Automatic Upright Adjustment of Photographs with Robust Camera Calibration

Hyunjoon Lee, *Member, IEEE,* Eli Shechtman, *Member, IEEE,* Jue Wang, *Senior Member, IEEE,* and Seungyong Lee, *Member, IEEE*

Abstract—Man-made structures often appear to be distorted in photos captured by casual photographers, as the scene layout often conflicts with how it is expected by human perception. In this paper we propose an automatic approach for straightening up slanted man-made structures in an input image to improve its perceptual quality. We call this type of correction *upright adjustment*. We propose a set of criteria for upright adjustment based on human perception studies, and develop an optimization framework which yields an optimal homography for adjustment. We also develop a new optimization-based camera calibration method that performs favorably to previous methods and allows the proposed system to work reliably for a wide range of images. The effectiveness of our system is demonstrated by both quantitative comparisons and qualitative user study.

Index Terms—Upright adjustment, perspective correction, photo aesthetics enhancement, single image camera calibration.

1 Introduction

A large portion of consumer photos contain man-made structures, such as urban scenes with buildings and streets, and indoor scenes with walls and furniture. However, photographing these structures properly is not an easy task. Photos taken by amateur photographers often contain slanted buildings, walls, and horizon lines due to improper camera rotations, as shown in Fig. 1. On the contrary, our visual system always expects tall man-made structures to be straight-up, and horizon lines to be parallel to our eye level. This conflict leads us to a feeling of discomfort when we look at a photo containing slanted structures.

Assuming the depth variations of the scene relative to its distance from the camera are small, correcting a slanted structure involves a 3D rotation of the image plane. We call this type of correction *upright adjustment*, since its goal is to make man-made structures straight up as expected by human perception. Similar corrections have been known as *keystoning* and *perspective correction*, which can be achieved by manually warping the image using existing software, or during capture using a special Tilt-Shift lens. However, the target domain of these tools is mostly facades of buildings, while our upright adjustment method does not explicitly assume specific types of objects in the scene. In addition, manual correction not only requires special skills, but also becomes tedious when we need to process hundreds of photos from a trip.

In this paper, we propose a fully automatic system for upright adjustment of photos. To the best of our knowledge, our system is the first one that automatically handles this kind of correction, although there have been several papers dealing with sub-problems of our framework. Our system introduces several novel technical contributions: (1) we propose various criteria to quantitatively measure the perceived quality of man-made structures, based on previous studies on human perception; (2) following the criteria, we propose an energy minimization framework to compute an optimal homography that can effectively minimize the perceived distortion of slanted structures; and (3) we propose a new camera calibration method which simultaneously estimates vanishing lines and points as well as camera parameters, and is more accurate and robust than the state-of-the-art. Although not designed to, our system is robust enough to handle some natural scenes as well (see Fig. 1e). We evaluate the system comprehensively through both quantitative comparisons and qualitative user study. Experimental results show that our system works reliably on a wide range of images without the need for user interaction.[1]

1.1 Related work

Photo aesthetics and composition

Automatic photo aesthetics evaluation tools [2], [3], [4], [5], [6] and composition adjustment systems [7], [8] have been proposed recently, which include various criteria for aesthetics and composition quality of photographs. The evaluation criteria are based on not only simple image statistics (*e.g.* color, sharpness) but semantic information such as saliency and main subjects of images. We propose a set of new criteria specific to the uprightness of man-made structures, based on well-known studies in human perception. Our method is based on an objective function that quantifies these criteria, and thus could potentially be

1. A shorter version of this paper appeared in CVPR 2012 [1].

used to enhance previous aesthetic evaluation methods with an additional uprightness score.

Automatic and manual photo correction

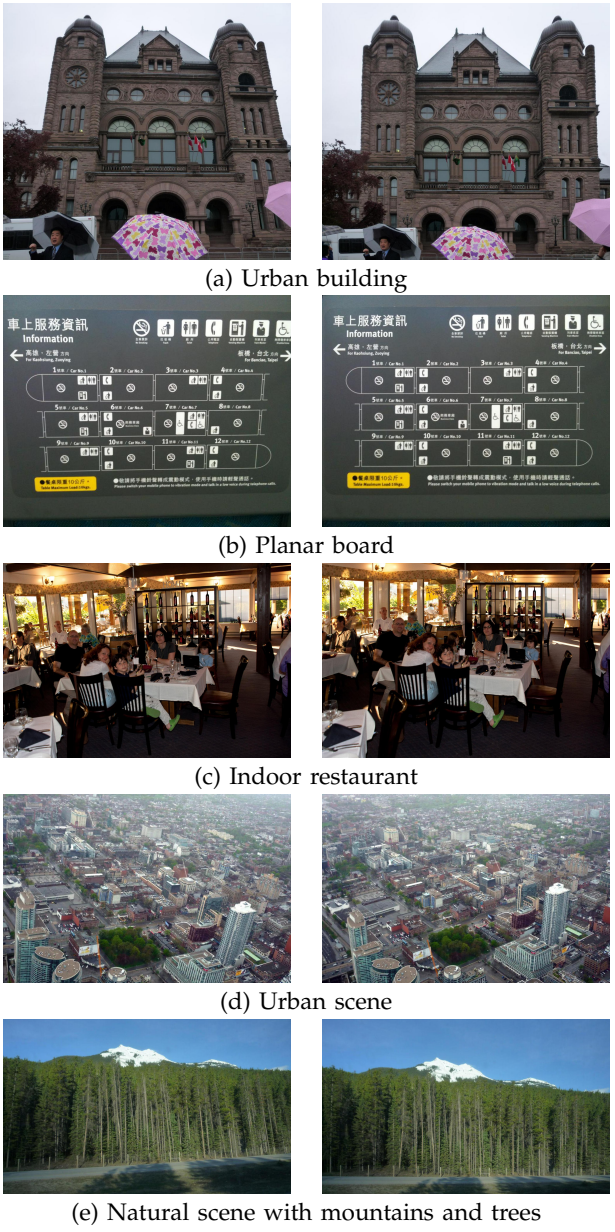
A few previous methods [7], [8], [9] addressed automatic enhancement of the geometric properties of photos. However, these methods use 2D techniques like cropping or retargeting of photos and do not address perspective distortions. Gallagher [10] proposed an automatic method to adjust the in-plane rotation of an image. Commercial software such as Adobe Photoshop provides manual adjustment tools, such as lens correction, 3D image plane rotation, and cropping. Professional photographers sometimes use a specialized Tilt-Shift lens for adjusting the orientation of the plane of focus and the position of the subject in the image for correcting slanted structures and converging lines. Both solutions require sophisticated interaction that is likely to be hard for a non-exert user. Carroll *et al*. [11] proposed a manual perspective adjustment tool based on geometric warping of a mesh, which is more general than a single homography used in this paper, but requires accurate manual control.

Camera calibration

Calibrating intrinsic parameters and orientation of a camera from a single image is a well-studied problem. Coughlan and Yuille [12] introduced the "Manhattan world" assumption to calibrate the camera parameters. They used raw edge pixels as their primitives for the calibration, but most following approaches [13], [14], [15] used line segments instead. These methods commonly apply a two-step approach: an orthogonal set of vanishing points and lines are first extracted from the input image, and then used to calibrate the camera. Recent methods [16], [17], on the other hand, use a unified optimization framework and could get highly accurate calibration results under the Manhattan world assumption. Beside line based methods, a few methods have been proposed recently that utilize underlying repeated patterns in an image for the calibration [18], [19]. However, such methods are limited to recover local geometry of a planar scene although their results can be used as the input of a more general framework as demonstrated in [18].

2 FORMULATION

Upright adjustment can be performed by transforming an input photo. We assume no depth information is available for the input photo, and thus use a homography to transform it for upright adjustment. A more complex transformation could be adopted, e.g., content-preserving warping [11]. However such a transformation contains more degrees of freedom, and therefore requires a large amount of reliable constraints which should be fulfilled with user interaction or additional information about the scene geometry. A homography provides a reasonable amount of control to achieve visually plausible results in most cases, especially for man-made structures.



(a) Urban building

(b) Planar board

(c) Indoor restaurant

(d) Urban scene

(e) Natural scene with mountains and trees

Fig. 1. Various examples of upright adjustment of photos. (left) original; (right) our result.

A given image can be rectified with a homography matrix using the following equation [14]:

$$\mathbf{p}' = \mathbf{H}\mathbf{p} = \mathbf{K}(\mathbf{K}\mathbf{R})^{-1}\mathbf{p}, \tag{1}$$

where $\mathbf{p}$ and $\mathbf{p}'$ represent a position and its reprojection in the image, respectively. $\mathbf{K}$ and $\mathbf{R}$ are intrinsic parameter and orientation matrices of the camera, respectively:

$$\mathbf{K} = \begin{pmatrix} f & 0 & u_0 \\ 0 & f & v_0 \\ 0 & 0 & 1 \end{pmatrix} \text{ and } \mathbf{R} = \mathbf{R}_\psi \mathbf{R}_\theta \mathbf{R}_\phi, \tag{2}$$

where $\mathbf{R}_\psi$, $\mathbf{R}_\theta$, and $\mathbf{R}_\phi$ are rotation matrices with angles ψ, θ, and ϕ along the x, y, and z axes, respectively.

Although rectification results are useful for other applications, they are often visually unpleasing (see Fig. 11). For upright adjustment, the rectified result is reprojected with a new set of camera parameters to enhance its perceptual quality, and our homography is modified as follows:

$$\mathbf{p}' = \mathbf{H}\mathbf{p} = \mathbf{K}_1 \left\{ \mathbf{R}_1 (\mathbf{K}\mathbf{R})^{-1} \mathbf{p} + \mathbf{t}_1 \right\}, \quad (3)$$

where

$$\mathbf{K}_1 = \begin{pmatrix} f_{1x} & 0 & u_1 \\ 0 & f_{1y} & v_1 \\ 0 & 0 & 1 \end{pmatrix}, \quad (4)$$

$$\mathbf{R}_1 = \mathbf{R}_{\psi_1} \mathbf{R}_{\theta_1} \mathbf{R}_{\phi_1} \text{ and } \mathbf{t}_1 = [t_{1x}\ t_{1y}\ 0]^T.$$

Compared to Eq. (1), Eq. (3) contains a new intrinsic parameter matrix $\mathbf{K}_1$ with additional 3D rotation $\mathbf{R}_1$ and translation $\mathbf{t}_1$. This reprojection model implies reshooting of the rectified scene using another camera placed at a possibly different position with novel orientation. We also allow this new camera to have different focal lengths in horizontal and vertical directions (Sec. 5.2).

In Sec. 3 we describe our camera calibration algorithm that estimates $\mathbf{K}$ and $\mathbf{R}$. We then propose our upright adjustment criteria and optimization framework for the estimation of $\mathbf{K}_1, \mathbf{R}_1$ and $\mathbf{t}_1$ in Sec. 4.

3 Camera Calibration

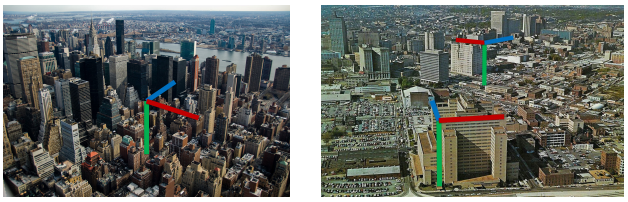
Calibration of camera parameters from a single image is a highly ill-posed problem. Several priors were utilized in previous approaches, such as the Manhattan world assumption. In this section we first explain the set of calibration priors we use in our method. We then formulate our calibration method as a maximum a-posteriori (MAP) estimation, and finally present our optimization algorithm. The effectiveness of our priors as well as the robustness and accuracy of our method are presented in Sec. 5.1.

3.1 Calibration priors

Scene priors

Manhattan world assumption is the most common prior in single image camera calibration methods. It assumes the existence of three dominant orthogonal directions in the scene, which are called "Manhattan directions" (Fig. 2a). By extracting those directions, the reference world coordinate axes can be recovered and the camera parameters can be calibrated [13], [14].

Despite of the effectiveness of the Manhattan world model, in some cases a scene can have multiple orthogonal directions that do not align, such as two groups of buildings with a non-right angle between their horizontal directions (*e.g.*, see Fig. 2b). This case was first introduced by Schindler *et al.* [20] and is often referred as the "Atlanta world" assumption. Recently Tretyak *et al.* [21] used the model to accurately estimate the eye-level in complex scenes.



(a) Manhattan (b) Atlanta

Fig. 2. Manhattan and Atlanta world assumptions. Manhattan world assumption assumes a single dominant orientation, while multiple orientations can exist under Atlanta world assumption.

In this paper, we adopt a similar prior. We assume that the input image has a dominant orthogonal frame, with additional horizontal directions sharing the same vertical direction. This extended model fits well most man-made environments [21], and enhances performance and robustness of our calibration algorithm.

Camera priors

Most previous algorithms utilize priors on the intrinsic parameter matrix $\mathbf{K}$ of the camera [21], [22]. The assumption is that the focal length in pixel dimension of the camera is the same as the width of the image and the center of projection is the image center, so that:

$$\mathbf{K} = \begin{pmatrix} f & 0 & u_0 \\ 0 & f & v_0 \\ 0 & 0 & 1 \end{pmatrix} \sim \begin{pmatrix} W & 0 & c_x \\ 0 & W & c_y \\ 0 & 0 & 1 \end{pmatrix}, \quad (5)$$

where W is the image width and (c_x, c_y) is the image center, respectively.

For the prior on external camera orientation $\mathbf{R}$, we adopt one based on human tendency that people tends to align their camera with the principal axes of the world. Under this assumption rotation angles of the orientation matrix $\mathbf{R}$ should be small so that

$$\{\psi, \theta, \phi\} \sim 0 \text{ where } \mathbf{R} = \mathbf{R}_\psi \mathbf{R}_\theta \mathbf{R}_\phi. \quad (6)$$

3.2 Calibration formulation

As most previous calibration methods, we use line segments as the basic primitives for calibration. From the input image, we extract a set of line segments $\mathbf{L}$, using the method of Von Gioi *et al.* [23] in a multi-scale fashion [21]. We store each line segment $\mathbf{l}_i$ with its two end points $\mathbf{p}_i$ and $\mathbf{q}_i$ in the projective plane $\mathbb{P}^2$.

Once line segments are extracted, we calibrate camera parameters $\mathbf{K}$ and $\mathbf{R}$. To utilize our calibration priors, we also extract Manhattan directions $\mathbf{M}$ and

additional horizontal vanishing points $\mathbf{A}$ during calibration, where

$$\mathbf{M} = [\mathbf{v}_x\ \mathbf{v}_y\ \mathbf{v}_z] \quad \text{and} \quad \mathbf{A} = [\mathbf{v}_{a_1}\ \mathbf{v}_{a_2}\ \cdots\ \mathbf{v}_{a_k}], \tag{7}$$

and $\mathbf{v}$ representing a vanishing point in $\mathbb{P}^2$. The joint probability of $\mathbf{K}$, $\mathbf{R}$, $\mathbf{M}$, and $\mathbf{A}$ with respect to $\mathbf{L}$ can be formulated as follows:

$$\begin{aligned} p(\mathbf{K},\mathbf{R},\mathbf{M},\mathbf{A}|\mathbf{L}) &\propto p(\mathbf{L}|\mathbf{K},\mathbf{R},\mathbf{M},\mathbf{A})p(\mathbf{K},\mathbf{R},\mathbf{M},\mathbf{A}) \\ &= p(\mathbf{L}|\mathbf{M},\mathbf{A})p(\mathbf{K},\mathbf{R},\mathbf{M},\mathbf{A}) \\ &= p(\mathbf{L}|\mathbf{M},\mathbf{A})p(\mathbf{M},\mathbf{A}|\mathbf{K},\mathbf{R})p(\mathbf{K},\mathbf{R}) \\ &= p(\mathbf{L}|\mathbf{M},\mathbf{A})p(\mathbf{M},\mathbf{A}|\mathbf{K},\mathbf{R})p(\mathbf{K})p(\mathbf{R}), \end{aligned} \tag{8}$$

with the assumption that $\mathbf{K}$ and $\mathbf{R}$ are independent of $\mathbf{L}$ and are also independent of each other. By taking a log, Eq. (8) turns into the following energy function:

$$E_{K,R,M,A|L} = E_K + E_R + E_{M,A|K,R} + E_{L|M,A}. \tag{9}$$

For the computation of $E_{L|M,A}$, we utilize our scene priors. Under the Manhattan world assumption, triplets of vanishing points that represent more line segments are preferred. Furthermore, the union of $\mathbf{M}$ and $\mathbf{A}$ should contain as many as possible line segments as vanishing lines. We therefore formulate our energy function as follows:

$$E_{L|M,A} = \lambda_{L_m} \sum_i d_m(\mathbf{M}, \mathbf{l}_i) + \lambda_{L_a} \sum_i d_m(\mathbf{M} \cup \mathbf{A}, \mathbf{l}_i),$$

where $\mathbf{l}_i$ represents a line segment. $d_m(\cdot)$ measures the minimum distance between a set of vanishing points $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$ and a line segment $\mathbf{l}$, as follows:

$$d_m(\mathbf{V}, \mathbf{l}) = \min\left\{d(\mathbf{v}_1, \mathbf{l}), d(\mathbf{v}_2, \mathbf{l}), \ldots, d(\mathbf{v}_k, \mathbf{l})\right\}.$$

$d(\cdot)$ is for measuring distance between a vanishing point and a line, we use the definition in [22]:

$$d(\mathbf{v}, \mathbf{l}) = \min\left(\frac{|\mathbf{r}^T\mathbf{p}|}{\sqrt{r_1^2 + r_2^2}},\ \delta\right), \tag{10}$$

where $\mathbf{p}$ and $\mathbf{q}$ are two end points of $\mathbf{l}$ and

$$\mathbf{r} = \left(\frac{\mathbf{p}+\mathbf{q}}{2}\right) \times \mathbf{v} = [r_1\ r_2\ r_3]^T.$$

δ is the given maximum error value for which we used 1.75 in our implementation. We set λ_{L_m} and λ_{L_a} to 0.01 and 0.02, respectively.

E_K and E_R are related to camera priors. From Eqs. (5) and (6), we have:

$$E_K = \lambda_f \left(\frac{\max(W,f)}{\min(W,f)} - 1\right)^2 + \lambda_c \left\|\mathbf{c}_p - \mathbf{c}_I\right\|^2$$

and

$$E_R = \lambda_\psi \psi^2 + \lambda_\theta \theta^2 + \lambda_\phi \phi^2.$$

For E_K, we set λ_f as 0.04 and λ_c as $(4/W)^2$. For E_R, the three rotation angles are not weighted equally. Particularly, we found that the prior for ϕ (z-axis rotation) should be stronger to enforce eye-level alignment. We thus use $[\lambda_\psi, \lambda_\theta, \lambda_\phi] = [3/\pi, 2/\pi, 6/\pi]^2$.

To compute $E_{M,A|K,R}$, we assume $E_{M|K,R}$ and $E_{A|K,R}$ can be computed independently so that

$$E_{M,A|K,R} = E_{M|K,R} + E_{A|K,R}.$$

Then, if $\mathbf{K}$ and $\mathbf{R}$ are known $\mathbf{M}$ can be estimated as:

$$\mathbf{M} = [\mathbf{v}_x\ \mathbf{v}_y\ \mathbf{v}_z] = (\mathbf{KR})\mathbf{I}_3,$$

where $\mathbf{I}_3 = [\mathbf{e}_x\ \mathbf{e}_y\ \mathbf{e}_z]$ is the identity matrix. Using this property, we formulate $E_{M|K,R}$ as follows:

$$E_{M|K,R} = \lambda_M \sum_{i\in\{x,y,z\}} \left[\cos^{-1}\left\{\mathbf{e}_i^T \frac{(\mathbf{KR})^{-1}\mathbf{v}_i}{\|(\mathbf{KR})^{-1}\mathbf{v}_i\|}\right\}\right]^2,$$

where λ_M is set as $(48/\pi)^2$ in our experiments. $\mathbf{A}$ represents horizontal directions, and should be perpendicular to $\mathbf{e}_y$. We thus formulate $E_{A|K,R}$ as:

$$E_{A|K,R} = \lambda_A \sum_i \left[\cos^{-1}\left\{\mathbf{e}_y^T \frac{(\mathbf{KR})^{-1}\mathbf{v}_{a_i}}{\|(\mathbf{KR})^{-1}\mathbf{v}_{a_i}\|}\right\} - \frac{\pi}{2}\right]^2,$$

where $\mathbf{v}_{a_i}$ represent a horizontal vanishing point. We set λ_A as $(24/\pi)^2$.

Dealing with missing vanishing points

When we estimate $\mathbf{M}$, we cannot always find all three vanishing points. Our energy terms should be able to handle this case for robustness. For $E_{M|K,R}$, we set the energy to be zero for a missing vanishing point, assuming that the point is located at the position estimated using $\mathbf{K}$ and $\mathbf{R}$. For $E_{L|M,A}$, we let $d(\mathbf{v}_{miss}, \mathbf{l}_i)$ always be δ for all $\mathbf{l}_i$ where $\mathbf{v}_{miss}$ represent a missing vanishing point in $\mathbf{M}$.

3.3 Calibration process

With the energy terms defined above, we can use an iterative approach to find the solution.

In the iteration, we alternatingly optimize $\mathbf{K}$ and $\mathbf{R}$, $\mathbf{M}$, and $\mathbf{A}$. If we fix $\mathbf{M}$ and $\mathbf{A}$, we can optimize Eq. (9) with respect to $\mathbf{K}$ and $\mathbf{R}$ by:

$$\arg\min_{\mathbf{K},\mathbf{R}}\ E_K + E_R + E_{M,A|K,R}. \tag{11}$$

Similarly, optimization of $\mathbf{M}$ and $\mathbf{A}$ can be achieved by solving the followings:

$$\arg\min_{\mathbf{M}}\ E_{M,A|K,R} + E_{L|M,A}, \tag{12}$$

$$\arg\min_{\mathbf{A}}\ E_{M,A|K,R} + E_{L|M,A} \tag{13}$$

while fixing other parameters.

For optimizing $\mathbf{K}$ and $\mathbf{R}$, our implementation uses `fminsearch` in Matlab [24]. On the other hand, optimizations of $\mathbf{M}$ and $\mathbf{A}$ are still hard since $E_{L|M,A}$ truncates distances to δ as defined in Eq. (10) and the size of $\mathbf{A}$ is unknown. To solve Eqs. (12) and (13), we use a discrete approximation inspired by [21]. From the line segments $\mathbf{L}$, we hypothesize a large set of

vanishing points $\mathbf{V} = [\mathbf{v}_1\ \mathbf{v}_2\ \ldots\ \mathbf{v}_n\ \mathbf{v}_{miss}]$, where each element is computed as the intersection point of two randomly selected lines except for $\mathbf{v}_{miss}$, representing the missing vanishing point. We set $n = 2000$ in our implementation. Optimizing $\mathbf{M}$ and $\mathbf{A}$ thus becomes selecting vanishing points from $\mathbf{V}$ to minimize energies in Eqs. (12) and (13).

To optimize $\mathbf{M}$, for each element of $\mathbf{M} = [\mathbf{v}_x\ \mathbf{v}_y\ \mathbf{v}_z]$ we find a vanishing point in $\mathbf{V}$ that minimizes the energy while retaining the other two elements. For optimizing $\mathbf{A}$, we use a greedy approach. Specifically, we select a vanishing point from $\mathbf{V}$ one by one that minimize Eq. (13) and add it to $\mathbf{A}$ until the energy does not decrease.

The iterative optimization process requires good initial values to work properly. In order to make initials of $\mathbf{M}$, we first select a small subset $\mathbf{V}_c = \{\mathbf{v}_{c_1}, \mathbf{v}_{c_2}, \ldots, \mathbf{v}_{c_k}\}$ from $\mathbf{V}$ that is the "closest to all lines" in the following way:

$$\underset{\{\mathbf{v}_{c_1},\ldots,\mathbf{v}_{c_k}\}}{\arg\min} \sum_i \min\left\{d(\mathbf{v}_{c_1}, \mathbf{l}_i), \ldots, d(\mathbf{v}_{c_k}, \mathbf{l}_i)\right\},$$

where we set $k = 9$ in our implementation. We then also add $\mathbf{v}_{miss}$ into $\mathbf{V}_c$, too. For each triplet of vanishing points in $\mathbf{V}_c$, we optimize initial $\mathbf{K}$ and $\mathbf{R}$ letting $\mathbf{M}$ as the triplet and $\mathbf{A}$ as empty, and then optimize initial $\mathbf{A}$. Then we optimize all the variables using Eqs. (11)-(13) and evaluate Eq. (9). Finally, $\mathbf{K}$, $\mathbf{R}$, $\mathbf{M}$, and $\mathbf{A}$ with the minimum energy are used as our calibration results. Although initial $\mathbf{V}_c$ may not contain all Manhattan directions, the missing directions can be detected from $\mathbf{V}$ while optimizing Eq. (12) in the iterative optimization process. Optimizing $\mathbf{K}$, $\mathbf{R}$, $\mathbf{M}$, and $\mathbf{A}$ for all possible triplets might be computationally expensive. Thus we use some early termination strategies for speedup.

After the calibration process, we determine the vanishing lines for each vanishing point in $\mathbf{M}$. Three pencils of vanishing lines, $\mathbf{L}_x$, $\mathbf{L}_y$, and $\mathbf{L}_z$, are obtained from $\mathbf{L}$ by:

$$\mathbf{L}_i = \{\mathbf{l} \in \mathbf{L} \mid d(\mathbf{v}_i, \mathbf{l}) < \delta\},\ i \in \{x, y, z\},$$

where $d(\cdot)$ is the distance function defined in Eq. (10).

Utilizing external information

Our MAP formulation can be reformulated to various forms to utilize additional information provided by the user or camera manufacturer. For example, one can fix focal length or center of projection during the calibration if they are given. In our experiments, we could obtain better calibration results if we fix known parameters although we did not put them in our paper. In our calibration method, we detect additional horizontal vanishing points $\mathbf{A}$ but they can be ignored if the scene strictly follows the Manhattan world assumption. In this case, $E_{L|M,A}$ and $E_{M,A|K,R}$ become $E_{L|M}$ and $E_{M|K,R}$, respectively, and the calibration is performed without detection of additional horizontal vanishing points.

4 Adjustment Optimization

In this section, we derive and minimize an energy function for our upright adjustment formulated in Sec. 2. As defined in Eq. (3), we adjust an input photo using a homography matrix $\mathbf{H}$, and we show how to optimize $\mathbf{K}_1$, $\mathbf{R}_1$ and $\mathbf{t}_1$. For this we first define perceptual criteria that adjusted results should satisfy. Then we derive our energy function based on the criteria to optimize the new camera parameters.

4.1 Adjustment Criteria

Scenes with well-structured man-made objects often include many straight lines that are supposed to be horizontal or vertical in the world coordinates. Our proposed criteria reflect these characteristics.

Picture frame alignment

When looking at a big planar facade or a close planar object such as a painting, we usually perceive it as orthogonal to our view direction, and the horizontal and vertical object lines are assumed to be parallel and perpendicular to the horizon, respectively. When we see a photo of the same scene, the *artificial* picture frame imposes significant alignment constraints on the object lines, and we feel discomfort if the object line directions are not well aligned with the picture frame orientation [25], [26]. Figs. 1a and 1b show typical examples. It is also important to note that such an artifact becomes less noticeable as the misalignments of line directions become larger, since in that case we begin to perceive 3D depths from a slanted plane.

Eye level alignment

The eye level of a photo is the 2D line that contains the vanishing points of 3D lines parallel to the ground [27]. In a scene of an open field or sea, the eye level is the same as the horizon. However, even when the horizon is not visible, the eye level can still be defined as the connecting line of specific vanishing points. It is a well-known principle in photo composition that the eye level or horizon should be horizontal [26]. The eye level alignment plays an important role in upright adjustment especially when there exist no major object lines to be aligned to the picture frame. In Fig. 1d, the invisible eye level is dominantly used to correct an unwanted rotation of the camera.

Perspective distortion

Since we do not usually see objects outside our natural field of view (FOV), we feel an object is distorted when the object is pictured as if it is out of our FOV [25], [27]. We can hardly see this distortion in ordinary photos, except those taken with wide-angle lenses. However, such distortion may happen if we apply a large rotation to the image plane, which corresponds to a big change of the camera orientation. To prevent this from happening, we explicitly constrain perspective distortion in our upright adjustment process.

Image distortion

When we apply a transformation to a photo, image distortion cannot be avoided. However, human visual system is known to be tolerant to distortions of rectangular objects, while it is sensitive to distortions of circles, faces, and other familiar objects [25]. We consider this phenomenon in our upright adjustment to reduce the perceived distortions in the result image as much as possible.

4.2 Energy terms

Picture frame alignment

For major line structures of the scene to be aligned with the picture frame, all vanishing lines corresponding to x- and y-directions should be horizontal and vertical in a photo, respectively. That is, vanishing lines in $\mathbf{L}_x$ and $\mathbf{L}_y$ should be transformed to horizontal and vertical lines by a homography $\mathbf{H}$, making vanishing points $\mathbf{v}_x$ and $\mathbf{v}_y$ placed at infinity in x- and y-directions, respectively.

Let $\mathbf{l}$ be a vanishing line, and $\mathbf{p}$ and $\mathbf{q}$ be two end points of $\mathbf{l}$. Then, the direction of the transformed line $\mathbf{l}'$ is:

$$\mathbf{d} = \frac{\mathbf{q}' - \mathbf{p}'}{\|\mathbf{q}' - \mathbf{p}'\|},$$

where

$$\mathbf{p}' = \frac{\mathbf{H}\mathbf{p}}{\mathbf{e}_z^T \mathbf{H}\mathbf{p}} \quad \text{and} \quad \mathbf{q}' = \frac{\mathbf{H}\mathbf{q}}{\mathbf{e}_z^T \mathbf{H}\mathbf{q}}.$$

$\mathbf{e}_z = [0\ 0\ 1]^T$ is used to normalize homogeneous coordinates. We define the energy term as:

$$E_{pic} = \lambda_v \sum_i w_i \left(\mathbf{e}_x^T \mathbf{d}_{y_i}\right)^2 + \lambda_h \sum_j w_j \left(\mathbf{e}_y^T \mathbf{d}_{x_j}\right)^2, \quad (14)$$

where $\mathbf{d}_{y_i}$ is the direction of the transformed line $\mathbf{l}'_{y_i}$ of a vanishing line $\mathbf{l}_{y_i}$ in $\mathbf{L}_y$. $\mathbf{e}_x = [1\ 0\ 0]^T$, and $\mathbf{e}_x^T \mathbf{d}_{y_j}$ is the deviation of $\mathbf{l}'_{y_i}$ from the vertical direction. $\mathbf{d}_{x_i}$ is defined similarly for a vanishing line $\mathbf{l}_{x_j}$ in $\mathbf{L}_x$, and $\mathbf{e}_y = [0\ 1\ 0]^T$ is used to measure the horizontal deviation.

In Eq. (14), the weight w for a line $\mathbf{l}$ is the original line length before transformation, normalized by the calibrated focal length f, i.e., $w = \|\mathbf{q} - \mathbf{p}\|/f$. The weights λ_v and λ_h are adaptively determined using initial rotation angles, as the constraint of picture frame alignment becomes weaker as rotation angles get bigger. We use:

$$\lambda_v = \exp\left(-\frac{\psi^2}{2\sigma_v^2}\right) \quad \text{and} \quad \lambda_h = \exp\left(-\frac{\theta^2}{2\sigma_h^2}\right), \quad (15)$$

where ψ and θ are calibrated rotation angles along x- and y-axes respectively. σ_v and σ_h are parameters to control the tolerances to the rotation angles. We fix them as $\sigma_v = \pi/12$ and $\sigma_h = \pi/15$ in our implementation.

Eye-level alignment

The eye-level in a photo is determined as a line connecting two vanishing points $\mathbf{v}_x$ and $\mathbf{v}_z$ [27]. Let $\mathbf{v}'_x$ and $\mathbf{v}'_z$ be the transformed vanishing points:

$$\mathbf{v}'_x = \frac{\mathbf{H}\mathbf{v}_x}{\mathbf{e}_z^T \mathbf{H}\mathbf{v}_x} \quad \text{and} \quad \mathbf{v}'_z = \frac{\mathbf{H}\mathbf{v}_z}{\mathbf{e}_z^T \mathbf{H}\mathbf{v}_z}.$$

Our objective is to make the eye-level horizontal, and the energy term is defined as:

$$E_{eye} = \left(\sum_i w_i + \sum_j w_j\right) \left(\mathbf{e}_y^T \mathbf{d}_e\right)^2,$$

where $\mathbf{d}_e = (\mathbf{v}'_z - \mathbf{v}'_x) / \|\mathbf{v}'_z - \mathbf{v}'_x\|$, and w_i and w_j are weights used in Eq. (14). Since eye-level alignment should be always enforced even when a photo contains lots of vanishing lines, we weight E_{eye} by the sum of line weights to properly scale E_{eye} with respect to E_{pic}.

Perspective distortion

Perspective distortion of a cuboid can be measured using Perkins's law [25], as illustrated in Fig. 3. To apply it, we have to detect corner points that are located on vertices of a cuboid. We first extract points where the start or end points of vanishing lines from two or three different axes meet. We then apply the mean-shift algorithm [28] to those points to remove duplicated or nearby points. We also remove corner points with too small corner angles. Fig. 4 shows a result of this method.

We use the extracted corner points to measure perspective distortion under Perkins's law. For each corner point, we draw three lines connecting it to the three vanishing points. We then measure angles between the three lines to see if Perkins's law is violated or not:

$$^{\forall}\mathbf{c}_i, \quad \min\left(\alpha_{i_1}, \alpha_{i_2}, \alpha_{i_3}\right) > \frac{\pi}{2} \quad (16)$$

where $\mathbf{c}_i$ represents a corner point. We only considers fork junctures, since arrow junctures can be transformed to fork junctures by flipping the direction of an edge.

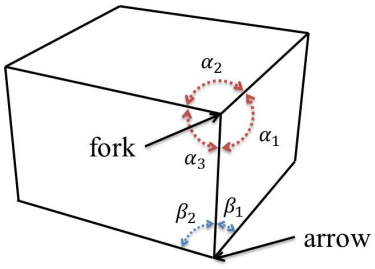


Fig. 3. Perkins's law. Vertices of a cube can be divided into two categories; fork and arrow junctures. For a fork juncture, α_1, α_2, and α_3 should be greater than $\pi/2$. For an arrow juncture, both β_1 and β_2 should be less than $\pi/2$, and sum of the two angles should be greater than $\pi/2$. Vertices that violate the above conditions will not be perceived as vertices of a cube to the viewer.

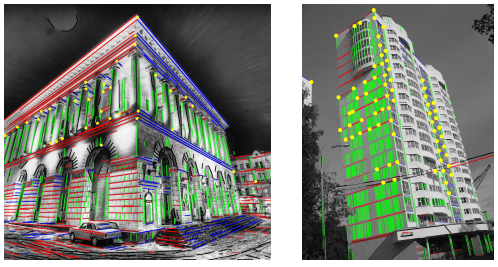


Fig. 4. Results of our corner point extraction. Extracted points are marked as yellow dots.

Image distortion

To accurately measure image distortion, we should detect circles and other important features in the input photo, which is a hard problem. We instead use an approximation in our system.

We first detect low-level image edges using Canny detector [29], then remove edge pixels that are nearby straight lines. Assuming the remaining edge pixels are from curved lines that could be originated from some features (see Fig. 5), we measure distortions of these pixels using the following Jacobian measure:

$$E_{reg} = \lambda_r \sum_i \left\{ \det\left(J(\mathbf{p}_i)\right) - 1 \right\}^2,$$

where $\mathbf{p}_i$ is a remaining edge pixel, $J(\cdot)$ is the Jacobian matrix, and $\det(\cdot)$ is the determinant. Jacobian matrix of a pixel $\mathbf{p}$ is discretely computed; let $\mathbf{q}$ and $\mathbf{r}$ be two neighbor pixels of $\mathbf{p}$, so that $\mathbf{p} = (x, y)^T$, $\mathbf{q} = (x+1, y)^T$ and $\mathbf{r} = (x, y+1)^T$. Then the Jacobian matrix of $\mathbf{p}$ under a homography $\mathbf{H}$ is approximated as:

$$J(\mathbf{p}) = \begin{bmatrix} \left(\frac{\mathbf{Hq}}{\mathbf{e}_z^T \mathbf{Hq}} - \frac{\mathbf{Hp}}{\mathbf{e}_z^T \mathbf{Hp}} \right)^T \\ \left(\frac{\mathbf{Hr}}{\mathbf{e}_z^T \mathbf{Hr}} - \frac{\mathbf{Hp}}{\mathbf{e}_z^T \mathbf{Hp}} \right)^T \end{bmatrix}.$$

This energy increases when non-rigid transforms are applied to the pixels causing distortions of features. For λ_r, we used a small value 10^{-4}.

Fig. 5. Feature detection: (left) original; (right) detected curved edge pixels. Some important features have been detected, such as human heads and letters, which should not be distorted.

Focal length difference

Our reprojection model for a homography allows different focal lengths along x- and y-axes for more natural results. However, we do not want the two lengths to differ too much. To enforce this property, we define the following energy:

$$E_{focal} = \lambda_f (f_{1x} - f_{1y})^2,$$

where we set $\lambda_f = (4/f)^2$ in our implementation.

4.3 Energy function minimization

Combining all the energy terms, the energy function we want to minimize for upright adjustment becomes:

$$\arg\min_{\mathbf{H}} \; E_{pic} + E_{eye} + E_{reg} + E_{focal} \qquad (17)$$
$$\text{subject to Eq. (16)}.$$

We have 9 unknowns to be optimized: $\mathbf{K}_1$, $\mathbf{R}_1$ and $\mathbf{t}_1$ are consist of f_{1x}, f_{1y}, u_1, v_1, ψ_1, θ_1, ϕ_1, t_x and t_y, as defined in Eq. (2). However, u_1 and v_1 simply shift the result image after the transformation, and we set $u_1 = u_0$ and $v_1 = v_0$. We thus optimize Eq. (3) with respect to 7 parameters. To initialize the variables, we use $f_{1x} = f_{1y} = f$, $\psi_1 = 0$, $\theta_1 = 0$, $\phi_1 = -\phi$, and $t_x = t_y = 0$, where f and ϕ are values obtained by camera calibration.

This energy function is non-linear and cannot be solved in a closed form. In practice, we use the numerical approach using `fmincon` in Matlab to minimize the energy function [30], [31]. Although global optimum is not guaranteed, this approach works quite well in practice.

5 Results

We implemented our algorithms using Matlab. For experiments, we used a PC with Intel Core i7 CPU (no multi-threading) and 6GB RAM. The processing time is reported in Table 1. Camera calibration took the largest portion of the time, followed by homography optimization and feature point detection. In our implementation, we downsized the input image to about 1M pixels for calibration and computing the homography $\mathbf{H}$, and applied the computed $\mathbf{H}$ to the original. All parameters were fixed in our experiments; Table 2 summarizes parameter values we used.

TABLE 2
Parameters of our method and the values used in our experiments. All the examples are produced with the same parameter set. Refer to the text for more explanation of the parameters.

Parameter	Value	Explanation
	Line segment detection	
Image size	1,280	Maximum image size in pixel
	Camera calibration	
λ_f, λ_c	$0.04, (10/W)^2$	E_K
$\lambda_\psi, \lambda_\theta, \lambda_\phi$	$(4/\pi)^2, (3/\pi)^2, (6/\pi)^2$	E_R
λ_M	$(24/\pi)^2$	$E_{M\|K,R}$ and $E_{A\|K,R}$
λ_L	0.01	$E_{L\|M}$
λ_A	0.02	$E_{L\|A}$
δ	2	Maximum distance between $\mathbf{l}$ and $\mathbf{v}$
n	2,000	Number of hypothesized vanishing points in $\mathbf{V}$
k_c	9	Number of elements in $\mathbf{V}_c$
	Homography optimization	
σ_v, σ_h	$\pi/12, \pi/15$	Adaptive weights for E_{pic}
λ_r	10^{-4}	E_{reg}
λ_f	$(4/f)^2$	E_{focal}

TABLE 1
Processing time for each component of our method. We measured the mean and standard deviation of the processing times for the images used in our user study. Time for image transform is not included since it only depends on the image size.

Component	Time (sec)
Line segment detection	0.59 ± 0.10
Camera calibration	3.39 ± 1.61
Corner & curved edges detection	1.06 ± 0.21
Homography optimization	1.03 ± 0.97
Total	$\mathbf{6.46 \pm 1.96}$

5.1 Evaluation of our camera calibration method

We compared our calibration method with two state-of-the-art techniques - Tardif [22] and Tretyak *et al*. [21], and using two datasets - York Urban [15] and Eurasian Cities [21]. For the results of Tardif [22] and Tretyak *et al*. [21], we used authors' implementations.

Fig. 6 shows a comparison with Tardif [22] using the accuracy of the estimated focal length. Fig. 7 shows comparisons with Tretyak *et al*. [21] using the accuracy of the estimated eye-level[2]. We also report our results using the Manhattan world assumption (no $E_{A|K,R}$, and $E_{L|M,A}$ becomes $E_{L|M}$) for comparison. Eye-level estimation is important due to the high sensitivity of the human perception to such misalignment. Our method achieved better eye-level accuracy results for the York Urban dataset which follows well the Manhattan assumption. Our method (the Atlanta version) showed better results also for the Eurasian Cities dataset which contains images not following the Manhattan assumption. The advantage comes mostly from our camera and scene priors.

Effect of the Atlanta world assumption

The Atlanta world assumption is less restrictive than the Manhattan world assumption and can be applied to most images with man-made structures. Tretyak *et al*. [21] used this assumption to estimate eye-level and zenith, and reported highly accurate results for the eye-level estimation. In our calibration method, we utilize this assumption with the two energy terms: $E_{A|K,R}$ and $E_{L|M,A}$. By plugging the Atlanta assumption in our calibration method we could estimate camera parameters more accurately and robustly. To show the effectiveness of our method, we applied our method to some challenging examples, as shown in Fig. 10, which suggests that our scene prior leads to more accurate calibration result.

Robustness to random initialization

We use a randomized approach for calibration and can get different camera parameters from the same image, some are less accurate than others. In practice, we found this to be less of an issue as our method is quite stable under the Atlanta world assumption on a wide variety of images we have tested; Fig. 9 illustrates the robustness of the proposed method.

Effect of the soft prior for the center of projection

Most previous calibration methods assume that the center of projection of the image is known [17], [21], [22]. In our method, we use a soft prior rather than using a fixed point. Although accurate estimation of the center of projection is impractical [17], our prior

2. The method of Tretyak *et al*. does not assume the Manhattan world and estimates the eye-level only, so we could not compare other quantity produced by our method, such as vanishing points and camera parameters.

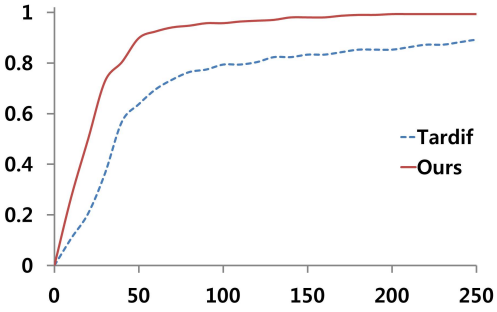


Fig. 6. Cumulative histograms of the errors in focal length estimation in the York Urban dataset. (x-axis) focal length error in pixels from the ground truth; (y-axis) proportion of the images in the dataset.

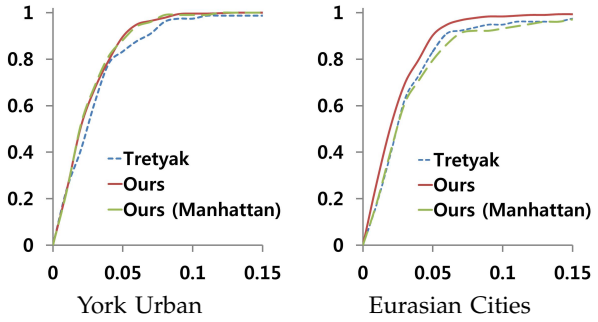


Fig. 7. Cumulative histograms of the errors in eye-level estimation. The red lines show results of our full method (with the Atlanta assumption). The green dashed lines show a degenerate version of our method with the Manhattan world assumption (no additional horizontal directions are detected). (x-axis) eye-level estimation error; (y-axis) proportion of the images in the date set. See [21] for the details of the error metric.

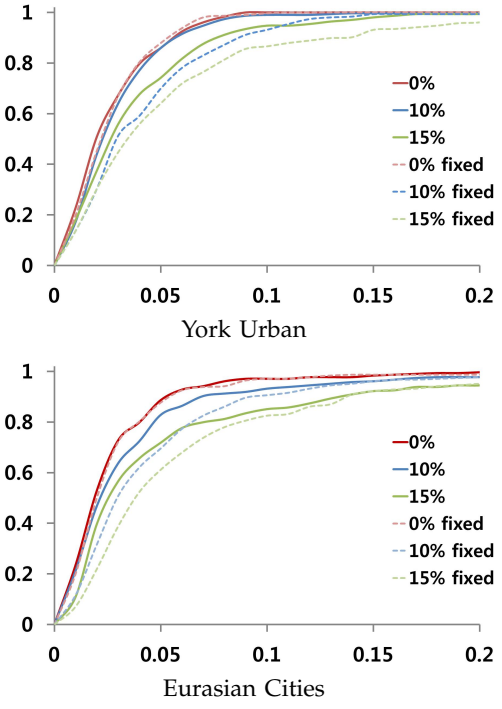


Fig. 8. Robustness of eye-level estimation to image cropping (eccentric center of projection). For each curve, we cropped the images to move their center of projections by 0, 10 and 15% of their width (or height, we took the maximum). The solid lines represent our method (soft center of projection prior), while the dotted lines represent our method where we fixed the center of projections to the image center, as commonly done by other method. Axes are as in Fig. 7. The soft prior is clearly advantageous in case of a crop.

can leverage the estimation error while providing more robustness for the estimation of focal length and eye level.

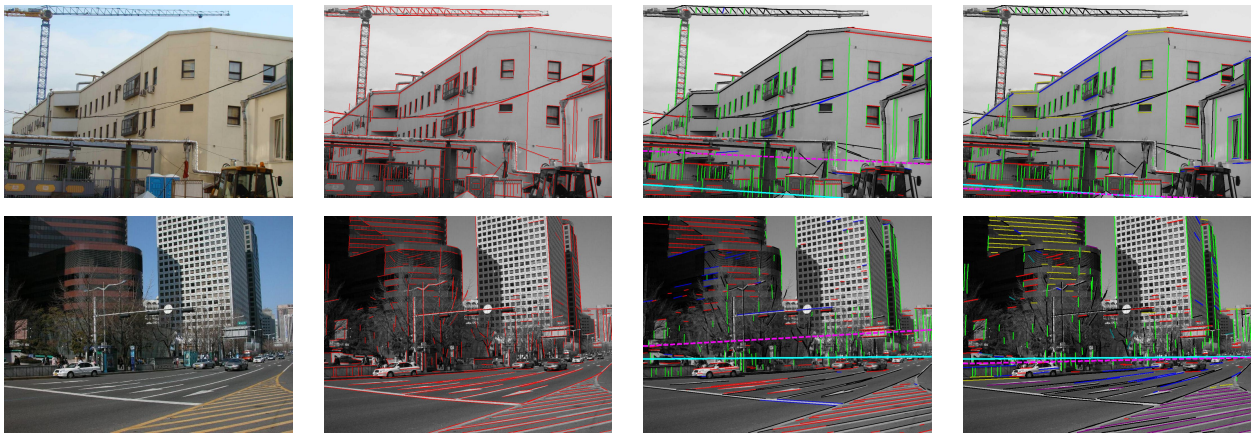
To verify the robustness of our soft prior, we conducted an experiment as follows: For both datasets, we randomly cropped images in each dataset. We assumed the center of projection is originally the image center so that in a cropped image the center of projection moves from its ground truth value. To maximize the amount of offset, we cropped either the top or bottom part of images (and left or right) but not both. Then we estimated camera parameters with cropped images with (1) using our soft prior and (2) using a fixed center of projection, set as the center of a cropped image. Fig. 8 shows estimation results; we could obtain more robust results with our soft prior.

5.2 Effects of upright adjustment criteria

Picture frame alignment is important for photos of big planar objects, such as facades of buildings and billboards. If picture frame alignment dominates other criteria, the adjustment result becomes similar to simple image rectification. However, its effect should diminish as the rotation angles of the camera increase, otherwise it will lead to undesirable distortion. Our system automatically handles this problem with the adaptive weight scheme (Eq. (15)) as well as the perspective and image distortion criteria, generating a better result. Fig. 11 shows the effectiveness of our automatic algorithm.

Eye-level alignment is always important, and its importance becomes more noticeable as the effect of picture frame alignment gets weaker (Fig. 1d). Perspective distortion control prevents too strong adjustment that could make objects in the image appear distorted (Fig. 12). We found that allowing the focal lengths in x- and y-directions to slightly deviate with Eq. (2), resulting in a small aspect ratio change, is often useful to ease the perspective distortion. We also found that artists do similar adjustments manually to make their results feel more natural in Keystone correction of photos[3].

3. http://youtu.be/QkG241258FE



(a) Original (b) Line segments (c) Manhattan world (d) Atlanta world

Fig. 10. Calibration results under Manhattan (c) and Atlanta (d) world assumptions. Ground truth eye levels are drawn with solid cyan lines and their estimates are drawn with dashed pink lines. In (c), non-orthogonal directions are detected. In contrast, three orthogonal directions are correctly detected in (d) and the eye level is accurately estimated with the help of additional horizontal directions (solid magenta and yellow lines).

Original 20% rectified 50% rectified 100% rectified Our result

Fig. 11. Comparison with rectification results. Eye-levels are corrected for all but original images. In the top row, 100% rectified result looks better than that of 20% rectified. However for the example in the bottom row 100% rectified result has too much distortion. Our optimization method can handle both images with no user interaction.

5.3 User study

To objectively evaluate the proposed system, we conducted a user study on Amazon Mechanic Turk. We used 40 pairs of original and adjusted images for the user study where 50 independent participants were asked to select the preferred image from each pair.

Setting

To ensure the quality of the user study results, we implemented two sanity check steps to filter out "bad" results produced by careless users. Firstly, we measured the decision time for each user on each pair of images, and if it is too short (less than one second), we asked the user to re-examine the images more carefully. Secondly, during the user study, for every 4 pairs of images we showed a duplicated pair that has already been shown before (10 pairs of duplication in total). For each duplicated pair we flipped the order of the two images compared to the first time it was shown. If a user made inconsistent decisions on more than half of these duplicate pairs, his/her results were rejected.

Effect of cropping

After upright adjustment, we have to crop the result photo to remove blank regions near the image boundary caused by homography projection (Fig. 13a). It thus makes the original and the result photos to have different aspect ratios, scales and contents. Since these differences may affect users' preference, we conduct two user studies by taking into account the cropping issue.

In the first study, we first crop out the blank boundary regions in the result photo (Fig. 13a). Then the original photo is cropped/resized to have the same size as the cropped result photo. When we crop/resize the original photo, we maximize the overlapping region with the cropped result photo. To compute the overlapping region, we use the back-projection of the cropped result onto the original photo (Fig. 13c). Fig. 14(a) shows the preference for each image in this

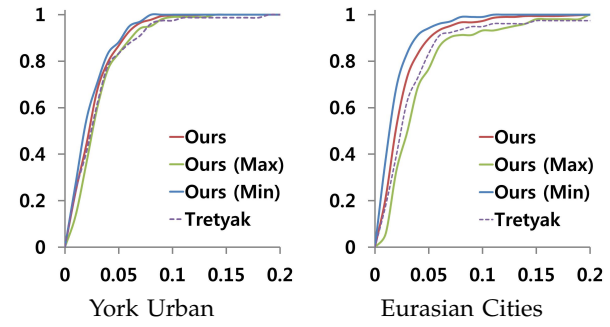


Fig. 9. Robustness of eye-level estimation to random initialization. We calibrated each image in the two datasets five times, each with different random initializations, and present plots for the average error (red, from Fig. 7), as well as the minimum (blue) and maximum (green) errors. Green lines represent the cumulative histograms of the minimum error values among the five for each image, and blue lines represent those of the maximum values. Axes are as in Fig. 7. Our method is quite stable w.r.t random initialization.

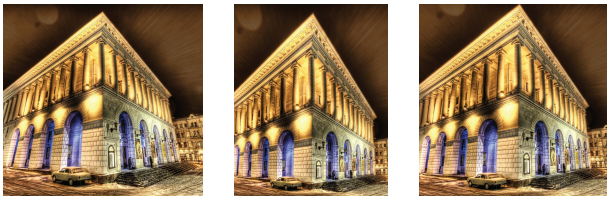
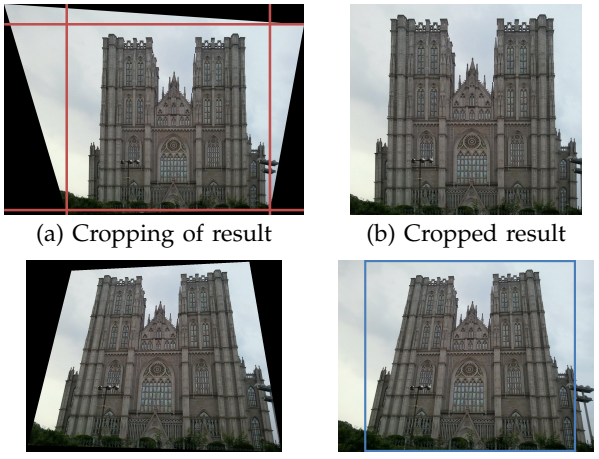


Fig. 12. Perspective distortion control. (left) original; (middle) w/o perspective distortion constraint; (right) w/ the constraint.

study; on average our result was preferred in 80.2% of comparisons.

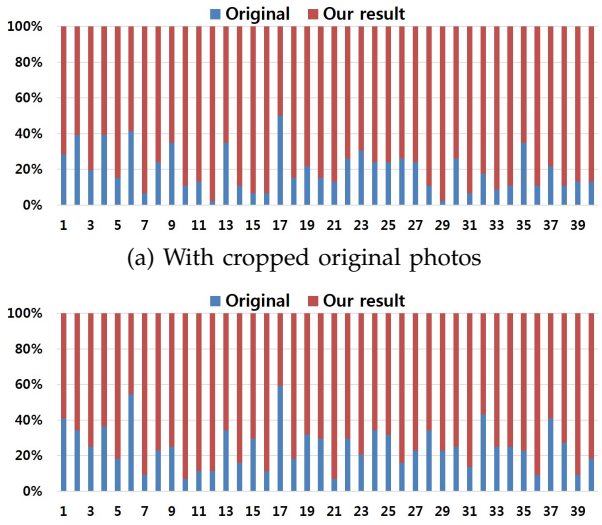
In the second study, all the setting are the same except that the original photos are not cropped, thus they have different sizes and aspect ratios from the result photos. The results of this study are shown in



(a) Cropping of result (b) Cropped result

(c) Back-projection of (b) (d) Cropping of original

Fig. 13. Cropping of the original and result photos.



(a) With cropped original photos

(b) With uncropped original photos

Fig. 14. User study result. (x-axis) photo index; (y-axis) preference percentage.

Fig. 14(b); the average preference was 74.5% in this case. The ratio is lower than that of the first user study, as in some examples some users preferred the uncropped original photos which have more content than the cropped result photos. Nevertheless, the overall preference ratio of our method is still high in this case, indicating that the proposed upright adjustment can largely improve the perceptual quality of these photos in general.

Discussion

To verify the effectiveness of both picture frame and eye-level alignments, we divided the images used in the user study into three categories: (1) "Picture frame" images that can be corrected by applying picture frame alignment (tilt) alone; (2) "Eye-level" images that can be corrected by eye-level alignment (rotation) alone; and (3) "Both" images that can be corrected by applying both alignments. The average preference in each category is shown in Table 3. It shows that users clearly preferred our results in all three categories. It also suggests that users preferred our results much more in category 2 and 3, where rotations were applied.

Fig. 15 shows three examples (6, 17, 32) where users did not clearly prefer our results. For photo 6 and 17, people preferred original photos that look more natural than our results. For photo 32, other photo aesthetics criteria such as sharpness and content coverage affected users' decisions.

Limitations

Our system uses a single homography to correct a photo under the uniform depth assumption for a scene. Although this assumption typically does not hold, in practice our method generates satisfying

(a) Photo 6 (b) Photo 17 (c) Photo 32

Fig. 15. Results not highly preferred in the user study. Cropped original images are shown in this figure.

TABLE 3
Relationship between alignment type and user preference. Preferences in parentheses are results from uncropped original photos.

Type	Photo IDs	Preference
Pic. frame	1, 3, 4, 5, 6, 10, 13, 17, 22, 23, 24, 40	70.5 (68.6)%
Eye level	8, 11, 15, 18, 19, 28, 29, 32, 33, 35, 37	84.0 (70.9)%
Both	2, 7, 9, 12, 14, 16, 20, 21, 25, 26, 27, 30, 31, 34, 36, 38, 39	83.5 (81)%

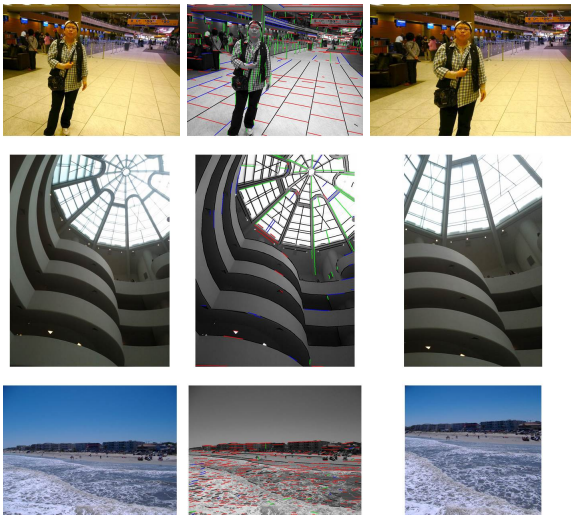


Fig. 16. Examples of failure cases. Top: the human face and body in the input image are distorted in the upright adjustment process. Middle: the scene does not satisfy the Atlanta world assumption and the camera calibration method produces wrong parameters, leading to a distorted adjustment. Bottom: camera calibration fails due to a large amount of false line segments, and the result image has too much distortion.

results for a wide range of images, due to the robustness of perspective perception [25]. However, for human faces or other important features in a photo, the adjusted result may contain noticeable distortion. For camera calibration, although the Atlanta world assumption used as our scene prior holds well in general, it sometimes can be broken and the calibration algorithm may fail to estimate accurate camera parameters. Finally, we use line segments as basic primitives for our method, and it might not produce a good result if the detected line segments contain a large amount of ourliers. Fig. 16 shows some examples of these failure cases.

6 Conclusion

We proposed an automatic system that can adjust the perspective of an input photo to improve its visual quality. To achieve this, we first defined a set of criteria based on perception theories, then proposed an optimization framework for measuring and adjusting the perspective. Experimental results demonstrate the effectiveness of our system as an automatic tool for upright adjustment of photos containing man-made structures.

As future work, we plan to incorporate additional constraints to avoid perspective distortions on faces or circles. We also plan to improve the calibration method by considering more image features used in recent state-of-the-art methods [32], [33]. Extending our method to video will be an interesting future research direction.

Acknowledgments

We would like to thank Patrick Theiner and Trey Ratcliff for the use of their photographs in this paper. This work was supported in part by Industrial Strategic Technology Development Program of KEIT (KI001820), IT/SW Creative Research Program of NIPA (NIPA-2012-H0503-12-1008), and Basic Science Research Program of NRF (2012-0008835).

References

[1] H. Lee, E. Shechtman, J. Wang, and S. Lee, "Automatic upright adjustment of photographs," in *Proc. CVPR*, 2012, pp. 877–884.

[2] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *Proc. ECCV*, 2006, pp. 288–301.
[3] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *Proc. CVPR*, 2006, pp. 419–426.
[4] Y. Luo and X. Tang, "Photo and video quality evaluation: Focusing on the subject," in *Proc. ECCV*, 2008, pp. 386–399.
[5] S. Dhar, V. Ordonez, and T. L. Berg, "High level describable attributes for predicting aesthetics and interestingness," in *Proc. CVPR*, 2011, pp. 1657 –1664.
[6] L. Yao, P. Suryanarayan, M. Qiao, J. Wang, and J. Li, "Oscar: On-site composition and aesthetics feedback through exemplars for photographers," *IJCV*, vol. 96, pp. 353–383, 2012.
[7] L. Liu, R. Chen, L. Wolf, and D. Cohen-Or, "Optimizing photo composition," *Computer Graphic Forum*, vol. 29, pp. 469–478, 2010.
[8] L.-K. Wong and K.-L. Low, "Saliency retargeting: An approach to enhance image aesthetics," in *Proc. WACV*, 2011, pp. 73–80.
[9] J. Park, J. Lee, and Y. Tai, "Modeling photo composition and its application to photo re-arrangement," in *Proc. ICIP*, 2012.
[10] A. Gallagher, "Using vanishing points to correct camera rotation in images," in *Proc. Computer and Robot Vision*, 2005, pp. 460–467.
[11] R. Carroll, A. Agarwala, and M. Agrawala, "Image warps for artistic perspective manipulation," *ACM TOG*, vol. 29, p. 1, 2010.
[12] J. M. Coughlan and A. L. Yuille, "Manhattan World: Compass direction from a single image by bayesian inference," in *Proc. ICCV*, 1999, pp. 941–.
[13] J. Kosecka and W. Zhang, "Video compass," in *Proc. ECCV*, 2002, pp. 476–490.
[14] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004, ch. 8, pp. 213–229.
[15] P. Denis, J. H. Elder, and F. J. Estrada, "Efficient edge-based methods for estimating Manhattan frames in urban imagery," in *Proc. ECCV*, 2008, pp. 197–210.
[16] F. M. Mirzaei and S. I. Roumeliotis, "Optimal estimation of vanishing points in a Manhattan world," in *Proc. ICCV*, 2011.
[17] H. Wildenauer and A. Hanbury, "Robust camera self-calibration from monocular images of manhattan worlds," in *Proc. CVPR*, 2012, pp. 2831–2838.
[18] Z. Zhang, A. Ganesh, X. Liang, and Y. Ma, "Tilt: Transform invariant low-rank textures," *IJCV*, vol. 99, pp. 1–24, 2012.
[19] D. Aiger, D. Cohen-Or, and N. J. Mitra, "Repetition maximization based texture rectification," *Computer Graphic Forum*, vol. 31, pp. 439–448, 2012.
[20] G. Schindler and F. Dellaert, "Atlanta world: An expectation maximization framework for simultaneous low-level edge grouping and camera calibration in complex man-made environments," in *Proc. CVPR*, 2004, pp. 203–209.
[21] E. Tretyak, O. Barinova, P. Kohli, and V. Lempitsky, "Geometric image parsing in man-made environments," *IJCV*, pp. 1–17, 2011.
[22] J.-P. Tardif, "Non-iterative approach for fast and accurate vanishing point detection," in *Proc. ICCV*, 2009, pp. 1250–1257.
[23] R. von Gioi, J. Jakubowicz, J.-M. Morel, and G. Randall, "LSD: A fast line segment detector with a false detection control," *IEEE PAMI*, vol. 32, pp. 722–732, 2010.
[24] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright, "Convergence properties of the nelder–mead simplex method in low dimensions," *SIAM J. on Optimization*, vol. 9, pp. 112–147, 1998.
[25] M. Kubovy, *The Psychology of Perspective and Renaissance Art*. Cambridge University Press, 2003.
[26] M. Freeman, *The Photographer's Eye: Composition and Design for Better Digital Photos*. Focal Press, 2007.
[27] J. D'Amelio, *Perspective Drawing Handbook*. Dover Publications, 2004.
[28] D. Comaniciu and P. Meer, "Mean Shift: A robust approach toward feature space analysis," *IEEE PAMI*, vol. 24, pp. 603–619, 2002.
[29] J. Canny, "A computational approach to edge detection," *IEEE PAMI*, vol. 8, pp. 679–698, 1986.
[30] *MATLAB Optimization Toolbox*. The MathWorks Inc., 2010.
[31] R. A. Waltz, J. L. Morales, J. Nocedal, and D. Orban, "An interior algorithm for nonlinear optimization that combines line search and trust region steps," *Math. Program.*, vol. 107, no. 3, pp. 391–408, 2006.
[32] Z. Zhang, Y. Matsushita, and Y. Ma, "Camera calibration with lens distortion from low-rank textures," in *Proc. CVPR*, 2011, pp. 2321–2328.
[33] A. Vedaldi and A. Zisserman, "Self-similar sketch," in *Proc. ECCV*, 2012.

Hyunjoon Lee received his B. Tech. degree in Computer Science and Engineering from POSTECH, South Korea in 2005. He is currently a PhD student in Computer Science and Engineering at POSTECH. His research interests include image manipulation, scene understanding, non-photorealistic rendering, and computer graphics.

Eli Shechtman is a Senior Research Scientist at Adobe Research. He received the B.Sc. degree in electrical engineering (magna cum laude) from Tel-Aviv University in 1996. Between 2001 and 2007 he attended the Weizmann Institute of Science where he received with honors his M.Sc. and Ph.D. degrees in Applied Mathematics and Computer Science. In 2007 he joined Adobe and started sharing his time as a post-doc with the University of Washington, Seattle. He was awarded the Weizmann Institute Dean prize for M.Sc. students, the J.F. Kennedy award (highest award at the Weizmann Institute) and the Knesset (Israeli parliament) outstanding student award. He received the best paper award at ECCV 2002 and a best poster award at CVPR 2004. His research interests include image and video processing, computational photography, object recognition and patch-based analysis and synthesis. He is a member of the IEEE and the ACM.

Jue Wang is a Senior Research Scientist at Adobe Research. He received his B.E. (2000) and M.Sc. (2003) from Department of Automation, Tsinghua University, Beijing, China, and his Ph.D (2007) in Electrical Engineering from the University of Washington, Seattle, WA, USA. He received Microsoft Research Fellowship and Yang Research Award from University of Washington in 2006. He joined Adobe Research in 2007 as a research scientist. His research interests include image and video processing, computational photography, computer graphics and vision. He is a senior member of IEEE and a member of ACM.

Seungyong Lee is a professor of computer science and engineering at the Pohang University of Science and Technology (POSTECH), Korea. He received the PhD degrees in computer science from the Korea Advanced Institute of Science and Technology (KAIST) in 1995. From 1995 to 1996, he worked at the City College of New York as a postdoctoral research associate. Since 1996, he has been a faculty member of POSTECH, where he leads the Computer Graphics Group. From 2003 to 2004, he spent a sabbatical year at MPI Informatik in Germany as a visiting senior researcher. From 2010 to 2011, he worked at Adobe Research in Seattle as a visiting professor. His current research interests include image and video processing, non-photorealistic rendering, 3D surface reconstruction, and graphics applications. He is a member of the IEEE Computer Society.