# Scene Geometry from Several Partly Cloudy Days

Nathan Jacobs, Scott Workman
Computer Science
University of Kentucky
{jacobs,smwork3}@cs.uky.edu

Richard Souvenir
Computer Science
UNC Charlotte
souvenir@uncc.edu

***Abstract*—We describe new methods for estimating the geometry of an outdoor scene from either (1) a single calibrated camera or (2) a network of arbitrary, uncalibrated cameras. Our methods do not require camera motion nor overlapping fields of view. We use simple geometric constraints based on appearance changes caused by cloud shadows and combine constraints from multiple days. We describe a linear method for calibrated cameras and a nonlinear method for an uncalibrated imaging system. We define these geometric constraints, describe new algorithms, and demonstrate the features of these algorithms on real and synthetic scenes.**

## I. Introduction

Knowledge of the 3D geometry of a scene can be useful in many applications in outdoor camera networks; however, standard approaches to estimating scene geometry that rely on camera motion [12] or controlled lighting [4] are not possible for images obtained from static outdoor cameras. In this domain, most approaches use either photometric cues from changes in the sun position [1], [3] to provide estimates of surface orientation or geometric cues based on cloud motion [6] to give direct constraints on scene shape. In this work, we focus on the cloud cue and generalize the problem setting to incorporate constraints provided by multiple days (with independent cloud motion directions). This eliminates the ambiguity inherent in only having a single cloud motion direction and enables partial shape estimation without camera calibration.

We assume that the dominant appearance variation in the scene is due to the shadows cast by moving clouds. Since our camera is static, each pixel has an intensity time series that reflects the pattern of clouds that passed between the sun and the imaged scene point. Since clouds translate due to cloud motion, nearby pixels that are inline with the wind will have a related, but temporally offset, intensity time series. We use signal processing methods to estimate this temporal delay and use the set of temporal delays between all pixels in the scene as the only image measurements to our algorithms. We define a geometric relationship between the scene shape, these temporal delays and the known cloud motion. Figure 1 provides an overview of our approach.

We focus on two common problem settings: one with a fully calibrated camera and one with an uncalibrated and general imaging model, which is suitable for very diverse camera networks. In both cases, we take as input video from a static camera observing a fixed-geometry scene. In the first problem setting, we derive a linear least-squares formulation that allows us to quickly solve for a full 3D scene model, with known metric scale. In the second setting, we are only able to reconstruct the 2D scene layout due to the ambiguity introduced by the orthographic lighting of the sun. The key *contributions* of this work include:

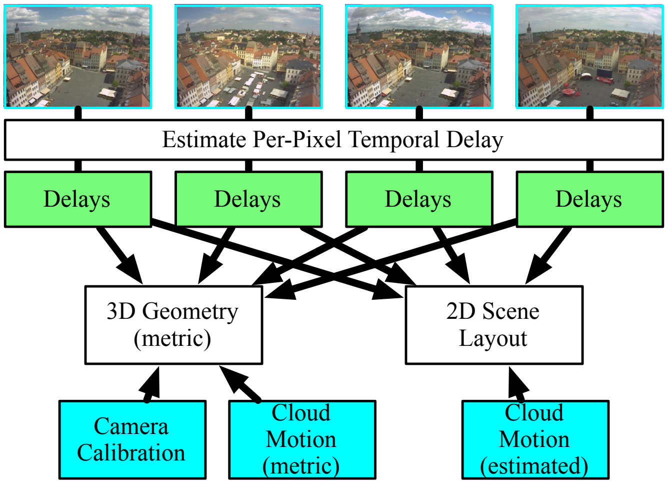


Fig. 1. Our methods combine per-pixel temporal delay estimates from multiple videos with information about the cloud motion and camera calibration to estimate scene geometry.

- describing the geometric relationship between scene geometry and cloud motion across multiple days (Section II);
- introducing a method for solving for scene geometry with known cloud motion and camera geometry (Section III-C); and
- introducing a method for solving for scene geometry with a general imaging model (Section III-D).

### A. Related Work

Several cues and approaches have been proposed for estimating shape information from video of an outdoor scene captured by a static camera [1], [2], [6], [9], [11]. These approaches all rely on strong assumptions about the scene, including clear sky conditions [1], [2], [9] or constant albedo [1], [9]. The most commonly used cues are based on appearance changes due to sun motion, such as photometric changes [1], [9] and cast shadows geometry [2]. Such approaches require clear sky conditions, or a method for filtering out cloudy conditions. Unlike our approach, these methods are currently restricted to estimating surface normals [1], [9], which would be challenging to convert into depth estimates in a general outdoor environment, or only provide depth estimates around shadow casters [2]. An additional limitation of all of these approaches is that the scale of the scene is ambiguous.

Methods that use moving cloud shadows to estimate scene geometry [6], [7] provide direct constraints on scene depth. This previous work introduces the delay-based constraint, which relates the distance between pairs of pixels to the rate of cloud motion and the time of transit (i.e., "distance equals rate times time"). In the case where the video framerate and the cloud velocity is known, this constraint provides a metric estimate of the scene geometry. Our work extends this line of research in several directions. For the case of a single calibrated camera, we show how to extend the depth-from-delay cue to incorporate constraints from multiple days. We show how this overcomes the inherent ambiguity in the constraints for a single day and show how additional days increase the robustness to various types of errors. In addition, we show how the need for camera calibration can be eliminated. For the case of an uncalibrated camera, or camera network, we solve for a 2D scene layout by incorporating multiple days of data.

## II. Problem Statement

Our goal is to estimate the shape of an outdoor scene given a collection of short videos (e.g. 20 minutes) of the scene captured by a static outdoor camera or camera network. We propose to use sunlight attenuation due to moving clouds to define geometric constraints on the scene layout. In this section, we define the image formation model and core assumptions.

### A. Image Formation Model

The brightness, $I(p,t,d)$, of a pixel, $p$, at particular time/frame, $t$, in a video captured on day, $d$, is a function of static scene geometry, time-varying lighting conditions, and camera properties, such as the focal length, location and orientation. We define a simple image formation model:

$$I(p,t,d) \approx \rho^1(p,d)S(p,t,d) + \rho^0(p,d). \tag{1}$$

This model assumes that the albedo terms, $\rho^0(p,d), \rho^1(p,d)$, are constant per pixel for each video, but that there exists a time-varying cloud shadow term, $S(p,t,d) \in \mathcal{R}$, which depends on both the direction of the sun, the cloud motion, and the light attenuation due to passing clouds. In this model, we assume that appearance variations over short periods of time, i.e., in a single video, are due only to sunlight attenuation caused by moving cloud shadows.

The cloud-shadow term has interesting geometric properties that both hinder and aid the ability to estimate scene geometry. Let $x_p$ be the 3D location of the point imaged by pixel $p$, and the cloud attenuation mask, $S(x_p,t,d)$, be the source of the time-varying cloud shadow term, $S(p,t,d)$. Assuming that the sun is an orthographic light source, all points along the sun direction, $\vec{L}_d$, will be similarly attenuated by the same set of clouds. In other words, $S(x_p,t,d) \approx S(x_p + \lambda\vec{L}_d, t, d)$ which implies that there is a family of pixels that are similarly attenuated at the same time. This implies that without camera calibration information, it is only possible to estimate a 2D scene layout, relative to the sun direction.

The key property that allows for scene structure estimation is the translational motion of clouds, which causes the cloud-shadow term to have a particular form. Pairs of nearby 3D points that are directly in-line with the cloud motion often have very similar patterns of intensity changes, just temporally offset, due to the motion of clouds. That is, $S(x,t,d) \approx S(x + w\Delta t, t + \Delta t, d)$ where $w$ is the cloud velocity. In the direction orthogonal to the cloud velocity, points are affected by a different set of clouds, but, in general, the time series of attenuations of nearby points are still very similar.

### B. Geometric Implications

We now formulate a constraint that describes the relationship between distance, rate (cloud velocity) and time [6]. Given two 3D points, $x_p, x_q$, directly in line with the cloud motion vector $w_d$, we define the following relation:

$$(x_p - x_q) = \lambda_{dqp} w_d$$

where $\lambda_{dqp}$ describes the time for the wind to move from $x_q$ to $x_p$. For another point, $x_r$, not directly in line with $x_p$ and $x_q$, this relation will not hold. However, we can define the following relation:

$$w_d^\intercal (x_p - x_r) = \lambda_{drp} w_d^\intercal w_d \tag{2}$$

which projects the displacement onto the wind direction vector. For a matrix of world points, $X = [x_1, x_2, \dots, x_N]$, we can generalize to the following set of constraints:

$$w_d^\intercal X H = w_d^\intercal w_d \Lambda_d$$

where $H = \mathcal{R}^{n \times K}$ is a matrix with elements in $\{0,-1,1\}$ such that $XH$ computes differences between pairs of columns of $X$ (to represent the left-hand side of (2)), and $\Lambda_d \in \mathcal{R}^{1\times K}$ is a matrix of temporal delays. The value of $K$ represents the number of pairs of pixels; in the simplest case, $K = \frac{N(N-1)}{2}$, where all possible pairs of pixels are considered. It is also possible, as demonstrated in Section III-A to estimate a single globally consistent set of delays in advance, and $K = N - 1$. Figure 2 provides a visual depiction of these constraints, and further shows how there is a fundamental ambiguity when only one wind direction is available.

In the remainder of this paper, we describe two methods for estimating the scene geometry, $X$, given temporal delay estimates, $\{\Lambda_d\}$, and known wind directions, $\{w_d\}$, from multiple days and present results on real and synthetic datasets.

## III. Methods

Given a set of videos recorded by a stationary imaging system on multiple partly cloudy days with linearly independent cloud motion directions, we translate the constraints defined in the previous section into a reconstruction error based framework for estimating scene geometry. Given the temporal delays, $\Lambda_d$, for a collection of timesteps, $\{d\}$, we define an objective function that relates the delays, $\{\Lambda_d\}$, and wind velocity, $\{w_d\}$, at each timestep, with the 3D layout of the scene, $X$:

$$f(X, \{w_d\}) = \sum_{d=1}^{D} \| w_d^\intercal X H - w_d^\intercal w_d \Lambda_d \|_2^2. \tag{3}$$

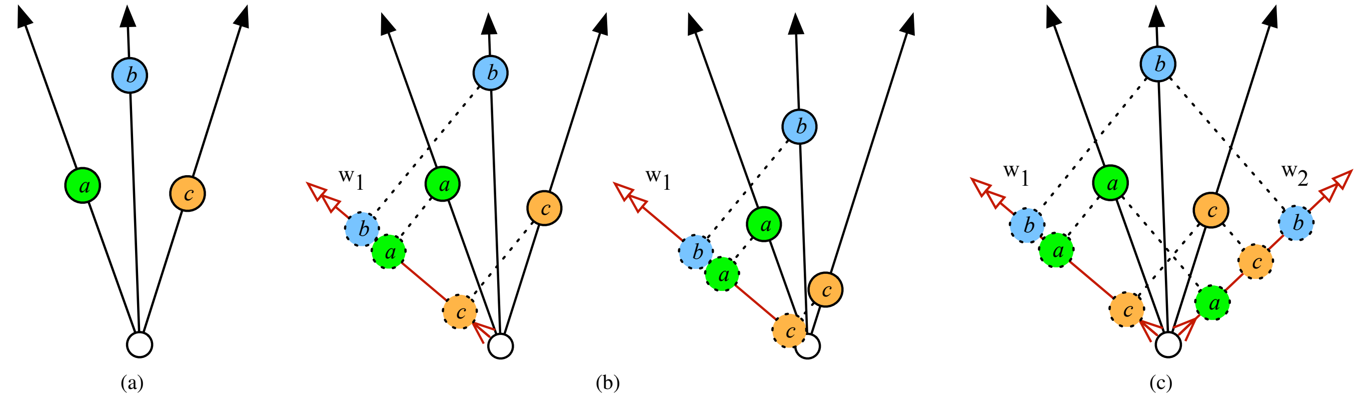


Fig. 2. A visual depiction of the ambiguity inherent from data from a single time period. (a) A 2D view of three points imaged by a camera. (b) For a single wind vector, $w_1$, there are many possible point configurations that lead to the same temporal delays (relative displacements along $w_1$). (c) By considering an additional wind vector, $w_2$, we eliminate the ambiguity in the point locations.

In the remainder of this work, we show how to use this framework to solve for shape in two specific scenarios. We begin both methods by estimating the temporal delays, $\Lambda_d$, for each video and converting this into a minimal set of temporal delays. We then obtain estimates of the cloud motion vectors, $\{w_d\}$, using one of several strategies. Given this information, we then estimate the scene geometry using one of two methods: (1) a linear method that requires a known camera geometry, or (2) a nonlinear method that makes almost no assumptions on the imaging system.

### A. Estimating Temporal Delays

On each day, $d \in \{1, \dots, D\}$, we assume the motion of the clouds is dominated by a single wind velocity vector, $w_d = [u_d, v_d, 0]^\mathsf{T}$, where the third coordinate is zero as it is assumed the wind is traveling parallel to the ground. This means there is a fixed temporal delay for each pair of pixels. For each video, we compute the temporal delay between all pairs of pixels, $\Lambda_d$, by comparing the pixel intensity time series for each pair using a two-stage process [6]. The first stage measures delay using pairwise cross correlation to obtain an initial, integer estimate of $\lambda_{dpq}$ and correlation, $\rho_{dpq}$. For sub-frame accuracy, this estimate is refined by choosing the maximum of a quadratic model of correlation given delay for a small window around the original estimate, $\lambda_{dpq}$.

Figure 3 shows the estimated delays, $\lambda_{dpq}$, and the delay-corrected correlation, $\hat{\rho}_{dpq}$, for an individual pixel (marked in red) compared to all other pixels in the scene from a typical outdoor video. In this scene, the isocontours of the delay are largely horizontal, except for near buildings; this is because the wind is blowing almost directly toward the camera. The values of $\hat{\rho}_{dpq}$ decrease as the distance increases away from the line defined by the direction of the wind.

For a set of $N$ pixels, there are $N(N-1)/2$ unique temporal delay estimates. However, due to the transitive relationship between temporal delays these are not linearly independent. In the noise-free case, there are only $N-1$ independent temporal delays. We convert the full set of temporal delay estimates into a minimal set of delays using linear least squares. This significantly reduces the computational cost of the subsequent methods, but is not explicitly required.

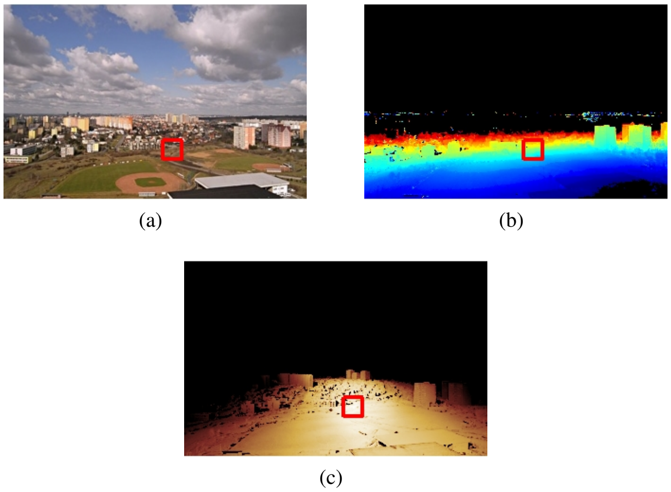


Fig. 3. Temporal delay and delay-corrected correlation for one pixel, (b) shows the temporal delay estimate and (c) shows the delay-corrected correlation values (images reproduced with permission from [6]).

### B. Estimating Cloud Motion

The exact velocity of the clouds is generally not available and must be estimated. There are multiple possible methods for estimating the cloud motion that are useful in different situations. First, given the approximate location and time that an image was captured, the cloud motion can be estimated by querying a weather database for the ground wind velocity. When available, these estimates are often very close to the true cloud motion. In practice, the errors are typically independent, so additional days of data can be used to minimize the variance. Second, when a collocated calibrated camera that sees a substantial portion of the sky is available, the cloud motion direction can be estimated using existing methods [8]. This gives the wind direction(s), and, if the cloud height can be estimated, the cloud velocity. Third, if the camera observes a sufficient portion of a ground plane, and it is possible to rectify the image [10], an affine, or metric, estimate of the cloud velocities can be obtained by computing the image-space cloud-shadow velocity (which is straightforward given the temporal delay estimates) and then rectifying these estimates.

For the types of video captured from static outdoor webcams, the third approach is the most widely applicable, so we adopt it for the experimental evaluation.

### C. Scenario 1: Calibrated Projective Camera

In the simplest scenario, with a calibrated projective camera, the 3D world-rays, $r_p$, that correspond to each pixel are known. Therefore, the scene geometry is $X = RZ$ where $R = [r_1, \ldots, r_n]$ is the matrix of pixel rays and $Z$ is a diagonal matrix of depths, $z$. Starting from (3), this problem setting results in a linear least-squares problem of the following form:

$$\begin{aligned} f(z) &= \sum_{d=1}^{D} \|w_d^\mathsf{T} RZH - w_d^\mathsf{T} w_d \Lambda_d\|_2^2 && (4)\\ &= \sum_{d=1}^{D} \|z^\mathsf{T} \hat{R}_d H - w_d^\mathsf{T} w_d \Lambda_d\|_2^2 && (5)\\ &= \sum_{d=1}^{D} \|z^\mathsf{T} Q_d - V_d\|_2^2 && (6)\\ &= \mathrm{Tr}(z^\mathsf{T}\mathbf{Q} - \mathbf{V})^\mathsf{T}(z^\mathsf{T}\mathbf{Q} - \mathbf{V}) && (7)\\ \frac{\partial f}{\partial z} &= (z^\mathsf{T}\mathbf{Q} - \mathbf{V})\mathbf{Q}^\mathsf{T} && (8)\\ z^\star &= \mathbf{V}\mathbf{Q}^\dagger && (9) \end{aligned}$$

where $\hat{R}_d$ is a diagonal matrix with the elements of $w_d^\mathsf{T} R$ along the diagonal, $\mathbf{Q} = [Q_1 \ldots Q_D]$, and $\mathbf{V} = [V_1 \ldots V_D]$. To address the "sun-direction ambiguity", the pixel rays are projected along the corresponding sun direction onto the horizontal plane that passes through the image center [6]. This simple approach allows the equations above to remain unchanged, except for replacing the pixel rays with their ground-plane projections.

In the case of a known camera calibration, the only unknowns are the depths along the pixels rays. Therefore, even for a 3D scene, there are only $N$ unknowns. The measured temporal delays give a set of $N - 1$ independent constraints, which is insufficient to solve for an unambiguous scene model. See Figure 2 for a visual representation of this fundamental ambiguity. We can overcome this by adding constraints from a day with an independent wind direction vector. Ideally, the other wind vector would be orthogonal, but it suffices for it to not be in the same or opposite direction. Two days of data provide $2(N - 1)$ constraints, which results in a solution that is much less sensitive to errors in temporal delays. However, for small numbers of days, the solution is very sensitive to errors in the cloud motion estimates. These observations are demonstrated with numerical simulations described in Section IV.

### D. Scenario 2: General Imaging System

Here, we consider a general case in which we only know that the camera (or cameras) are static. We assume that each pixel has an arbitrary and unknown location and orientation, as in the raxel imaging model [5]. This scenario subsumes a variety of scenarios, including an individual projective camera with unknown calibration or a network of such cameras.

Despite the general nature of this imaging system, it is possible to solve for the partial scene geometry, $X$, (where $X$ is now a 2D, ground-plane projection, along the sun direction, of the full 3D model) without any need for calibration by optimizing the following objective function:

$$f(X) = \sum_{d=1}^{D} \|w_d^\mathsf{T} XH - w_d^\mathsf{T} w_d \Lambda_d\|_2^2. \qquad (10)$$

We fix one of the point locations to be the origin, and minimize this using nonlinear optimization with the following gradient:

$$\frac{\partial f}{\partial X} = 2\sum_{d=1}^{D} w_d (w_d^\mathsf{T} XH - w_d^\mathsf{T} w_d \Lambda_d) H^\mathsf{T}. \qquad (11)$$

For the uncalibrated general imaging case, there are $3N$ unknowns in the full 3D scene model. $N$ unknown parameters are eliminated by projecting all points along the corresponding sun vector onto the ground plane. Additionally, two more are eliminated by fixing the origin at one of our pixels, which leaves $2(N - 1)$ unknowns. Therefore, the 2D scene layout of the scene can be estimated using constraints from (at least) two days with independent cloud motions. Adding additional days will likely reduce the error, but it will not enable us to recover the scene height or the absolute position.

## IV. Results

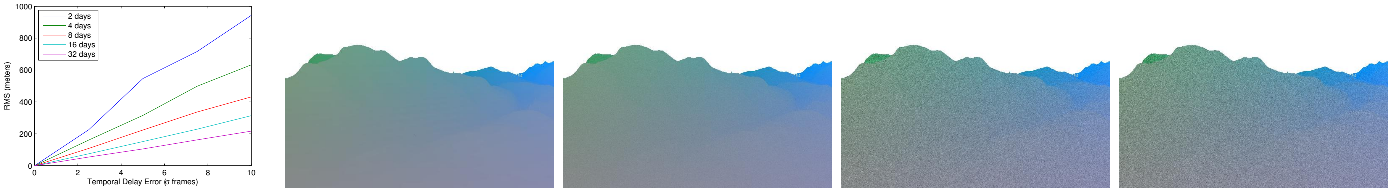
### A. Synthetic Data

To evaluate the sensitivity of our methods to errors, we perform experiments using a known 3D model that represents a landscape with a small number of hills. In these experiments, we assume that the camera is calibrated and use the approach described in III-C to estimate the scene geometry. Figure 4b (left) shows a false-color image that represents the 3D geometry of the scene, with the red, blue and green channels, respectively, representing the x, y and z coordinates of the imaged point location. For reference, the most distant point in the scene is approximately 10km from the camera. In all cases, we randomly generate cloud motions and explicitly compute the temporal delay. Cloud motion was selected so that it takes, on average, 60 frames of video for a cloud to pass through the scene. This results in a set of $N-1$ constraints for each virtual video.

The first experiment measures the effect of errors in the temporal delay calculation. We introduce varying amounts of independent, zero-mean, Gaussian error into the temporal delay estimates. As depicted in Figure 4, no error in temporal delay results in perfect reconstructions. As the error increases, the reconstruction quality degrades approximately linearly, but adding additional days of data reduces the impact of errors in the temporal delay estimates. In our second experiment, shown in Figure 5, noise was added to the wind direction estimates. From these results, it can be seen that even relatively small angular errors leads to large errors in the scene geometry estimates. This issue is exacerbated with less different days of data.

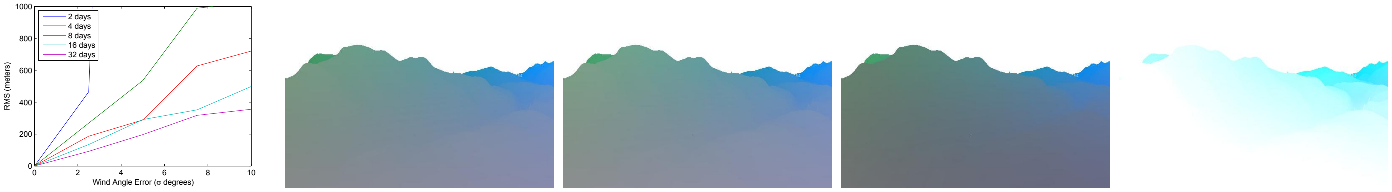
### B. Qualitative Evaluation

Figure 6 shows the results of a case study of the proposed methods on real images. Videos were captured at noon, each approximately 30 minutes long at 1Hz, on several different



(a) (b)

Fig. 4. Sensitivity to errors in temporal delay estimates on synthetic data. (a) Adding additional days of data reduces errors in scene reconstruction due to errors in temporal delay estimates. (b) The scene reconstructions that correspond to four points along the "4 days" line in (a). The reconstruction at right has the highest error, however it still reflects the general structure of the scene.



(a) (b)

Fig. 5. Sensitivity to errors in cloud motion direction on synthetic data. (a) As in Figure 4, adding additional days reduces errors but the errors are much more significant. (b) The scene reconstructions that correspond to four points along the "2 days" line in (a). The reconstruction at right has the highest error and does not clearly reflect the general structure of the scene.

days from two different static outdoor webcams. We investigate the most challenging problem setting, when the location of the camera is unknown and calibration information is not available. Temporal delays are computed for each frame and pixels for which a sufficient number of accurate temporal delay estimates cannot be obtained are filtered. The cloud velocities are estimated using a planar region on the ground near the camera (a straightforward linear problem given the global temporal delays) and rectified using a standard technique [10]. We then solve for shape using the unconstrained imaging model (Section III-D). This model only provides an estimate of the 2D ground layout, which we represent as a false-color image for each dimension. Important features of the scene geometry can be captured using this approach. For example, in Figure 6, the model captures the planar region of the airport and the depth discontinuities in the building scene. We show the resulting 2D layout as scatter plots in Figure 7. Geometric features, such as the shed in the foreground of the airport scene, can be seen clearly in these plots.

## V. Conclusion

We presented new methods for estimating the geometry of an outdoor scene using geometric assumptions on natural scene variations due to cloud motion. The cloud motion cue does not require any camera motion or manual scene manipulation and is therefore suitable for a wide variety of settings, including outdoor webcams. The estimated scene geometry information, whether full 3D or 2D scene layout, could be used in many applications, including video surveillance and environmental monitoring. In this work, we focus on the constraints provided by multiple cloud motion directions. This enabled us to remove the ambiguity inherent in recent, related work. In addition, we address shape estimation for general uncalibrated camera networks, which has not been previously addressed. We evaluate the effectiveness of all approaches on real and synthetic scenes. An important area for future work is in further validating these approaches on real data and incorporating constraints provided by other shape estimation cues.

## References

[1] Austin Abrams, Christopher Hawley, and Robert Pless. Heliometric stereo: Shape from sun position. In *ECCV*, 2012. 1

[2] Austin Abrams, Kylia Miskell, and Robert Pless. The episolar constraint: Monocular shape from shadow correspondence. In *CVPR*, 2013. 1, 2

[3] J. Ackermann, F. Langguth, S. Fuhrmann, and M. Goesele. Photometric stereo for outdoor webcams. In *CVPR*, 2012. 1

[4] C Chen and A Kak. Modeling and calibration of a structured light scanner for 3-d robot vision. In *Robotics and Automation. Proceedings. 1987 IEEE International Conference on*, volume 4, pages 807–815. IEEE, 1987. 1

[5] Michael D Grossberg and Shree K Nayar. The raxel imaging model and ray-based calibration. *IJCV*, 2005. 4

[6] Nathan Jacobs, Austin Abrams, and Robert Pless. Two cloud-based cues for estimating scene structure and camera calibration. *PAMI*, 2013. 1, 2, 3, 4

[7] Nathan Jacobs, Brian Bies, and Robert Pless. Using cloud shadows to infer scene structure and camera calibration. In *CVPR*, 2010. 2

[8] Nathan Jacobs, Mohammad Islam, and Scott Workman. Cloud motion as a calibration cue. In *CVPR*, 2013. 3

[9] Fabian Langguth, Jens Ackermann, Simon Fuhrmann, and Michael Goesele. Photometric stereo for outdoor webcams. In *CVPR*, 2012. 1

[10] David Liebowitz and Andrew Zisserman. Metric rectification for perspective images of planes. In *CVPR*, pages 482–488. IEEE, 1998. 3, 5

[11] K. Sunkavalli, F. Romeiro, W. Matusik, T. Zickler, and H. Pfister. What do color changes reveal about an outdoor scene? In *CVPR*, 2008. 1

[12] Shimon Ullman. The interpretation of structure from motion. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 203(1153):405–426, 1979. 1

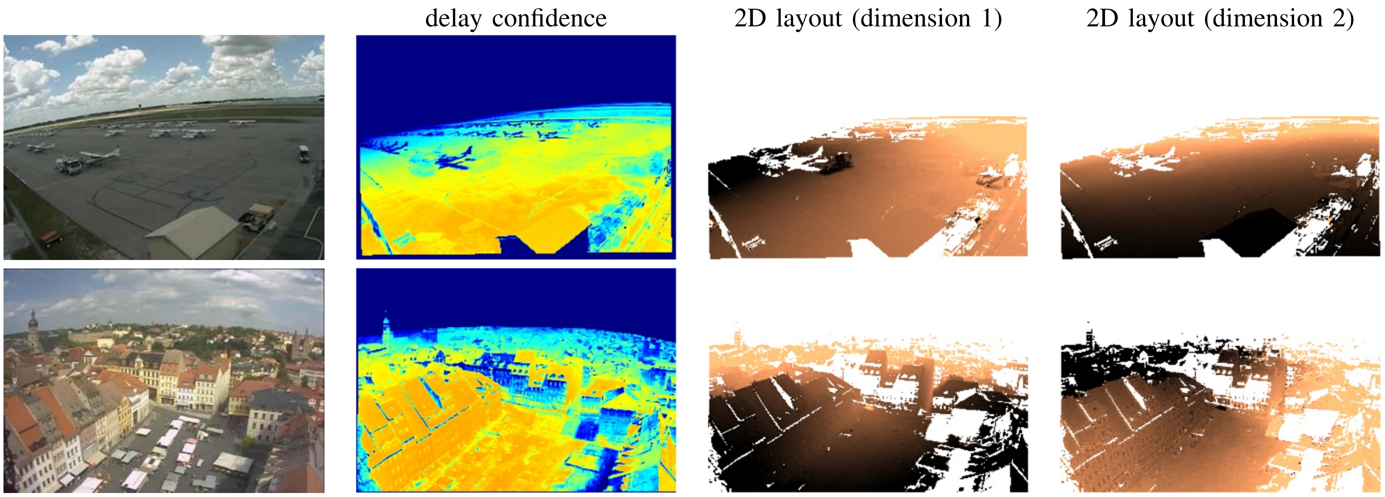


Fig. 6. Two examples of estimating the shape of scene using the unconstrained imaging model (III-D). The average confidence is dark blue in regions where the temporal delay cannot be estimated (such as in shadows), and orange in regions with high confidence. The false-color 2D layout images represent the 2D scene shape in regions where there was sufficient confidence in the temporal delay estimates.

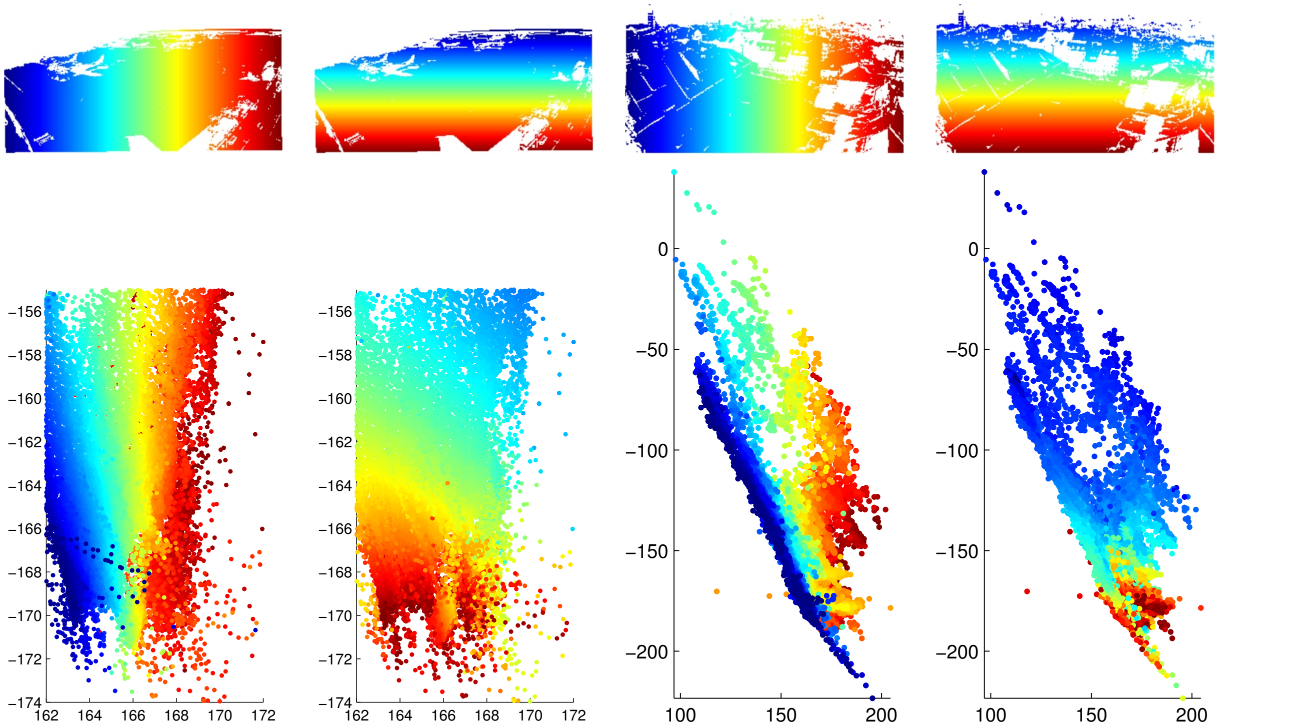


Fig. 7. (top) False-color images, from the scenes in Figure 6, that are color-coded by the $x$ (left) and $y$ (right) image coordinates. (bottom) The corresponding 2D scene layout for each of the colored pixels in the images, with the same color coding. Although not explicitly solved for, the camera location would be near the bottom, with the view frustum expanding upward. The large hole in the middle represents regions that are either always in shadow or are not visible from the camera location.