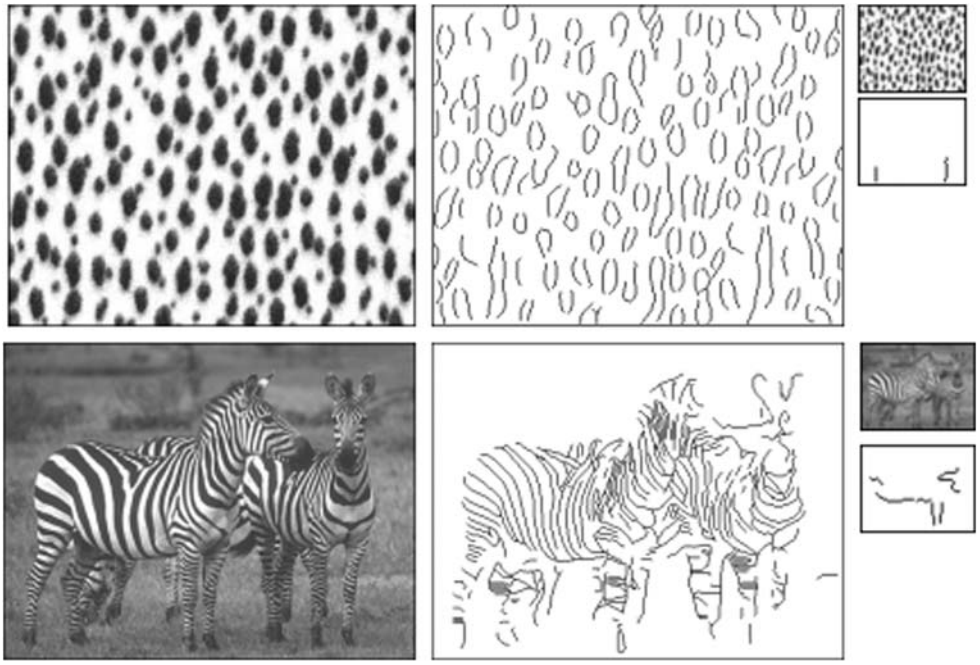

Fig. 9 Catastrophic texture-texton transition occurs when a large amount of primitives of similar sizes disappear (or appear) collectively

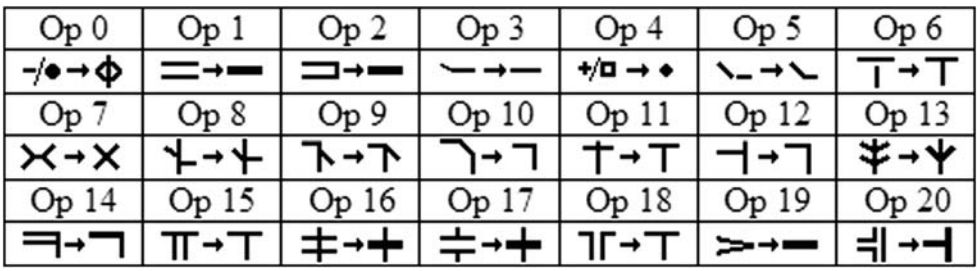

Fig. 10 Twenty graph operators identified from down-scaling image pyramids. For each operator, the left graph turns into the right graph when the image is down-scaled

These operators explain the gradual loss of details (in red), for example, a cross shrinks to a dot, a pair of parallel lines merges into a bar (ridge), and so on. The operators are reversible depending on the upward or downward scaling.

Type 3: Catastrophic texture-texton transition. At certain critical scale, a large number of similar size primitives may disappear (or appear reversely) simultaneously. Figure 9 shows two examples—cheetah dots and zebra stripe patterns. At high resolution, the edges and ridges for the zebra stripes and the cheetah blobs are visible, when the image is scaled down, we suddenly perceive only structureless textures. This transition corresponds to a significant model switching and we call it the catastrophic texture-texton transition. Another example is the leaves in Fig. 1.

In summary, a sketch pyramid represents structural changes reversibly over scales which are associated with appearance changes. Each concept, such as a blob, a cross, a parallel bar only exists in a certain scale range (lifespan) in a sketch pyramid.

4 A Generative Model for Perceptual Scale-Space

In natural images, as image structures are highly statistical, they may not exactly follow the ideal PDE model assumptions. In this section, we formulate a generative model for the perceptual scale-space representation and pose the inference problem in a Bayesian framework.

We denote a Gaussian pyramid by $\mathbf{I}[0, n] = (\mathbf{I}_0, \ldots, \mathbf{I}_n)$. The perceptual scale-space representation, as shown in Fig. 3, consists of two components—a sketch pyramid denoted by $\mathbf{S}[0, n] = (\mathbf{S}_0, \ldots, \mathbf{S}_n)$, and a series of graph grammar rules for perceptual transitions denoted by $\mathbf{R}[0, n-1] = (\mathbf{R}_0, \mathbf{R}_1, \ldots, \mathbf{R}_{n-1})$. Our objective is to define a joint probability $p(\mathbf{I}[0, n], \mathbf{S}[0, n], \mathbf{R}[0, n-1])$ so that an optimal sketch pyramid and perceptual transitions can be computed though maximizing the joint posterior probability $p(\mathbf{S}[0, n], \mathbf{R}[0, n-1]|\mathbf{I}[0, n])$. This is different from computing each sketch level $\mathbf{S}_k$ from $\mathbf{I}_k$ independently. The latter may cause "flickering" effects such as the disappearance and reappearance of an image feature across scales.

4.1 Formulation of a Single Level Primal Sketch

Following the discussion in Sect. 2.2, the generative model for primal sketch is a joint probability of a sketch $\mathbf{S}$ and an image $\mathbf{I}$,

$$p(\mathbf{I}, \mathbf{S}; \Delta_{\text{sk}}) = p(\mathbf{I}|\mathbf{S}; \Delta_{\text{sk}})p(\mathbf{S}).$$

The likelihood is divided into a number of primitives and textures,

$$p(\mathbf{I}|\mathbf{S}; \Delta_{\text{sk}}) \propto \prod_{k=1}^{N_{\text{sk}}} \exp\left\{-\sum_{(u,v)\in\Lambda_{\text{sk},k}} \frac{(\mathbf{I}(u,v) - B_k(u,v))^2}{2\sigma_o^2}\right\} \times \prod_{j=1}^{N_{\text{nsk}}} p(\mathbf{I}_{\Lambda_{\text{nsk},j}} | \mathbf{I}_{\Lambda_{\text{sk}}}; \eta_j).$$

$\mathbf{S} = \langle V, E\rangle$ is an attribute graph. V is a set of primitives in $\mathbf{S}$

$$V = \{B_k, k = 1, 2, \ldots, N_{\text{sk}}\}.$$

E denotes the connectivity for neighboring structures,

$$E = \{e = \langle i, j\rangle : B_i, B_j \in V\}$$

The prior model $p(\mathbf{S})$ is an inhomogeneous Gibbs model defined on the attribute graph to enforce some Gestalt properties, such smoothness, continuity and canonical junctions:

$$p(\mathbf{S}) \propto \exp\left\{-\sum_{d=0}^{4} \zeta_d N_d - \sum_{\langle i,j\rangle\in E} \psi(B_i, B_j)\right\},$$

where the N_d is the number of primitives in $\mathbf{S}$ whose degree of connectivity is d. ζ_d is the parameter that controls the number of primitives N_d and thus the density. In our experiments, we choose $\zeta_0 = 1.0$, $\zeta_1 = 5.0$, $\zeta_2 = 2.0$, $\zeta_3 = 3.0$, $\zeta_4 = 4.0$. The reason we give more penalty for terminators is that the Gestalt laws favor closure and continuity properties in perceptual organization. $\psi(B_i, B_j)$ is a potential function of the relationship between two vertices, e.g. smoothness and proximity. A detailed description is referred to (Guo et al. 2007).

4.2 Formulation of a Sketch Pyramid

Because of the intrinsic perceptual uncertainty in the posterior probability (see (3)), the sketch pyramid $\mathbf{S}_k, k = 0, 1, \ldots, n$ will be inconsistent if each level is computed independently. For example, we may observe a "flickering" effect when we view the sketches from coarse-to-fine (see Fig. 16b).

To ensure monotonic graph transitions and consistency of a sketch pyramid, we define a common set of graph operators

$$\Sigma_{\text{gram}} = \{\mathcal{T}_{\emptyset}, \mathcal{T}_{dn}, \mathcal{T}_{me2r}, \ldots\}.$$

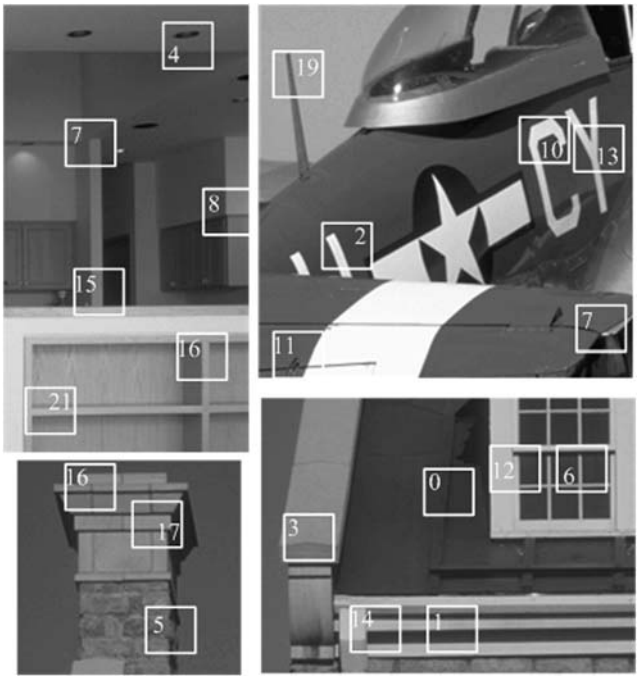

Fig. 11 Each square marks the image patch where the graph operator occurs, and each number in the squares correspond to the index of graph operators listed in Fig. 10

They stand, respectively, for null operation (no topology change), death of a node, merging a pair of step-edges into a ridge, etc. Figure 10 shows a graphical illustration of twenty down-scale graph operators (rules), which are identified and learned through a supervised learning procedure in Sect. 5. Figure 11 shows a few examples in images by rectangles where the operators occur.

Each rule $\gamma_n \in \Sigma_{\text{gram}}$ is applied to a subgraph g_n with neighborhood ∂g_n and replaces it by a new subgraph g_n'. The latter has smaller size for monotonicity, following the Proposition 4.

$$\gamma_n : g_n|\partial g_n \to g_n'|\partial g_n, \quad |g_n'| \le |g_n|.$$

Each rule is associated with a probability depending on its attributes,

$$\gamma_n \sim p(\gamma_n) = p(g_n \to g_n'|\partial g_n), \quad \gamma_n \in \Sigma_{\text{gram}}.$$

$p(\gamma_n)$ will be decided through supervised learning described in Sect. 5.

As discussed previously, transitions from $\mathbf{S}_k$ to $\mathbf{S}_{k+1}$ are realized by a sequence of $m(k)$ production rules $\mathbf{R}_k$,

$$\mathbf{R}_k = (\gamma_{k,1}, \gamma_{k,2}, \ldots, \gamma_{k,m(k)}), \quad \gamma_{k,i} \in \Sigma_{\text{gram}}.$$

The order of the rules matters and the rules constitute a path in the space of sketch graphs from $\mathbf{S}_k$ to $\mathbf{S}_{k+1}$.

The probability for the transitions from $\mathbf{S}_k$ to $\mathbf{S}_{k+1}$ is,

$$p(\mathbf{R}_k) = p(\mathbf{S}_{k+1} \,|\, \mathbf{S}_k) = \prod_{i=1}^{m(k)} p(\gamma_{k,i}).$$

A joint probability of the scale-space is

$$p(\mathbf{I}[0,n], \mathbf{S}[0,n], \mathbf{R}[0,n-1]) = \prod_{k=0}^{n} p(\mathbf{I}_k|\mathbf{S}_k; \Delta_{\mathrm{sk}}) \cdot p(\mathbf{S}_0) \cdot \prod_{k=0}^{n-1} \prod_{j=1}^{m(k)} p(\gamma_{k,j}), \tag{5}$$

4.3 A Criterion for Perceptual Transitions

A central issue for computing a sketch pyramid and associated perceptual transitions is to decide which structure should appear at which scale. In this subsection, we shall study a criterion for the transitions. This is posed as a model comparison problem in the Bayesian framework.

By induction, suppose $\mathbf{S}$ is the optimal sketch from $\mathbf{I}$. At the next level, image $\mathbf{I}_{\mathrm{sm}}$ has decreased resolution, and so $\mathbf{S}_{\mathrm{sm}}$ has less complexity following Proposition 4. Without loss of generality, we assume that $\mathbf{S}_{\mathrm{sm}}$ is reduced from $\mathbf{S}$ by a single operator γ.

$$\mathbf{S} \stackrel{\gamma}{\Longrightarrow} \mathbf{S}_{\mathrm{sm}}$$

we compute the ratio of the posterior probabilities.

$$\delta(\gamma) \triangleq \log \frac{p(\mathbf{I}_{\mathrm{sm}}|\mathbf{S}_{\mathrm{sm}})}{p(\mathbf{I}_{\mathrm{sm}}|\mathbf{S})} + \lambda_\gamma \log \frac{p(\mathbf{S}_{\mathrm{sm}})}{p(\mathbf{S})} \tag{6}$$

$$= \log \frac{p(\mathbf{S}_{\mathrm{sm}}|\mathbf{I}_{\mathrm{sm}})}{p(\mathbf{S}|\mathbf{I}_{\mathrm{sm}})}, \quad \text{if } \lambda_\gamma = 1. \tag{7}$$

The first log-likelihood ratio term is usually negative even for a good choice of $\mathbf{S}_{\mathrm{sm}}$, because a reduced generative model will not fit an image as well as the complex model $\mathbf{S}$. However, the prior term $\log \frac{p(\mathbf{S}_{\mathrm{sm}})}{p(\mathbf{S})}$ is always positive to encourage simpler models.

Intuitively, the parameter λ_γ balances the model fitting and the model complexity. As we know in Bayesian decision theory, a decision may not be only decided by the posterior probability or coding length, it is also affected by some cost function (not simply $0-1$ loss) related to perception. The cost function is summarized into λ_γ for each $\gamma \in \Sigma_{\mathrm{gram}}$.

- $\lambda_\gamma = 1$ corresponds to the Bayesian (MAP) formulation, with $0-1$ loss function.
- $\lambda_\gamma > 1$ favors applying the operator γ earlier in the down-scaling process, and thus the simple description $\mathbf{S}_{\mathrm{sm}}$.
- $\lambda_\gamma < 1$ encourages "hallucinating" features when they are unclear.

Therefore, γ is accepted, if $\delta(\gamma) \rangle 0$. More concretely, a graph operator γ occurs if

$$\begin{aligned} &\log \frac{p(\mathbf{I}_{\mathrm{sm}}|\mathbf{S}_{\mathrm{sm}})}{p(\mathbf{I}_{\mathrm{sm}}|\mathbf{S})} + \lambda_\gamma \log \frac{p(\mathbf{S}_{\mathrm{sm}})}{p(\mathbf{S})} > 0, \\ &\log \frac{p(\mathbf{I}|\mathbf{S}_{\mathrm{sm}})}{p(\mathbf{I}|\mathbf{S})} + \lambda_\gamma \log \frac{p(\mathbf{S}_{\mathrm{sm}})}{p(\mathbf{S})} < 0. \end{aligned} \tag{8}$$

The transitions $\mathbf{R}_k$ between $\mathbf{S}_k$ and $\mathbf{S}_{k+1}$ consist of a sequence of such greedy tests. In the next section, we learn the range of the parameters λ_γ for each $\gamma \in \Sigma_{\mathrm{gram}}$ from human experiments.

5 Supervised Learning of Parameters

In this section, we learn a set of the most frequent graph operators Σ_{gram} in image scaling and learn a range of parameter λ_γ for each operator $\gamma \in \Sigma_{\mathrm{gram}}$ through simple human experiments. The learning results will be used to infer sketch pyramids in Sect. 6.

We selected 50 images from the Corel image database. The content of these images covers a wide scope: natural scenes, architectures, animals, human beings, and man-made objects. Seven graduate students with and without computer vision background were selected randomly from different departments as subjects. We provided a computer graphics user interface (GUI) for the 7 subjects to identify and label graph transitions in the 50 image pyramids. The labeling procedure is as follows. First, the software will load a selected image $\mathbf{I}_0$, and build a Gaussian pyramid $(\mathbf{I}_0, \mathbf{I}_2, \ldots, \mathbf{I}_n)$. Then the sketch pursuit algorithm (Guo et al. 2003a) is run on the highest resolution to extract a primal sketch graph, which is then manually edited to fix some errors to get a perfect sketch $\mathbf{S}_0$.

Next, the software builds a sketch pyramid upwards by generating sketches simply by zooming out the sketch at the level below (starting with $\mathbf{S}_0$) one by one till the coarsest scale. The subjects will search across both the Gaussian and sketch pyramids to label places where they think graph editing is needed, e.g. some sketch disappears, or a pair of double edge sketches may be replaced by a ridge sketch. Each type of transition corresponds to a graph operator or a graph grammar rule. All these labeled transitions are automatically saved across all scale and for the 50 images.

The following are some results from this process.

Learning Result 1: frequencies of Graph Operators Figure 10 shows the top 20 graph operators which have been applied most frequently in the 50 images among the 7 subjects. Figure 12 plots the relative frequency for the 20 operators. It is clear that the majority of perceptual transitions correspond to operator 1, 2, and 10, as they are applied to the most frequently observed and generic structures in images.

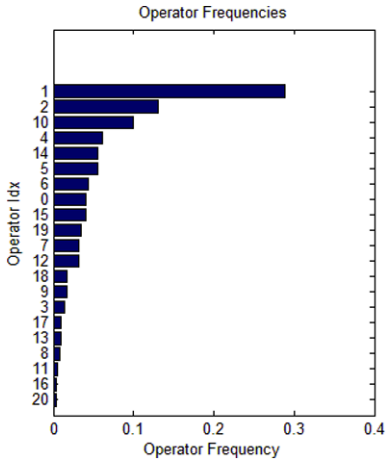

Fig. 12 The frequencies of the top 20 graph operators shown in Fig. 10

Learning Result 2: Range of λ_γ The exact scale where a graph operator γ is applied varies a bit among the human subjects. Suppose an operator γ occurs between scales $\mathbf{I}$ and $\mathbf{I}_{\text{sm}}$, then we can compute the ratios.

$$a_1 = -\log \frac{p(\mathbf{I}_{\text{sm}}|\mathbf{S}_{\text{sm}})}{p(\mathbf{I}_{\text{sm}}|\mathbf{S})}, \qquad b_1 = \lambda_\gamma \log \frac{p(\mathbf{S}_{\text{sm}})}{p(\mathbf{S})},$$

$$a_2 = -\log \frac{p(\mathbf{I}|\mathbf{S}_{\text{sm}})}{p(\mathbf{I}|\mathbf{S})}, \qquad b_2 = \lambda_\gamma \log \frac{p(\mathbf{S}_{\text{sm}})}{p(\mathbf{S})}.$$

By the inequalities in (8), we can determine an interval for λ_γ

$$\frac{a_1}{b_1} < \lambda_\gamma < \frac{a_2}{b_2}.$$

The interval above is for a specific occurrence of γ and it is caused by finite levels of a Gaussian pyramid. By accumulating the intervals for all instances of the operator γ in the 50 image and the 7 subjects, we obtain a probability (histogram) for γ. Figure 13 shows the cumulative distribution functions (CDF) of λ_γ for the top 20 operators listed in Fig. 10.

Learning Result 3: Graphlets and Composite Graph Operators Often, several graph operators occur simultaneously at the same scale in a local image structure or subgraph of a primal sketch. Figure 14(a) shows some typical subgraphs where multiple operators happen frequently. We call these subgraphs *graphlets*. By using the "*Apriori Algorithm*" (Agrawal and Srikant 1994), we find the most frequently associated graph operators in Fig. 14(b). We call these *composite operators*.

Figure 15 shows the frequency counts for the graphlets and composite graph operators. Figure 15(a) shows that a majority of perceptual transitions involves sub-graphs with no more than 5 primitives (nodes). Figure 15(b) shows that the frequency for the number of graph operators involved in each composite operator. We include the single operator as a special case for comparison of frequency.

In summary, the human experiments on the sketch pyramids set the parameters λ_γ, which will decide the threshold of transitions. In our experiments, it is evident that human vision has two preferences in comparison with the pure maximum posterior probability (or MDL) criterion (i.e. $\lambda_\gamma = 1$, $\forall \gamma$).

- Human vision has a strong preference for simplified descriptions. As we can see that in Fig. 13, $\lambda_\gamma > 1$ for most operators. Especially, if there are complicated structures, human vision is likely to simplify the sketches. For example, λ_γ goes to the range of (Cootes et al. 1998; Gauch and Pizer 1993) for operators No. 13, 14, 15.
- Human vision may hallucinate some features, and delay their disappearance, for example, $\lambda_\gamma < 1$ for operator No. 4.

These observations become evident in our experiments in the next section.

6 Upwards-Downwards Inference and Experiments

In this section, we briefly introduce an algorithm that infers hidden sketch graphs $\mathbf{S}[0, n]$ upwards and downwards across scales using the learned models $p(\lambda_\gamma)$ for each grammar rule. Then we show experiments of computing sketch pyramids.

6.1 The Inference Algorithm

Our goal is to infer consistent sketch pyramids from Gaussian image pyramids, together with the optimal path of transitions by maximizing a Bayesian posterior probability,

$$\begin{aligned}(\mathbf{S}[0,n], \mathbf{R}[0,n-1])^* &= \arg\max p(\mathbf{S}[0,n], \mathbf{R}[0,n-1]|\mathbf{I}[0,n]) \\ &= \arg\max \prod_{k=0}^{n} p(\mathbf{I}_k|\mathbf{S}_k; \Delta_k) \cdot p(\mathbf{S}_0) \prod_{k=1}^{n} \prod_{j=1}^{m(k)} p(\gamma_{k,j}).\end{aligned}$$

Our inference algorithm consists of three stages.

Stage I: Independent sketching. We first apply the primal sketch algorithm (Guo et al. 2003a, 2007) to image $\mathbf{I}_0$ at the bottom of a Gaussian pyramid to compute $\mathbf{S}_0$. Then we compute $\mathbf{S}_k$ from $\mathbf{I}_k$ using $\mathbf{S}_{k-1}$ as initialization for $k = 1, 2, \ldots, n$. As each level of sketch is computed independently by MAP estimation, the consistency of the sketch pyramid is not guaranteed. Figure 16(b) shows the sketch pyramid where we observe some inconsistencies in the sketch graph across scales.

Step II: Bottom-up graph matching. This step gives the initial solution of matching sketch graph $\mathbf{S}_k$ to $\mathbf{S}_{k+1}$. We

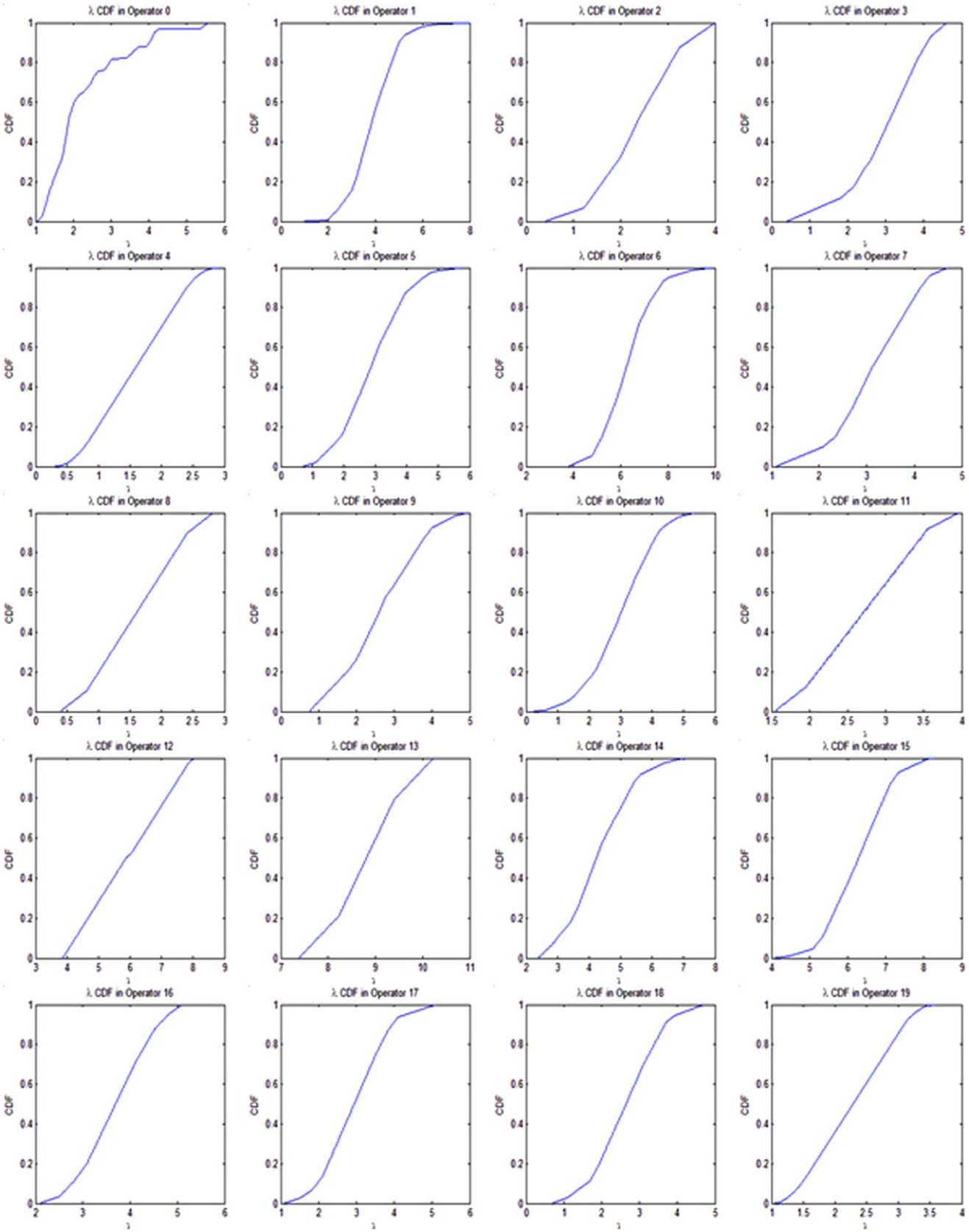

Fig. 13 CDFs of λ for each graph grammar rule or operator listed in Fig. 10

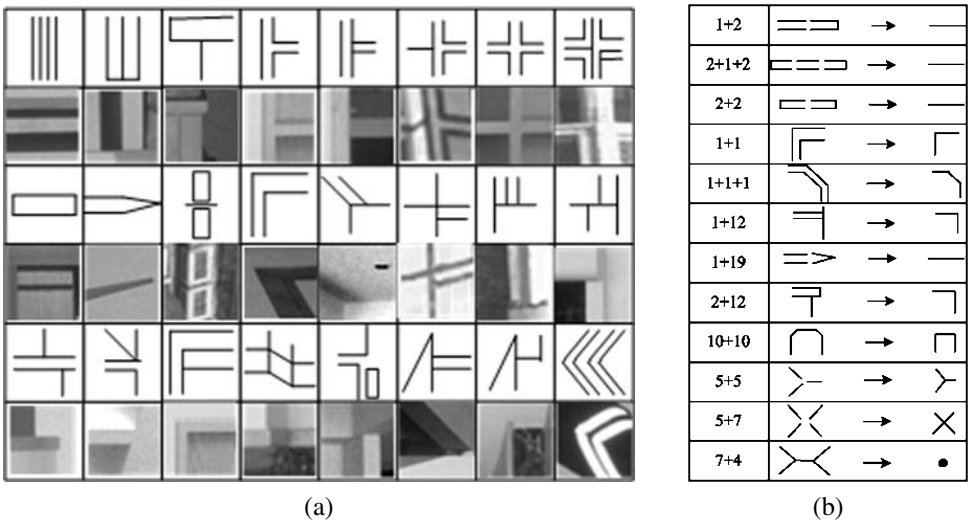

Fig. 14 (**a**) Frequently observed local image structures. The line drawings above each image patch are the corresponding subgraphs in its sketch graph. (**b**) The most frequently co-occurring operator sets. The numbers in the *left column* are the indices to the top 20 operators listed in Fig. 10. The diagrams in the *right column* are the corresponding subgraphs and transition examples

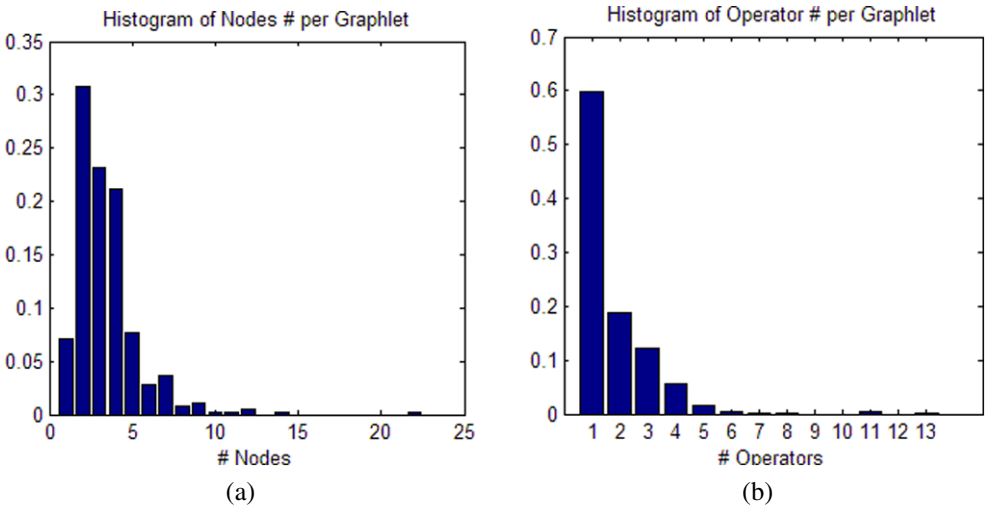

Fig. 15 Frequency of the graphlets and composite operators. (**a**) Histogram of the number of nodes per *graphlet* where perceptual transitions happen simultaneously. (**b**) Histogram of the number of operators applied to each "graphlet"

adopt standard graph matching algorithm discussed in (Zhu and Yuille 1996; Klein et al. 2001) and (Wang and Zhu 2004) to match attribute sketch graphs across scales. Here, we briefly report how we compute the matching in a bottom-up approach.

A sketch as an attribute node in a sketch graph (as shown in Fig. 16) has the following properties.

1. Normalized length **l** by its scale.
2. Shape **s**. A set of control points connectively define the shape of a sketch.
3. Appearance **a**. (Pixel intensities of a sketch.)
4. Degree **d**. (Number of connection at each ends of a sketch.)

In the following, we also use **l**, **s**, **a**, **d** as functions on sketch node i, i.e., each returns the corresponding feature. For example, $\mathbf{l}(i)$ tells the normalized length of the i'th sketch node by its scale.

A match between the i'th sketch node at scale k (denoted as v_i) and the j'th sketch node at scale $k+1$ (denoted as v_j) is defined as a probability:

$$P_{match}[v_i, v_j] = \frac{1}{Z}\exp\left\{-\frac{(\mathbf{l}(i)-\mathbf{l}(j))^2}{2\sigma_c^2} - \frac{(\mathbf{s}(i)-\mathbf{s}(j))^2}{2\sigma_s^2} - \frac{(\mathbf{a}(i)-\mathbf{a}(j))^2}{2\sigma_a^2} - \frac{(\mathbf{d}(i)-\mathbf{d}(j))^2}{2\sigma_d^2}\right\}$$

where $\sigma_\cdot$'s are the variances of the corresponding features. This similarity measurement is also used in the following graph editing part to compute the system energy. When matching $\mathbf{S}_k = (v_i(k), i = 1, \ldots, n)$ and $\mathbf{S}_{k+1} = (v_i(k+1), i = 1, \ldots, n)$, where n is the larger number of sketches

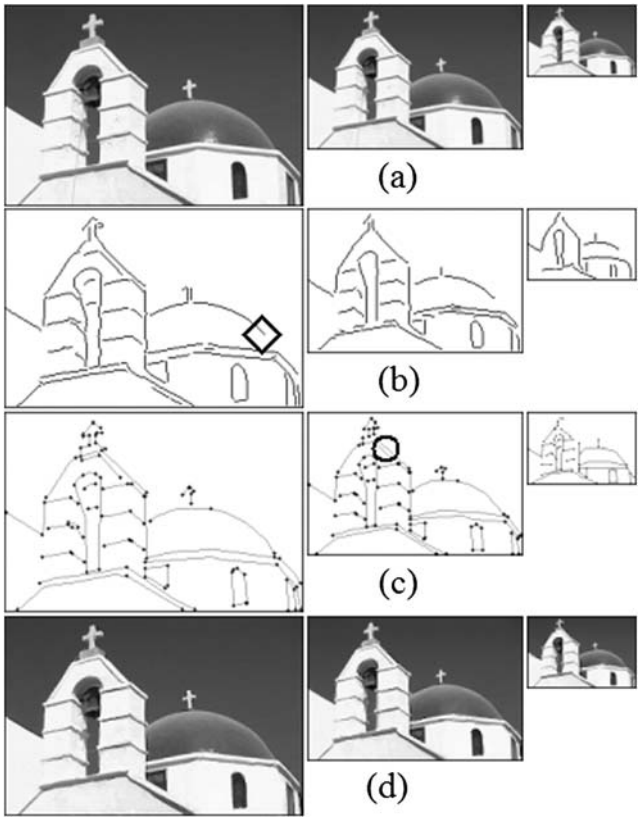

Fig. 16 A church image in scale-space. (**a**) Original images across scales. The largest image size is 241 × 261. (**b**) Initial sketches computed independently at each level by algorithm. (**c**) Improved sketches across scales. The *dark dots* indicate end points, corners and junctions. (**d**) Synthesized images by the sketches in (**c**). The *symbols* mark the perceptual transitions

in either of the two graphs, it is reasonable to allow some sketches in $\mathbf{S}_k$ map to null, or multiple sketches in $\mathbf{S}_k$ map to a same sketch in $\mathbf{S}_{k+1}$, and vice versa. Thus, the similarity between graph $\mathbf{S}_k$ and $\mathbf{S}_{k+1}$ is defined as a probability:

$$P[\mathbf{S}_k, \mathbf{S}_{k+1}] = \prod_{i=1}^{n} P_{match}[v_i(k), v_i(k+1)].$$

As the sketch graphs at two adjacent levels are very similar and well aligned. Finding a good match is not difficult. The graph matching result are used as an initial match to feed into the following Markov chain Monte Carlo (MCMC) process.

Step III: Matching and editing graph structures by MCMC sampling. Because of the intrinsic perceptual uncertainty in the posterior probability, and the huge and complicated solution space for hidden dynamic graph structures, we have to adopt the MCMC reversible jumps (Green 1995) to match and edit the computed sketch graphs both upwards and downwards iteratively in scale-space. In another word, these reversible jumps, a.k.a. graph operators, are used as a computation mechanism to infer the hidden dynamic graph structures and to find the optimal transition paths so as to pursue globally optimal and perceptually consistent primal sketches across scales.

Our Markov chain consists of twenty pairs of reversible jumps (listed in Fig. 10) to adjust the matching of adjacent graphs in a sketch pyramid based on the initial matching results in Step II, so as to achieve a high posterior probability. These reversible jumps correspond to the grammar rules in Σ_{gram}. Each pair of them is selected probabilistically and they observe the detailed balance equations. Each move in the Markov chain design is a reversible jump between two states A and B realized by a Metropolis-Hastings method (Metropolis et al. 1953).

For clarity, we put the description of these reversible jumps to Appendix 2.

6.2 Experiments on Sketch Pyramids

We successfully applied the inference algorithm on 20 images to compute their sketch pyramids in perceptual scale-space. In this subsection, we show some of them to illustrate the inference results.

Inference Result 1: Sketch Consistency Figure 16(c) shows examples of the inferred sketch pyramids obtained by the MCMC sampling with the learned graph grammar rules. We compare the results with the initial bottom-up sketches (b) where each level is computed independently and in a greedy way. The improved results show consistent graph matching across scales.

Inference Result 2: Compactness of Perceptual Sketch Pyramid In Fig. 17, we compare the inferred sketch graphs in column (c) with the sketch graph obtained by only applying death operator in column (b). We can see that the sketch graphs inferred from perceptual scale-space is more compact and close to human perception.

Figure 18 compares the compactness of the sketch pyramid representation obtained by applying learned perceptual graph operators against that rendered by applying only simple death operators. The compactness is measured by the number of sketches. Images at scale level 1 has the highest resolution, containing maximum number of sketches in the sketch graph, and the number of sketches at this scale level is normalized to 1.0. When scaling down, the sketch nodes are getting shorter and shorter, till they finally disappear, where the death operator applies. Thus the higher the scale level, the fewer the sketch nodes. The dashed line shows the relative number of sketches at each scale level with only the simple death operator applied when scaling down. The solid line shows the relative number of sketches at each scale level by applying perceptual operators learned from human subjects. The number of sketches indicates the coding length

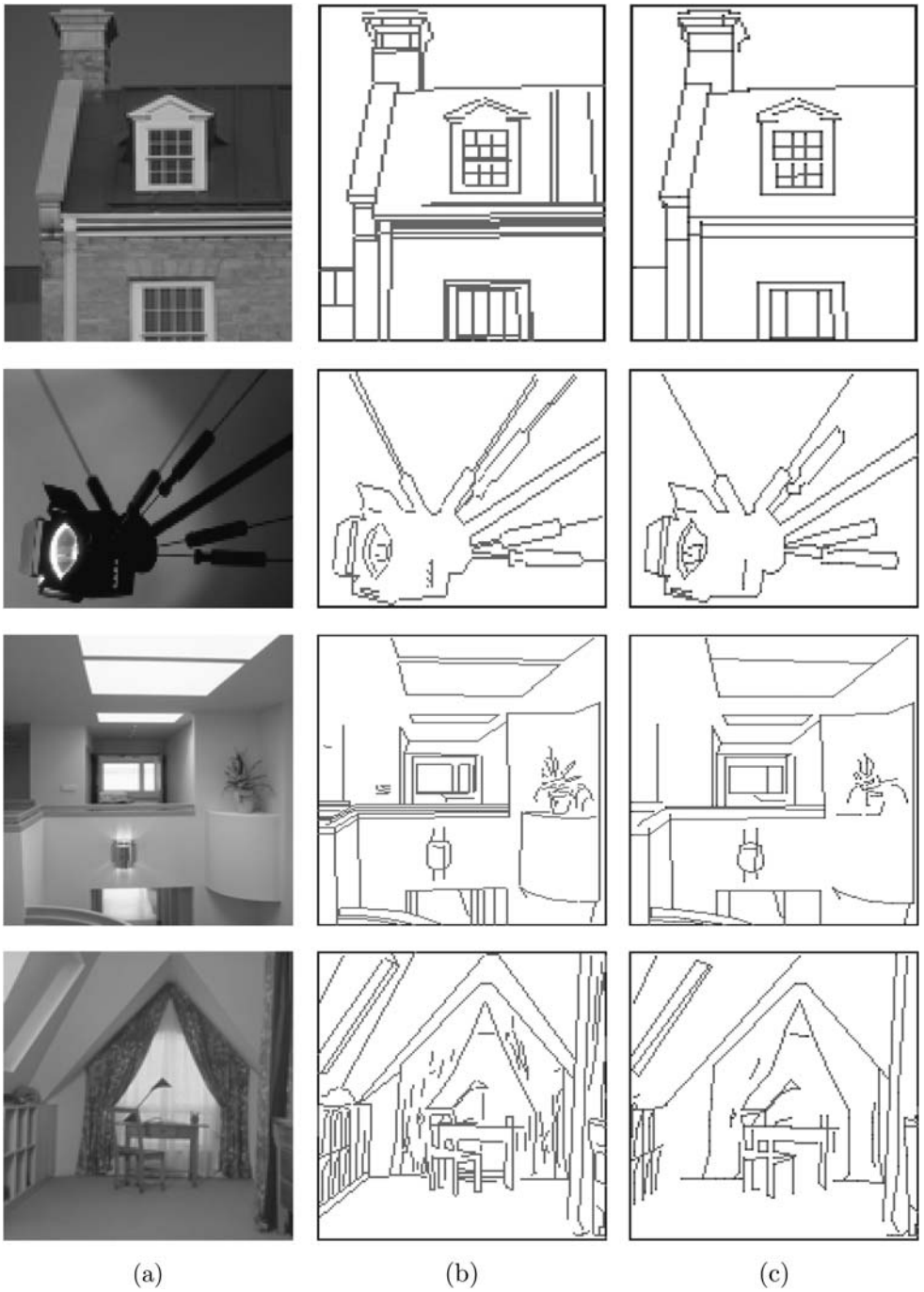

Fig. 17 Sketch graph comparison between applying simple death operator and applying learned perceptual graph operators. (**a**) Original images. (**b**) Sketch graphs at scale level 7 with only death operator applied. (**c**) Sketch graphs at scale level 7 with learned perceptual graph operator applied

of sketch graphs. As human beings always prefer simpler models without perceptual loss, the sketch pyramid inferred with learned perceptual graph operators is a more compact representation.

In summary, from the above inference results, it seems that the inferred perceptual sketch pyramid matches the transitions in human perception experiment well. By properly addressing the perceptual transition issue in perceptual scale-space, we expect performance improvements in many vision applications. For example, by explicitly modeling these discrete jumps, we can finally begin to tackle visual scaling tasks that must deal with objects and features that look fundamentally different across scales. In object recognition, for example, the features that we use to recognize an object from far away may belong to a different class than those we use to recognize it at a closer distance. The perceptual scale-space will allow us to connect these scale-variant features in a probabilistic chain. In the following section, we show an applications based on a inferred perceptual sketch pyramid.

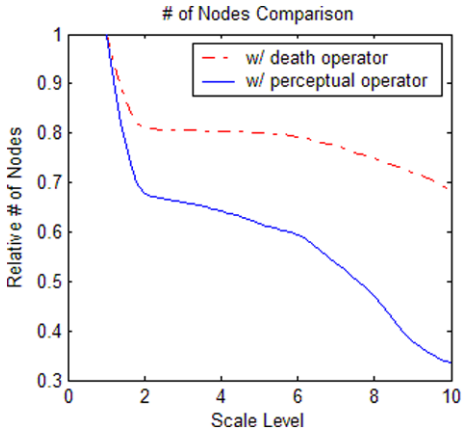

Fig. 18 A comparison of relative node number in sketch graphs at each scale between applying simple death operators and applying the learned perceptual graph operators. Images at scale level 1 has the highest resolution, containing maximum number of nodes in the sketch graphs, which is normalized to 1.0. The *dashed line* shows the number of sketches at each scale level when simple death operator is applied. The *solid line* shows the number of sketches after applying perceptual operators learned from human subjects

Fig. 19 A tour over a Gaussian pyramid. Visiting the decomposed quad-tree nodes in a sketch pyramid is an efficient way to automatically convey a large image's informational content

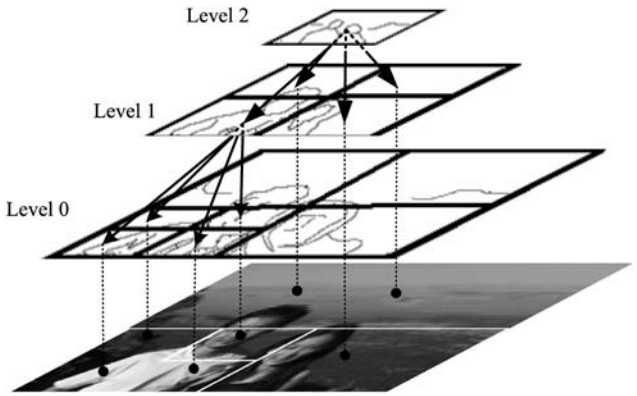

Fig. 20 A Gaussian pyramid is partitioned into a quad-tree. We only show the nodes which are visited by our algorithm. During the "tour", a quad-tree node is visited if and only if its sketch sub-graph expands from the level above, which indicates additional semantic/structural information appears

7 An Application—Adaptive Image Display

Because of improving resolution of digital imaging and use of portable devices, recently there have been emerging needs, as stated in Xie et al. (2003), for displaying large digital images (say $\Lambda = 2048 \times 2048$ pixels) on small screens (say $\Lambda_o = 128 \times 128$ pixels), such as in PDAs, cellular phones, image icon display on PCs, and digital cameras. To reduce manual operations involved in browsing within a small window, it is desirable to display a short movie for a "tour" that the small window Λ_o flying through the lattice Λ, so that most of the *information* in the image is displayed in as few frames as possible. These frames are snapshots of a Gaussian pyramid at different resolutions and regions. For example, one may want to see a global picture at a coarse resolution and then zoom in some interesting objects to view the details.

Figure 19 shows a demo on a small PDA screen. The PDA displays the original image at the upper-left corner, within which a window (white) indicates the area and resolution shown on the screen. To do so, we decompose a Gaussian pyramid into a quadtree representation shown in Fig. 20. Each node in the quadtree hierarchy represent a square region of a constant size. We say a quadtree node is "visited" if we show its corresponding region on the screen. Our objective is to design a visiting order of some nodes in the quadtree. The quadtree simplifies the representation but causes border artifacts when an interesting object is in the middle and has to be divided into several nodes.

With this methodology, two natural questions will arise for this application.

1. *How do we know what objects are interesting to users?* The answer to this question is very subjective and user dependent. Usually people are more interested in faces and texts (Xie et al. 2003), which requires face and text detection. We could add this function rather easily to the system as there are off-the-shelf code for face and text detection working reasonably well. But it is beyond the scope of this paper, which is focused on low-middle level representation of generic images.
2. *How do we measure the information gain when we zoom in an area?* There are two existing criteria: one is the focus of attention models (Ma et al. 2002), which essentially favors areas of high intensity contrast. A problem with this criterion is that we may be directed to some

boring areas, for example smooth (cartoon like) regions where few new structures will be revealed when the area is zoomed-in. The other is to sum up the Fourier power of the Laplacian image at certain scale over a region. This could tell us how much signal power we gain when the area is zoomed in. But the frequency power does not necessarily mean structural information. For example, we may zoom into the forest (texture) in the background (see Fig. 21).

We argue that a sketch pyramid together with perceptual transitions provide a natural and generic measure for the information gains when we zoom in an area or visit a node in the quadtree.

In computing a perceptual sketch pyramid, we studied the inverse process when we compute operators from high-resolution to low-resolution. A node v at level k corresponds to a sub-graph $\mathbf{S}_k(v)$ of sketches, and its children at the higher resolution level correspond to $\mathbf{S}_{k-1}(v')$. The information gain for this split is measured by

$$\delta(v) = -\log_2 \frac{p(\mathbf{S}_{k-1}(v'))}{p(\mathbf{S}_k(v))}. \tag{9}$$

It is the number of new bits needed to describe the extra information when graph expands. As each node in the quad-tree has an information measure, we can expand a node in a sequential order until a threshold τ (or a maximum number of bits M) is reached. Figure 21 shows results of the quad-tree decomposition and multi-resolution image reconstruction. The reconstructed images show that there is little perceptual loss of information when each region is viewed at its determined scale.

The information gain measure in (9) is more meaningful than calculating the power of bandpass Laplacian images. For example, as shown in Fig. 4(b), a long sharp edge in an image will spread across all levels of the Laplacian pyramid, and thus demands continuous refining in the display if we use the absolute value of the Laplacian image patches. As shown in Fig. 5, in contrast, in a sketch pyramid, it is a single edge and will stop at certain high level. As a result, the quad-tree partition makes more sense to human beings in the perceptual sketch pyramid than in the Laplacian of Gaussian pyramid, as shown in Figs. 21 and 22.

To evaluate the quad-tree decomposition for images based on the inferred perceptual sketch pyramids, we design the following experiments to quantitatively verify the perceptual sketch pyramid model and its computation.

We selected 9 images from the Corel image database and 7 human subjects with and without computer vision background. To get a fair evaluation of the inference results, we deliberately chose 7 different persons from the 7 graduate students who had done the perceptual transition labeling in the learning stage. Each human subject was provided with a computer graphical user interface, which allowed them to partition the given high-resolution images into quad-trees as shown in Fig. 24. The depth ranges of the quad-trees were specified by the authors in advance. For example, for the first three images in Fig. 24, the quad-tree depth range was

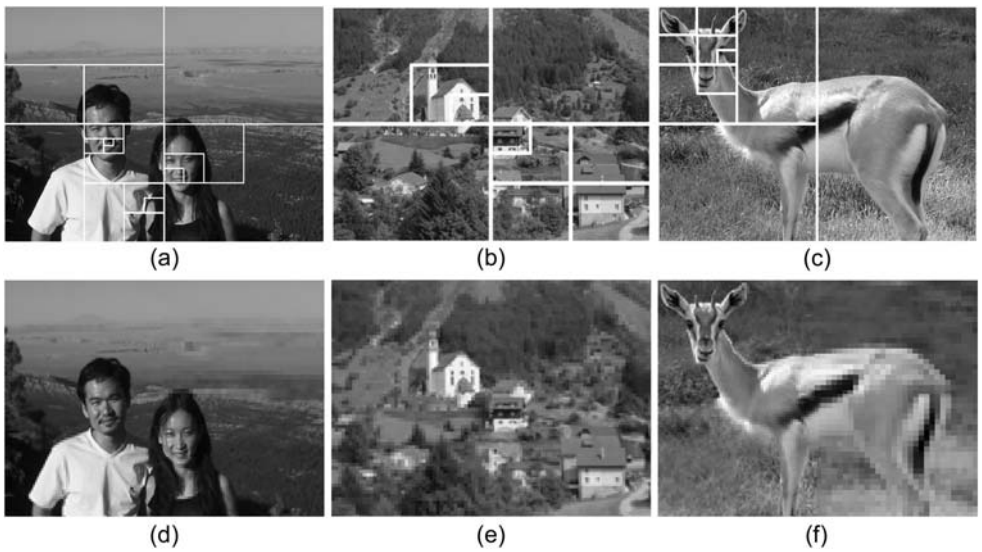

Fig. 21 (**a**–**c**) Show three examples of the tours in the quad-trees. The partitions correspond to regions in the sketch pyramid that experience sketch graph expansions. If the graph expands in a given partition, then we need to increase the resolution of the corresponding image region to capture the added structural information. (**d**–**f**) Represent the replacement of each sketch partition with an image region from the corresponding level in the Gaussian pyramid. Note that the areas of higher structural content are in higher resolution (e.g. face, house), and areas of little structure are in lower resolution (e.g. landscape, grass)

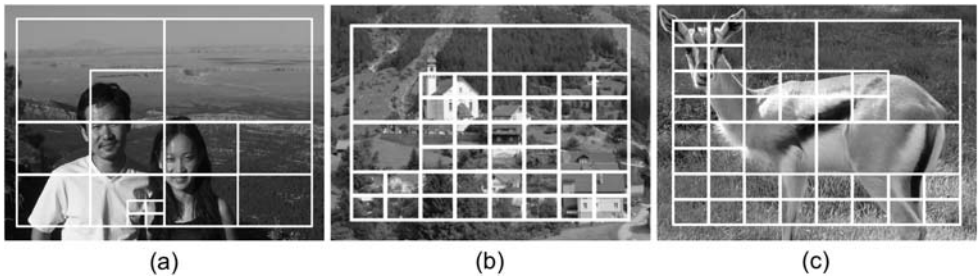

Fig. 22 For comparison with the corresponding partitions in the sketch pyramid (Fig. 21), a Laplacian decomposition of the test images are shown. The outer frame is set smaller than the image size to avoid Gaussian filtering boundary effects. The algorithm greedily splits the leaf nodes bearing the most power (sum of squared pixel values in the node of the Laplacian pyramid I_k^+). As clearly evident, the Laplacian decomposition does not exploit the perceptually important image regions in its greedy search (e.g. facial features) instead focusing more on the high frequency areas

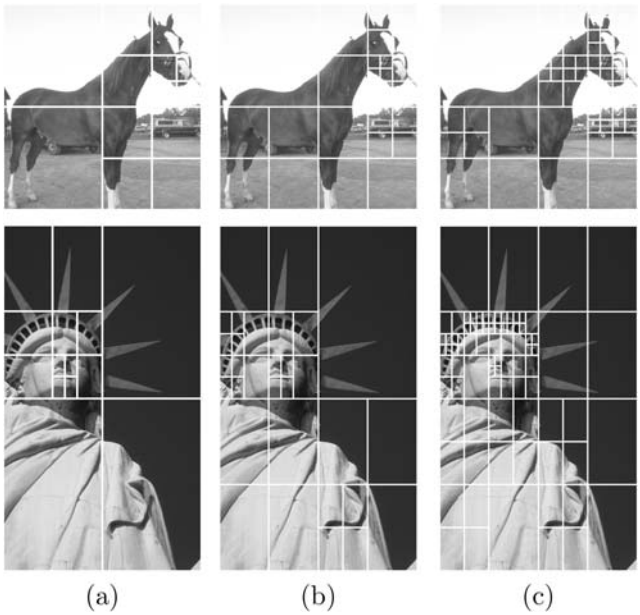

Fig. 23 Comparison of the quad-trees partitioned by the human subjects and the computer algorithm—Part II. The computer performance is within the variation of human performance. (**a**) The computer algorithm partitioned quad-tree. (**b**) & (**c**) Two examples of human partitioned quad-tree

Table 1 Comparison table of quad-tree partition performance on 9 images between human subjects and the computer algorithm based on inferred perceptual image pyramids. This table shows that for 7 out of 9 images, the computer performance is within the variation of human performance

Image ID	Human partition STD	Computer partition error
54076	0.231265	0.202402
77016	0.140912	0.092913
180088	0.246501	0.154938
234007	0.239068	0.208406
404028	0.357495	0.325235
412037	0.178133	0.162965
street	0.279815	0.263395
197046*	0.168721	0.244858
244000*	0.245081	0.369920

set to 1 to 4 levels, 1 to 4 levels and 2 to 5 levels, respectively. Then, the human subjects partitioned the given images into quad-trees based on the *information* distributed on the images according to their own perception and within the specified depth ranges.

After collecting the experiment results from human subjects, we compared the quad-tree partition performance by human subjects and that by our computer algorithm based on inferred perceptual sketch pyramids. The quantitative measure was computed as follows. Each testing image was split into quad-tree by the 7 human subjects. Consequently, each pixel of the image was assigned 7 numbers, which were the corresponding depth of the quad-tree partitioned by the 7 human subjects at the pixel. In Table 1, the middle column is the average standard deviation (STD) of quad-tree depth per pixel of each image partitioned by the human subjects. It tells the human performance variation. For each image, we take the average depth of the human partition as the "truth" for each pixel. The right column shows the computer algorithm's partition error from the "truth". This table shows that in 7 out of 9 images, the computer performance is within the variation of human performance.

Figure 24 and Fig. 23 compare the partition results between computer algorithm and human subjects. In these figures, the first column shows the quad-tree partitions of each image by our computer algorithm. The other two columns are two sample quad-tree partitions by the human subjects. In Fig. 24, our computer algorithm's performance is within the variation of human performance. Figure 23 shows the two exceptional cases. However, from the figure, we can see that the partitions by our computer algorithm are still very reasonable.

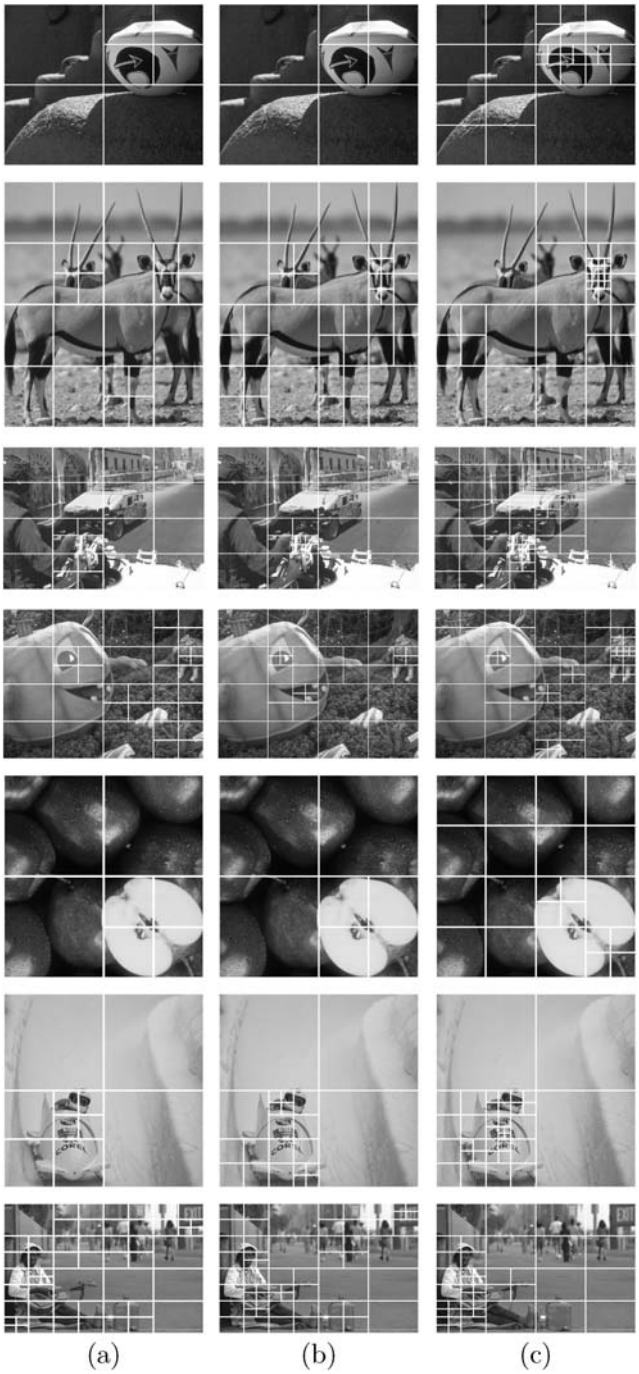

Fig. 24 Comparison of the quad-trees partitions by the human subjects and the computer algorithm. The computer performance is within the variation of human performance. (**a**) The computer algorithm partitioned quad-tree. (**b**) & (**c**) Two examples of human partitioned quad-tree

8 Summary and Discussion

In this paper, we propose a perceptual scale-space representation to account for perceptual jumps amid continuous intensity changes. It is an augmentation to the classical scale-space theory. We model perceptual transitions across scales by a set of context sensitive grammar rules, which are learned through a supervised learning procedure. We explore the mechanism for perceptual transitions. Based on inferred sketch pyramid, we define *information gain* of an image across scales as the number of extra bits needed to describe graph expansion. We show an application of such an information measure for adaptive image display.

Our discussion is mostly focused on the mild perceptual transitions. We have not discussed explicitly the mechanism for the catastrophic texture-texton transitions, which is referred to in a companion paper (Wu et al. 2007). In future work, it shall also be interesting to explore the link between the perceptual scale-space to the multi-scale feature detection and object recognition (Lowe 2004; Kadir and Brady 2001), and applications such as super-resolution, and tracking objects over a large range of distances.

Acknowledgements This work was supported in part by NSF grants IIS-0707055 and IIS-0413214, and an ONR grant N00014-05-01-0543. We thank Siavosh Bahrami for his contribution in generating the scale sequence in Fig. 2 and assistant in the PDA experiment. We also thank Ziqiang Liu for his programming code and Dr. Xing Xie and Xin Fan at Microsoft Research Asia for discussions on the adaptive image display task. We also thank the support of two Chinese National 863 grants 2006AA01Z121 and 2007AA01Z340 for the work at Lotus Hill Institute which provides some dataset (Yao et al. 2007).

Appendix 1: Interpreting the Information Scaling Proposition

This appendix provides a brief proof and interpretation of Proposition 4, following the derivations in (Wu et al. 2007). Let W be a general description of the scene following a probability $p(W)$ and it generates an image by a deterministic function $\mathbf{I} = g(W)$. $\mathbf{I} \sim p(\mathbf{I})$. Because $\mathbf{I}$ is decided by W, we have $p(W, \mathbf{I}) = p(W)$. So,

$$p(W|\mathbf{I}) = \frac{p(W, \mathbf{I})}{p(\mathbf{I})} = \frac{p(W, \mathbf{I})}{p(\mathbf{I})}. \tag{10}$$

Then by definition

$$\mathcal{H}(p(W|\mathbf{I})) = -\sum_{W,\mathbf{I}} p(W, \mathbf{I}) \log p(W|\mathbf{I}), \tag{11}$$

$$= -\sum_{W,\mathbf{I}} p(W, \mathbf{I}) \log \frac{p(W)}{p(\mathbf{I})}, \tag{12}$$

$$= \mathcal{H}(p(W)) - \mathcal{H}(p(\mathbf{I})). \tag{13}$$

Now, if we have a reduced image $\mathbf{I}_{\text{sm}} = \text{Re}(\mathbf{I})$, for example, by smoothing and subsampling, then we have

$$\mathcal{H}(p(W|\mathbf{I}_{\text{sm}})) = \mathcal{H}(p(W)) - \mathcal{H}(p(\mathbf{I}_{\text{sm}})). \tag{14}$$

So,

$$\mathcal{H}(p(W|\mathbf{I}_{\text{sm}})) - \mathcal{H}(p(W|\mathbf{I})) = \mathcal{H}(p(\mathbf{I})) - \mathcal{H}(p(\mathbf{I}_{\text{sm}})) \tag{15}$$

As $\mathbf{I}_{\text{sm}}$ is a reduced version of $\mathbf{I}$, suppose both $\mathbf{I}_{\text{sm}}$ and $\mathbf{I}$ are discretized properly. Then $\mathcal{H}(p(\mathbf{I})) \geq \mathcal{H}(p(\mathbf{I}_{\text{sm}}))$. The conclusion follows.

This proposition is quite intuitive. When we zoom out, some information will be lost for the underlying description W and thus they become less perceivable.

Appendix 2: The Reversible Jumps for Graph Editing Operators

The 20-graph editing operators are used to edge the sketch graphs so that they are matched in graph structures and the remaining differences are the transitions. Because the graph matching needs to consider information propagation, they are made reversible. Reversibility in MCMC is similar to back-tracing in heuristic searches in computer science.

Consider a reversible move between two states A and B by applying an operator back and forth.

We design a pair of proposal probabilities for moving from A to B, with $q(A \to B) = q(B|A)$, and back with $q(B \to A) = q(A|B)$. The proposed move is accepted with probability, according to the well-known Metropolis-hastings method,

$$\alpha(A \to B) = \min\left(1, \frac{q(A|B) \cdot p(B|\mathbf{I}^{\text{obs}}[1,\tau])}{q(B|A) \cdot p(A|\mathbf{I}^{\text{obs}}[1,\tau])}\right).$$

These MCMC moves simulate a Markov chain with invariant probability $p(\mathbf{S}[0,n], \mathbf{R}[0,n-1]\,|\mathbf{I}[0,n])$. Each probability model in (5), including the image photometric model $p(\mathbf{I}_k|\mathbf{S}_k)$, the primal sketch geometric model $p(\mathbf{S}_k)$ and the graph grammar rule model $p(\gamma_i)$. These probability models are used in this inference process when sampling from the posterior probability. The design of reversible jumps is very similar to the design in (Wang and Zhu 2004; Guo et al. 2007). Due to the page limit, we only introduce one pair of Markov chain moves—split/merge (Operator 7 in Fig. 10). The moves are illustrated in Fig. 25 and they are jump processes between two states A and B, where

$$\begin{aligned} A &= (n, \mathbf{S} = \langle (V_-, v_j), (E_-, e_{i,j}) \rangle) \\ &\rightleftharpoons (n-1, \mathbf{S}' = \langle V_-, E_- \rangle) = B, \end{aligned}$$

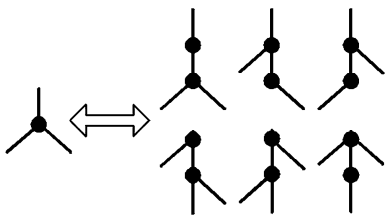

Fig. 25 Split/merge graph operation diagram. A vertex can be split into two vertices with one of six edge configurations

where n is the number of sketches in sketch graph $\mathbf{S}$. V_- and E_- denote the unchanged sketch set and edge set, respectively. $e_{i,j}$ is the edge between sketches v_i and v_j, and v_j is the sketch disappeared after merging. We define the proposal probabilities as follows

$$q(A \to B) = q_{s/m} \cdot q_m \cdot q(i) \cdot q(j),$$

$$q(B \to A) = q_{s/m} \cdot q_s \cdot q'(i) \cdot q(\textit{pattern}).$$

$q_{s/m}$ is the probability for selecting this split/merge move among all possible graph operations. q_m and q_s is the probability to choose either split or merge, respectively, where $q_m + q_s = 1$. $q(i)$ is the probability of selecting v_i as the anchor vertex for the other vertex to merge into, which is usually set to $1/n$. $q(j)$ is the probability to choose v_j from v_i's neighbors, which is set to be inversely proportional to the distance between v_i and v_j. When proposing a split move, $q'(i)$ is the probability to choose v_i. It is assumed to be uniform among those qualified vertices. When a sketch with m edges is split, there are $1/(2^m - 2)$ ways for two vertices to share these m edges. Therefore, $q(\textit{pattern})$ is set to be $1/(2^m - 2)$.

References

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of the international conference on very large data bases*.

Ahuja, N. (1993). A transform for detection of multiscale image structure. In Proceedings of the computer vision and pattern recognition (Vol. 15–17, pp. 780–781).

Chan, T., & Shen, J. (2001). Local inpainting model and TV-inpainting. *SIAM Journal on Applied Mathematics*, *62*(3), 1019–1043.

Cootes, T. F., Edwards, G. J., & Taylor, C. J. (1998). Active appearance models. In *Proceedings of the European conference on computer vision*.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of Optical Society of America*, *4*(12), 2379–2394.

Gauch, J., & Pizer, S. (1993). Multiresolution analysis of ridges and valleys in grey-scale images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *15*(6), 635–646.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*, 711–732.

Guo, C., Zhu, S. C., & Wu, Y. (2003a). A mathematical theory of primal sketch and sketchability. In *Proceedings of the international conference on computer vision*.

Guo, C. E., Zhu, S. C., & Wu, Y. N. (2003b). Modeling visual patterns by integrating descriptive and generative models. *International Journal of Computer Vision*, *53*(1), 5–29.

Guo, C. E., Zhu, S. C., & Wu, Y. N. (2007). Primal sketch: integrating texture and structure. *Computer Vision and Image Understanding*, *106*(1), 5–19.

Julesz, B. (1981). Textons, the elements of texture perception, and their interactions. *Nature*, *290*, 91–97.

Kadir, T., & Brady, M. (2001). Saliency, scale and image description. *International Journal of Computer Vision*.

Klein, P., Sebastian, T., & Kimia, B. (2001). Shape matching using edit-distance: an implementation. In *SODA* (pp. 781–790).

Koenderink, J. J. (1984). The structure of images. *Biological Cybernetics*.

Lifshitz, L., & Pizer, S. (1990). A multiresolution hierarchical approach to image segmentation based on intensity extrema. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *12*(6), 529–540.

Lindeberg, T. (1993). Detecting salient blob-like image structures and their scales with a scale-space primal sketch: a method for focus-of-attention. *International Journal of Computer Vision*, *11*(3), 283–318.

Lindeberg, T. (1994). *Scale-space theory in computer vision*. Dordrecht: Kluwer Academic.

Lindeberg, T. (1998a). Feature detection with automatic scale selection. *International Journal of Computer Vision*, *30*(2), 77–116.

Lindeberg, T. (1998b). Edge detection and ridge detection with automatic scale selection. *International Journal of Computer Vision*, *30*(2), 117–154.

Lindeberg, T., & Eklundh, J.-O. (1992). The scale-space primal sketch: construction and experiments. *Image and Vision Computing*, *10*, 3–18.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*.

Ma, Y. F., Lu, L., Zhang, H. J., & Li, M. J. (2002). An attention model for video summarization. *ACM Multimedia*, *12*.

Mallat, S. (1998). *A wavelet tour of signal processing*. New York: Academic Press.

Mallat, S., & Zhang, Z. (1993). Matching pursuit in a time-frequency dictionary. *IEEE Transactions on Signal Processing*, *41*, 3397–3415.

Marr, D. (1983). *Vision*. New York: Freeman.

Metropolis, N., Rosenbluth, M., Rosenbluth, A., Teller, A., & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, *21*, 1087–1092.

Mumford, D. B., & Gidas, B. (2001). Stochastic models for generic images. *Quarterly of Applied Mathematics*, *59*(1), 85–111.

Olsen, O., & Nielsen, M. (1997). Multi-scale gradient magnitude watershed segmentation. In *Proceedings of the international conference on image analysis and processing*. *Lecture notes in computer science*, Florence (pp. 6–13). Springer: Berlin.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*, 607–609.

ter Haar Romeny, B. M. (1997). *Front-end vision and multiscale image analysis: introduction to scale-space theory*. Dordrecht: Kluwer Academic.

Ruderman, D. L. (1994). The statistics of natural images. *Network*, *5*, 517–548.

Shokoufandeh, A., Dickinson, S., Jonsson, C., Bretzner, L., & Lindeberg, T. (2002). On the representation and matching of qualitative shape at multiple scales. In *Proceedings of the 7th European conference on computer vision* (pp. 759–775), Copenhagen.

Simoncelli, E. P., Freeman, W. T., Adelson, E. H., & Heeger, D. J. (1992). Shiftable multi-scale transforms. *IEEE Transactions on Information Theory*, *38*(2), 587–607.

Sporring, J., Nielsen, M., Florack, L., & Johansen, P. (1996). *Gaussian scale-space*. Dordrecht: Kluwer Academic.

Srivastava, A., Lee, A. B., Simoncelli, E. P., & Zhu, S. C. (2003). On advances in statistical modeling of natural images. *Journal of Mathematical Imaging and Vision*, *18*(1), 17–33.

Wang, Y., & Zhu, S. C. (2004). Modeling complex motion by tracking and editing hidden Markov graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Wang, Y., Bahrami, S., & Zhu, S. C. (2005). Perceptual scale space and its applications. In *Proceedings of the international conference on computer vision*.

Witkin, A. P. (1983). Scale space filtering. In *International joint conference on AI*. Palo Alto: Kaufman.

Wu, Y. N., Guo, C. E., & Zhu, S. C. (2007). From information scaling of natural images to regimes of statistical models. *Quarterly of Applied Mathematics*.

Xie, X., Liu, H., Ma, W. Y., & Zhang, H. (2003). Browsing large pictures under limited display sizes. *IEEE Transactions on Multimedia*.

Xu, Z. J., Chen, H., & Zhu, S. C. (2005). A high resolution grammatical model for face representation and sketching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, San Diego, June 2005.

Yao, Z. Y., Yang, X., & Zhu, S. C. (2007). Introduction to a large scale general purpose groundtruth dataset: methodology, annotation tool, and benchmarks. In *6th international conference on EMMCVPR*.

Zhu, S. C. (1999). Embedding gestalt laws in Markov random fields—a theory for shape modeling and perceptual organization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *21*(11), 1170–1187.

Zhu, S. C., & Yuille, A. L. (1996). FORMS: a flexible object recognition and modeling system. *International Journal of Computer Vision*, *20*(3), 187–212.

Zhu, S. C., Wu, Y. N., & Mumford, D. B. (1997). Minimax entropy principle and its applications to texture modeling. *Neural Computation*, *9*, 1627–60.