EUROGRAPHICS 2007 / D. Cohen-Or and P. Slavík
(Guest Editors)

Volume 26 (2007), Number 3

Diorama Construction From a Single Image

J. Assa and L. Wolf [1]

[1]School of Computer Science, Tel Aviv University

Abstract

Diorama artists produce a spectacular 3D effect in a confined space by generating depth illusions that are faithful to the ordering of the objects in a large real or imaginary scene. Indeed, cognitive scientists have discovered that depth perception is mostly affected by depth order and precedence among objects. Motivated by these findings, we employ ordinal cues to construct a model from a single image that similarly to Dioramas, intensifies the depth perception. We demonstrate that such models are sufficient for the creation of realistic 3D visual experiences.
The initial step of our technique extracts several relative depth cues that are well known to exist in the human visual system. Next, we integrate the resulting cues to create a coherent surface. We introduce wide slits in the surface, thus generalizing the concept of cardboard cutout layers. Lastly, the surface geometry and texture are extended alongside the slits, to allow small changes in the viewpoint which enriches the depth illusion.

Categories and Subject Descriptors (according to ACM CCS): I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism; I.4.8 [Image Processing and Computer Vision]: Depth cues

1. Introduction

Presenting a realistic view of a large scene in a small compact space has been the main goal of dioramas since their invention in the 19th century. Diorama techniques have included placing object-models at small depth changes, and using occlusion, parallax and light changes to reinforce the illusion of viewing a larger space. The goal of this work is to create a similar illusion from a single input image, i.e., to automatically construct a model in which depth can be realistically perceived by the observer.

The construction of a 3D scene from a single image, is a highly challenging problem mainly due to the obvious view/space ambiguities and the sparseness of the details that can be utilized to reconstruct the scene. Over the last few years, various effective techniques have been applied to this problem including, among others 'Tour Into The Picture' [HAA97] and 'Automatic Photo Popup' [HEH05].

While attempting to produce a model which resembles the original 3D scene, these methods often fit a simplified 3D model, such as piecewise planar models. Our technique focuses on a general model that provides depth sensation, but does not reconstruct the original 3D structure. I.e., we simplify the reconstruction problem by considering only the essential elements that assist the image depth perception, by allowing to change viewpoint and experience the parallax and occlusions between the various image objects.

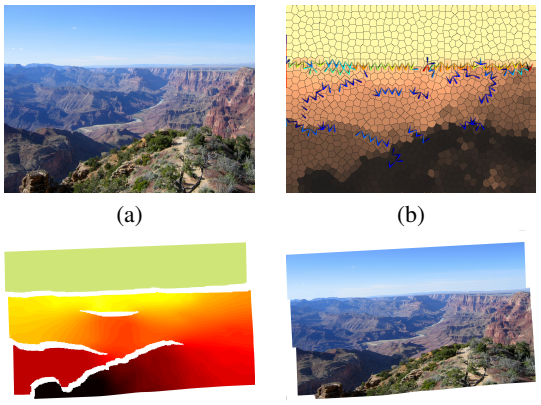

Figure 1: *Diorama construction from a single image. The original image (a); An example of the extracted relative cues (b), shown as arrow heads pointing further away from the viewer; The generated surface with slits (c); and a synthesized Diorama model (d).*

We scan the image for cues which convey *relative* depth

differences between the various objects in the image, and construct a depth-map by using quadratic optimization which is constrained by *inequalities*. This is in contrast to previous work that focus on absolute depth and constraints of the form of equalities. We build a piecewise smooth depth map surface, into which we insert wide slits expressing depth discontinuities. An example of a reconstructed surface with slits is shown in Figure 1 (c). To create a novel-view, we expand the surface geometry and texture beyond the slit edges. The resulting views express realistically the occlusions and parallax effects between the different parts of the surface.

More specifically, the contributions herein include:

- Describing a semi-automatic technique for building a model from a single image that conveys a depth experience to the viewer. This is accomplished despite the fact that an accurate and complete 3D model cannot be reproduced due to the inherent ambiguities and the sparse set of available depth cues.
- Presenting an analysis of the image and extracting a set of depth cues. While the basic ideas behind many of those cues are not new, we construct novel solutions which take advantage of our use of ordinal and sparse constraints.
- Introducing a novel optimization scheme for integrating relative depth cues and building a coherent slitted model.
- Enhancing the model by extending its geometry and texture to complete occluded regions and allow to perceive a continuous view for small viewpoint changes.

2. Related Work

Pictorial relief is the three-dimensional spatial impression obtained when one looks at a two-dimensional picture. Cues which contribute to the pictorial relief have been used by artists throughout time in paintings and sculptures. In 1822, the term *Diorama* was coined by Louis Daguerre, for a 3D scene replica showing historical events or nature scenes. The Diorama was designed to give the viewer an impression of full 3D, although the set is highly confined in its dimensions. Diorama models, which were later massively adopted in museums and in art, are based on making use of pictorial relief cues such as depth from lighting and shading [HG68].

Human perception of depth has been also studied by cognitive scientists. Koenderink and his colleges [KvDKT01, CM84] studied how perceptual relief can be measured and which aspects of it are especially robust against day-to-day intra observer variations. They argued that only aspects of the *partial* depth order and precedence in infinitesimal regions are actually used as stable cues. Their work had motivated our attempt to detect these ordinal cues in the image and to exhibit them in the resulting model.

The construction of an accurate 3D model from 2D pictures is an active research topic. Classical structure-from-motion approaches for 3D reconstruction, usually require two or more images of the scene. Over the last few years, additional techniques were introduced, which use special apparatus or stimuli to gain depth information, for example, close/far focused images, and using structured light. In the following review, we focus on studies for 3D modeling based on one casual image.

The work of Horry *et al.* [HAA97] is one of the early attempts in constructing 3D scene from a single image. Their 'Tour Into The Picture' technique extracts the 3D scene model from a 2D picture and makes a "walk through" animation of the 2D picture. In general, 'Tour Into The Picture' employs a spidery mesh over the picture to obtain a pseudo-3D scene model. This simple planes-based mesh is constructed from realization of the perspective through vanishing points and requires tedious user interaction. Their technique was extended by others [KS02], and recently Boulanger *et al.* [BBP06] suggested an automatic way for camera calibration that would allow to automatically detect vanishing points, by analyzing lines found in the image.

Sturm and Maybank [SM99] assembled the objects within the scene, by examining their corners, edges and planes for perspective constraints. Shesh and Chen [SC05], followed the edges of buildings to recover their shapes as cubes. Both of these works rely on human assistance and work especially well on man-made scenes, where the object facets and their correspondence are often simple and regular. For natural landscape images, such methods are not appropriate.

The system proposed by Zhang *et al.* [ZDPSS01] has significantly influenced our work. They propose a system which enables manual construction of free-form, texture-mapped, 3D scene models from a single painting or photograph. It allows users to specify a set of sparse constraints, including absolute depth, depth discontinuities, and absolute normal directions, and produces 3D models at interactive rates. Similar methods for manual mesh construction and modifications were also suggested by Kang [Kan98]. Our method, which is based upon constraints of a different nature, is geared towards an automatic model extraction.

Recently, Prasad *et al.* [PZF06] demonstrated a novel method for the construction of a surface from its apparent contour and potentially other types of information such as surface normals. They demonstrate remarkable results in the analysis of simple objects scenes. Extending this method to scenes with multi-objects such as casual pictures is not straightforward.

Statistical techniques were also used to infer depth from a single image. Han and Zhu [HZ03] proposed a two levels scheme in which objects are recognized and their interrelations are inferred from a set of training scenes with some degree of human assistance. Although they introduce both man made objects and natural objects such as plants, their model is limited to a predefined set of objects and relations. A more recent work by Hoiem *et al.* [HEH05], employed machine learning techniques for recovering a 3D 'contextual frame' of an image, containing the image major sur-

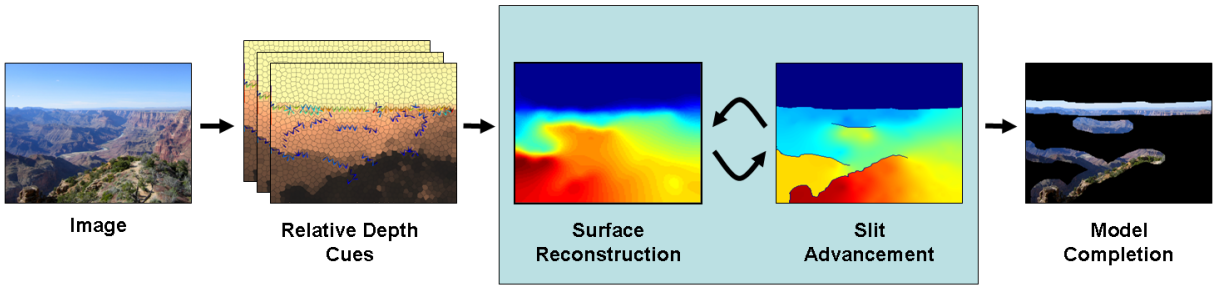

Figure 2: *An overview of our technique. Given an image, we analyze it using a set of relative depth cues. Next, we iteratively construct a surface with slits. This surface is completed alongside the slits, to allow for novel-view synthesis.*

face components, by classifying each pixel as being a part of the ground plane, the sky, vertical surface, and other sub classes. They demonstrated excellent results in recovering scenes composed on such elements. Subsequently, Saxena *et al.* [SCN06] used a Markov Random Field to model the depth relations among image patches, and learn the underlying statistical distributions from a large training set.

3. Overview

The general framework of our method is illustrated in Figure 2. It can be broken into three main stages: cues analysis, cues integration and the diorama construction.

The **cue analysis** (Section 4) consists of the extraction of ordinal depth constraints from cues such as texture and focus differences. Next, in the **cue integration** stage (Section 5) we integrate the cues into a depth map that satisfies the depth constrains. Slits are introduced along paths of large depth discontinuities, and the process is repeated iteratively. Finally, in the **diorama construction** stage (Section 6) , we allow the creation of novel-views by extending the image and the model geometry to make up for regions which are exposed by the introduction of the slits.

The depth cues which are used in our system are sparse, and are applied on image regions. Therefore, as a preliminary step, we segment the given image in two different levels. First, we apply a high level segmentation into **major segments**. Typically an image is segmented into approximately 10-20 major segments. In most cases these segments will result in coherent and smooth surface pieces. The second segmentation is much finer - it breaks the image into small image **superpixels** each consisting of regions of similar texture or color [RM03]. Using superpixels reduces the complexity of the input image and improves the coherence of the results . We have examined various segmentation algorithms for both segmentation levels including the ones described in [FH04, RM03], but no significant differences in our generated dioramas were observed.

The boundary between adjacent major segments defines

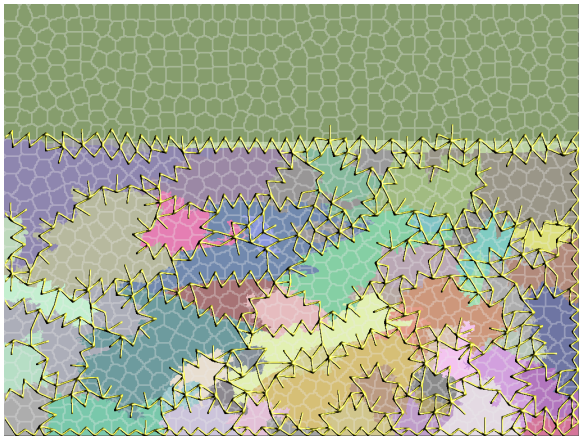

Figure 3: *An example of the major-segments map (colored patches) superimposed with the superpixel map (gray tiles), and their corresponding stitches (yellow lines).*

a list of **stitches** that connect pairs of adjacent superpixels which are not part of the same major segment. The stitches, (Figure 3), reduce the complexity of the following steps.

As a last preliminary stage we define the diorama backplane. Although this is optional, it improves the depth perception of distant objects. Since we focus on outdoor scenes, in most cases we select the sky as the back region, by using the simple sky detection algorithm of [LE02]. For other types of scenes or cases where the sky exhibits large color gradients, we select the backplane manually.

4. Cue Analysis

In his work, Cutting studies a comprehensive list of cues that influence the depth perception. He argues that most of the monocular cues are global in the sense that they cannot be observed by examining small neighborhoods, and that they generate mostly ordinal information with large amount of

ambiguities. It is the overall organization of the image, he claims, and the combination of the cues, that create the pictorial relief of the viewed image [CM84]. We follow these studies by integrating what Cutting refers to as "traditional pictorial cues": Partial occlusion between objects (interposition), height in visual field, relative size and density (texture gradient), aerial perspective (known today as environmental cue or atmospheric scattering). Additionally, we use depth from focus as a cue, following [MBA*96, WLGW01, PK06]. Currently, we do not include perspective and shading/lighting constraints directly, since their extraction in unconstrained scenery images is extremely challenging.

4.1. Partial Occlusion

Partial Occlusion (interposition) is one of the most perceptually-dominant depth cues. Here, we propose a straight-forward occlusion detector that is based on filtering many pairs of horizontal image edges and checking whether these two edges are the boundaries of an occluding object. In the case of an occlusion we can expect the appearance within these boundaries to vary considerably from the appearances outside. We can also expect the appearance just outside of the boundaries of the occluding object to be similar on both sides. This is illustrated in Figure 4.

To model appearances we define visual signatures $C_{sig}(p,H)$ of a point p on a line H in the following way. We consider a line perpendicular to H at p of length 70 pixels in each direction. We group nearby points of similar colors to generate a sequence of color clusters. The sequence of colors clusters along the line encodes the rough color structure of the line.

Every straight horizontal line H in the image crosses edges of major-segments (as defined in section 3) in points $e_1 \dots e_n$. Our algorithm enumerates over all pairs $\{(e_k,e_l), e_k < e_l\}$ and selects those for which the outer signatures at points $e_k - \varepsilon$ and $e_l + \varepsilon$ are similar and non uniform, while the inner signatures are different. More formally, we locate cases where the following hold:

$$\frac{Length(C_{sig}(e_k-\varepsilon,H) \cup C_{sig}(e_l+\varepsilon,H))}{d_{edit}\left(C_{sig}(e_k-\varepsilon,H), C_{sig}(e_l+\varepsilon,H)\right)} > Th_1$$
$$d_{edit}\left(C_{sig}(e_k-\varepsilon,H), C_{sig}(e_k+\varepsilon,H)\right) > Th_2$$
$$d_{edit}\left(C_{sig}(e_l-\varepsilon,H), C_{sig}(e_l+\varepsilon,H)\right) > Th_2$$

In the formula above, ε is taken as the largest edge width (usually considered as 15 pixels), and Th_1, Th_2 are thresholds used to control the number of cues generated. Edit distance is used to compare two sequences, in which a constant cost is used for insertion and deletions, and a variable cost is defined for modification based distance between colors.

We apply this approach on a grid of horizontal lines for several image rotation. To reduce the number of false-positives we use only points which exhibit the required properties in several nearby rotations. We assume that a persistent detection always indicates an occluding object. Note that this assumption sometimes fail since holes or painted textures may also be detected. We therefore include partial occlusion constrains only if they are validated by other constraints.

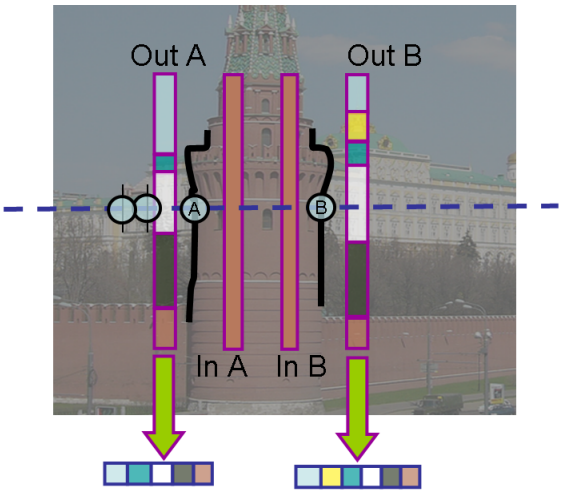

Figure 4: *Partial occlusion cue analysis. We look for similar color appearances outside an edge pair.*

4.2. Texture Analysis

In images with large homogeneous regions, depth difference can be detected reliably using texture gradients [MR97]. However, it is often found to be impractical for our purposes due to non-homogeneous major segments and high level of noise. Instead, we focus on more rough cues that can be analyzed from smaller regions. Namely, that similar textures found in different scaled-down versions of two different regions may indicate that such regions have the same texture at different depths.

To describe texture appearances, we employ the textons of [MBLS01] as a basic building block. We apply a filter bank composed of Gabor filters in different orientations and scales. We then cluster the filter responses into a small set of prototype response-vectors called textons and examine their histograms within small patches. Such texton-histograms are also computed for various scaled-down versions of the image. We detect scale differences, by comparing the local texton distribution in one scale to the texton distributions in another. A match (across the image and between scales) indicates a scaling effect that we attribute to perspective. We validate each match by repeating the comparison for scaled down versions of the matching regions as shown in Figure 6. We search for depth differences among all superpixels participating in stitches and in major segments' centers. We use as relative cues only patches with sufficient superpixels support, and minimal contradicting evidance.

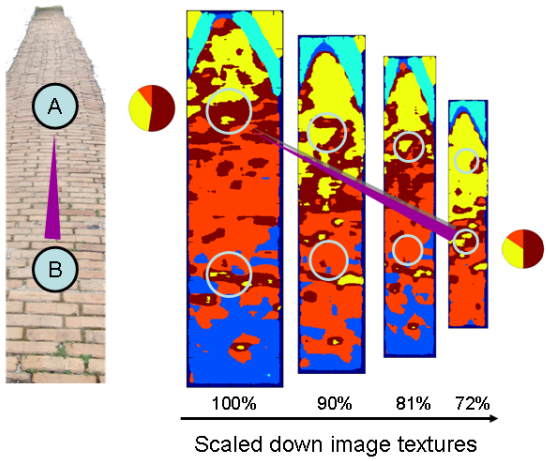

Figure 5: *Texture analysis. We create scaled-down versions of the image shown on the left, and examine the texture histogram similarities between different points at different scales. In this example, for visualization purposes, we clustered the texton histograms space into 6 groups each marked with different color. As shown, at a certain scale ratio the texton histograms in regions A and B becomes similar.*

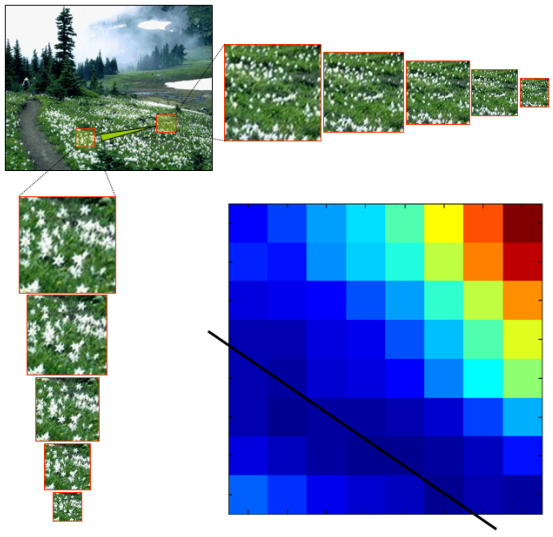

Figure 6: *Validation of relative depth. The histogram distances between several scaled down versions of two regions are measured and shown in the middle affinity matrix representation. Blue rectangles indicate possible matches between histograms in different scales. The black line illustrates our verification technique which examines the difference also in smaller scales.*

4.3. Depth from Focus

The amount of blur on an object provides a remarkable depth sensation. We use two complementary methods for extracting depth information from the image depth of field (*DOF*). Also, because DOF measurements are potentially unreliable, we compute these for every super-pixel, but only consider depth constraints for stitches between major segments, if and only if there is a high level of agreement between the stitches.

The first DOF measurement is inspired by the work of Park and Kim [PK06], and it examines variances in neighboring pixels. We define η as the set of neighboring pixels for each point and I_c as the intensity map for the color channel c. We calculate $\tilde{m} = \underset{p\in\eta}{Mean}(I_c(p))$. Next, we calculate an effective blur measure for each pixel:

$$I_{DOF_1} = \underset{c\in channels}{Max}\left(\frac{1}{|\eta|}\sum_{p\in\eta}(I_c(p)-\tilde{m}_c)^4\right) \qquad (1)$$

The second method examines a larger neighborhood by applying a 'Difference of Gaussians' (DoG) operator to quantify the amount of blur in the various layers of the image. We use Gaussians with $\sigma_1 = 5$ and $\sigma_2 = 0.5$ and supports of size 20×20 and 10×10 respectively. Note that more elaborated methods exist, such as the work of Wang *et al.* [WLGW01]. On our image set, the simple DOG methods had proved to be sufficient.

To construct a depth cue from either DOF measurement, we assume the image sharpest focus is on the nearest image object. This assumption holds for many natural images. A more elaborate model, which uses the sharpness of the edge between the pairs of objects to determine their order [Mat97] was examined but had failed to deliver consistent reliable results.

4.4. Atmospheric Scattering

Atmospheric scattering is the phenomena of light degradation over the atmosphere. There is a simple, albeit non-linear, relationship between the radiance of an image at any given wavelength and the distance between object and viewer. To use this cue, we apply a method similar to the one suggested by [NN02]. In their work, they recover the ratio of depth of three locations, assuming that the viewed image intensity is homogeneous, and with a known value of the sky intensity L_s (sky intensity is considered as the area in which objects are indistinguishable). Given 3 regions with similar texture in different depths d_1, d_2, d_3, and corresponding image luminance L_1, L_2, L_3, the following holds:

$$\frac{d_1-d_2}{d_2-d_3} = \frac{Log\left(\frac{L_1-L_s}{L_2-L_s}\right)}{Log\left(\frac{L_2-L_s}{L_3-L_s}\right)} \qquad (2)$$

We found the RGB blue channel luminance to provide a

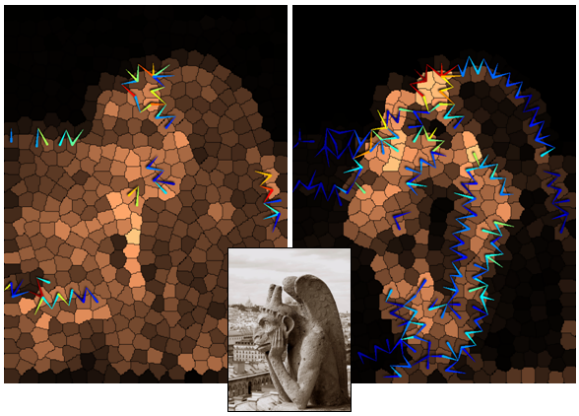

Figure 7: *'Depth from Focus' cues. The resulting cues are presented on top of superpixels heat map - darker superpixels indicate low depth of focus. The left image shows results for method I_{DOF_1} and the right one for DoG. The source image is shown on the middle.*

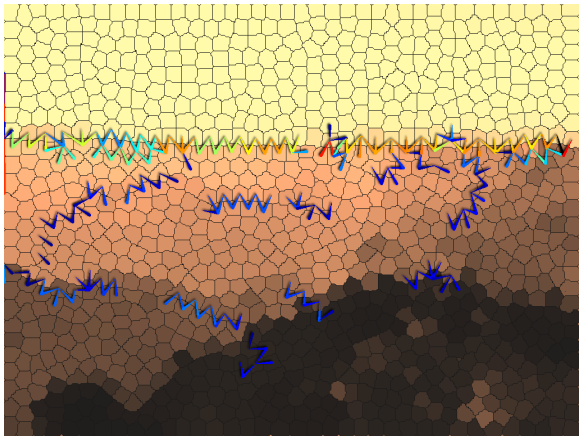

Figure 8: *Atmospheric scattering cues. This cue detects differences in superpixel average luminance in a smoothed blue channel of the image.*

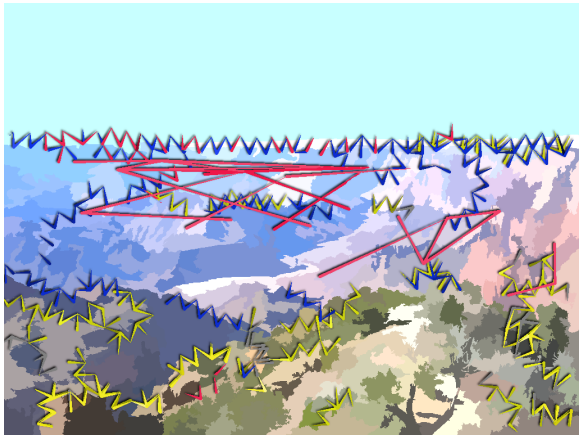

Figure 9: *Retrieved cues. The cues gathered for this example are shown in different colors: Red indicates depth of field, Blue shows scattering cues, and yellow represents texture. Arrow directions indicate depth (arrow base is closer)*

good estimate of the scattering effect. To allow automatic analysis of this cue, we make the following assumptions: The sky intensity is measured from the defined Diorama backplane region, or assumed to be the highest luminance value and the average luminance value is considered to be associated with the average-depth point in the scene. Cue detection is being done by examining differences between superpixels with uniform luminance participating in stitches. Relative information can be retrieved from Eq. 2 where L_1 and L_2 are acquired as the luminance of the superpixels and L_3 is the average luminance. To reduce the number of false-positives, we smooth the blue channel noise while retaining its edges, by applying bilateral filtering [TM98], and compare only superpixels with uniform luminance.

4.5. Height in Visual Field

Psychophysical studies have shown that lower objects in the visual field are perceived nearer than higher ones [CM84]. We apply this cue between a pixel and the three pixels beneath it (applying to pairs of top and bottom pixels results in vertical lines artifacts). At occlusions boundaries this cue is not reliable, therefore we do not apply it at the edges of major segments.

5. Cue Integration

Cue integration in the human vision has been the subject of many studies. Although some advances were made, the synthesis method still remains unknown. Some studies had observed that the human vision cue preference is adaptive – each cue's reliability weight is proportional to its previous success in resolving the correct depth [Tod04]. This suggests the use of learning techniques in order to determine, based on similar images, the relevance of each type of depth cue. Learning methods, however, require a large training set which we do not currently have. Instead, we assign each cue a priority based on its stability in preliminary experiments and provide the user with an option not to use a specific type of cue on an input image. The assigned priorities are used to resolve contradictions. We employ the following order: Atmospheric scattering > defocus > partial occlusion > texture > height in the visual field. The user ability to veto a certain type of cue is the main aspect in which our system is semi-automatic rather than automatic. In the supplementary material we specify when such intervention was used.

Even if all cues are appropriate for a given image, our set

of constraints is usually still sparse, and the reconstruction problem is ill-posed. To reduce the space of possible solutions, we seek a surface with the following properties:

- It is piecewise smooth, where we encourage continuity anywhere except for along a minimal number of slits.
- It has as little depth differences as possible, i.e., we give preference to surfaces that are parallel to the image plane.
- Its maximal depth, defined as the depth from the frontal objects to the background region, is limited.

To generate the surface, we apply several iterations of surface construction and slit creation phases. In each iteration we measure the total surface gradient energy. The process halts when the surface is sufficiently smooth, or after a couple of iterations.

5.1. Surface Construction

The basic construction of the smooth surface is done using quadratic thin-plate energy, in a similar manner to [ZDPSS01, PZF06]. There are two major differences, though: (1) Our constraints are mainly inequalities and (2) We minimize the first spatial derivatives of depth as well as the second derivatives. The first difference allows the introduction of ordinal cues, and the second promotes surfaces that are as parallel to the image plane as possible. Specifically, we compute depth $d_{i,j}$ for each point in the image by minimizing the sum of the following energy score over all pixels (i,j) not on slits :

$$[(d_{i-1,j}-2d_{i,j}+d_{i+1,j})^2+(d_{i,j-1}-2d_{i,j}+d_{i,j+1})^2+ \\ (d_{i,j-1}-d_{i,,j+1})^2+(d_{i-1,j-1}-d_{i+1,,j+1})^2+ \\ (d_{i-1,j}-d_{i+1,,j})^2+(d_{i-1,j+1}-d_{i+1,,j-1})^2]$$

The minimization is performed subject to several constraints. First, for every point (i,j) on the back-plane (see Section. 3) we set $d_{i,j}=0$. Second, we have several constrains for points (i,j) not on image slits: $\left(d_{i,j}-d_{i,j+1}\right) < MaxGradient$, $\left(d_{i,j}-d_{i+1,j}\right) < MaxGradient$, $d_{i,j} < MaxDepth$, where $MaxGradient$ and $MaxDepth$ are constants controlling the generated diorama depth. Third, for each ordinal cue between image-points a and b we set $d_a+\text{diff} < d_b$, where diff is the minimal difference in their depth, determined by the magnitude of the cue, e.g., it is proportional to the scale difference in the texture cue, and to the amount of scattering the atmospheric cue. Lastly, based on the height in visual field cue, we define the following term for each pixel which is not on the border of a major segment: $d_{i,j} > \frac{1}{3}\left(d_{i-1,j+1}+d_{i,j+1}+d_{i+1,j+1}\right)$.

Sometimes, the set of constraints is not feasible, i.e., it contains contradictions. To resolve those cases we construct a directed graph with one vertex per superpixel and one edge per cue. A contradiction would manifest itself as a cycle in this graph. We therefore detect cycles and remove the cue of the cycle-edge with the lowest priority. Another step we take to make the above computation more robust is to examine the resulting depth map, and remove cues that are near points where the depth gradient is excessively high.

5.2. Slit formation

In this stage we create the surface slits in areas of high depth gradient. The slits should match the edges of the image objects and should be as continuous and smooth as possible. To establish such properties, we have implemented a curve evolution algorithm which considers the existing strong edges in the image, the depth gradient and existing slits, by mimicking a known physical model of crack propagation.

The likelihood of a point to participate in a crack depends on external pulling forces (E), its local strength (S) and the location of nearby cracks, through a stress function σ. Specifically, this likelihood grows with the following **tear potential**: $(S+\sigma)E$.

By crack propagation theory [Ric68] directional stress σ_x, σ_y at distance r and angle θ from the crack tip is calculated as:

$$\sigma_x = \frac{C\sqrt{a}}{\sqrt{2\Pi r}}\cdot\cos\frac{\theta}{2}\cdot\left(1+\sin\frac{\theta}{2}\cdot\sin\frac{3\theta}{2}\right)$$
$$\sigma_y = \frac{C\sqrt{a}}{\sqrt{2\Pi r}}\cdot\cos\frac{\theta}{2}\cdot\left(1-\sin\frac{\theta}{2}\cdot\sin\frac{3\theta}{2}\right)$$

where a is the current crack length, and C is a material dependent constant. Such formulas are only accurate starting some distance from the crack-tip. We therefore, in accordance with crack propagation theory, inhibit the calculated stress in a small region near the tip (we inhibit for $r<5$). For simplicity, we consider $\sigma = ||\sigma_x, \sigma_y||$.

We define the strength of each image point as inversely proportional to its edge magnitude, computed via the compass image operator [RT01]. We therefore promote slit creation on object boundaries. An example of such edge map is shown in Figure 10. As a tearing force we take the magnitude of the gradient of the depth map, recovered as in Section 5.1. A high gradient in the depth field may indicate the need to tear the surface at that location.

We extend existing slits or create new ones using a simple greedy process. At each step we select the point p with the highest tear potential. We connect it to the closest slit tip, if nearby, or create a new crack. In both cases p becomes a tip-end. When the last updated crack becomes longer than 100 pixels, we terminate the crack advancement for this cycle and recompute the depth map. This re-computation is required since once cracks (i.e., slits) are introduced the tear potential becomes invalid.

The introduction of slits reduces the depth map gradients in the next cycle. The iterative two step process, of computing the depth map and of introducing slits terminates once the total magnitude of gradients in the depth map drops below a certain threshold.

6. Model Extension And Diorama Viewer

Modifications to the image view-point can create prominent depth illusions by using motion parallax. The main difficulty

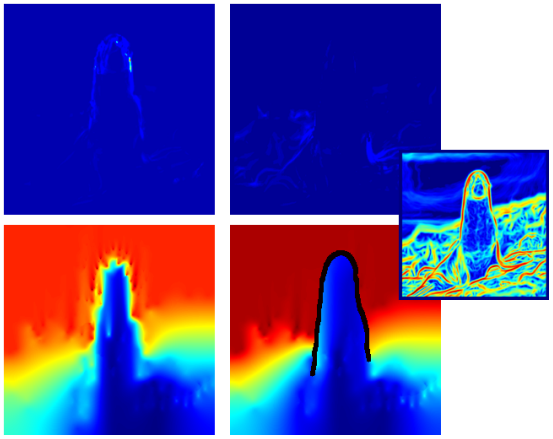

Figure 10: *Slit creation. The computed model is slitted shown in the lower depth maps, (initial -left and result- right). The slit advancement is guided by the surface tension (depth gradient) shown in the upper heat maps (initial -left and result- right), and the image edges (middle right).*

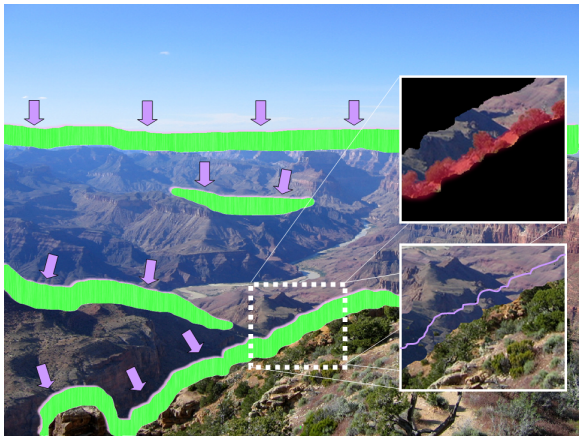

Figure 11: *Surface completion alongside the slits. We complete surfaces that are occluded in original viewpoint (shown in green, example in the lower closeup) by extending the image and the geometry in the direction of the purple arrows. We also extend the occluding surface, matting the slit edge, (shown in red in the upper closeup).*

in creating such novel-views given the original image is the need to present information that is occluded in the original view but exposed in the new one. To overcome this problem we extend both the occluding surface and the occluded one alongside each slit edge. Figure 11 shows an example.

We expand the geometry and texture of the occluded object by using image completion techniques. The maximal required completion surface width along the edge, referred to as the required strip, can be easily calculated given the desired viewpoint change and its distance to the edge. Since we require both the geometry of the strip and its texture to reflect the occluded object, we utilize a variant of a fragment based image completion [DCOY03], in which we encourage the selection of fragments that belong to the occluded object. This is done by considering weights for each fragment candidate that are proportional to their depth similarity, i.e., a large difference in depth reduces the likelihood for selecting a fragment. We extend the surface geometry in a similar manner, by using the calculated depth map as the source for the selected fragments. The resulting extended surface is therefore constructed by elements of the original image texture and geometry, creating a plausible result. Alternative surface completion methods include [SACO04].

To enhance the realism of the slit, we use the Poisson Matting technique [SJTS04] to smooth the occluding object jagged boundaries. We define a narrow band (6-10 pixels wide) over the slit, as the 'unknown region' and the foreground and background as the two edge sides.

7. Results

Our method was tested on a multitude of casual photographs of various subjects including panoramic views of both natural and urban locations and closeups on objects such as monuments, statues and animals, several are presented in Figure 12. The Dioramas depth illusion can be viewed in the submitted supplementary material. In general, we found our technique suited for images of outdoor scenes with minimal regular patterns or straight lines. Straight parallel lines and straight angles can be distorted by our technique, creating disturbing artifacts, since we do not currently enforce such constraints. Diorama creation in its current form is therefore complementary to other methods specializing in urban scenes [HAA97, HEH05], or alternately can be extended by integrating algorithms for detecting 3D model from image edges [PZF06].

The presented results were created mostly in Matlab. The Mosek [ART03] solver was used for the optimization problem of Section 5.1. At its current form the complete end-to-end process of generating the model takes about half an hour for a 2 megapixels image, however we estimate that at least an order of magnitude can be gained from optimizing the code. The main computational hurdle is the computation of cues such as partial occlusion and texture analysis, which perform better in high resolutions. Carefully subsampling the image, may improve running times.

8. Conclusions and Future work

We propose a semi-automatic way to generate 3D models that can induce a realistic depth perception. Since the input is

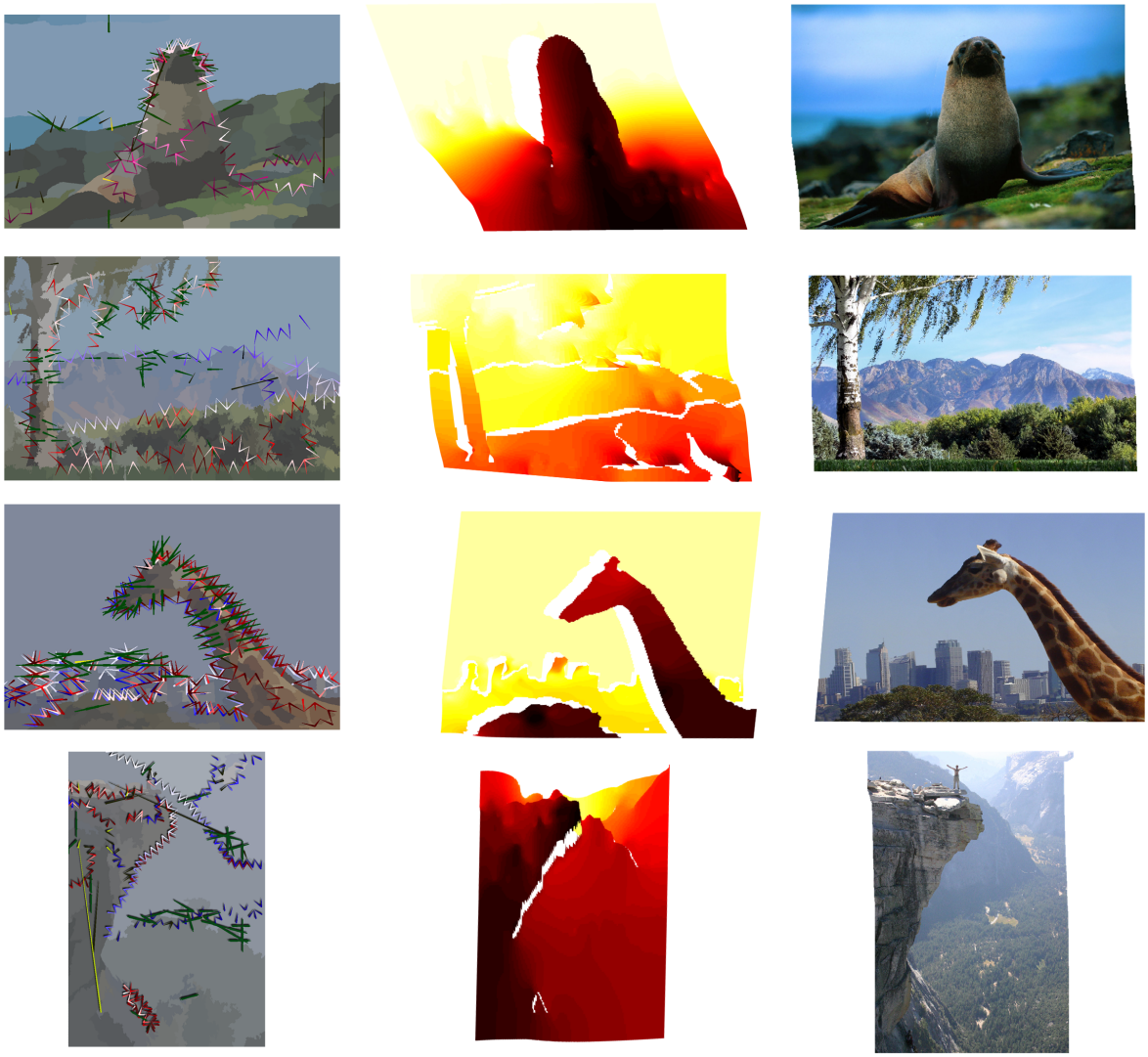

Figure 12: *Several examples of generated models: The detected relative depth cues (left); the generated depth map (middle); and the resulting diorama from an oblique viewpoint (right). For the diorama illusion examine the supplementary video*

a single casually taken photograph, such Diorama visualizations can be integrated into image sharing applications or image directories to enrich the users' experience. The method can also be used to create realistic stereo pairs from a single image that can be printed using lenticular printers and serve as a cheap substitute to current dioramas. 3D images and videos for digital 3D displays can also be created.

This work can be extended in several directions: Our current implementation is only semi-automatic and near-perfect results may require some minimal user intervention which modifies the relative strength of the various cues according to the image characteristics. This task can be automated by using machine learning techniques, so that the a combined solution may not require any user intervention over the full spectrum of input images. In addition, a hybrid method which relies on both relative and quantitative cues can prove to be more robust.

Recovering only specific aspects of the 3D scene while not attempting to build a full reconstruction of the original, provides new possibilities in redefining view based modeling quality. The quality of the result can be judged on how realistic it looks, instead of measuring how close it is to the originating scene. This quality definition opens a door to new reconstruction algorithms, in a similar manner that non-photo

realistic rendering (NPR) opens new possibilities of rendering models while conveying some of the original model attributes. Additionally, our current method uses parallax and occlusion as a mean to highlight 3D perception. This can be extended to additional effects such as lighting and shadowing, thus increasing the 3D perception by the viewer.

9. Acknowledgments

We would like to thank Daniel Cohen-Or for his assistance and ideas. LW is supported by the Israel Science Foundation (grants No. 1440/06, 1214/06), and by the Colton Foundation. We also thank the following Flickr web site users for allowing us to reproduce their photographs, under Creative Commons licenses: winkyintheuk, Pierre Pouliquin, jotor, ama photography, jespahjoy, Argenberg.

References

[ART03] ANDERSEN E. D., ROOS C., TERLAKY T.: On implementing a primal-dual interior-point method for conic quadratic optimization. *Math. Programming 95*, 2 (2003).

[BBP06] BOULANGER K., BOUATOUCH K., PATTANAIK S.: Atip: A tool for 3d navigation inside a single image with automatic camera calibration. In *EG UK Theory and Practice of Computer Graphics 2006* (2006).

[CM84] CUTTING J. E., MILLARD R. T.: Three gradients and the perception of flat and curved surfaces. *J Exp Psychol Gen 113*, 2 (June 1984), 198–216.

[DCOY03] DRORI I., COHEN-OR D., YESHURUN H.: Fragment-based image completion. *ACM Trans. Graph. 22*, 3 (2003), 303–312.

[FH04] FELZENSZWALB P. F., HUTTENLOCHER D. P.: Efficient graph-based image segmentation. *Int. J. Comput. Vision 59*, 2 (2004), 167–181.

[HAA97] HORRY Y., ANJYO K.-I., ARAI K.: Tour into the picture: using a spidery mesh interface to make animation from a single image. In *ACM SIGGRAPH* (1997), pp. 225–232.

[HEH05] HOIEM D., EFROS A. A., HEBERT M.: Geometric context from a single image. In *International Conference of Computer Vision (ICCV)* (October 2005), IEEE.

[HG68] HELMUT, GERNSHEIM A.: *L.J.M. Daguerre: The History of the Diorama and the Daguerreotype*. Dover Pub., 1968.

[HZ03] HAN F., ZHU S.-C.: Bayesian reconstruction of 3d shapes and scenes from a single image. In *Int. Workshop on Higher-Level Knowledge in 3D Modeling and Motion Analysis* (2003).

[Kan98] KANG S.: *Depth painting for image-based rendering applications*. Tech. report, CRL, Compaq Research Lab, 1998.

[KS02] KANG H. W., SHIN S. Y.: Tour into the video: image-based navigation scheme for video sequences of dynamic scenes. In *Proc. ACM symp. on Virtual reality software and technology* (New York, NY, USA, 2002), ACM Press, pp. 73–80.

[KvDKT01] KOENDERINK J. J., VAN DOORN A. J., KAPPERS A. M., TODD J. T.: Ambiguity and the 'mental eye' in pictorial relief. *Perception 30*, 4 (2001), 431–448.

[LE02] LUO J., ETZ S. P.: A physical model-based approach to detecting sky in photographic images. *Image Processing, IEEE Transactions on 11*, 3 (2002), 201–212.

[Mat97] MATHER G.: The use of image blur as a depth cue. *Perception 26*, 9 (1997), 1147–1158.

[MBA*96] MARSHALL J. A., BURBECK C. A., ARIELY D., ROLLAND J. P., MARTIN K. E.: Occlusion edge blur: a cue to relative visual depth. *Intl. J. Opt. Soc. Am. A 13*, 4 (1996).

[MBLS01] MALIK J., BELONGIE S., LEUNG T., SHI J.: Contour and texture analysis for image segmentation. *International Journal of Computer Vision V43*, 1 (June 2001), 7–27.

[MR97] MALIK J., ROSENHOLTZ R.: Computing local surface orientation and shape from texture for curved surfaces. *Int. J. Comput. Vision 23*, 2 (June 1997), 149–168.

[NN02] NARASIMHAN S. G., NAYAR S. K.: Vision and the atmosphere. *Int. J. Comput. Vision 48*, 3 (2002), 233–254.

[PK06] PARK J., KIM C.: Extracting focused object from low depth-of-field image sequences. In *Visual Communications and Image Processing* (2006), vol. 6077, SPIE.

[PZF06] PRASAD M., ZISSERMAN A., FITZGIBBON A. W.: Single view reconstruction of curved surfaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (June 2006), vol. 2, pp. 1345–1354.

[Ric68] RICE J.: Mathematical analysis in the mechanics of fracture. In *Fracture*, Luxmoore H., (Ed.). Academic Press, 1968.

[RM03] REN X., MALIK J.: Learning a classification model for segmentation. In *Proc. 9th Int'l. Conf. Computer Vision* (2003).

[RT01] RUZON M. A., TOMASI C.: Edge, junction, and corner detection using color distributions. *IEEE Trans. Pattern Anal. Mach. Intell. 23*, 11 (2001), 1281–1295.

[SACO04] SHARF A., ALEXA M., COHEN-OR D.: Context-based surface completion. *ACM Trans. Graph. 23*, 3 (2004).

[SC05] SHESH A., CHEN B.: Peek-in-the-pic: Architectural scene navigation from a single picture using line drawing cues. In *Pacific Graphics 2005* (Oct 2005).

[SCN06] SAXENA A., CHUNG S. H., NG A.: Learning depth from single monocular images. In *Advances in Neural Information Processing Systems 18*, Weiss Y., Schölkopf B., Platt J., (Eds.). MIT Press, Cambridge, MA, 2006, pp. 1161–1168.

[SJTS04] SUN J., JIA J., TANG C.-K., SHUM H.-Y.: Poisson matting. *ACM Transactions on Graphics 23*, 3 (2004).

[SM99] STURM P., MAYBANK S.: A method for interactive 3d reconstruction of piecewise planar objects from single images. In *British Machine Vision Conference* (1999), pp. 265–274.

[TM98] TOMASI C., MANDUCHI R.: Bilateral filtering for gray and color images. In *Int. Conf. Computer Vision* (1998).

[Tod04] TODD J. T.: The visual perception of 3d shape. *Trends Cogn Sci 8*, 3 (March 2004), 115–121.

[WLGW01] WANG J. Z., LI J., GRAY R. M., WIEDERHOLD G.: Unsupervised multiresolution segmentation for images with low depth of field. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 23*, 1 (2001), 85–90.

[ZDPSS01] ZHANG, DUGAS-PHOCION G., SAMSON J. S., SEITZ S. M.: Single view modeling of free-form scenes. In *IEEE Conf. Computer Vision and Pattern Recognition* (2001).