Vision, Modeling, and Visualization (2016)
M. Hullin, M. Stamminger, and T. Weinkauf (Eds.)

PEEP: Perceptually Enhanced Exploration of Pictures

Marco Agus [1,2], Alberto Jaspe Villanueva [1], Giovanni Pintore [1], Enrico Gobbetti [1]

[1] CRS4, Visual Computing Group, Italy – http://www.crs4.it/vic/
[2] King Abdullah University of Science and Technology (KAUST), Visual Computing Center (VCC), Thuwal 23955-6900, Saudi Arabia

Abstract

We present a novel simple technique for rapidly creating and presenting interactive immersive 3D exploration experiences of 2D pictures and images of natural and artificial landscapes. Various application domains, ranging from virtual exploration of works of art to street navigation systems, can benefit from the approach. The method, dubbed PEEP, is motivated by the perceptual characteristics of the human visual system in interpreting perspective cues and detecting relative angles between lines. It applies to the common perspective images with zero or one vanishing points, and does not require the extraction of a precise geometric description of the scene. Taking as input a single image without other information, an automatic analysis technique fits a simple but perceptually consistent parametric 3D representation of the viewed space, which is used to drive an indirect constrained exploration method capable to provide the illusion of 3D exploration with realistic monocular (perspective and motion parallax) and binocular (stereo) depth cues. The effectiveness of the method is demonstrated on a variety of casual pictures and exploration configurations, including mobile devices.

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

The enormous proliferation of digital cameras and photo-sharing websites and services, such as Instagram, Google Photos, Facebook, and Twitter, has led during the last years to an exponentially growing number of pictures created, captured and shared over the Internet [LPC*15]. The emergence of computational photography and digital imaging techniques is also radically changing the way people captures and shares images, since the vast majority of shared pictures are nowadays heavily processed after optical recording. Digital processing is increasingly expected to work "automagically", and is not limited to in-camera filters for deblurring, color correction, and artistic effects, but also covers off-line automatic post-processing methods, such as Google's *Auto-Awesome* image enhancement feature, which automatically processes images uploaded to the cloud to create novel contents, such as panoramic views and short looping videos. In this context, providing tools to interactively and dynamically explore landscape photographs in order to create engaging viewing experiences is a very active area of multimedia research. The illusion of movement can be created, for instance, by adding stochastic motion textures to a scene containing passive elements, such as sea-waves, smoke and foliage to mimic the motion effect of their response to natural forces [CGZ*05], or by synthesizing the motion of clouds in a picture's sky [JC16]. Vivid and interesting interactive visuals can be obtained by associating a 3D structure to the still image, in order to provide additional depth cues, e.g., for stereo viewing, and/or permit a virtual 3D exploration of the 2D world. However, most of the methods attempt depth estimation and 3D reconstruction, and, to reach this goal, they have to use further ground-truth information coming from training data [HEH05, FNPS15], 3D point clouds [KNC*08], or photo collections [ZGW*14]. To cope with the complexity of faithful 3D estimation, we shift the focus from the extraction of a realistic 3D scene model from still images to the definition of a minimal image-dependent 3D structure capable to provide enough depth cues to emulate the qualitative experience of 3D immersion and navigation in the imaged landscape.

Our work focuses on exploiting the perceptual characteristics of the human visual system [Erk15] in the detection of perspective angles and in the depth discrimination to generate novel believable views for single zero point and one point perspective pictures representing without requiring a costly and difficult detailed depth estimation (see Sec 3). We present here a simple method which, starting from a single picture with a dominant 0- or 1-point perspective, automatically computes a simple parametric 3D representation matching with the perspective information contained inside the photo. The obtained representation can then be interactively and naturally explored in a 3D way, while providing perceptually consistent depth cues to the observers. Our main novel contributions are the following:

- an effective fast automatic method for fitting a simplified 3D parametric representation of the perspective space to a zero-point or one-point picture (see Sec. 4); the method exploits a multilevel

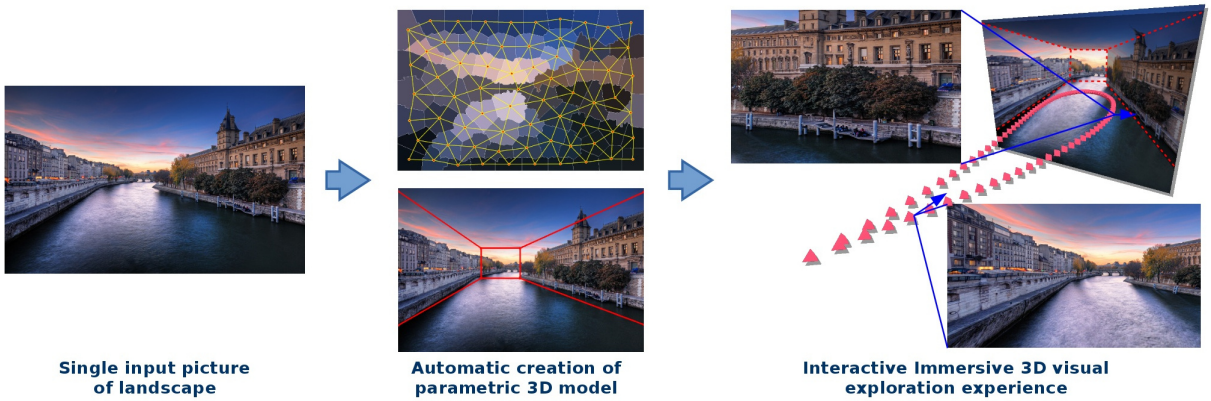

Figure 1: ***PEEP overview.*** *Taking as input a single image with zero or one vanishing points, such as those representing natural or urban landscapes, we automatically create a simple but perceptually consistent parametric 3D representation of the viewed space, which is used to create interactive and immersive 3D experiences with realistic depth cues.*

superpixel representation of the image for supervised labeling by a graph-cut energy minimization scheme;

- a constrained perceptually-motivated interactive 3D exploration method of landscape scenes enriched with a parametric representation of the linear perspective cues (see Sec. 5); the method constrains the viewer within the allowable space defined by the original 2D image, and is capable to provide the illusion of 3D exploration with realistic monocular (perspective and motion parallax) and binocular (stereo) depth cues.

To the best of our knowledge, this is the first system exploiting just the linear perspective cues for creating virtual 3D exploration experiences starting from still images. We tested the PEEP system on various devices and displays for a variety of natural and artificial landscape scenes (see Sec. 6).

2. Related work

Providing realistic immersive 3D exploration of still pictures requires the combination of image analysis techniques with user interfaces for enabling constrained inspection. We discuss here the works which most closely relate to our method.

Constructing 3D from single image. The construction of an accurate 3D model from 2D pictures is a very active research topic. Classical structure-from-motion approaches for 3D reconstruction, usually require two or more images of the scene: we refer readers to the survey from Seitz et al. [SCD*06]. Shape-from-shading methods [JMBB15] recover the depth of objects under the assumption of Lambertian surfaces. Our work focuses on single-image techniques: to this end, the method Tour Into The Picture [HAA97] represents the seminal work in this field and the first attempt to extract a 3D scene model from a 2D picture, by manually fitting a spidery mesh over the picture to obtain a pseudo-3D scene model. The technique was successively extended by others, to work with panoramic images [KPAS01], or to automatically derive vanishing points [BBP06], or to automatically recognize building edges and corners for architecture landscapes [SC05]. All of these methods rely on human assistance and manual work and are mostly suited for regular man-made scenes, but not for natural images and they still suffer from evident artifacts during navigation. More recently, various methods have been proposed to improve 3D reconstruction, by integrating additional information, like laser scanned 3D data [KNC*08], or training set data for labeling scene parts [HEH05] or for estimating scene depth [SSN09], or large scale photo collections [FNPS15,RNR*16]. All these methods obtain excellent results, but they need access to further external data, and they are not practical for implementation on mobile architectures. However, instead of considering precise metric 3D reconstruction, other methods consider the human perceptual system to construct realistic 3D scenes. In this category, Assa and Wolf [AW07] studied depth perception order between objects to develop a method for semi-automatic diorama generation by generalizing the concept of cardboard cutout layers. The method works well when there is clear depth separation between foreground and background, and it can be considered orthogonal to our technique. Depth ordering has been also recently exploited by Zhao et al. [ZHC15] for constructing layered scene models from single hazy images of outdoor scenes, and by Zeng et al. [ZCW*15] for hallucinating a rough approximation of the scene's 3D model by starting from a segmented image and using a number of simple depth and occlusion cues and shape priors. Our method starts from different perceptual considerations with respect to Assa et al. [AW07], and it constructs a perceptually plausible parametric 3D scene representation of perspective pictures without the need of further prior information. The resulting scenes can be interactively explored in a constrained workspace in various setups.

Constrained exploration. Since the genesis of computer graphics, researchers and engineers have been tackling the problem of efficient and intuitive exploration of 3D scenes. According to the application domain, the characteristics of the system, the input device, and the scene to explore, different camera controls can be designed. For an extensive survey on the subject, we refer the reader to Jankowski et al. [JH15]. Classic camera motion control ranges from unconstrained navigation such as the virtual trackball [Sho92], to various form of user assistance, such as constrained navigation over the surface of objects [KKS*05, MBB*14, Bou14], automated viewpoint computation [RU14], and camera path planning [LC15]. However, most of these approaches require that the user has direct control over the visualization space, but this solution can be

ineffective when the visualization area is very small, like it happens in smartphones or tablets. To solve this problem, Decle and Hachet [DH09] proposed an indirect method based on strokes for moving 3D objects in a touch screen mobile phone. Our technique generalizes the latter method by employing a constrained interaction based on effective view parametrization [LC15] and able to define an interaction workspace inside which the users can explore the enhanced scene in a perceptually plausible way, and create meaningful camera paths with a limited number of inputs.

3. Perceptual motivation

Our work was mainly influenced by the recent discoveries in perception applied to 3D understanding of the environment [Erk15]. In general, since scenes are 3D distributions of matter while images are only 2D distributions of colors, extracting the 3D structure of an image consists of creating a depth interpretation, commonly called *pictorial relief* [KvDKT01], which was largely exploited by artists throughout time in paintings and pictures. However, the qualitative experience of pictorial space under normal viewing is different from that obtained when the original real scene is viewed with both eyes [VVD14], and the deficiency of pictorial depth perception is usually attributed to the lack of parallax information. The goal of this paper is exactly to tackle this deficiency by building from pictures 3D representations perceptually consistent and able to provide the missing parallax cues. To this end, we were inspired by the reverspective artworks from Patrick Hughes [DP13], which are paintings drawn on 3-D surfaces that are usually composed of pointed (or truncated) pyramids in reverse perspective, in a way that the painted perspective cues compete against the 3-D surface geometry resulting in bistable 3-D shapes. It has been shown that viewers report perceiving either the veridical structure or the reverse (illusory) depth structure, and the current explanation is that we perceive the illusory motion as the only real-world scenario (distal stimulus) that could have created the perspective and parallax transformations (proximal stimulus) that stimulate our visual systems [RG10]. Thus, we made the hypothesis that a 3D frustum based parametric representation of pictures could be exploited to enhance depth understanding during common interactive explorations.

4. Scene construction

Given a generic image, with zero-point perspective or one-point perspective, representing a natural or urban environment (see examples in figure 1), our method automatically builds a 3D parametric representation which can be interactively explored. The rest of the section details the proposed scene model, and the method for automatic computation.

4.1. Scene parametrization

Inspired by [RG10], we considered a 3D scene parametrization consisting of an hollow inverted frustum, composed by six planes, arranged in a way to match with all the perspective information contained in the image and to provide enough cues for effective pictorial relief and depth understanding. The full scene 3D parameterization is composed by a perspective projection matrix P, a viewing matrix V, and a set of planes Π_i, which can be used for constructing a

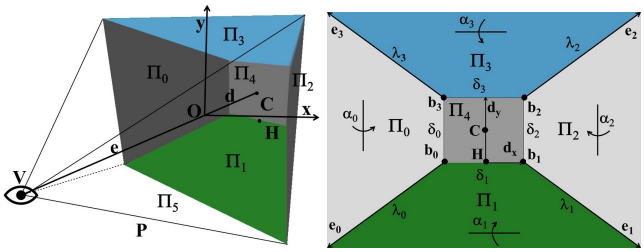

Figure 2: ***Scene parametrization.*** *Left: our 3D parameterization is composed by a perspective projection matrix P, a viewing matrix V, and a set of 6 planes Π_i . Right: the corresponding 2D representation is composed by four line segments $\lambda_i = (b_i, e_i)$ obtained by projecting the intersections between the scene planes, according to the projection and view matrix.*

texture mapped tessellation of the original image (see figure 2 left). The corresponding 2D representation is instead composed by four line segments $\lambda_i = (b_i, e_i)$, with $i = 1..4$, obtained by projecting the intersections between the scene planes (see figure 2 right). The 2D representation can be simplified by imposing that the origins b_i of the line segments λ_i are constrained to be the corners of a rectangle, in a way that the corresponding horizon portion plane is orthogonal to the viewing direction and parallel to the original image plane(as shown in figure 2 left). With this simplification, the overall number of parameters for defining the 2D projection of planes reduces to eight. Although perspective matrix P and viewing matrix V should be derived with camera calibration methods [Zha00], in our case the goal is to build an user centered perceptually plausible representation, optimized for ego-motion simulation. In this regard we assume a perspective transform P similar to a representative camera (in most cases, digital cameras in mobile phones have horizontal field of view of 54.4 degrees), and the view matrix V with eye position in $E = (0,0,e)$, and directed along negative z axis. Furthermore, we assume the Horizon plane Π_4 with the normal along the positive z axis and equation $z = -d$, depending on the desired finite scene depth, and the image plane Π_5, containing the origin of the system and with equation $z = 0$. Thus, the rest of the planes Π_i can be parametrized by two parameters each: the angle α_i with respect to the horizon plane, and the distance δ_i to the intersection C of the viewing direction with respect to the horizon plane Π_4 and computed with respect to the expected horizon point (assumed to be at position $H = (h_x, h_y, -d)$): from figure 2 right, $\delta_0 = h_x - d_x, \delta_1 = h_y, \delta_2 = h_x + d_x, \delta_3 = h_y + d_y$. Thus, the scene can be completely parametrized with eight parameters $\Pi_i = f(\alpha_i, \delta_i)$, corresponding to the eight parameters defining the 2D line segments λ_i which are the projections of scene planes over the image.

4.2. Automatic scene fitting and model optimization

The method for fitting the image content to the parametrized scene description represented by the planes $\Pi_i = f(\alpha_i, d_i)$ is based on a supervised image labeling technique aimed at subdividing the input picture in a limited number of significant regions. In our case, similarly to [HEH05], we consider three main regions: the top part, representing sky in outdoor and roof in indoor pictures, the middle part, representing the environment to be mapped in the horizon and the lateral planes, and the bottom part, representing the ground, be it the floor in indoor picture as well as terrain or water surfaces

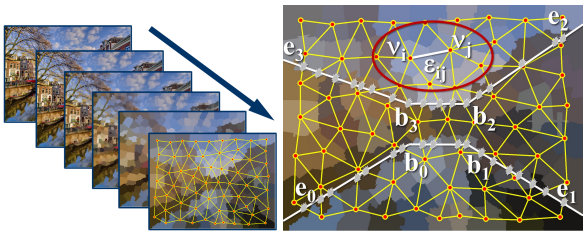

Figure 3: ***Automatic scene fitting.*** *Left: a multiresolution superpixel representation is obtained by repeated application of the SLIC clustering algorithm, and a hierarchy of graphs is built on it. Right: the scene representation is fitted by employing a graph-cut strategy for minimizing a penalty function computed along the edges of the graph.*

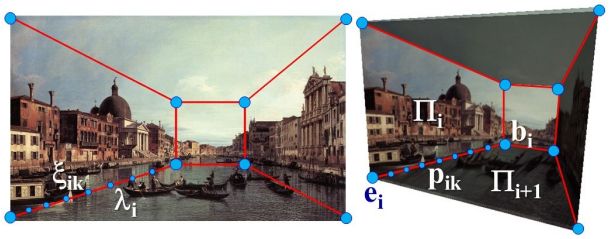

Figure 4: ***Semiautomatic correction.*** *The scene model can be tuned by modifying the 2D positions of 8 control points. The system automatically computes the optimal scene parameters matching the scene plane intersections to the selected line segments.*

in outdoor pictures. Hence, the goal of the scene fitting is to find the parameters for 3D planes Π_i, whose image projected segments λ_i provide the best fit with respect to the rough subdivision of the image (top, environment and bottom).

Image analysis and automatic model optimization. For the image analysis, in order to reduce the computation workload, we consider a multiresolution superpixel representation, obtained by repeated application of the SLIC clustering algorithm [ASS*12] (see figure 3 left). For each level l of superpixel sets, we construct a graph $\mathcal{G}_l = (N, E)$ representation in which each node $\nu_i = (p_i, c_i, \sigma_i) \in N$ is represented by the corresponding superpixel (the node attributes are the centroid p_i, the average color c_i and the color variance σ_i of the superpixel region), and each edge $\varepsilon_{ij} \in E$ connects two superpixels ν_i, ν_j having at least a pixel of contact. From these graph representations , we can fit the scene parameterization planes Π_i, by employing a graph-cut strategy [RKB04] for minimizing a penalty function computed along the edges of the graph. Specifically, given a parameter vector $\mathbf{x} = (\alpha_0, d_0, \alpha_1, d_1, \alpha_2, d_2, \alpha_3, d_3, \alpha_4, d_4)$ representing the 3D scene planes Π_i, the corresponding image projected intersection segments λ_i (see section 4.1) can be used for deriving a cut $\mathcal{C}(\mathbf{x}, \mathcal{G}_l)$ from the graph $\mathcal{G}_l$, as indicated in figure 3 right. The cut is obtained by removing from the original graph all the edges intersecting the set of segments $(e_0,b_0),(b_0,b_1),(b_1,e_1),(e_3,b_3),(b_3,b_2),(b_2,e_2)$ in a way to subdivide the image in three separate parts: bottom part (or ground), middle part (environment), top part (sky). Thus, the fitting problem consists of solving the following optimization problem:

$$\mathbf{x} = \underset{\mathbf{x}}{\operatorname{argmin}} \sum_{\varepsilon_{ij} \in \mathcal{C}(\mathbf{x}, \mathcal{G}_l)} E_p(\nu_i, \nu_j) \tag{1}$$

where the penalization function $E_p(\nu_i, \nu_j)$ is defined as follows:

$$E_p(\nu_i, \nu_j) = \frac{\|c_i - c_j\|^2}{(1 + \Omega(c_i, \sigma_i, c_j, \sigma_j)) \cdot (1 + \|p_i - p_j\|^2)}, \tag{2}$$

where p and c are respectively the position and the mean color value of a superpixel, and $\Omega(c_i, \sigma_i, c_j, \sigma_j)$ is an estimation of the volume intersection of the color space portions inside the superpixels represented by the gaussians $\mathcal{N}(c_i, \sigma_i)$ and $\mathcal{N}(c_j, \sigma_j)$ (for reducing processing complexity we considered overlapping intervals). The optimal solution of Eq. 1 is found by exploring the 8-dimensional parameter space using a stochastic global solver based on Bilbro and Snyder's tree annealing algorithm [BS91], which allows to minimize non-linear scalar fields over complex shaped search spaces.

Optional semi-automatic correction. According to the proposed parametric scene description, it is also possible for the user to interactively tune the scene model, by correcting the positions of the control points defining the segments λ_i, as indicated in Fig. 4 left, and automatically derive the parameters describing the scene in a way that the perspective projections match with the segments defined by the control points(see figure 4 right). To this end, we consider a uniform sampling ξ_{ik} of the segments λ_i, and employ the Levenberg–Marquardt algorithm (LMA) [Lou04], to minimize the least square differences between the sampled points and the projected ones, according to the scene parametrization $\mathbf{x}$. Specifically, the optimization problem assumes the following form:

$$\mathbf{x} = \underset{\mathbf{x}}{\operatorname{argmin}} \sum_i \sum_k (\xi_{ik} - PV\, p_{ik})^2 \tag{3}$$

where

$$p_{ik} = \cap_k(\Pi_i, \Pi_{i+i}) = b_i + \frac{k}{K-1}(e_i - b_i) \tag{4}$$

are obtained by sampling the image projected scene planes intersection segments (b_i, e_i) with the same number K of points used for sampling the user generated control segments λ_i (see Fig. 4 right).

5. Constrained exploration

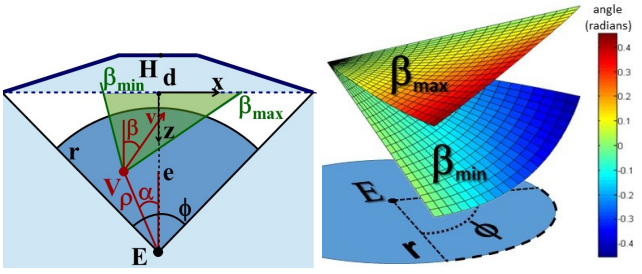

Figure 5: ***Constrained exploration.*** *Left: the workspace for exploration is defined by the starting view point, the horizon position and the field of view. The view matrix can be parametrized with three values: a distance* ρ*, and two angles* α,β*, for representing the view point V and the view direction* $\hat{v}$*. Right: the 3D representation of the constraints for the viewing direction angle* β*.*

In order to explore the reconstructed scene in a way to perceive the original picture in immersive way, we developed a constrained interaction method which enable the observers to navigate into the image in 3D without recognizing the geometry shape of the reconstructed scene, and in a way to boost coherent monocular depth

cues like perspective and motion parallax. To this end, we considered the recent results in automatic viewpoint computation [LC15], and we modeled the viewing transform according to the *lookat* metaphor, in which the up vector is fixed to the vertical direction, and the interaction consists of changing the observer position V and the view direction $\hat{v} = \hat{v}(\beta)$ (see figure 5).

Figure 6: ***Stroke based animations.*** *Various interactive and meaningful animation paths can be easily created by interpolating between few control points created with simple stroke inputs.*

Exploration workspace. Given a 3D reconstruction of the image pictorial space, for obtaining meaningful and realistic explorations, it is important that in any moment of the interaction the observer field of view is always kept inside a limited area, in a way that eventual scene distortions are negligible and that the geometry describing the scene is not perceived. This imposes constraints to the exploration workspace, which can be described as the feasible region for the view position V and the viewing direction $\hat{v}$. To this end, we considered a parametrization similar to the one proposed by Lino et al. [LC12], and indicated in figure 5 left. Since the exploration goal is to exploit the horizontal parallax, given the initial viewing position E, and the field of view ϕ, we parametrized the view matrix with three values (ρ, α, β), where ρ is the distance between the current viewpoint and the initial viewpoint E, α is the horizontal angle defining the viewpoint, and β is the horizontal angle of viewing direction $\hat{v}$. According to this parametrization, it is possible to easily define the viewing region: the viewer position V is constrained inside the circle sector with center E and angle ϕ ($0 \leq \rho \leq r$ and $-\frac{\phi}{2} \leq \alpha \leq \frac{\phi}{2}$), where r is the maximum distance from the initial viewpoint and it should be chosen in order to avoid that the scene geometry shape is perceivable (see section 6). Furthermore, for each view position $V(\rho, \alpha)$, considering the figure 5 it is possible to define constraints for the view direction $\hat{v}$ ($\beta_{min} \leq \beta \leq \beta_{max}$, see 3D color mapped representations in figure 5 right), where $\beta_{min} = \beta_c - \delta\beta$ and $\beta_{max} = \beta_c + \delta\beta$, with:

$$\beta_c = \frac{\tan^{-1}\left(\frac{-e\tan\frac{\phi}{2}-\rho\sin\alpha}{e-\rho\cos\alpha}\right)+\tan^{-1}\left(\frac{e\tan\frac{\phi}{2}-\rho\sin\alpha}{e-\rho\cos\alpha}\right)}{2}$$

$$\delta\beta = \frac{\tan^{-1}\left(\frac{e\tan\frac{\phi}{2}-\rho\sin\alpha}{e-\rho\cos\alpha}\right)-\tan^{-1}\left(\frac{-e\tan\frac{\phi}{2}-\rho\sin\alpha}{e-\rho\cos\alpha}\right)-\phi}{2}.$$

Figure 5 right shows the 3D color mapped representation of the surfaces $\beta_{min} = \beta_{min}(V)$ and $\beta_{max} = \beta_{max}(V)$.

Interaction control. Given a constrained workspace defined by the parametrization (ρ, α, β), various interaction metaphors can be defined, which map user inputs to the parameters defining the view positions, and various kind of exploration effects can be performed (zooming, panning, hovering). In our system, since the goal is to provide effective and realistic depth cues by exploiting zooming and parallax motions on touch based mobile interfaces, we designed a stroke based interaction metaphor, which constructs closed loop animations starting from simple user generated strokes (see figure 6). Specifically, user-generated strokes are sampled to compute a limited number of control points $C_k = (\rho_k, \alpha_k, \beta_k)$ inside the viewing workspace, and these control points are interpolated to create smooth close-loop parallax animations starting from the initial viewpoint. Examples of explorations and animations obtained with PEEP can be viewed in the accompanying video and snapshots are shown in figure 7 and 8.

Figure 8: ***Live mobile exploration.*** *The PEEP stroke animation controller has been integrated on an interactive Android mobile viewer.*

6. Results

Input image	Construction times	Result
	*Size* : 1800x1200 t_S : 5.8s, t_C : 11.7s t_M : 38.9s, SC: N	
	*Size* : 4000x3000 t_S : 26.7s, t_C : 32.5s t_M : 32.9s, SC: N	
	*Size* : 1536x1151 t_S : 5.5s, t_C: 11.6s t_M : 44.8s, SC: N	
	*Size* : 1400x920 t_S : 4.0s, t_C: 9.8s t_M : 32.6s, SC: N	
	*Size* : 2688x1520 t_S : 11.3s, t_C : 17.2s t_M : 32s, SC: Y	
	*Size* : 900x594 t_S: 4.5s, t_C : 10.6s t_M : 35.6s, SC: Y	

Table 1: ***PEEP construction statistics:*** *we tested our reconstruction method on a variety of casual pictures representing urban, natural and indoor environments. The multiresolution graph construction times t_S and the total construction times t_C are shown together with time t_M employed by Make3D [SSN09] and the outputs for a variety of examples.*

We implemented all the methods of the PEEP pipeline in C++. We tested our automatic fitting method on a Linux laptop equipped with a 8 Intel Core i7-4700HQ 2.4GHz CPU Processor, 8GB RAM and a NVidia GeForce GT 730M. For testing and evaluating the PEEP components we considered a variety of casual pictures representing urban, natural and indoor environments, and compared with respect to Make3D [SSN09] and automatic photo pop-up [HEH05].

Figure 7: ***Examples of interaction animations.*** *The PEEP stroke animation controller enables users to easily create perceptually significant exploration loops. Complete interaction examples can be viewed in the accompanying video.*

Construction. The PEEP automatic construction method complexity depends on the size of the input image, the number of superpixels employed for the optimization, and the number of stochastic optimization iterations. Figure 9 shows the graph construction times in seconds as function of image size (with varying number of levels L), and the global optimization times as function of number of iterations (with varying number of superpixels $S = 2^{16-2(L-1)}$). The statistics of automatic construction results on various example pictures, two zero-point and three one-point perspective, are shown in table 1: for each picture, image size (width and height), the multiresolution graph construction time (t_S) and the total construction time t_C are reported, with the total time employed by Make3D (t_M). Furthermore, the original picture as well as the matching scene intersection lines are shown. The following parameters were used: the maximum number of SLIC superpixels $S_M = 2^{16}$, the SLIC reduction factor $f = 16$, the number of levels employed for multiresolution graph construction $L = 3$, and the maximum number of optimization iterations $I = 50K$. Only in few cases, especially in zero-point perspective images and in general in pictures with no evident color separations, the semiautomatic correction was necessary (indicated with SC).

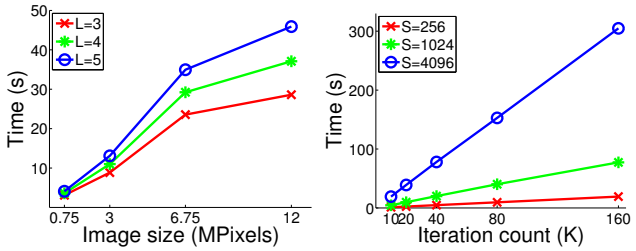

Figure 9: ***PEEP construction.*** *Left: graph construction times as function of image size (with varying number of levels L). Right: global optimization times as function of number of iterations (with varying number of superpixels $S = 2^{16-2(L-1)}$)*

Exploration. With respect to the interactive exploration component, various viewers have been implemented to explore the enhanced pictures. Accompanying video shows a variety of exploration animations on zero-point and one-point perspective pictures created with our system and rendered on various platforms, ranging from desktop system to mobile setups, to stereo VR mobile setups. The mobile interactive viewer has been implemented on Android with Trolltech Qt library, and it employs an indirect stroke animation controller, able to easily create meaningful and realistic interactive picture close-loop explorations with simple strokes. Figures 7 and 8 shows snapshots of explorations performed by our system to exploit motion parallax on desktop and mobile platforms.

Figure 10: ***Comparison with Make3D [SSN09]:*** *on top the 3D reconstructions, and on bottom correspondent interaction snapshots.*

Figure 11: ***Comparison with automatic photo pop-up [HEH05]:*** *on top the 3D reconstructions, and on bottom correspondent interaction snapshots.*

A qualitative comparison with Make3D and automatic photo pop-up (see figures 10 and 11) showed that even if depth maps obtained with [SSN09] and [HEH05] are metrically more accurate, interactive explorations performed with PEEP result to be perceptually more plausible (see the accompanying video). Furthermore, most casual

users were convinced to see real 3D videos when observing short animations recorded with PEEP.

Evaluation. In order to demonstrate the effectiveness of the method, we also carried out a preliminary perceptual evaluation to quantify whether and how much the exploration of pictures is improved by employing the enhanced PEEP representation. We involved 20 naive subjects (15 males and 5 females, with ages ranging from 28 to 51 years),with normal or corrected-to-normal vision, and we asked them to perform perceptual tests to analyze the two main PEEP parameters normalized with respect to the eye distance e: the scene depth $\frac{d}{e}$, and the maximum interaction distance $\frac{r}{e}$. Specifically , we employed a typical 2 force-choice (2FC) design comparing the exploration of flat unmodified pictures with the exploration of scenes at various depths levels $\frac{d}{e}$ in exponential scale (namely 0.0125, 0.025, 0.05, 0.1, 0.2 and 0.4). The setup consisted of a standard desktop PC equipped with an 8-core CPU Intel i7-3820 and a 4K resolution monitor, while the trials consisted of horizontally splitting the viewport in two parts and presenting the same precomputed 10 seconds animation representing a zoom exploration or a motion parallax exploration on the unmodified picture and on the scene exhibiting depth, randomly rendered on the left or on the right side. Subjects, after the animation, had to rapidly select in which part of the screen they perceived more immersive animation. Since we expected a significant effect in depth discrimination due to the different kind of animation, we considered different stimulus levels $\frac{d}{e}$ for the zoom animation (namely 0.05, 0.1, 0.2 and 0.4), and for the motion parallax animation (namely 0.0125, 0.025, 0.05, 0.1). For reducing eventual bias due to stress, we limited the total number of trials to 80, 40 for zooming, and 40 for motion parallax, in order to have an average total test duration of around 20 minutes per subject. Depth discrimination results (see 12 left for hit rates

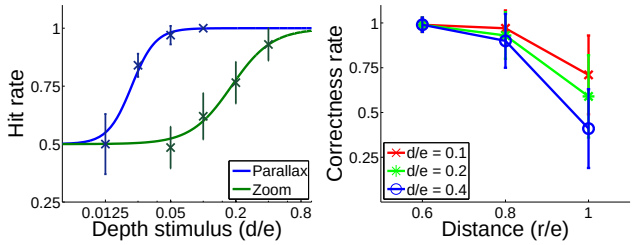

Figure 12: ***Perceptual evaluation.*** *Left: depth discrimination rates (standard error bars and psychometric fit) obtained with zoom and motion parallax animations. Right: correctness judgment rates (standard error bars) with respect to interaction distance obtained with motion parallax animations at varying scene depths ($\frac{d}{e} = 0.1, 0.2, 0.4$).*

together with psychometric fits obtained by employing the psignifit package [FHW11]) showed that the depth discrimination is dramatically dependent from the presence of motion parallax (even scenes with depths $\frac{d}{e}$ under 0.05 were perceived as immersive), while the effect is reduced in case of zooming actions in which the motion parallax is not present. ANOVA revealed a significant effect for both interaction types ($p \approx 0.0$ with $F(1,78) = 263.01$ for zoom animations and $p \approx 0.0$ with $F(1,78) = 217.89$ for parallax animations). Successively, we carried out tests for the derivation of the maximum interaction distance r (see section 5): the same 20 subjects were asked to perform 30 trials, with three stimuli distance values ($\frac{r}{e}$ varying between 0.6 , 0.8, and 1.0), and three different scene depths ($\frac{d}{e} = 0.1, 0.2, 0.4$), with the aim to investigate the eventual existence of a correlation between the two parameters $\frac{r}{e}$ and $\frac{d}{e}$. In this case, subjects observed precomputed 10 seconds animations representing motion parallax explorations reaching the maximum distance r, and they had to rapidly judge if they perceived them correctly or not. Figure 12 right shows mean correctness judgment rates and standard error bars on a linear scale. As expected, results showed that image quality is perceived degraded as the interaction approaches the borders of the 3D parametric representation (see figure 12 right), and two-way ANOVA revealed significant effects either with respect to interaction distance ($p \approx 0.0$ with $F(1,176) = 183.8$) and scene depth ($p < 0.001$ and $F(1,176) = 15.75$). From the above analysis, it is possible to accept a distance level $\frac{r}{e}$ lower then 0.7 as a safe bound for obtaining convincing and low-artifact explorations.

7. Conclusions

We have presented a generic framework for generating visually plausible 3D explorations of a single input landscape image. To cope with the complexity of 3D structure estimation from still images, we have introduced a light-weight automatic method to quickly recover just a simplified perspective structure from a typical landscape scene, and shown how the method works on a variety of images. The method, by design, is limited to scenes with a dominant 0- or 1-point perspective. It will therefore partially or totally fail if this assumption is not met, as, e.g., on scenes containing mirrors, multiple camera-facing layers, or no particular perspective views. It should be noted, however, that the chosen domain is typical of many landscape images and, therefore, the applicability domain is very large. Our simplified representation of viewed space consists only of an inverted frustum bound by 5 planes and is therefore capable to capture only the dominant perspective structure of the scene. Our results and tests show, however, that such a simple structure is sufficient to provide adequate perceptual information to generate a compelling illusion of 3D exploration with realistic monocular (perspective and motion parallax) and binocular (stereo) depth cues, even if the small-scale structure of the scene is not recovered. Such a simple approach leads to very simple and efficient tools for structure extraction from a single still image, as well as for virtual 3D exploration in constrained environments, e.g., on web browsers or mobile devices. Since the scene parametric model is generated without further prior information in short times, it can be directly embedded in mobile systems, in addition to being a candidate for automatic image enhancement in cloud-based environments. A preliminary perceptual evaluation revealed that the representation combined with constrained parallax and zooming exploration is able to provide convincing interactions for a significant distance and depth range. We plan to better explore and quantify the extent of the plausible perceptual workspace in future work. Besides improving the range of applicability of the method for handling foreground objects, we are currently looking at exploring the capabilities of our method in different application areas, including artwork exploration, physical renderings similar to the artworks of Patrick Hughes, or street view navigation.

Acknowledgments. This work was partially supported by the Sardinian Regional Authorities under projects VIGEC and Vis&VideoLab, as well as by the King Abdullah University of Science and Technology (KAUST).

References

[ASS*12] ACHANTA R., SHAJI A., SMITH K., LUCCHI A., FUA P., SUSSTRUNK S.: Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Patt. Anal. Machine Intell. 34*, 11 (2012), 2274–2282. 4

[AW07] ASSA J., WOLF L.: Diorama construction from a single image. In *Comp. Graph. Forum* (2007), vol. 26, pp. 599–608. 2

[BBP06] BOULANGER K., BOUATOUCH K., PATTANAIK S.: Atip: A tool for 3d navigation inside a single image with automatic camera calibration. *EG UK theory and practice of computer graphics 15* (2006). 2

[Bou14] BOUBEKEUR T.: Shellcam: Interactive geometry-aware virtual camera control. pp. 4003–4007. 2

[BS91] BILBRO G., SNYDER W.: Optimization of functions with many minima. *IEEE Trans. on Systems, Man and Cybernetics 21*, 4 (1991), 840–849. 4

[CGZ*05] CHUANG Y.-Y., GOLDMAN D. B., ZHENG K. C., CURLESS B., SALESIN D. H., SZELISKI R.: Animating pictures with stochastic motion textures. In *ACM Trans. Graph.* (2005), vol. 24, pp. 853–860. 1

[DH09] DECLE F., HACHET M.: A study of direct versus planned 3d camera manipulation on touch-based mobile phones. In *Proc. MobileHCI* (2009), ACM, pp. 32–35. 3

[DP13] DOBIAS J. J., PAPATHOMAS T. V.: Recovering 3-d shape: Roles of absolute and relative disparity, retinal size, and viewing distance as studied with reverse-perspective stimuli. *Perception 42*, 4 (2013), 430–446. 3

[Erk15] ERKELENS C. J.: The perspective structure of visual space. *i-Perception 6*, 5 (2015), 10.1177/2041669515613672. 1, 3

[FHW11] FRÜND I., HAENEL N. V., WICHMANN F. A.: Inference for psychometric functions in the presence of nonstationary behavior. *Vision 11*, 6 (2011). 7

[FNPS15] FLYNN J., NEULANDER I., PHILBIN J., SNAVELY N.: Deepstereo: Learning to predict new views from the world's imagery. *arXiv preprint arXiv:1506.06825* (2015). 1, 2

[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 *Proc. ACM SIGGRAPH* (1997), pp. 225–232. 2

[HEH05] HOIEM D., EFROS A. A., HEBERT M.: Automatic photo pop-up. *ACM Trans. Graph. 24*, 3 (2005), 577–584. 1, 2, 3, 5, 6

[JC16] JHOU W.-C., CHENG W.-H.: Animating still landscape photographs through cloud motion creation. *IEEE Trans. Multimedia 18*, 1 (2016), 4–13. 1

[JH15] JANKOWSKI J., HACHET M.: Advances in interaction with 3d environments. In *Computer Graphics Forum* (2015), vol. 34, pp. 152–190. 2

[JMBB15] JU Y. C., MAURER D., BREUSS M., BRUHN A.: Direct variational perspective shape from shading with cartesian depth parametrisation. *arXiv preprint arXiv:1505.06163* (2015). 2

[KKS*05] KHAN A., KOMALO B., STAM J., FITZMAURICE G., KURTENBACH G.: Hovercam: interactive 3D navigation for proximal object inspection. In *I3D* (2005), ACM, pp. 73–80. 2

[KNC*08] KOPF J., NEUBERT B., CHEN B., COHEN M., COHEN-OR D., DEUSSEN O., UYTTENDAELE M., LISCHINSKI D.: Deep photo: Model-based photograph enhancement and viewing. In *ACM Trans. Graph.* (2008), vol. 27, p. 116. 1, 2

[KPAS01] KANG H. W., PYO S. H., ANJYO K.-I., SHIN S. Y.: Tour into the picture using a vanishing line and its extension to panoramic images. In *Computer Graphics Forum* (2001), vol. 20, pp. 132–141. 2

[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. 3

[LC12] LINO C., CHRISTIE M.: Efficient composition for virtual camera control. In *Computer Animation* (2012), pp. 65–70. 5

[LC15] LINO C., CHRISTIE M.: Intuitive and efficient camera control with the toric space. *ACM TOG 34*, 4 (2015), 82. 2, 3, 5

[Lou04] LOURAKIS M.: Levmar: Levenberg-marquardt nonlinear least squares algorithms in C/C++. www.ics.forth.gr/ lourakis/levmar/, Jul. 2004. 4

[LPC*15] LI X., PHAM T.-A. N., CONG G., YUAN Q., LI X.-L., KRISHNASWAMY S.: Where you Instagram?: Associating your instagram photos with points of interest. In *Proc. ACM Inform. and Knowledge Management* (2015), pp. 1231–1240. 1

[MBB*14] MARTON F., BALSA RODRIGUEZ M., BETTIO F., AGUS M., JASPE VILLANUEVA A., GOBBETTI E.: Isocam: Interactive visual exploration of massive cultural heritage models on large projection setups. *ACM JOCCH* (2014), 12:1–12:24. 2

[RG10] ROGERS B., GYANI A.: Binocular disparities, motion parallax, and geometric perspective in patrick hughes's âĂŸreverspectivesâĂŹ: theoretical analysis and empirical findings. *Perception 39*, 3 (2010), 330–348. 3

[RKB04] ROTHER C., KOLMOGOROV V., BLAKE A.: Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM Trans. Graph.* (2004), vol. 23, pp. 309–314. 4

[RNR*16] REMATAS K., NGUYEN C., RITSCHEL T., FRITZ M., TUYTELAARS T.: Novel views of objects from a single image. *arXiv preprint arXiv:1602.00328* (2016). 2

[RU14] RANON R., URLI T.: Improving the efficiency of viewpoint composition. *IEEE Trans. Vis. Comput. Graph. 20*, 5 (2014), 795–807. 2

[SC05] SHESH A., CHEN B.: Peek-in-the-pic: Architectural scene navigation from a single picture using line drawing cues. In *Proc. Pacific Graphics* (2005). 2

[SCD*06] SEITZ S. M., CURLESS B., DIEBEL J., SCHARSTEIN D., SZELISKI R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proc. IEEE CVPR* (2006), vol. 1, pp. 519–528. 2

[Sho92] SHOEMAKE K.: Arcball: a user interface for specifying three-dimensional orientation using a mouse. In *Proc. Graphics interface* (1992), pp. 151–156. 2

[SSN09] SAXENA A., SUN M., NG A. Y.: Make3d: Learning 3d scene structure from a single still image. *IEEE Trans. Patt. Anal. Machine Intell. 31*, 5 (2009), 824–840. 2, 5, 6

[VVD14] VOLCIC R., VISHWANATH D., DOMINI F.: Reaching into pictorial spaces. In *IS&T/SPIE Electronic Imaging* (2014), pp. 901413–901413. 3

[ZCW*15] ZENG Q., CHEN W., WANG H., TU C., COHEN-OR D., LISCHINSKI D., CHEN B.: Hallucinating stereoscopy from a single image. *Computer Graphics Forum 34*, 2 (2015), 1–12. 2

[ZGW*14] ZHANG C., GAO J., WANG O., GEORGEL P., YANG R., DAVIS J., FRAHM J.-M., POLLEFEYS M.: Personal photograph enhancement using internet photo collections. *IEEE Trans. Vis. Comput. Graph. 20*, 2 (2014), 262–275. 1

[Zha00] ZHANG Z.: A flexible new technique for camera calibration. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 22*, 11 (2000), 1330–1334. 3

[ZHC15] ZHAO L., HANSARD M., CAVALLARO A.: Pop-up modelling of hazy scenes. In *Proc. ICIAP*. 2015, pp. 306–318. 2