Position Determines Perspective: Investigating Perspective Distortion for Image Forensics of Faces

Bo Peng [1,3], Wei Wang*[1,2], Jing Dong*[1,4] and Tieniu Tan [1]
[1] Center for Research on Intelligent Perception and Computing, Institute of Automation, CAS
[2] State Key Laboratory of Cryptology [3] University of Chinese Academy of Sciences
[4] State Key Laboratory of Information Security
{bo.peng, wwang, jdong, tnt}@nlpr.ia.ac.cn

Abstract

This paper points out a new telltale trace – the characteristic of perspective distortion (CPD), for the image forensics of faces. The perspective distortion is determined by the position of image shooting, and it is often overlooked when creating a forgery, which results in the inconsistency between the claimed camera parameters and the CPD in the face image. To investigate this consistency problem, we cast it to the consistency between the claimed camera intrinsic parameters and the estimated ones from the CPD. Our parameter estimation approach is based on geometric observations that are related to CPD, like facial landmarks and contours. We analyze the estimation uncertainty caused by indeterminacy of observation to obtain a more reliable forensic decision. Experiments on synthetic datasets and real forgery examples demonstrate the effectiveness of the proposed method.

1. Introduction

> *"Sometimes two people could look at the same picture and see different images.*
> *Position determines perspective."*
>
> — Nikki Turner

Images with faces are ubiquitous in daily life, as they frequently appear in all kinds of media like TV, newspaper and social network. They also play an important role for biometric identification. People rely on these face images to trust news reports about public figures or to identify the identities of other people. Hence, the authenticity of face images is vitally important, and vicious forgeries towards faces are particularly harmful. Unfortunately, people are often easily fooled by forgery face images, if they only focus on the figures appeared in the image content or the identification of those faces. To spot forgeries, we need to concentrate on the subtle and unnoticeable traces. In this paper, we reveal a new telltale trace, which is the inconsistency between claimed camera parameters and the characteristic of perspective distortion (CPD).

Photos by Dan Vojtěch, copyright 2016.
Figure 1. Different perspective distortions caused by different shooting positions. From left to right, the shooting position pulled farther away, and the focal length also increases correspondingly to make the size of head almost constant. The bottom numbers stand for the used focal lengths.

A well-known perspective distortion is that nearer components seem larger while farther ones seem smaller. More interestingly, the perspective distortion is related to the position of shooting, as the maxim goes: "position determines perspective". As photos [1] showing in Fig. 1, the face shot at a closer distance demonstrates more prominent perspective distortion, where it is more foreshorten and has larger nose compared with the ones shot farther. This phenomenon is used by photographers to achieve different artistic effects they want to express. It is also inspiring to us, since we can investigate the anomaly that a photo does not appear to be shot by the claimed camera. For example, a forger may falsely claim that the 200mm photo in Fig. 1 is shot by a camera with 20mm focal length for bad purpose, such as the scenario of recaptured photo. We note that for the 20mm camera to shoot an image with the same size of head, the

*Corresponding authors

[1] http://www.danvojtech.cz/blog/2016/07/amazing-how-focal-length-affect-shape-of-the-face/

shooting position has to be much nearer, resulting in more distortion than expected in the image. Hence, we can successfully claim this forgery.

To investigate the aforementioned inconsistency, we cast it into the problem of checking the consistency between claimed intrinsic parameters and those estimated from the CPD. The estimation is based on some geometric observations that reflect the CPD visually, which include the facial landmarks and contours. As we know, camera parameter estimation from only a single image is ill-posed. Hence, we need to build 3D face models. There exist 3D morphable models [1] which are statistical shape models for human faces that may serve this purpose. For serious forensic scenarios, like court of law and police investigation, we resort to high-precision 3D scanning devices for obtaining 3D face models. The indeterminacy in image observations of facial landmarks leads to estimation uncertainty. To analyze the uncertainty for each specific estimation, we use both random perturbation and theoretical approximation strategies. Experimental results verify the efficacy of the proposed estimation method and also show its forensic application to the detection of recaptured or spliced face images.

The main contributions of this work are:
1. We point out the phenomenon of inconsistent CPD as a new telltale trace for image forensics of faces.
2. In the forensic estimation of camera parameters, we employ an extra contour constraint in addition to traditional landmarks to tackle the problem of lacking precise feature points on faces.
3. To evaluate estimation uncertainty for more reliable forensic decision, we explore both random perturbation and theoretical approximation strategies and further show their accordance experimentally.

The rest of this paper is organized as follows. In Sec. 2, we briefly review some related scene based forensic methods. Sec. 3 gives some insights about the perspective distortions for forensics. Detailed descriptions of the proposed camera parameter estimation approach are presented in Sec. 4. In Sec. 5, estimation uncertainty is analyzed and forensic consistency measure is proposed. We then discuss some experimental results in Sec. 6, and finally conclusions are drawn in Sec. 7.

2. Related Work

There have been different forensic methods focusing on a variety of traces like the artifacts of copy-move [5], JPEG compression [19] and camera sensor noise [13]. In this section, we mainly review some more related scene based forensics methods. The scene traces that can be used for image forensics include height ratios, planar metrics, lighting environment, physical stability and maybe more. In [7, 22], the authors propose to estimate the height ratio of people from image using cross ratio property and compare it to the assumed known actual height ratio for forensics. Metric measurements on planar surfaces are made possible by estimating the planar homography and then rectifying it based on known planar shapes like polygons [8] and characters [3]. The work [11, 17, 16] estimates the 3D lighting environments using approximate 3D face models and Lambertian reflection model, and then decide the consistency in 3D lighting. Similarly, the 3D eyeball model is used in [9] together with specular reflection on the eyes to estimate the lighting direction for forensics. Approximate 3D models of the imaged objects are also used in [4, 18], where the authors recover the scene of the famous *Lee Harvey Oswald backyard photo* and verify the consistency in multiple clues like the lighting, the object size and the physical stability.

We can see that the scene traces for image forensics all lie in the subtle and unnoticeable aspects of the image, or the human visual system is not accurate enough to calculate their consistencies. Our proposed trace of different characteristics of perspective distortion is also a valuable trace of this kind that has potential for image forensics. It should also be noted that most scene based forensic methods require some levels of prior knowledge about the imaged scene as our method does.

Since the proposed method estimates the camera intrinsic parameters for forensics, we also review the related work in [20, 10, 12, 15, 14]. The authors in [20] propose to estimate the skewness parameter for detecting re-projected videos. In [10], planar homography is estimated using images of eye circles, and then the position of 2D principal point is estimated assuming known focal length. However, images of eye circles are often in low resolution and partially occluded. In a following-up work [12], the authors approximately treat the face as a planar surface, and use feature points on it to estimate the homography. The work in [15] employs three groups of mutually orthogonal straight lines to estimate the 3D principal point position in images containing mirrors, and the work in [14] applies similar idea to detect image cropping. Our work is different from these methods in that we explore a new trace of perspective distortion in face images.

3. Perspective Distortions of Cameras

This work uses the perspective pinhole camera model to describe the transformation from a 3D model point $\mathbf{X}$ to its 2D image point $\mathbf{x}$:

$$\mathbf{x} = P\mathbf{X} = K[R|\mathbf{t}]\mathbf{X} \tag{1}$$

Here, the points are in homogeneous coordinate form, P denotes the projection matrix, and $K, R, \mathbf{t}$ are respectively the camera intrinsic matrix, the rotation matrix and the translation vector. More specifically, the rotation matrix can be parameterized by three Eulerian angles (α, β, γ), and

$\mathbf{t} = [t_x, t_y, t_z]^{\mathrm{T}}$. The intrinsic matrix in its full form is:

$$K = \begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

where f_x, f_y are respectively the focal lengths in x and y pixel units, (c_x, c_y) is the position of the camera principal point and s denotes the skewness of pixel. Since modern camera sensors usually have square pixel units, f_x and f_y approximately equal and s is approximately zero. Thus, a simplified intrinsic matrix is parameterized as:

$$K = \begin{bmatrix} f & 0 & c_x \\ 0 & f & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

By using the simplified camera intrinsic model (3), the camera projection in Eqn. (1) can be written as:

$$\mathbf{x} = P(\boldsymbol{\theta})\mathbf{X} \tag{4}$$

where $\boldsymbol{\theta} = [f, c_x, c_y, \alpha, \beta, \gamma, t_x, t_y, t_z]^{\mathrm{T}}$ is the vector of 9 unknown camera parameters. We denote the intrinsic parameters $[f, c_x, c_y]^{\mathrm{T}}$ as $\boldsymbol{\theta}^{in}$ and the extrinsic parameters $[\alpha, \beta, \gamma, t_x, t_y, t_z]^{\mathrm{T}}$ as $\boldsymbol{\theta}^{ex}$. The projection function with known 3D model point is also denoted as:

$$\mathbf{x} = g(\boldsymbol{\theta}; \mathbf{X}) \tag{5}$$

The perspective distortion on an object is the warping phenomenon of the object's image that differs from the object's normal geometric appearance [2]. It is dependent on the position of shooting. To our knowledge, there is no established quantitative definition for perspective distortion. In the following, we simply express it in the form of relative component size to explain the phenomenon in Fig. 1 in a simplified setting. Fig. 2 shows the imaging of a head seen from above. Using the law of camera projection, we can obtain that:

$$\frac{d_1}{d_2} = \frac{f\frac{w_1}{z_1}}{f\frac{w_2}{z_2}} = \frac{w_1}{w_2}(1 + \frac{\delta_z}{z_1}) \tag{6}$$

We can see that, since w_1, w_2, δ_z, which respectively represent the width of the nose, the width between two ears, and the depth between nose and ears, are constant measures for a given neutral face, the ratio between the image widths of nose and ears is only related to the inverse of the distance from the head to the camera z_1. As a result, a imaged face has relatively smaller nose (compared to the distance between ears) when it is far away from the camera, and a larger nose when it is close. We also verified that this perspective distortion is directly related to the shooting position z_1, and indirectly related to the camera intrinsic parameter f if we want to keep the size of the head image constant as in Fig. 1. As a matter of fact, besides the perspective distortion related to the object's position in the depth direction, there also exists another distortion related to that in the lateral direction relative to the camera principal axis. A common example is that faces in the periphery of images are often more stretched laterally. Because of the limit of space, we do not discuss this lateral distortion in more details.

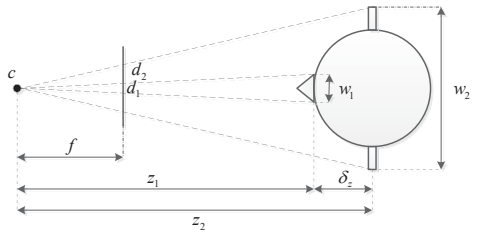


Figure 2. The imaging of a head seen from above.

When a forger recaptures or splices a face image, the CPD is usually overlooked. However, the inconsistency between claimed camera intrinsic parameters and the distortion in the image is still there, just like the example described in Sec. 1. Since this inconsistency is not directly measurable, we cast it into the inconsistency between the claimed camera intrinsics and those estimated ones from the CPD. In practice, we use some geometric features that visually reflect the CPD for the estimation, which includes facial landmarks and contours.

4. Camera Parameter Estimation

There already exist classic camera calibration methods in the literature like the Gold Standard method [6]. However, they cannot perfectly deal with the estimation problem for face images. Classic camera calibration has good estimation results, because it uses a precise 3D calibration object with black and white checkerboard patterns on it, where accurate positions of both 2D and 3D feature points can be easily obtained. While in our case, the facial landmarks are semantically defined and are lack of texture, hence they cannot be accurately localized to obtain a good estimation result. To tackle this problem, we propose to further refine the estimation result using the observation of 2D contour points, which can be more precisely localized. More details are described in the following.

4.1. Estimation with Landmarks

Following the classic camera calibration method in [6], we first describe the estimation method using known correspondences between 2D and 3D facial landmarks. The 2D facial landmarks are automatically detected using the SDM method in [21], and we only keep a part of the detected land-

[2] https://en.wikipedia.org/wiki/Perspective_distortion_(photography)

marks which are more robust and recognizable as shown in Fig. 3. We also manually pick the landmarks on tips of ears, earlobes and the chin when they are visible in the image. The landmark positions on the known 3D face model are manually picked and also shown in Fig. 3.

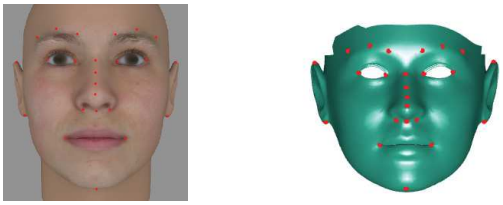


Figure 3. The facial landmark positions on the 2D image (left) and the 3D model (right).

Given the corresponding 2D and 3D facial landmark coordinates, the 3x4 projective matrix P in Eqn. (1) can be solved using the Gold Standard method in [6], which comprises of a Direct Linear Transform (DLT) step for minimizing algebraic error and a Levenberg-Marquardt step for refining geometric error. The readers are referred to [6] for more details about the Gold Standard camera calibration method. After the calculation of the projective matrix P, we can use RQ-decomposition to factor P into $K, R, \mathbf{t}$. However, the intrinsic matrix K obtained this way is in its full form as (2) which is not desired. Because errors in the localization of 2D and 3D facial landmarks can result in large camera skewness s and big difference in f_x and f_y, representing a highly unlikely estimation of camera model.

Starting from the parameters estimated by the Gold Standard method, we further gently pull s to zero and f_x, f_y to each other by minimizing the following regularized geometric error:

$$E_{land}(\tilde{\boldsymbol{\theta}}) = \sum_{i=1}^{N_l} d(\mathbf{x}_i, P(\tilde{\boldsymbol{\theta}})\mathbf{X}_i)^2 + w_s s^2 + w_f (f_x - f_y)^2 \tag{7}$$

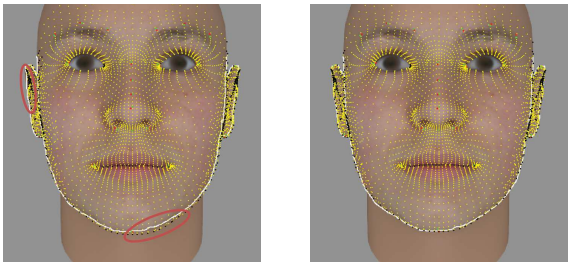
Here N_l is the number of landmarks. $\tilde{\boldsymbol{\theta}}$ is the augmented version of $\boldsymbol{\theta}$ in (4), since it contains intrinsic parameters in the full form K. The Euclidean distance between two points is represented by $d(,)$, and w_s and w_f are respectively the regularization weights on s and $(f_x - f_y)$. We gradually strengthen the weights (w_s, w_f) in multiple iterations to gently pull s to zero and f_x, f_y to each other while in the same time keeping the data error at a relatively low level. When s is sufficiently close to zero and f_x, f_y are sufficiently close to each other, we clamp s to zero and initialize f at $(f_x + f_y)/2$ and minimize the following geometric error:

$$E_{land}(\boldsymbol{\theta}) = \sum_{i=1}^{N_l} d(\mathbf{x}_i, P(\boldsymbol{\theta})\mathbf{X}_i)^2 \tag{8}$$

Both the cost functions in (7) and (8) are minimized using Levenberg-Marquardt algorithm.

4.2. Refinement by Contours

As has been explained, the error-prone landmark positions alone are not adequate to constrain the camera parameters, and we propose to further refine the estimation result using contours. Our contour points are defined as the occluding contours of the face, as shown by the white points in Fig. 4. We denote the set of observed 2D contour points as $\mathcal{C}_2$. They usually lie around the face and ears, and also around the nose when the face is not in frontal poses. The estimation process using contour points is more complicated than that using facial landmarks in that 3D contour points can change positions with respect to camera parameters and correspondences between 2D and 3D contour points are not predefined. To tackle these problems, we use the Iterative Closest Point (ICP) algorithm to update 2D-3D correspondences and estimate the parameters iteratively.



(a) Before refinement (b) After refinement

Figure 4. The effect of contour refinement (please see the online electronic version and zoom in for details). The white points are the user annotated contour points around the face and ears. Yellow points represent the projected vertices of the 3D face model under the current camera parameters. Red and green points are the 2D facial landmarks and the projected positions of 3D ones respectively. Black points represent the projections of 3D contour points. The areas emphasized by the red circles in (a) show some misalignments before contour refinement.

More specifically, in each iteration, the 3D contour points at the current camera parameters are first determined. We define the occluding contours as those points whose normal directions are perpendicular to their viewing rays in the camera coordinate frame, i.e.:

$$\mathcal{C}_3 = \{\mathbf{X}_i |\ \mathbf{X}_i \in \mathcal{V}, 0 \leq (R\mathbf{n}_i)^{\mathrm{T}} \cdot [R|\mathbf{t}]\mathbf{X}_i < \epsilon\} \tag{9}$$

where $\mathcal{C}_3$ represents the set of 3D occluding contour points, $\mathcal{V}$ is the set of all vertices on the 3D face model, and $\mathbf{n}_i$ is the normal vector at $\mathbf{X}_i$. After the 3D contour points $\mathcal{C}_3$ are located, we project them to the image plane using $P(\boldsymbol{\theta})$, which are shown by the black points in Fig. 4 and denoted as the set $\hat{\mathcal{C}}_2$. Then, for each observed 2D contour point in $\mathcal{C}_2$, we assign the closest point in $\hat{\mathcal{C}}_2$ as its corresponding point. In this way, we can find its 3D corresponding point in $\mathcal{C}_3$. To exclude potential outlier matches, we discard the correspondences whose distance is greater than a threshold.

With the contour correspondences updated, we now refine the parameters by minimizing the following combined cost function:

$$E_{totle}(\boldsymbol{\theta}) = E_{cont}(\boldsymbol{\theta}) + \lambda E_{land}(\boldsymbol{\theta}) \tag{10}$$

where the contour cost function $E_{cont}(\boldsymbol{\theta})$ is:

$$E_{cont}(\boldsymbol{\theta}) = \sum_{i=1}^{N_c} d(\mathbf{c}_i, P(\boldsymbol{\theta})\mathbf{C}_i)^2 \tag{11}$$

where N_c is the number of valid contour correspondences, $\mathbf{c}_i$ is a valid point in $\mathcal{C}_2$, and $\mathbf{C}_i$ is its corresponding point in $\mathcal{C}_3$. The weight λ in (10) is set to 0.1 normalized by number of points, i.e. $\lambda = 0.1N_c/N_l$. The combined cost function (10) is also minimized by Levenberg-Marquardt algorithm.

5. Uncertainty Analysis and Consistency Measure

Evaluating the uncertainty of a decision is very import for a forensic tool. For our estimation method, a major uncertainty source is the localization of 2D and 3D facial landmarks. Since these landmarks are defined semantically, their locations are not precise. Although we further refine the estimation result using contour points, the refinement itself is initialized by the landmark result and can be affected by the initialization. In this work, we analyze the estimation uncertainty caused by facial landmark uncertainty. Instead of considering the uncertainties in both 2D and 3D landmarks, we treat the 3D landmarks as golden standard or without errors, and cast all errors to 2D landmarks. We describe two ways to evaluate the estimation uncertainty, i.e. random perturbation and theoretical approximation.

The random perturbation method for evaluating estimation uncertainty is inspired by [15, 2]. We randomly perturb the positions of 2D facial landmarks, and then estimate the camera parameters multiple times using the perturbed landmarks. If we denote our maximum likelihood estimator as $\hat{\boldsymbol{\theta}} = h(\{\mathbf{x}_i\}; \{\mathbf{X}_i\})$, where $h(\circ)$ is a complicated function with no analytic form, then the random perturbation method estimate the uncertainty of parameters by repeatedly evaluating $h(\circ)$ at randomly sampled $\{\mathbf{x}_i\}$ in an uncertainty zone. Here, we model the uncertainty in 2D landmark positions as Gaussian distributions centered at the initially localized positions with a standard deviation of σ. We set σ as the root of mean squared error (rmse) after the minimization of E_{land} in (8) using the initially localized landmarks. The 2D facial landmarks are independently drawn from this Gaussian N_s times, and each time the estimation process is re-run with the currently drawn landmarks. An example of estimates obtained in this way is shown in Fig. 5. We can see that the uncertainty zone of estimation tightly covers the groundtruth value for this example.

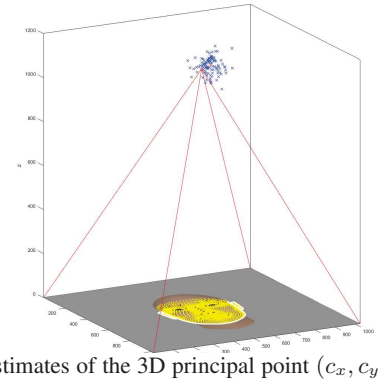


Figure 5. Estimates of the 3D principal point (c_x, c_y, f) obtained by 100 times of random perturbation. The red circle is the groundtruth position, blue crosses represent the estimates, and the red cross is the estimate without perturbation (zoom in for details).

Another way to evaluate the uncertainty in estimated parameters is by theoretical approximation [6]. By approximating the projection function $\{\mathbf{x}_i\} = g(\boldsymbol{\theta}; \{\mathbf{X}_i\})$ using first order affine function in a vicinity of the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ and back propagating the covariance matrix of $\{\mathbf{x}_i\}$, the covariance matrix of $\hat{\boldsymbol{\theta}}$ is obtained as [6]:

$$\Sigma_{\boldsymbol{\theta}} = (J^{\mathrm{T}}\Sigma_{\mathbf{x}}^{-1}J)^{-1} \tag{12}$$

Here, the notations $\Sigma_{\boldsymbol{\theta}}, \Sigma_{\mathbf{x}}$ are respectively the covariance matrices of estimated parameters and 2D points, and J denotes the Jacobian matrix of $\{\mathbf{x}_i\}$ in terms of $\boldsymbol{\theta}$. Here, $\Sigma_{\mathbf{x}}$ is set to a diagonal matrix, and the diagonal elements are the mean squared error (mse) after the minimization of E_{land} using the initially localized landmarks. This theory only applies to fixed 2D and 3D correspondences like landmarks, and it is essentially an approximation. Thus, we mainly employ the random perturbation method in the experiments, and just use the theoretical method as a comparison.

In the following, we describe the consistency measure between the estimated intrinsics and the claimed one. To account for the estimation uncertainty, we use the random perturbation method and obtain a set of estimated intrinsic parameters $\{\hat{\boldsymbol{\theta}}_i^{in}\}$. We then measure the Mahalanobis distance between $\{\hat{\boldsymbol{\theta}}_i^{in}\}$ and the parameter of the claimed camera $\boldsymbol{\theta}^{in}$ as follows:

$$D(\{\hat{\boldsymbol{\theta}}_i^{in}\}, \boldsymbol{\theta}^{in}) = \sqrt{(\boldsymbol{\theta}^{in} - \boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{\theta}^{in} - \boldsymbol{\mu})} \tag{13}$$

where $\boldsymbol{\mu}$ and Σ are respectively the mean and covariance matrix of the estimate set $\{\hat{\boldsymbol{\theta}}_i^{in}\}$. The Mahalanobis distance normalizes the absolute distance by the uncertainty or covariance of the estimates, assuming the estimates are in Gaussian distribution. Hence it is more suitable for distance measurement in our forensic application. After the calculation of D, we compare it with an experimentally determined threshold D_t to decide the consistency between the estimated parameter and the claimed one.

6. Experiments

In this section, we experimentally show the accuracy and uncertainty of the proposed estimation method. The reasons behind these results are also analyzed. Finally, we also show possible applications in image forensics.

6.1. Estimation under Different Imaging Distances

We first test the method's accuracy in the scenario as Fig. 1, where we adjust both the distance and focal length parameters to make the size of the head almost constant in the image. Six synthetic images generated this way are shown in Fig. 6. The resolution of these images is 1024x1024 pixels. The relation between field of view (FOV) and focal length is $FOV = 2\arctan(w/2/f)$, where w is 1024 here. Hence, large FOV is equivalent to small focal length.

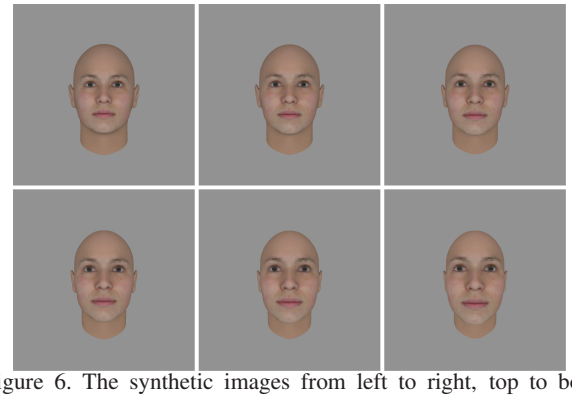


Figure 6. The synthetic images from left to right, top to bottom are generated using cameras with field of view (FOV) being: $20°, 30°, 40°, 50°, 60°, 70°$. The distances are respectively: 1336, 879, 647, 505, 408, 336 in millimeters.

The estimation results for these 6 images with and without the contour refinement step are shown in Fig. 7. The boxplots are produced using 1000 estimates of random perturbation. We can see that the groundtruth values of intrinsic parameters all lie within the zone of our estimations, and some of them lie in the middle half of the estimations. By comparing the intervals of estimations with and without contour refinement, we can verify that contour refinement can effectively reduce the uncertainty interval of the estimation. Without contour refinement, the estimation intervals for focal length of adjacent images will intersect. While using contour refinement, these intervals can be separated. This is clearly preferable for the forensic method to be more discriminative. We also note that for most images, using contour refinement can get a more accurate estimation, since the central line of boxplot is more close to the groundtruth, especially for the estimations of f and c_y.

Another very interesting and prominent observation in Fig. 7 is that the estimation uncertainty becomes smaller and smaller with the distance gets closer, or with the focal

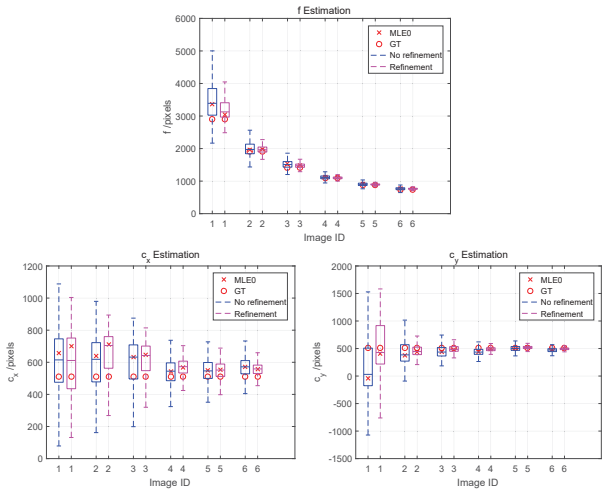


Figure 7. Box plots for focal length and principal point estimation results on each image. "MLE0" is the estimate without perturbation, and "GT" stands for the groundtruth value.

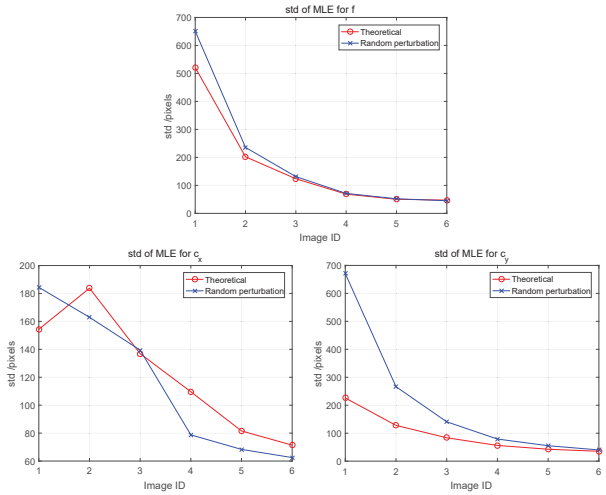


Figure 8. Standard deviation of estimates without contour refinement for each image.

length gets smaller. To explain this intuitively and informally, this is because when the head is closer to the camera, a small variation in the distance will cause relatively larger differences in the CPD. Recall from Eqn. (6) that the distortion is related to the inverse of distance. Hence, smaller distance has larger derivative and is more sensitive to distance variation. Note that the focal length is varied proportionally in the same time to make the size of head almost constant. As a result, when the head is nearer, the estimation uncertainty for focal length is smaller. In Fig. 8, we also compare the estimation uncertainties obtained by random perturbation to those obtained by theoretical calculation. Here we just show the uncertainties of estimation without contour refinement, since the theoretical method only applies to fixed correspondences. We can see that the trend of theo-

retical results agrees well with that of random perturbation, especially for the estimation of focal length. The theoretical result also shows that the estimation is more uncertain when the head is farther away. Note that the theoretical result is not exactly the same with random perturbation, because it is a first order approximation.

6.2. Estimation with Partial Observation

We have shown the estimation performance on images in Fig. 6, where the facial observations are complete. However, in some cases, we only have partial face data due to the occlusion. For example, the ears are sometimes covered by one's hair, or one of the ears is not visible in non-frontal poses. To show the influence of partial observation on the estimation performance, we apply the proposed method to images in Fig. 6 again, but omit the landmark and contour observations on both ears. The results can be seen in Fig. 9. As expected, the comparison shows that partial observation will result in larger uncertainty and worse accuracy. The estimation intervals for focal length intersect with each other among different images, and some estimations for principal point even cannot cover the groundtruth values. Through this, we can find that full observation is very import for the performance of the estimation method. More image observations lead to better accuracy and certainty. Also, the observations on ears are crucial, because with observations on the ears, the total extent of observations in the depth direction is extended.

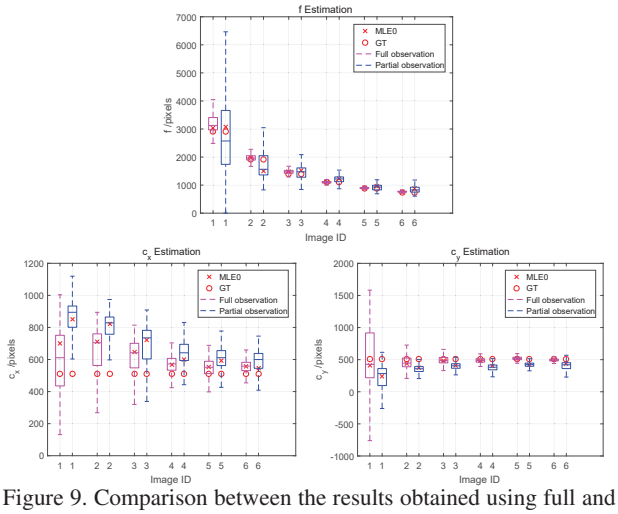


Figure 9. Comparison between the results obtained using full and partial observations.

6.3. Determination of Distance Threshold D_t

To determine the distance threshold D_t as described in Sec. 5, we created a larger synthetic dataset comprising of 100 images. Three example images are shown in Fig. 10. These images are rendered with random camera field of view from $20°$ to $70°$, the pose and position of the head are also random, and the distance between eyes ranges from 100 to 150 pixels. The image resolution is also 1024x1024. On some images, the observations on the ears are not available due to self-occlusion. For all the images, we manually annotate the contours around the face, ears and nose.

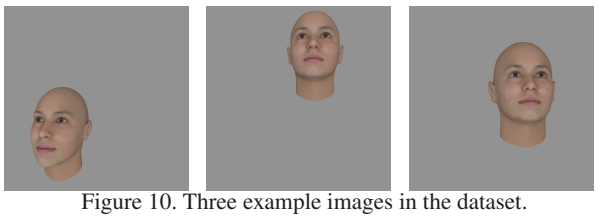
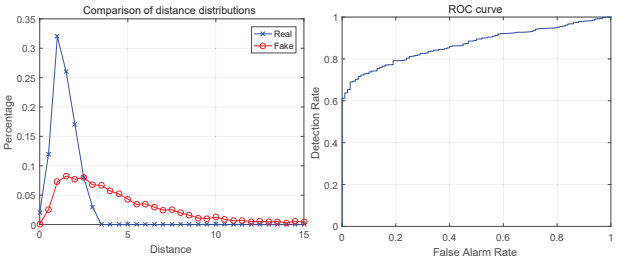


Figure 10. Three example images in the dataset.

We then estimate the camera parameters 1000 times for each image. The certainty levels or covariances of these estimations are apparently different from image to image, because as shown by the previous two subsections, the different distances, poses and observations have influence on estimation certainty. Here, we show the statistics of Mahalanobis distances (13) on this dataset. The distributions of Mahalanobis distances between estimates and the claimed intrinsic parameters are shown in Fig. 11 (a). In this figure, the "Real" curve stands for the distance between estimates and the corresponding groundtruth intrinsics. As a comparison, for each image, we also calculate the distance between its estimates and the groundtruth intrinsics of each of the other 99 images. We refer to the curve calculated this way as "Fake", because it can be imagined as an image claimed to be shot by a different camera. Note that for this dataset, we have 100 real distances and $100 \times 99 = 9900$ fake distances, and we present the histograms in percentage in Fig. 11 (a). The "Fake" curve actually has a long tail all the way to 78.4, which is clipped for better visualization.



(a) (b)

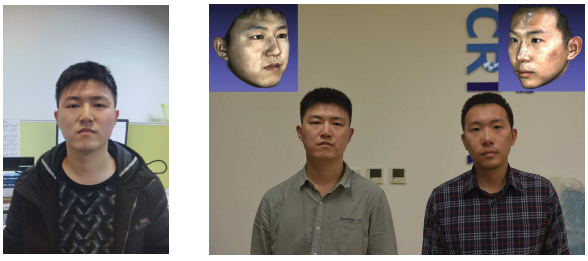
Figure 11. Estimation results on 100 synthetic images. (a) The distributions (histograms) of distances between estimates and the claimed camera intrinsic parameters. (b) ROC curve for (a).

The ROC curve for detecting the "Fake" images is shown in Fig. 11 (b) with an AUC value being 87%. In preference for a low false alarm rate (FAR), we choose the distance threshold D_t as the threshold at 1% FAR, which is **3**. Under this threshold, 37.6% of the fake distances will be false-

ly accepted as real. However, this false acceptance rate can be further decreased if we also consider the scenario of face splicing. Because the splicing of a face to another image often translates the position of the face, causing the estimated principal point to shift away from the image center [10]. Thus, the difference between estimates and claimed intrinsics also has large variation in (c_x, c_y) domain apart from in the focal length direction, leading to the "Fake" curve shifting rightwards in Fig. 11 (a). Another way to decrease the false acceptance rate is to improve the estimation accuracy and lower its uncertainty by precisely employing more image observation types, which is a future research direction.

6.4. Examples of Forensic Application

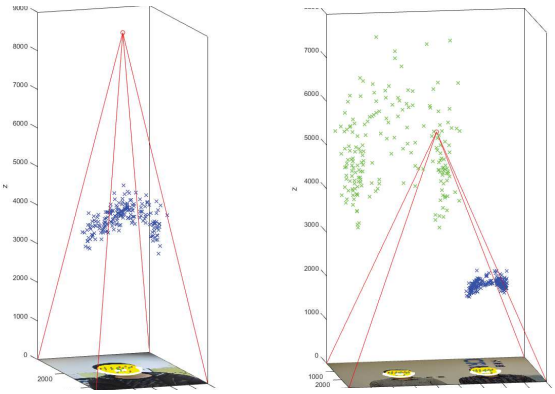
We then show two examples of our proposed method applied to the forensic detection of image recapture and face splicing. The questioned images are shown in Fig. 12, where (a) is an image captured by an iPhone 5S camera and then recaptured by a NIKON D750 camera, and (b) is a composition of two faces shot by the NIKON D750 at two different distances. The person on the left side (ID1) is original in the image shot at a farther distance with larger focal length, while the person on the right side (ID2) is spliced in and originally shot at a closer distance with smaller focal length. For the two involved persons, we obtained their 3D face models in the same neutral expressions using a high precision 3D face scanner, which are also shown in Fig. 12.



(a) Recapture (b) Face splicing

Figure 12. Two example questioned images and the scanned 3D models of the two involved faces. The person in the right side of (b) is spliced from another image.

The results of our estimation method with 200 times of random perturbation are shown in Fig. 13. For the recaptured image, the Mahalanobis distance between the estimates and the claimed intrinsic parameter is 20.4, which is much larger than the threshold 3. Thus, we can correctly decide this image as inconsistent with the claimed camera. For the spliced image, the distance between ID1 estimates and the extracted intrinsics is 1.8 which is lower than D_t and deemed to be real, while the distance for ID2 is 35.0 and can be decided as a spliced part. Note that in Fig. 13 (b), the uncertainties for the two estimations are different because of different imaging distances as discussed in Subsection 6.1. We can see that the proposed method makes correct decisions on these examples, verifying its applicability and effectiveness for potential forensic usages.



(a) Recapture (b) Face splicing

Figure 13. Estimation results for the two example images. The red circles in two images are the "groundtruth" 3D principal point positions extracted from the EXIF information. The green point cloud in (b) is the set of estimates for ID1 on the left, and the blue cloud is for ID2 on the right.

7. Conclusion

This paper proposes the characteristic of perspective distortion (CPD) as a novel trace for image forensics of faces. The CPD is usually overlooked by a forger when falsely claiming an image to a camera with inconsistent intrinsic parameters. To detect this inconsistency, we propose to estimate the camera parameters from the CPD using both facial landmarks and contours, and also evaluate the estimation uncertainty for decision making. Experimental results on synthetic data verify the efficacy of the method and give an insight on its impacting factors like the imaging distance and partial observations. We also showcase the effectiveness for potential forensic applications on real examples.

An issue of our method is the assumption of having exact 3D face models. This limits its application to scenarios where the involved people can be cooperative, like the court of law and police investigation. In this spirit, anti-spoofing in face recognition is also a potential application. For future research, a direction is to examine the dependency on 3D model accuracy and to explore more general models, e.g. [1]. More evaluations on real forgery data are also needed.

Acknowledgements

This work is funded by the National Natural Science Foundation of China (Grant No. 61502496, No. U1536120 and No. U1636201), Beijing Natural Science Foundation (Grant No. 4164102), and the National Key Research and Development Program of China (No. 2016YFB1001003).

References

[1] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '99, pages 187–194, New York, NY, USA, 1999. ACM Press/Addison-Wesley Publishing Co.

[2] T. Carvalho and H. Farid. Exposing photo manipulation from user-guided 3d lighting analysis. In *SPIE Symposium on Electronic Imaging*, pages 940902–940902–10, 2015.

[3] V. Conotter, G. Boato, and H. Farid. Detecting photo manipulation on signs and billboards. In *2010 IEEE International Conference on Image Processing*, pages 1741–1744. IEEE, 2010.

[4] H. Farid. A 3-d photo forensic analysis of the lee harvey oswald backyard photo. *Hanover, NH*, 2010.

[5] A. J. Fridrich, B. D. Soukal, and A. J. Lukáš. Detection of copy-move forgery in digital images. In *in Proceedings of Digital Forensic Research Workshop*. Citeseer, 2003.

[6] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[7] M. Iuliani, G. Fabbri, and A. Piva. Image splicing detection based on general perspective constraints. In *Information Forensics and Security (WIFS), 2015 IEEE International Workshop on*, pages 1–6. IEEE, 2015.

[8] M. K. Johnson and H. Farid. Metric measurements on a plane from a single image. *Dept. Comput. Sci., Dartmouth College, Tech. Rep. TR2006-579*, 2006.

[9] M. K. Johnson and H. Farid. Exposing digital forgeries through specular highlights on the eye. In *Information Hiding*, pages 311–325. Springer, 2007.

[10] M. K. Johnson and H. Farid. *Detecting Photographic Composites of People*, pages 19–33. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[11] E. Kee and H. Farid. Exposing digital forgeries from 3-d lighting environments. In *Information Forensics and Security (WIFS), 2010 IEEE International Workshop on*, pages 1–6. IEEE.

[12] E. Kee and H. Farid. Detecting photographic composites of famous people. Technical Report TR2009-656, Department of Computer Science, Dartmouth College, 2009.

[13] J. Lukáš, J. Fridrich, and M. Goljan. Detecting digital image forgeries using sensor pattern noise. In *Electronic Imaging 2006*, pages 60720Y–60720Y. International Society for Optics and Photonics, 2006.

[14] X. MENG, S. NIU, R. YAN, and Y. Li. Detecting photographic cropping based on vanishing points. *Chinese Journal of Electronics*, 22(2):369–372, 2013.

[15] J. F. O'Brien and H. Farid. Exposing photo manipulation with inconsistent reflections. *ACM Transactions on Graphics*, 31(1):4:1–11, Jan. 2012. Presented at SIGGRAPH 2012.

[16] B. Peng, W. Wang, J. Dong, and T. Tan. Automatic detection of 3d lighting inconsistencies via a facial landmark based morphable model. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3932–3936, Sept 2016.

[17] B. Peng, W. Wang, J. Dong, and T. Tan. Optimized 3d lighting environment estimation for image forgery detection. *IEEE Transactions on Information Forensics and Security*, PP(99):1–1, 2016.

[18] S. Pittala, E. Whiting, and H. Farid. A 3-d stability analysis of lee harvey oswald in the backyard photo. *The Journal of Digital Forensics, Security and Law: JDFSL*, 10(3):87, 2015.

[19] W. Wang, J. Dong, and T. Tan. Exploring dct coefficient quantization effects for local tampering detection. *IEEE Transactions on Information Forensics and Security*, 9(10):1653–1666, 2014.

[20] W. Wang and H. Farid. *Detecting Re-projected Video*, pages 72–86. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[21] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 532–539, 2013.

[22] H. Yao, S. Wang, Y. Zhao, and X. Zhang. Detecting image forgery using perspective constraints. *Signal Processing Letters, IEEE*, 19(3):123–126, 2012.