CAMERAS AND GRAVITY: ESTIMATING PLANAR OBJECT ORIENTATION

Zhaoyin Jia, Andrew Gallagher, Tsuhan Chen

School of Electrical and Computer Engineering, Cornell University

ABSTRACT

Photography on a mobile camera provides access to additional sensors. In this paper, we estimate the absolute orientation of a planar object with respect to the ground, which can be a valuable prior for many vision tasks. To find the planar object orientation, our novel algorithm combines information from a gravity sensor with a planar homography that matches a region of an image to a training image (e.g., of a company logo). We demonstrate our approach with an iPhone application that records the gravity direction for each captured image. We find a homography that maps the training image to the test image, and propose a novel homography decomposition to extract the rotation matrix.

We believe this is the first paper to estimate absolute planar object orientation by combining the inertial sensor information with vision algorithms. Experiments show that our proposed algorithm performs reliably.

Index Terms— Image processing, Image motion analysis, Image detection

1. INTRODUCTION

Many mobile phones are equipped with inertial sensors, such as accelerometers and gyroscopes, to provide relatively accurate measurements of the phone's position, such as the orientation of the phone with respect to the ground. This information is usually used to improve user interaction, for example, showing an application window in landscape when a user rotates the mobile phone horizontally.

In this paper, we combine the mobile device's camera with a gravity sensor and show that the gravity sensor data can estimate not only the position of the phone itself, but also the orientation of a planar object captured by the camera. One example is shown in Fig. 1: given a planar target object to detect (in Fig. 1 (a), the **training view**) and the gravity direction with respect to the camera (Fig. 1 (b)), we match the object in the test image (in Fig. 1 (c), the **test view**), and estimate the orientation of the object with respect to the ground through a special homography decomposition, shown in Fig. 1 (d).

To our knowledge, we are the first to combine a vision algorithm (planar matching) with the inertial sensor to determine the orientation of the object. The orientation of an object is a useful prior for many vision tasks. For example, even with the same image shown in Fig. 1 (a), if the image is displayed vertically ($\theta = 90°$), it is more likely to be a painting hanging on the wall; if it is horizontal ($\theta = 0°$), then it may be a picture in a book on a table.

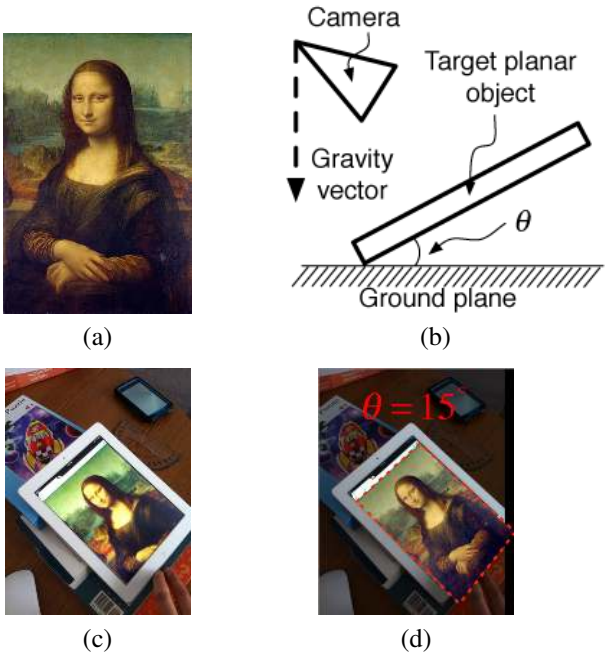


Fig. 1. (a) The training view (frontal, no distortion) of the target. (b) The testing setup. (c) The test image containing the target. (d) The homography from the training image to the testing image. We compute the absolute orientation of the planar object, e.g. $\theta = 15°$, indicating the angle between the target artwork and the ground plane.

Related work: Hoiem et. al [1] show that estimating surface orientation helps occlusion boundary detection, depth estimation and object recognition. Further applications of estimating the camera position, vanishing lines and the horizon are presented in [2]. This prior work proposes a learning-based approach to find the surface orientation, and roughly classify these surfaces into 'vertical' or 'horizontal'. Our algorithm predicts the object orientation more precisely in degree, and incorporates a gravity sensor with pixel data. In addition, there are other applications with different goals that combine inertial sensors with images for matching, photo enhancement and robotics, such as [3], [4] and [5], but none combine ho-

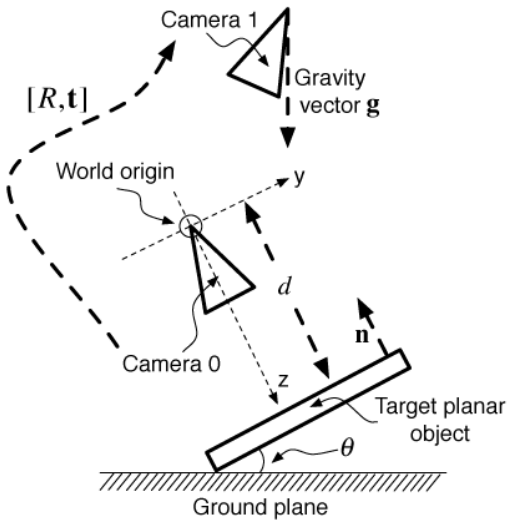


Fig. 2. Two views introduced by a homography with the gravity vector.

mography matching with gravity sensors to accurately measure planar surface absolute orientation.

Our work is closely related to homography decomposition presented in [6] [7] [8] and [9]. However, our setting and task is different. These algorithms use the camera intrinsic matrices for decomposition. In our case, although the mobile phone for testing can be calibrated, the training images containing targets to match have no camera information: they can be images downloaded from Internet, as shown in Fig. 1 (a). Our derivation shows that we can synthesize the camera matrix for the training view and get the target orientation, as long as the training image is frontal with no distortion, a usually valid assumption. We believe we are the first to combine this decomposition with a gravity sensor to estimate the planar object orientation.

2. ALGORITHM

We introduce our definition of the variables in Fig. 2: we have two cameras (camera 0, which captures the training target image, and camera 1, which captures the test image) related by a planar homography. The intrinsic matrices are K_0 for the training view, and K_1 for the test image; the world coordinate is defined as camera 0's axis, i.e., camera 0 has no rotation and zero translation. The planar object is placed at distant d with normal direction $\mathbf{n}$. The extrinsic matrix for camera 1 is $\left[\begin{array}{cc} R & \mathbf{t} \end{array}\right]$. The gravity vector $\mathbf{g}$ is provided in the camera 1's coordinates during testing. The goal is to find the projection of the surface normal $\mathbf{n}$ in camera 1's coordinates, and then compute its angle with the gravity vector $\mathbf{g}$. To do this, we must find the rotation matrix R from camera 0 to camera 1, despite the challenge that camera 0 has unknown internal parameters.

2.1. Two-view geometry with planar homography

For a homography H that maps points $\mathbf{x}_0$ in the image of camera 0 to points $\mathbf{x}_1$ in the image of camera 1, i.e., $\mathbf{x}_1 = H\mathbf{x}_0$, [6] shows that the induced planar homography H for the plane $[\mathbf{n}^T, d]^T$ between the two cameras is:

$$H = K_1(R - \frac{\mathbf{t}\mathbf{n}^T}{d})K_0^{-1}, \tag{1}$$

We define H^* as:

$$H^* = K_1^{-1}HK_0 = R - \frac{\mathbf{t}\mathbf{n}^T}{d}. \tag{2}$$

Unfortunately, we cannot directly decompose H into R and $\mathbf{t}$ to get the position of the target ([8] and [10]), because although we have the camera parameter K_1 for the testing view (camera 1), the training view's intrinsics, K_0, are unknown.

2.2. Depth and intrinsic matrix K_0

With unknown intrinsic matrix K_0, we make use of our assumptions about the training images 1) they are frontal views of a planar object; 2) the camera has zero-skew; 3) it also has square pixels. These assumptions hold for most planar targets, such as paintings, logos, book or CD covers online. Then K_0 becomes:

$$K_0 = \left[\begin{array}{ccc} f & 0 & c_x \\ 0 & f & c_y \\ 0 & 0 & 1 \end{array}\right]. \tag{3}$$

For any 3D point X that lies on the plane $[\mathbf{n}^T, d]^T$, since the camera is taking the frontal view of the plane object, then $X = [x, y, d, 1]^T$. By definition, camera 0 has identity rotation and zero translation, then the extrinsic matrix for camera 0 is $\left[\begin{array}{cc} I_{3\times3} & \mathbf{0} \end{array}\right]$. In homogeneous coordinates, the 2D projection $\mathbf{x}_0$ of 3D point X to camera 0 is :

$$\begin{aligned} \mathbf{x}_0 &= K_0 \left[\begin{array}{cc} I_{3\times3} & \mathbf{0} \end{array}\right] [x, y, d, 1]^T \\ &= \left[\begin{array}{ccc} 1 & & c_x \\ & 1 & c_y \\ & & 1 \end{array}\right] \left[\begin{array}{c} x \\ y \\ \frac{d}{f} \end{array}\right] \end{aligned} \tag{4}$$

We can rewrite K_0 as K_0^* :

$$K_0^* = \left[\begin{array}{ccc} 1 & 0 & c_x \\ 0 & 1 & c_y \\ 0 & 0 & 1 \end{array}\right] = \left[\begin{array}{cc} I_{2\times2} & \mathbf{c} \\ 0 & \end{array}\right], \tag{5}$$

where $\mathbf{c} = [c_x, c_y, 1]^T$, and represent the depth d as d^* :

$$d^* = \frac{d}{f}. \tag{6}$$

For a frontal view of a planar object, the depth and the focal length are related: we get the exact same image results

if we increase the focal length of the camera, and move the object further away. With this derivation, the equation for the homography H in (2)becomes :

$$H^* = K_1^{-1} H \begin{bmatrix} I_{2\times 2} & \mathbf{c} \\ 0 & \end{bmatrix} = R - \frac{f\mathbf{t}\mathbf{n}^T}{d}, \quad (7)$$

2.3. Decompose H to R

To decompose H^*, we define the vectors $\mathbf{u}$ and $\mathbf{v}$ as:

$$\mathbf{u} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \qquad \mathbf{v} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad (8)$$

and multiply them by H^* in (7). For $\mathbf{u}$, on the left side of (7) we have:

$$H^*\mathbf{u} = K_1^{-1} H \begin{bmatrix} I_{2\times 2} & \mathbf{c} \\ 0 & \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = K_1^{-1} H \mathbf{u}. \quad (9)$$

On the right side, since $\mathbf{n}^T\mathbf{u} = 0$, (remember, in camera 0 the frontal view of the planar object has $\mathbf{n} = [\mathbf{0}, \mathbf{0}, -\mathbf{1}]^\mathbf{T}$) we have:

$$(R - \frac{f\mathbf{t}\mathbf{n}^T}{d})\mathbf{u} = R\mathbf{u} \quad (10)$$

Therefore by combining (9) and (10) we have

$$K_1^{-1} H \mathbf{u} = R\mathbf{u}. \quad (11)$$

Following the same derivation for vector $\mathbf{v}$, we also have

$$K_1^{-1} H \mathbf{v} = R\mathbf{v}. \quad (12)$$

Since R is a rotation matrix and $[\mathbf{u}, \mathbf{v}, \mathbf{u} \times \mathbf{v}] = \mathbf{I_{3\times 3}}$, thus

$$[R\mathbf{u}, R\mathbf{v}, (R\mathbf{u}) \times (R\mathbf{v})] = R[\mathbf{u}, \mathbf{v}, \mathbf{u} \times \mathbf{v}] = R \\ = [K_1^{-1}H\mathbf{u}, K_1^{-1}H\mathbf{v}, (K_1^{-1}H\mathbf{u}) \times (K_1^{-1}H\mathbf{v})]. \quad (13)$$

In practice, due to the noise in computing H and K_1, the final rotation matrix R may not be orthogonal. We approximate R to the nearest orthogonal matrix R by the Frobenius norm, i.e., by taking SVD of the Eq. 13, $R = U\Sigma V^T$, and then the final $R = UV^T$.

Eq. 13 shows that, under our assumptions for the training images (frontal view, no distortion), the rotation matrix R is independent of the camera center $\mathbf{c}$ and the focal length f of camera 0, the depth d and the translation $\mathbf{t}$. Once we retrieve the homography H between the two views and know the intrinsic parameter K_1 for the testing camera 1, we can determine the rotation matrix R from camera 0 to camera 1.

2.4. Planar object orientation θ

During testing, the gravity direction $\mathbf{g}$ is in camera 1's coordinates. To compute the planar object orientation θ with respect to the ground, we compute the projection of plane's normal vector $\mathbf{n}$ in camera 1's projection, i.e., $R\mathbf{n}$, and calculate its angle to the gravity vector $\mathbf{g}$. This gives the orientation θ.

More formally, define angle α as by applying Eq. 13:

$$\alpha = \arccos(R\mathbf{n} \cdot \mathbf{g}), \quad (14)$$

$$\theta = \begin{cases} \alpha & , \quad 0 \le \alpha \le \frac{\pi}{2} \\ \pi - \alpha & , \quad \frac{\pi}{2} < \alpha < \pi \end{cases}. \quad (15)$$

Fig. 3. Sample frontal views of our planar object images tested.

(a) $\theta_{gt} = 72.5°$ (b) $\theta_{gt} = 0°, 30°$ respectively

Fig. 4. Experiment setting with (a) fixed object orientation, and different camera positions, and (b) gradually increasing object orientation from $0°$ to $75°$.

3. EXPERIMENTS

An iPhone 4 is used for the experiments. The camera is calibrated to find K_1, and the inertial sensor data is recorded while capturing images. We assume the coordinates of the gravity vector are aligned with the camera axis, and directly use the gravity vector from the inertial sensor as $\mathbf{g}$.

We use several different types of planar images, all downloaded, for experiments. We choose five famous paintings (Mona Lisa, The Girl with a Pearl Earring, Marriage of Arnolfini, The Music Lesson and Birth of Venus), a logo (Starbucks), and a CD cover (Beatles). We render the image on an iPad as the testing target, and capture the scene with the calibrated iPhone 4 camera to estimate the absolute orientation of the iPad. The ground truth angle of the planar object is manually measured by a protractor. The homography is computed through SIFT point matching [11] and RANSAC algorithm with DLT [6].

Three types of experiments are performed (see Fig. 4 for the first two): 1) we keep the planar object at a fixed angle with respect to the ground, and take images with different camera positions; 2) we keep the camera at a fix angle, and

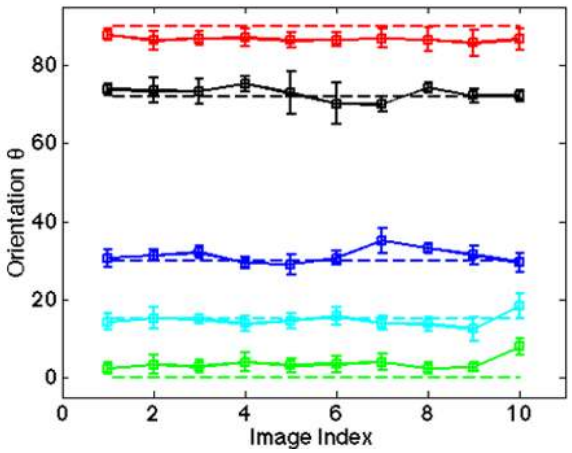


Fig. 5. Experiment result of different object orientations.

gradually rotate the planar object from $0°$ to $75°$ with respect to the ground; 3) we capture images in the real-world to show some qualitative results.

Fixed object, different camera positions: In this setting we keep the object at a fixed angle, and take images from different position. One example is shown in Fig. 4 (a). We take 10 images for each planar object at a certain angle. For each image, we compute homography 30 times using RANSAC, and compute the angle θ through each homography. The experiment results are shown in Fig. 5. The dashed lines indicates the ground-truth angle. Different colors indicate different images at a specific angle. Because RANSAC produces a different homography each time, the error bar shows the standard deviation of θ.

Our proposed algorithm predicts the angle of the planar object with small errors, that may be introduced by misalignment of the gravity vector axes, calibration of K_1, and especially by the quality of homography. One example is shown in Fig. 6. A good homography gives much smaller error in estimating the planar object's orientation.

Fixed camera, different object orientations: In another test we keep the camera roughly fixed, but gradually increase the orientation angle θ of the planar object from $0°$ to $75°$. One example is shown in Fig. 4 (b). The results are shown in Table 1. Our proposed algorithm still robustly estimates the object orientation in this case. For these two scenarios, overall 88% of the testing cases are within $5°$ from the ground truth, and 98% of them are within $10°$ from the ground truth.

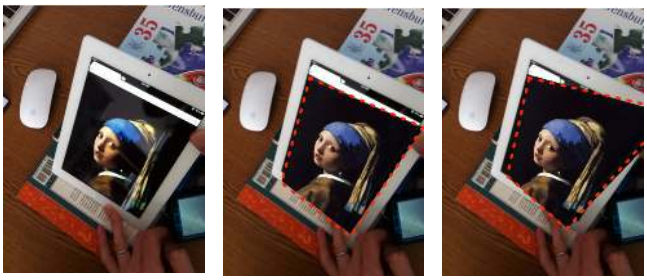
Real world example: We also test on real-world images with a Starbucks logo for qualitative results, shown in Fig. 7. Our proposed algorithm predicts the logo orientation θ accurately in different situations.

4. CONCLUSION

This paper estimates the absolute orientation of a planar object using the gravity sensor with a mobile camera. We made

gt	**0°**	**15°**	**30°**	**45°**	**60°**	**75°**
obj1-mean	**4.6°**	**17.1°**	**29.8°**	**44.3°**	**61.2°**	**76.8°**
obj1-var	2.4°	2.7°	3.0°	3.2°	4.3°	3.5°
obj2-mean	**7.7°**	**13.8°**	**29.0°**	**45.0°**	**56.8°**	**72.2°**
obj2-var	3.5°	4.4°	4.1°	2.4°	2.1°	1.8°

Table 1. Results when having one object rotating from $0°$ to $75°$. **gt** is the ground-truth angle. **obj1** is the Starbuck logo, and **obj2** is the Beatles CD cover. The table shows the mean and variance (var) for each image by sampling homographies with RANSAC.



(a) $\theta_{gt} = 42.5°$ (b) $\theta = 43.0°$ (c) $\theta = 55.1°$

Fig. 6. Homography quality affects the orientation θ. The ground-truth orientation (in (a)) is $42.5°$. A good homography in (b) has a better prediction than a worse one in (c).

weak assumptions about the training image, e.g., frontal view, zero skew and no distortion. During testing we estimated the rotation through a special homography decomposition, and calculated the projection of the normal vector of the planar object, and its angle between the gravity vector as the object orientation. Experiments showed that our proposed algorithm robustly predicts the object orientation in different scenarios.

Future applications can be achieved based on this algorithm, e.g., correcting the homography or improving object detection. Since we find the rotation matrix of the testing camera, our algorithm can also estimate the in-plane rotation of the planar object. This can be used for other applications, e.g. to predict if the object is up-side down.

(a) $\theta = 87.5°$ (b) $\theta = 30.7°$ (c) $\theta = 6.6°$

Fig. 7. Predicting the angle of Starbucks logo in real-world images: a) A Starbucks logo outside the Cafe. The ground-truth $\theta_{gt} = 90°$. b) A bag of coffee in hand. θ_{gt} should be in between of $0°$ and $90°$. c) A menu on the table. $\theta_{gt} = 0$. θ under each figure is the prediction from our algorithm.

5. REFERENCES

[1] D. Hoiem, A. A. Efros, and M. Hebert, "Closing the loop in scene interpretation," in *CVPR*, 2008.

[2] D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," *IJCV*, vol. 80, 2008.

[3] D Kurz and S Himane, "Inertial sensor-aligned visual feature descriptors," in *CVPR*, 2011.

[4] H Lee, E Shechtman, J Wang, and S Lee, "Automatic upright adjustment of photographs," in *CVPR*, 2012.

[5] J. Lobo and J. Dias, "Vision and inertial sensor cooperation using gravity as a vertical reference," *PAMI*, vol. 25, no. 12, 2003.

[6] A. Harltey and A. Zisserman, *Multiple view geometry in computer vision (2. ed.)*, Cambridge University Press, 2006.

[7] Y. Ma, S. Soatto, J. Kosecka, and S. Sastry, *An Introduction to 3D Vision*, Springer, 2003.

[8] O. D. Faugeras and F. Lustman, "Motion and structure from motion in a piecewise planar environment," *International Journal of Pattern Recognition and Artificial Intelligence*, 1988.

[9] Z. Zhang and A.R. Hanson, "3d reconstruction based on homography mapping," *Proc. ARPA96*, 1996.

[10] E. Malis and M. Vargas, "Deeper understanding of the homography decomposition for vision-based control," 2008.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, 2004.