EYE SPECULAR HIGHLIGHTS TELLTALES FOR DIGITAL FORENSICS: A MACHINE LEARNING APPROACH

Priscila Saboia, Tiago Carvalho, and Anderson Rocha

University of Campinas (Unicamp)
Campinas, SP, Brazil
{psaboia, tjose, anderson.rocha}@ic.unicamp.br

ABSTRACT

Among the possible forms of photographic fabrication and manipulation, there is an increasing number of composite pictures containing people. With such compositions, it is very common to see politicians depicted side-by-side with criminals during election campaigns, or even Hollywood superstars relationships being wrecked by allegedly affairs depicted in gossip magazines. Thinking about this problem, in this paper we analyze telltales obtained from highlights in the eyes of every person standing in a picture in order to decide whether or not those people were really together at the moment of such image acquisition. We validate our approach with a data set containing realistic photographic compositions, as well as authentic unchanged pictures. As a result, our proposed extension improves the classification accuracy of the state-of-art solution in more than 20%.

Index Terms— Composite Photographs of People, Digital Forensics, Eye Specular Highlights

1. INTRODUCTION

In every minute of our digital lives we are struck by an ever-growing flood of information. Drowned within such an amount of information, we have too much at stake to take everything we have access to as the sole truth. Once taken as genuine for granted, photographs are no more perceived as a "piece of truth".

With the advance of digital image processing and computer graphics techniques, it has never been so easy to manipulate images and, therefore, forge new realities. When such modifications are no longer innocent image adjustments and start implying legal threats to a society, it becomes imperative to devise and deploy efficient and effective approaches to detect such activities [1]. Unfortunately, most of times, these modifications seek to deceive viewers, change opinions or even affect how people perceive reality [1].

To keep the pace with the advances in digital image processing and computer graphics tools, forensics experts strive for developing modern and sophisticated tools to identify forgeries. However, in this "arms race", for each new forensics method developed, a new method to perform a more sophisticated forgery is developed as a counterpart. This leads forensics approaches to aim at detecting all possible tampering telltales present in a given image in order to undermine forgers.

We thank the São Paulo Research Agency (FAPESP) for funding the research under the Award 2010/05647-4. We also express our gratitude to M. Johnson and H. Farid for kindly providing us with their source code and promptly answering our questions.

Recently, some researchers have successfully presented forensics approaches exploring features such as compression artifacts [2], statistical descriptors [3], acquisition telltales [4, 5], and illumination inconsistencies [6, 7, 8]. Please refer to [1] for a comprehensive survey.

Approaches based on illumination inconsistencies are of particular interest since a perfect illumination adjustment in a digital composite is very difficult to obtain. Normally, a composition involves splicing together two or more images, each one potentially having a different illumination condition, hardening the forgery creation. Another advantage of this class of methods is that it can be used to analyze analog pictures [1].

Among all different possible forms of photo manipulation, there is an astonishing number of composites of people. With such montages it is very common to see politicians depicted side-by-side with criminals during election campaigns or even Hollywood superstars' relationships being wrecked by allegedly affairs depicted in gossip magazines. Thinking about this problem, Johnson and Farid [7] proposed a method to analyze eye highlights telltales of every person standing in a still picture depicting two or more people and to confront the analyzed clues to decide whether or not those people were together for real in the moment of image acquisition.

Although the authors presented promising results with their approach, in this paper we extend their work giving the forensics community a step further with respect to the detection of composite photographs of people. We extend the original features proposed by Johnson and Farid in [7] and propose to take all the full advantage of recent machine learning algorithms to improve the previous work.

We validate our approach with a data set comprising realistic photomontages as well as natural still images. The proposed extension improves the classification accuracy of Johnson and Farid's previous solution in more than 20%. Finally, envisioning the use of such a method in a forensics scenario, we also discuss some method's limitations and present future directions for further improvements.

2. STATE-OF-THE-ART

Forensics methods that analyze image lighting inconsistencies to reveal traces of digital tampering are promising given that it is difficult to match the different lighting conditions when creating a composite. Johnson and Farid [6] presented an approach for estimating the light source direction from a single image assuming some simplifying conditions: (1) the surface is Lambertian (it reflects light isotropically); (2) it has a constant reflectance value; (3) it is illuminated by a point light source infinitely far away. Johnson and Farid further extended this solution to complex lighting environments [8].

In Johnson and Farid [7], the authors present another technique

which also investigates lighting inconsistencies but this time for the particular case of composition (fakes) involving people. According to the authors, specular highlights that appear on the eye are a powerful cue to the shape, color, and location of the light source in the scene [7]. Inconsistencies in these light properties can be used as telltales for detecting tampering. The method is based on the fact that the position of a specular highlight is determined by the relative positions of the light source, the reflective surface of the eye, and the viewer (i.e., the camera). Roughly speaking, the method can be divided into three stages, as Fig. 1 depicts.

The first stage aims at estimating the direction of the light source for each eye present in the picture. The second stage (characterization) seeks to estimate the position of the light source based on the specular highlights present in the eyes and on the corresponding estimated directions of the light source. The calculated position of the light source is then used to calculate the angular error for each specular highlight (given by the angle between the estimated direction of the light source and the vector direction connecting the eye specular highlight position to the light source position). Finally, the third and final stage (decision) calculates the average angular error and use a classical hypothesis test with an 1% significance level to decide whether or not a given image under investigation is a composite.

The authors tested their technique for estimating the 3-D light direction on synthetic images of eyes that were rendered using the `PBRT` environment and with a few real images. The come out with a decision for a given image, the authors determine whether the specular highlights in an image are inconsistent considering only the average angular error and a classical hypothesis test.

In a forensic scenario, only the average angular error for deciding about inconsistencies might be rather limiting. In possession of an image for investigation, we can explore other important information to use in conjunction with the average angular error. We found out that the location of the viewer (e.g., camera) is also important, as well as other characteristics described in Section 3. Therefore, in this paper we extend Johnson and Farid's approach [7] so as to consider more discriminative features and have more confidence when deciding about the authenticity of an image under investigation. In addition, instead of using a classical hypothesis test (like [7]), we propose to analyze the calculated features with a two-class supervised machine learning classification approach. With the new proposed features and the decision rule, our approach improves the prior work in more than 20%.

3. OUR METHOD

In this work, we extend the method proposed by Johnson and Farid in [7] by using more discriminative features in the problem characterization stage and a supervised machine learning classifier in the decision stage. In this section, we review Stages 1 and 2 of Johnson and Farid's method [7], then present the new proposed features to be used along with their features and a new decision stage for the method which is able to analyze all the composed features at once.

The first stage consists of estimating the direction of the light source for each eye in the picture. The authors assume that the eyes are perfect reflectors and use the law of reflection:

$$\boldsymbol{L} = 2(\boldsymbol{V}^{\boldsymbol{T}}\boldsymbol{N})\boldsymbol{N} - \boldsymbol{V}, \tag{1}$$

where the 3-D vectors $\boldsymbol{L}$, $\boldsymbol{N}$ and $\boldsymbol{V}$ correspond to the direction to the light, the surface normal at the highlight, and the direction in which the highlight will be seen. Therefore, the light direction $\boldsymbol{L}$ can be estimated from the surface normal $\boldsymbol{N}$ and viewer direction $\boldsymbol{V}$ at a specular highlight. However, it is difficult to estimate these

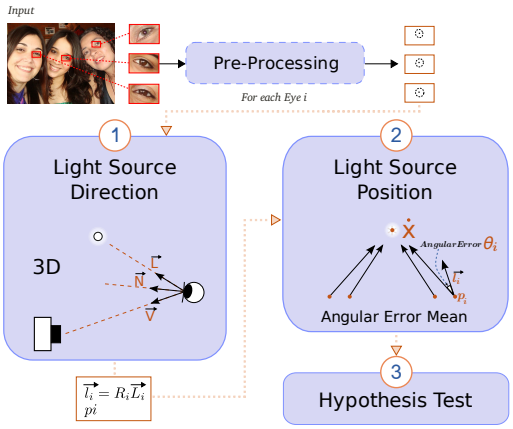

Fig. 1. Diagram depicting Johnson and Farid's three-stage approach [7].

two vectors in the 3-D space from a single 2-D image. In order to circumvent this difficulty it is possible to estimate a transformation matrix H that maps 3-D world coordinates to 2-D image coordinates by making some simplifying assumptions such as:

1. the limbus (the boundary between the sclera and iris) is modeled as a circle in the 3-D world system and as an ellipse in the 2-D image system;
2. the distortion of the ellipse with respect to the circle is related to the pose and position of the eye relative to the camera;
3. and points on a limbus are coplanar.

With these assumptions, H becomes a 3×3 planar projective transform, in which the world points $\boldsymbol{X}$ and image points $\boldsymbol{x}$ are represented by 2-D homogeneous vectors, $\boldsymbol{x} = H\boldsymbol{X}$. Then, the matrix H as well as the circle center point $C = (C_1\ C_2\ 1)^T$, and radius r (recall that C and r represent the limbus in world coordinates) are obtained by minimizing the error function:

$$E(\boldsymbol{P}; H) = \sum_{i=1}^{m} \min_{\hat{\boldsymbol{X}}} \|\boldsymbol{x}_i - H\hat{\boldsymbol{X}}_i\|^2, \tag{2}$$

where $\hat{\boldsymbol{X}}$ is on the circle parameterized by $\boldsymbol{P} = (C_1\ C_2\ r)^T$, and m is the total amount of data points in the image system. The matrix H is decomposed to obtain the matrix $\hat{H}$, representing the transformation of the world system in the camera system, and the matrix R representing the rotation between them.

The camera direction $\boldsymbol{V}$ and the surface normal $\boldsymbol{N}$ can then be calculated in the world system. $\boldsymbol{V}$ is $R^{-1}\boldsymbol{v}$, where $\boldsymbol{v}$ represents the direction of the camera in the camera system, and it can be calculated by $\boldsymbol{v} = -\boldsymbol{x}_c/\|\boldsymbol{x}_c\|$, where $\boldsymbol{x}_c$ is the center point of the limbus in the camera system obtained with $\boldsymbol{x}_c = \hat{H}C$. The surface normal $\boldsymbol{N}$ at a specular highlight is computed from a 3-D model of the human eye first proposed by [9]. Then, $\boldsymbol{N}$ is given by $\boldsymbol{N} = \boldsymbol{S} + \boldsymbol{V}$, where $\boldsymbol{S}$ represents the specular highlight in the world coordinate, measured with respect to the center of the limbus and to the human eye model. The first stage of the method in [7] is completed by calculating the light source direction $\boldsymbol{L}$ by replacing $\boldsymbol{V}$ and $\boldsymbol{N}$ in Eq 1. In order to compare light source estimates in the image system, the light source estimate is converted to camera coordinates: $\boldsymbol{l} = R\boldsymbol{L}$.

The second stage is based on the assumption that all estimated directions $\boldsymbol{l}_i$ converge toward the position of the light source, where

$i = 1, ..., n$ and n is the number of specular highlights in the picture. This position can be estimated by minimizing the error function

$$E(\boldsymbol{x}) = \sum_{i=1}^{n} \theta_i(\boldsymbol{x}), \tag{3}$$

where $\theta_i(\boldsymbol{x})$ represents the angle between the vector to the light source at position $\boldsymbol{x}$ and the estimated direction $\boldsymbol{l}_i$, at the i^{th} specular highlight $\boldsymbol{p}_i$. $\theta_i(\mathbf{x})$ is given by

$$\theta_i(\boldsymbol{x}) = \arccos\left(\boldsymbol{l}_i{}^T \frac{\boldsymbol{x} - \boldsymbol{p}_i}{||\boldsymbol{x} - \boldsymbol{p}_i||}\right). \tag{4}$$

Being $\dot{\boldsymbol{x}}$ the point representing the light source position obtained by Eq. 3, the angular error of the i^{th} specular highlight is $\theta_i(\dot{\boldsymbol{x}})$.

For an image that has undergone composition it is expected that the angular errors are higher than in pristine images. Based on this statement the authors apply a hypothesis test with the angular error average to identify whether or not the image under investigation contains is the result of a composition.

In this paper, we make the important observation that in the forensic scenario, beyond the angular error average, there are other important characteristics that must also be taken into account in the decision-making stage in order to further improve the quality of any eye-highlight-based detector.

Therefore, we first decide to take into account the standard deviation of angular errors (θ_i), given that even in the original images there is a non-null standard deviation. This is due the successive minimization of functions and simplification of the problem, adopted in the previous steps.

Another key feature is related to the position of the viewer (the device that captured the image). In a pristine image (one that is not a result of a composition) the camera position must be the same for all persons in the photograph, i.e., the estimated directions $\boldsymbol{v}$ must converge to a single camera position.

To find the camera position and take it into account, we minimize the following function

$$E(\boldsymbol{x}) = \sum_{i=1}^{n} \beta_i(\boldsymbol{x}), \tag{5}$$

where $\beta_i(\boldsymbol{x})$ represents the angle between the estimated direction of the camera $\boldsymbol{v}_i$ and direction of the vector pointing from the specular highlight $\boldsymbol{p}_i$ to the camera, calculated by

$$\beta_i(\boldsymbol{x}) = \arccos\left(\boldsymbol{v}_i{}^T \frac{\boldsymbol{x} - \boldsymbol{p}_i}{||\boldsymbol{x} - \boldsymbol{p}_i||}\right). \tag{6}$$

Considering $\ddot{\boldsymbol{x}}$ to be viewer position obtained by Eq. 5, the angular error of the i^{th} specular highlight is $\beta_i(\ddot{\boldsymbol{x}})$. In order to use this information in the decision-making stage, we can average all the available angular errors. In this case, it is also important to analyze the standard deviation of angular errors β_i.

Our extended approach now comprises four characteristics of the image instead of just one as the prior work we rely upon:

1. **(LME)** – mean of the angular errors θ_i, related to the light source $\boldsymbol{L}$;
2. **(LSE)** – standard deviation of the angular errors θ_i, related to the light source $\boldsymbol{L}$;
3. **(VME)** – mean of the ang. errors β_i, related to the viewer $\boldsymbol{V}$;
4. **(VSE)** – standard desviation of the angular errors β_i, related to the viewer $\boldsymbol{V}$.

In order to set forth the standards for a more general and easy to extend smart detector, instead of using a simple hypothesis testing in the decision stage, we turn to a supervised machine learning scenario in which we feed a Support Vector Machine classifier (SVM) or a combination of those with the calculated features.

4. EXPERIMENTS

Although the method proposed by Johnson and Farid in [7] has a great potential, the authors have validated their approach using mainly PBRT synthetic images which is rather limiting. In contrast, in this paper we perform our experiments using a data set comprising everyday still pictures typically with more than three megapixels in resolution. We acquired 120 images in which 60 images are normal (without any tampering or processing) and the other 60 images are composed by manipulated images (tampered with using splicing of people). The images always depict two or more people (Fig. 2).

(a) Pristine (No manipulation) (b) Fake

Fig. 2. Examples of data set images.

The experiment pipeline begins with the limbus point extraction for every person in every image. The limbus point extraction can be performed using a manual marker around the iris, or with an automatic method such as [10]. As this is not our primary focus in this paper, we used manual markers. Afterwards, we characterize the images, considering the features described in Section 3. Since one feature vector is extracted for one single image, we obtain 120 features vectors, that together compound a single data set. As some features in our proposed method rely upon non-linear minimization methods, which are initialized with random seeds, we can extract features using different seeds with no additional mathematical effort. Thus, we extract five feature vectors for each image. As a result, we obtain five data sets.

We then feed two-class classifiers with these features in order to achieve a final outcome. For this task, we use an out-of-the-box SVM with a standard RBF kernel. For a fair comparison, we perform 5-fold cross validation in all the experiments. We present results for a single classifier (a), and we also present results for a pool of five classifiers (one for each data set), analyzing an image in conjunction in a classifier-fusion fashion approach (b, c, d):

a. **Single Classifier (SC):** a single classifier fed with the proposed features to predict the class (pristine or fake).

b. **Classifier Fusion with Majority Voting (MV):** a new sample is classified by a pool of five classifiers. Each classifier casts for a class vote in a winner-takes-all approach.

c. **Classifier Fusion with OR Rule (One Pristine):** similar to MV except that the decision rule decides for non-fake if at least one classifier casts a vote in this direction.

d. **Classifier Fusion with OR Rule (One Fake):** similar to MV except that the decision rule decides for fake if at least one classifier casts a vote in this direction.

To show the behavior of each round compared with Johnson and Farid's approach we used an ROC curve, in which the y-axis (Sensitivity) represents the fake images correctly classified as fakes and the x-axis (1 - Specificity) represents pristine images incorrectly classified. Figure 3 shows the results for our proposed approach (with four different classifier decision rules) compared to the results of Johnson and Farid's approach. All the proposed classification decision-rules

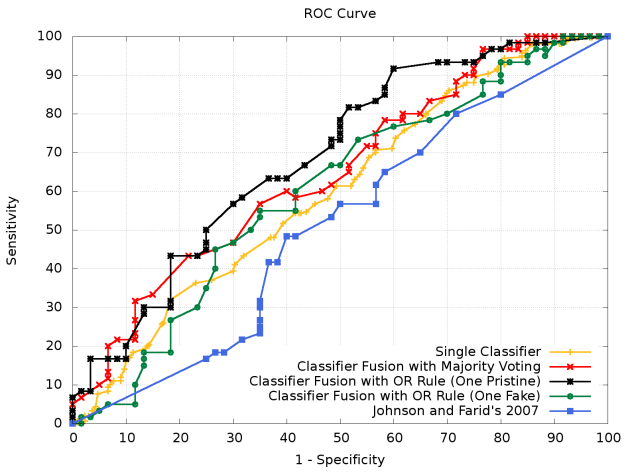

Fig. 3. Classification results for Johnson and Farid's method [7] compared to the methods we introduce in this paper.

perform better than the prior work we rely upon in this paper. This allows us to come up with two conclusions: first, the new proposed features indeed make difference and contribute for the final classification decision; and second, different classifiers can take advantage of different seeds used in the calculation of the features. Note that with 40% specificity, we detect 92% of fakes correctly while the prior work Johnson and Farid's prior work achieves $\cong 64\%$.

Another way to compare our approach to Johnson and Farid's one is to assess the classification behavior on the "Equal Error Rate" (EER) point. Table 1 shows this comparison.

The best proposed method (shaded cell) – *Classifier Fusion with OR Rule (One Pristine)* decreases the classification error in 21% when compared to Johnson and Farid's approach at the EER point. Even if we consider just a single classifier (no fusion at all), the proposed extension performs 7% better than the prior work we rely upon considering the ERR point.

5. CONCLUSIONS AND FUTURE WORK

Johnson and Farid's method [7] as well as our proposed extension to their work have a great potential for detecting composite photographs of people as long as there are visible eyes in the image under investigation.

In this paper, we extended Johnson and Farid's prior [7] in such a way we now derive more discriminative features for detecting traces of tampering in composite photographs of people and use the calculated features with powerful decision-making classifiers based on simple, yet powerful, combinations of the Support Vector Machines. The new features and the new decision-making process reduced the classification error in more than 20% (or in absolute value, 11%) when compared to the prior work. To validate our ideas, we have used a data set of real composite photographs

Table 1. Equal Error Rate – Four proposed approaches and the baseline [7]. Percentage (%) in terms of relative values.

	EER (%)	Imprv. over baseline (%)
Single Classifier	44	7
Fusion MV	40	15
Fusion One Pristine	**37**	**21**
Fusion One Fake	41	13
Johnson and Farid's	48	–

of people and images typically with more than three mega-pixels in resolution. We intend to make this data set open to the community (http://www.ic.unicamp.br/∼rocha/pub/communications.html).

It is worth noting, however, the classification results are still affected by some drawbacks in which we are now striving to circumvent. First of all, the accuracy of light direction estimation relies heavily on the camera calibration step. If the eyes are occluded by eyelids or are too small, the limbus selection becomes too difficult to accomplish, demanding an experienced user. Second, the focal length estimation method is often affected by numerical instabilities due to the starting conditions of the minimization function suggested in [7]. Some improvements would lead us to a better light direction estimation and higher success rate. Finally, we intend to take advantage of the focus information as a possible characteristic, since in an image without forgeries, the focus value must be very similar across different people in a scene. In our future work, we intend to evaluate the effectiveness of these features individually.

6. REFERENCES

[1] A. Rocha, W. Scheirer, T. E. Boult, and S. Goldenstein, "Vision of the unseen: Current trends and challenges in digital image and video forensics," *ACM Computing Surveys (CSUR)*, , no. To Appear, 2011.

[2] J. He, Z. Lin, L. Wang, and X. Tang, "Detecting doctored jpeg images via DCT coefficient analysis," in *ECCV (3)*, 2006, pp. 423–435.

[3] A.C. Popescu and H. Farid, "Exposing digital forgeries by detecting traces of re-sampling," *IEEE TSP*, vol. 53, no. 2, pp. 758–767, 2005.

[4] Z. Lint, R. Wang, X. Tang, and H-Y. Shum, "Detecting doctored images using camera response normality and consistency," in *IEEE CVPR*, 2005, pp. 1087–1092.

[5] Yu-Feng Hsu and Shih-Fu Chang, "Image splicing detection using camera response function consistency and automatic segmentation," in *IEEE ICME*, 2007, pp. 28–31.

[6] M.K. Johnson and H. Farid, "Exposing digital forgeries by detecting inconsistencies in lighting," in *ACM MM&SEC*, 2005.

[7] M.K. Johnson and H. Farid, "Exposing digital forgeries through specular highlights on the eye," in *IHW*, 2007.

[8] M.K. Johnson and H. Farid, "Exposing digital forgeries in complex lighting environments," *IEEE TIFS*, vol. 3, no. 2, pp. 450–461, 2007.

[9] M. Hogan, J. Alvarado, and J. Weddell, *Histology of the Human Eye*, W.B. Saunders Company, 1971.

[10] Z. He, T. Tan, Z. Sun, and X. Qiu, "Toward accurate and fast iris segmentation for iris biometrics," *IEEE TPAMI*, vol. 31, pp. 1670–1684, September 2009.