DETECTING PHOTOGRAPHIC COMPOSITES OF FAMOUS PEOPLE

Eric Kee and Hany Farid

Department of Computer Science
Dartmouth College
Hanover, NH 03755

ABSTRACT

Photos are commonly falsified by compositing two or more people into a single image. We describe how such composites can be detected by estimating a camera's intrinsic parameters. Differences in these parameters across the image are then used as evidence of tampering. Expanding on earlier work, this approach is more applicable to low-resolution images, but requires a reference image of each person in the photo as they are directly facing the camera. When considering composites of famous people, such a reference photo is easily obtained from an on-line image search.

Index Terms— Digital Image Forensics

1. INTRODUCTION

When *Star* magazine was unable to obtain a highly coveted photo of then rumored sweethearts Brad Pitt and Angelina Jolie, they simply created their own. The resulting magazine cover was created by compositing two separate photos of Pitt and Jolie to give the impression that they were together. Previous work has shown how to detect such composites using various geometric techniques. In [6], the authors described how a re-projection of a video or image causes a distortion in the camera skew. In [7], the authors detect inconsistencies in the inter-image transform from a pair of images of the same scene taken from different vantage points. In the most closely related work, the authors in [3] showed how photo compositing can change the principal point (the projection of the camera center onto the image plane). For each person in the image, the transformation from world to image coordinates was estimated, from which the principal point could be determined. Inconsistencies in the estimated principal point were then used as evidence of tampering. The primary drawback of this approach is that the world to image transformation was computed from the image of a person's eyes which are difficult to resolve in low resolution images.

Here we describe a variant of this earlier work that does not require resolving a person's eye. This approach relies on a set of co-planar facial features (e.g., the corners of the eyes, the base of the nose, the chin, etc.) which are easier to resolve in low-resolution images. The disadvantage of this approach is that it requires a reference image of the person as they are directly facing the camera. While this may not be practical in the most general sense, it is useful when analyzing photos of famous people, for which a reference photo is easily obtained from an on-line image search, or for any person or object for which a reference photo is available.

2. METHODS

Image formation can be modeled by an extrinsic rigid-body transformation from 3-D world to 3-D camera coordinates, and an intrinsic perspective projection from 3-D camera to 2-D image coordinates. The intrinsic transformation is described by two primary parameters: the camera focal length and principal point. Compositing two people into a single photo can lead to differences in these parameters.

In general, estimating a camera's intrinsic parameters requires multiple images from different vantage points, or an image of a 3-D object with known geometry (e.g., a cube) [2, 4]. In a forensic setting, neither of these options are reasonable. We will show how to estimate a camera's intrinsic parameters from facial features. To do so, we make several assumptions: (1) four or more co-planar features can be localized on the face; (2) a reference image is available which shows the selected features as they are directly facing the camera; and (3) the camera focal length is known (the focal length can easily be computed from image EXIF data).

2.1. Planar homography

The mapping between points in 3-D world coordinates to 2-D image coordinates can be expressed by the projective imaging equation $\vec{x} = P\vec{X}$, where the 3×4 matrix P embodies the

This work was supported by a gift from Adobe Systems, Inc., a gift from Microsoft, Inc., a grant from the National Science Foundation (CNS-0708209), and by the Institute for Security Technology Studies at Dartmouth College under grants from the Bureau of Justice Assistance (2005-DD-BX-1091) and the U.S. Department of Homeland Security (2006-CS-001-000001). Points of view or opinions in this document are those of the author and do not represent the official position or policies of the U.S. Department of Justice, the U.S. Department of Homeland Security, or any other sponsor.

projective transform, the vector $\vec{X}$ is a 3-D world point in homogeneous coordinates, and the vector $\vec{x}$ is a 2-D image point also in homogeneous coordinates. We consider a special case of this projective transform where all of the world points $\vec{X}$ lie on a single plane. In this case, the projective transform P reduces to a 3×3 planar projective transform H, also known as a homography:

$$\vec{x} = H\vec{X}, \quad (1)$$

where the world $\vec{X}$ and image points $\vec{x}$ are now represented by 2-D homogeneous vectors.

The planar homography H can be factored into a product of two matrices:

$$H = \lambda KM \quad (2)$$

where λ is a scale factor, K is the intrinsic matrix, and M is the extrinsic matrix. The intrinsic matrix is embodied by four parameters: the focal length f, aspect ratio α, skew s, and principal point (c_1, c_2). In modern day cameras, we can generally assume square pixels ($\alpha = 1$ and $s = 0$). With these assumptions, the intrinsic matrix takes the following form:

$$K = \begin{pmatrix} f & 0 & c_1 \\ 0 & f & c_2 \\ 0 & 0 & 1 \end{pmatrix}, \quad (3)$$

The 3×3 extrinsic matrix M embodies the transformation from world to camera coordinates:

$$M = \begin{pmatrix} \vec{r}_1 & \vec{r}_2 & \vec{t} \end{pmatrix}, \quad (4)$$

where $\vec{r}_1$ and $\vec{r}_2$ are the first two columns of the 3×3 rotation matrix[1], and $\vec{t}$ is a 3×1 translation vector, which combined, define the world to camera transformation.

2.2. Homography estimation

We briefly review how to estimate the planar homography H, Equation(1), from known world and image coordinates [2]. This estimation begins with a cross production formulation of Equation(1):

$$\begin{aligned} \vec{x} \times [H\vec{X}] &= 0 \\ \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \times \left[\begin{pmatrix} h_1 & h_2 & h_3 \\ h_4 & h_5 & h_6 \\ h_7 & h_8 & h_9 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \right] &= 0. \end{aligned} \quad (5)$$

This constraint is linear in the unknown elements of the homography h_i. A matched set of world $\vec{X}$ and image $\vec{x}$ coordinates appears to provide three constraints on the eight unknown elements of H (the homography is defined up to an unknown scale factor, reducing the number of unknowns from nine to eight). However, the constraints are not linearly independent. As such, this system provides only two constraints in the eight unknowns. Therefore, a total of four or more points with known world and image coordinates are required to estimate the homography[2]. From these points, standard least-squares techniques [2] can be used to solve for $\vec{h}$.

Since we are assuming that the world points $\vec{X}$ lie on a plane, their coordinates can be determined from a single image where the planar surface is coplanar with the camera sensor.

2.3. Homography factorization

Given the estimated planar homography H, we next factor H into a product of intrinsic and extrinsic matrices. More specifically, we are interested in the intrinsic camera parameters (the principal point (c_1, c_2)).

Recall that the homography can be expressed as a product of intrinsic and extrinsic matrices:

$$H = \lambda K \begin{pmatrix} \vec{r}_1 & \vec{r}_2 & \vec{t} \end{pmatrix}. \quad (6)$$

The camera's intrinsic components can be estimated by decomposing H according to Equation (6). It is straightforward to show that $\vec{r}_1 = \frac{1}{\lambda} K^{-1}\vec{h}_1$ and $\vec{r}_2 = \frac{1}{\lambda} K^{-1}\vec{h}_2$ where $\vec{h}_1$ and $\vec{h}_2$ are the first two columns of the matrix H. The constraint that $\vec{r}_1$ and $\vec{r}_2$ are orthogonal (they are columns of a rotation matrix) and have the same norm (unknown due to the scale factor λ) yields two constraints on the unknown matrix K:

$$\begin{aligned} \vec{r}_1^T \vec{r}_2 &= 0 \\ \vec{h}_1^T (K^{-T}K^{-1})\vec{h}_2 &= 0, \quad (7) \\ &\text{and} \\ \vec{r}_1^T \vec{r}_1 - \vec{r}_2^T \vec{r}_2 &= 0 \\ \vec{h}_1^T (K^{-T}K^{-1})\vec{h}_1 - \vec{h}_2^T (K^{-T}K^{-1})\vec{h}_2 &= 0. \quad (8) \end{aligned}$$

With only two constraints, it is possible to estimate the principal point (c_1, c_2) or the focal length f, but not both [8]. As such, we will assume a known focal length.

For notational simplicity we solve for the components of $Q = K^{-T}K^{-1}$, which contain the desired coordinates of the principal point and the assumed known focal length:

$$Q = \frac{1}{f^2} \begin{pmatrix} 1 & 0 & -c_1 \\ 0 & 1 & -c_2 \\ -c_1 & -c_2 & c_1^2 + c_2^2 + f^2 \end{pmatrix}. \quad (9)$$

In terms of Q, the first constraint, Equation (7), takes the form:

$$\begin{aligned} h_1h_2 + h_4h_5 - (h_2h_7 + h_1h_8)c_1 - (h_5h_7 + h_4h_8)c_2 \\ + h_7h_8(c_1^2 + c_2^2 + f^2) = 0, \quad (10) \end{aligned}$$

[1] The third column of the rotation matrix is given by the cross product of the first two columns: $\vec{r}_3 = \vec{r}_1 \times \vec{r}_2$

[2] The 3-D world and 2-D image coordinates should be translated so that their centroid is at the origin, and scaled isotropically so that the average distance to the origin is $\sqrt{2}$. This normalization improves stability of the homography estimation in the presence of noise [1].

Note that this constraint is a second-order polynomial in the coordinates of the principal point, which can be factored as follows:

$$(c_1 - \alpha_1)^2 + (c_2 - \beta_1)^2 \quad = \quad \gamma_1^2, \tag{11}$$

where:

$$\begin{aligned} \alpha_1 &= (h_2h_7 + h_1h_8)/(2h_7h_8), \\ \beta_1 &= (h_5h_7 + h_4h_8)/(2h_7h_8), \\ \gamma_1^2 &= \alpha_1^2 + \beta_1^2 - f^2 - (h_1h_2 + h_4h_5)/(h_7h_8). \end{aligned}$$

Similarly, the second constraint, Equation (8), takes the form:

$$\begin{aligned} h_1^2 + h_4^2 + 2(h_2h_8 - h_1h_7)c_1 + 2(h_5h_8 - h_4h_7)c_2 \\ -h_2^2 - h_5^2 + (h_7^2 - h_8^2)(c_1^2 + c_2^2 + f^2) = 0, \end{aligned} \tag{12}$$

or,

$$(c_1 - \alpha_2)^2 + (c_2 - \beta_2)^2 \quad = \quad \gamma_2^2, \tag{13}$$

where:

$$\begin{aligned} \alpha_2 &= (h_1h_7 - h_2h_8)/(h_7^2 - h_8^2), \\ \beta_2 &= (h_4h_7 - h_5h_8)/(h_7^2 - h_8^2), \\ \gamma_2^2 &= \alpha_2^2 + \beta_2^2 - (h_1^2 + h_4^2 - h_2^2 - h_5^2)/(h_7^2 - h_8^2) - f^2. \end{aligned}$$

Both constraints, Equations (11) and (13) are circles in the desired coordinates of the principal point c_1 and c_2, and the solution is the intersection of the two circles.

For certain homographies this solution can be numerically unstable. For example, if $h_7 \approx 0$ or $h_8 \approx 0$, the first constraint becomes numerically unstable. Similarly, if $h_7 \approx h_8$, the second constraint becomes unstable. In order to avoid these instabilities, an error function with a regularization term is introduced. We start with the following error function to be minimized:

$$E(c_1, c_2) \quad = \quad g_1(c_1, c_2)^2 + g_2(c_1, c_2)^2, \tag{14}$$

where $g_1(c_1, c_2)$ and $g_2(c_1, c_2)$ are the constraints on the principal point given in Equations (10) and (12), respectively. To avoid numerical instabilities, a regularization term is added to penalize deviations of the principal point from the image center $(0, 0)$ (in normalized coordinates). This augmented error function takes the form:

$$E(c_1, c_2) = g_1(c_1, c_2)^2 + g_2(c_1, c_2)^2 + \Delta(c_1^2 + c_2^2), \tag{15}$$

where Δ is a scalar weighting factor. This error function is a nonlinear least-squares problem, which can be minimized using a Levenberg-Marquardt iteration. The image center $(0, 0)$ is used as the initial condition for the iteration.

Note that if there is no relative rotation between the world and camera coordinate systems, then there is an inherent ambiguity between the world to camera translation in X and Y and the position of the principal point, and between the translation in Z (depth) and the focal length. As such, the factorization of the homography is not unique in the case of a fronto-parallel view of a face.

2.4. Photo Compositing

When creating a photo composite it is often necessary to translate, scale, rotate, etc. portions of an image. Such image-based manipulations can change the effective intrinsic camera parameters. Therefore, differences in these estimated parameters can be used as evidence of tampering. As we showed in [3], a translation in the image is equivalent to translating the camera's principal point. In homogeneous coordinates, translation is represented by multiplication with a matrix T: $\vec{y} = T\vec{x}$. With a horizontal and vertical translation in the image of (d_1, d_2), the mapping from world $\vec{X}$ to (translated) image coordinates $\vec{y}$ is:

$$\begin{aligned} \vec{y} &= THM\vec{X} \\ &= \lambda TKM\vec{X} \\ &= \lambda \begin{pmatrix} 1 & 0 & d_1 \\ 0 & 1 & d_2 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} f & 0 & c_1 \\ 0 & f & c_2 \\ 0 & 0 & 1 \end{pmatrix} M\vec{X} \\ &= \lambda \begin{pmatrix} f & 0 & c_1 + d_1 \\ 0 & f & c_2 + d_2 \\ 0 & 0 & 1 \end{pmatrix} M\vec{X}. \end{aligned} \tag{16}$$

That is, translation in the image is equivalent to translating the principal point. Similarly, other image-based transformations such as scaling can affect the camera intrinsic matrix parameters. As such, these transformations can be detected by estimating the principal point for different faces/objects in an image, from which any inconsistencies can be used as evidence of tampering.

3. RESULTS

We describe a set of simulations that demonstrate the efficacy of the proposed technique. In each case, the principal point is estimated as follows. A set of nine planar world coordinates are first specified. The world to camera rotation $\vec{\phi} = \begin{pmatrix} \phi_x & \phi_y & \phi_z \end{pmatrix}$ is randomly selected, where each rotation angle does not exceed $\pm 30°$, $\pm 30°$ and $\pm 15°$, respectively, and where the combined rotation is at least $10°$ (i.e., $\|\vec{\phi}\| \geq 10$). The world to camera translation $\vec{t}$ is selected randomly so that the world points project in their entirety onto a sensor of size 2×2 units (centered at $(0, 0)$), and such that the image of these points occupies 20% of the sensor width. The rotation and translation combine to yield the extrinsic matrix, Equation (4). The intrinsic and extrinsic matrices define the transformation H, Equation (2), from world $\vec{X}$ to image $\vec{x}$ coordinates, $\vec{x} = H\vec{X}$.

As described in Section 2.2, the homography H is estimated from corresponding world and image coordinates. Factoring this estimated homography requires the minimization of the error function $E(c_1, c_2)$ defined in Equation (15), where $\Delta = 0.1$.

noise	mean $\|\vec{c}\|$	std. dev. $\|\vec{c}\|$	$\|\vec{c}\| \leq 0.1$	$\|\vec{c}\| \leq 0.2$
0.5	0.013	0.007	100.00	100.00
1.0	0.027	0.016	100.00	100.00
2.0	0.056	0.029	91.67	100.00
3.0	0.083	0.043	66.22	99.51
4.0	0.109	0.053	46.50	94.64
5.0	0.134	0.064	32.80	85.90

Table 1. Shown are the mean and standard deviation for the estimated principal point norms averaged over $10,000$ random images. The actual principal point was $\vec{c} = (0\ 0)$. Each row corresponds to different amounts of additive image noise (in pixels). The last two columns give the percentage of estimates whose magnitude are 0.1 and 0.2 units from the origin.

3.1. Simulations

In the first set of simulations, the 3-D world points consisted of 9 planar points in the shape of a 3×3 square grid. The intrinsic matrix K, Equation (3), consisted of a focal length $f = 3.5$ and a principal point $c_1 = c_2 = 0$ (the image center). With this configuration, the principal point was estimated as described above. This entire process was repeated $10,000$ times. In order to simulate real-world conditions, the equivalent of 0.5 to 5.0 pixels of noise were added to the image coordinates. Shown in Table 1 are the mean and standard deviation for the estimated principal point norms, as well as the percentage of estimates that fell within varying distances to the origin. With small amounts of noise, the estimated principal points are highly accurate. Even with as much as 3 pixels of noise, nearly all estimates fall within a radius of 0.2 units from the origin.

In the second set of simulations, the intrinsic matrix K consisted of a focal length $f = 3.5$ and a principal point $c_1 = c_2 = 0.5$. With 2 pixels of noise, and averaged over $10,000$ random trials, the mean estimated principal point was $(0.336, 0.333)$ with a standard deviation of $(0.069, 0.068)$. The bias in the mean is due to the regularization term in Equation (15).

3.2. Simulations: forensics

In a forensic setting, we seek to determine if two faces in an image are consistent with a single intrinsic matrix. In this second set of simulations, images were generated with two sets of planar points (as described above). In one case the intrinsic matrix consisted of a focal length $f = 3.5$ and a principal point $c_1 = c_2 = 0$, and in the other case $f = 3.5$ and c_1, c_2 were selected randomly such that their norm was greater than 0.2. Two pixels of noise were added to the image coordinates. Shown in Fig. 1 are the estimated principal points from $10,000$ random trials, where the central black points

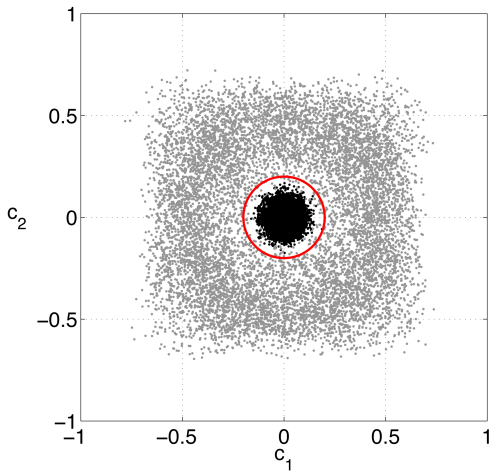

Fig. 1. Estimates of the principal point in normalized coordinates. The central black dots correspond to a principal point $c_1 = c_2 = 0$, and the gray dots correspond to principal points with magnitude greater than 0.2. The red circle is drawn at 0.2 units from the origin for reference.

(authentic) correspond to $c_1 = c_2 = 0$, and the gray points (inauthentic) correspond to principal points whose magnitude was greater than 0.2. With 100% of the "authentic" principal points having a magnitude less than 0.2, 99.01% of the "inauthentic" principal points had a magnitude greater than 0.2. With 4 pixels of noise, 94.27% of the authentic principal points had a magnitude less than 0.2, and 99.85% of the inauthentic points had a magnitude greater than 0.2. Even with relatively large amounts of noise, the false positive rate and detection accuracy are quite good. The false positive rate and detection accuracy can be controlled by adjusting the threshold on the expected difference in the principal point across an image.

3.3. Faces

Shown in Fig. 2 is a computer generated head rendered using the `pbrt` environment [5]. The virtual camera had a principal point of $c_1 = c_2 = 0$, a focal length[3] of $f = 3.5$mm on a 3×2 mm^2 sensor, and the image was rendered at a resolution of 3000×2000 pixels. Because this synthetic face was largely featureless, we texture-mapped the head with three "freckles" which were then used as feature points. Eight other images of this head were rendered by translating the head vertically and/or horizontally and rotating about the vertical and/or horizontal axis by $20°$. In each of these eight images, the corresponding features were selected manually, from which H was estimated and factored to determine the principal point. For

[3]The focal length f in mm is converted to normalized units as follows. Denote the image dimensions in pixels as N_x and N_y, the sensor size in mm as S_x and S_y. The conversion from mm to pixels is $p = (N_x/S_x + N_y/S_y)/2$. The normalized focal length is $pf/(\max(N_x, N_y)/2)$.

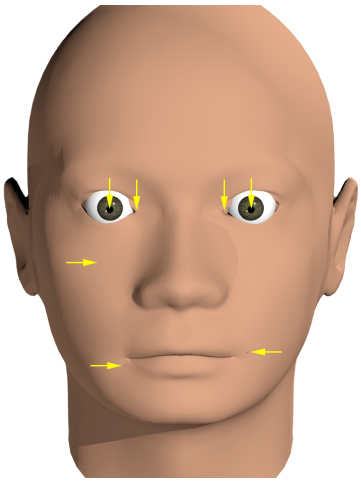

Fig. 2. A frontal-parallel face with seven feature points (yellow arrows).

the eight images, the estimated principal point norms were 0.01, 0.13, 0.18, 0.05, 0.06, 0.03, 0.09, and 0.07. With an average norm of 0.087, these estimates are well below our 0.2 threshold.

To simulate tampering, the head was translated to various locations and the principal point was estimated at each location. Shown in Fig. 3 are level curves denoting the deviation of the estimated principal point from the origin as a function of spatial position in the image. The inner curve denotes a distance of 0.2, and each subsequent curve denotes an increment of 0.1. With a threshold of 0.2, translations of the head outside of the inner-most curve can be detected as fake.

4. DISCUSSION

When creating a composite of two or more people it is often necessary to move a person in the image relative to their position in the original image. We have shown that this manipulation can be revealed by measuring inconsistencies in the camera principal point estimated from each person (or object). For each face, this technique estimates the world to image transformation from corresponding features on a reference image of the person directly facing the camera. This transformation is then factored to yield the desired principal point.

The major sensitivity with this technique is in localizing a sufficient number of feature points on the face. We are currently developing a method for automatic feature extraction that should improve the reliability of this technique. In addition, this technique could be used in conjunction with our earlier work that estimated the camera principal point from the image of a person's eye [3].

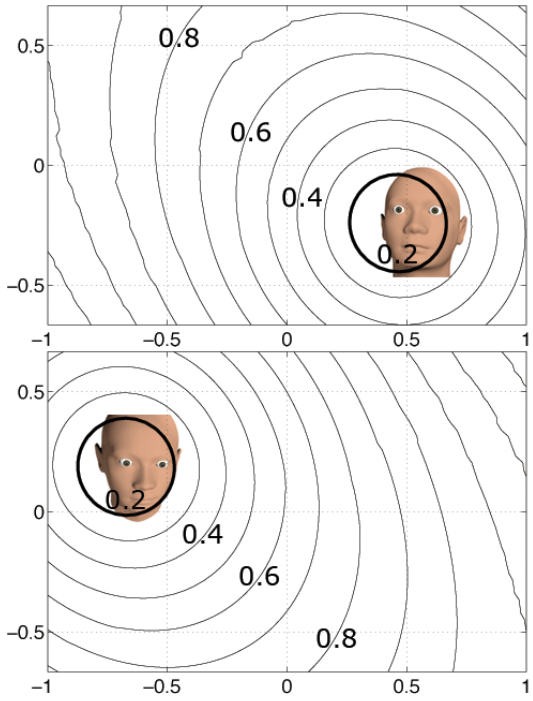

Fig. 3. The level curves show the deviation of the estimated principal point from the origin as a function of spatial position in the image. The values on the curve denote the principal point magnitude.

5. REFERENCES

[1] R. I. Hartley. In defense of the eight-point algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(6):580–593, 1997.

[2] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004.

[3] M. Johnson and H. Farid. Detecting photographic composites of people. In *6th International Workshop on Digital Watermarking*, Guangzhou, China, 2007.

[4] Y. Ma, S. Soatto, J. Kosecka, and S. Sastry. *An Invitation to 3D Vision: From Images to Geometric Models*. Springer Verlag, 2003.

[5] M. Pharr and G. Humphreys. *Physically Based Rendering: From Theory to Implementation*. Morgan Kaufmann, 2004.

[6] W. Wang and H. Farid. Detecting re-projected video. In *10th International Workshop on Information Hiding*, Santa Barbara, CA, 2008.

[7] W. Zhang, X. Cao, Z. Feng, J. Zhang, and P. Wang. Detecting photographic composites using two-view geometrical constraints. In *IEEE International Conference on Multimedia and Expo*, 2009.

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