DETECTING IMAGE SPLICING USING GEOMETRY INVARIANTS AND CAMERA CHARACTERISTICS CONSISTENCY

Yu-Feng Hsu and Shih-Fu Chang

Department of Electrical Engineering
Columbia University
{yfhsu,sfchang}@ee.columbia.edu

ABSTRACT

Recent advances in computer technology have made digital image tampering more and more common. In this paper, we propose an authentic vs. spliced image classification method making use of geometry invariants in a semi-automatic manner. For a given image, we identify suspicious splicing areas, compute the geometry invariants from the pixels within each region, and then estimate the camera response function (CRF) from these geometry invariants. The cross-fitting errors are fed into a statistical classifier. Experiments show a very promising accuracy, 87%, over a large data set of 363 natural and spliced images. To the best of our knowledge, this is the first work detecting image splicing by verifying camera characteristic consistency from a single-channel image.

1. INTRODUCTION

As computer technology advances, tampered photos become more and more popular, which can cause substantial social impact. Two famous examples include a 1994 TIME magazine cover image of O.J. Simpson's face whose skin was deliberately darkened and a photomontage on L.A. Times front page in 2003 showing a spliced soldier pointing his gun at a group of Iraqi people. The saying "seeing is believing" is no longer true in this digital world, and one would naturally ask whether the photo he/she receives is a real one.

In the past, people used active approaches to tackle digital image tampering. This was typically done by embedding watermarks in an unperceptible way. At the point of verification, the image is fed into an authentication engine. If the embedded watermark is successfully extracted, then the image is claimed authentic, otherwise, tampered.

However, in practice, very few images are created with watermarks. Under most circumstances active approaches fail because there is no watermark to detect. This gives rise to research activities in passive blind image authentication that handle images with no prior added hidden information.

Defining an authentic image itself is a challenging task. One needs to carefully draw the line between common image operations (eg. compression) and malicious attacks (eg. copy-and-paste of human figures in order to alter image semantics). There are two common properties that an authentic image must bear: natural scene quality and natural imaging quality [1]. The former relates to the consistency in lighting and reflection patterns of the scene, while the latter indicates an authentic image must be one that went through some image acquisition device. Therefore an image with inconsistency between the light direction and the shadows is not authentic because it fails to satisfy the natural scene quality, and an image generated by photomontage fails to meet the natural imaging quality since different parts of the image do not share consistent characteristics of imaging devices.

2. PREVIOUS WORK

To determine if an image is authentic or tampered, one can analyze the inconsistency within the image, eg. lighting or image source. Farid *et al* have developed techniques for spliced image detection by object lighting inconsistency [2]. Lin *et al* proposed a colinearity based spliced image detection method by observing the abnormality in the camera response functions (CRF) [3]. Memon *et al* used cross color channel features to detect whether two images came from the same camera [4]. Lukáš *et al* used pattern noise correlation to identify the camera source of an image [5].

Another image tampering approach is based on the statistical point of view. Such methods extracted visual features from natural images and attempted to model their statistical properties, which were then used to distinguish spliced from natural images. Ng *et al* applied bicoherence along with other features to detect spliced images [6]. Farid *et al* used wavelet features to classify natural images and computer graphics [7]. Ng *et al* also looked at a similar problem using geometric features motivated by physical models of natural images [1].

3. PROPOSED APPROACH

We propose an approach to distinguish authentic images from spliced ones. Although there can be countless possible definitions for 'authentic images', we restrict ourselves to those taken by the same camera. Our approach is mainly motivated by the intuition that authentic images must come from the same camera - inconsistency in camera characteristics among

different image regions naturally leads to detection of suspicious cases.

Our method is semi-automatic at this stage. While a user inspects an image, he/she may raise suspicion that some areas are produced by tampering. In such a case, he/she may label the image into three distinct regions: region from camera 1, region from camera 2, and the interface region between them.

We estimate the CRF from each region using geometry invariants and check if all these CRF's are consistent with each other using cross-fitting techniques. Intuitively, if the data from a certain region fits well to the CRF from another region, this image is likely to be authentic. Otherwise, if it fits poorly, then the image is very likely to be spliced. Finally, the fitting statistics are fed to a learning-based method (support vector machine) to classify the images as authentic or spliced.

3.1. Camera Response Function

CRF is one of the most widely used camera characteristics. It transforms irradiance values received by CCD sensors into brightness values that are output to film or digital memory. As different cameras have different response functions, CRF can serve as a natural feature that identifies the camera model.

In this paper, we will use the following convention:

$$R = f(r) \tag{1}$$

to denote the irradiance (r), the brightness (R), and the CRF (f). f can be as simple as a gamma transform:

$$R = f(r) = r^{\alpha} \tag{2}$$

Or a more general form, called the linear exponent model [8]:

$$R = f(r) = r^{\alpha+\beta r} \tag{3}$$

We will use the linear exponent model for CRF's in this paper.

3.2. Geometry Invariants

Geometry invariants are used in [8] to estimate the CRF from a single-channel image. Taking the first partial derivative of Eq. (1) gives us $R_x = f'(r)r_x,\ \ R_y = f'(r)r_y$. Taking the second derivative, we get

$$\begin{aligned} R_{xx} &= f''(r)r_x^2 + f'(r)r_{xx} \\ R_{xy} &= f''(r)r_x r_y + f'(r)r_{xy} \\ R_{yy} &= f''(r)r_y^2 + f'(r)r_{yy} \end{aligned} \tag{4}$$

All subscripts denote the derivative in that corresponding direction. Now suppose the irradiance is locally planar, i.e., $r = ax + by + c$, we have $r_{xx} = r_{xy} = r_{yy} = 0$. Then

$$\frac{R_{xx}}{R_x^2} = \frac{R_{xy}}{R_x r_y} = \frac{R_{yy}}{R_y^2} = \frac{f''(r)}{(f'(r))^2} = \frac{f''(f^{-1}(R))}{(f'(f^{-1}(R)))^2} \tag{5}$$

This quantity, denoted as $A(R)$, can be shown to be independent of the geometry of r. As shown in [8], if CRF is a gamma transform as in Eq. (2), then $A(R)$ is related to the gamma parameter as follows

$$A(R) = (\frac{\alpha - 1}{\alpha})R^{-1} \tag{6}$$

Define

$$Q(R) = \frac{1}{1 - A(R)R} \tag{7}$$

Then for gamma transform, $Q(R) = \alpha$, which carries the degree of nonlinearity of the CRF. When the CRF takes the linear exponent form, $Q(R)$ becomes

$$Q(R) = \frac{(\beta r \ln(r) + \beta r + \alpha)^2}{\alpha - \beta r} \tag{8}$$

From a region, we extract the points that satisfy the locally planar assumption, and compute their $Q(R)$'s. We then estimate the CRF in an iterated manner [8]. Each CRF is represented by its linear exponent parameters (α, β).

3.3. Cross-fitting

The idea of cross-fitting came from speaker transition detection [9], where several models are trained from different speakers and used to fit a certain segment to determine if it is a transition. Here we use the same framework with $Q(R)$ as the model representing a camera.

We divide the image into three regions: region 1 potentially from camera 1, region 2 potentially from camera 2, and region 3 near the suspicious splicing boundary.

For each region, we extract the points satisfying the locally planar assumption, compute their $Q(R)$'s, and estimate the CRF. With three regions, we get four sets of points and parameters:$\{R_k, Q_k(R)\}$ and (α_k, β_k) where $k \in \{0, 1, 2, 3\}$. $\{R_1, Q_1(R)\}$, $\{R_2, Q_2(R)\}$, and $\{R_3, Q_3(R)\}$ are points from regions 1, 2, and 3, respectively. $\{R_0, Q_0(R)\}$ is the combined set from the entire image. If regions 1 and 2 are indeed from different cameras, then $\{R_1, Q_1(R)\}$ should fit poorly to (α_2, β_2), and vice versa. Also, if region 3 really contains the splicing boundary, then either its parameters (α_3, β_3) will exhibit abnormality or $\{R_3, Q_3(R)\}$ will fit in a strange manner to (α_3, β_3), same with (α_0, β_0) and $\{R_0, Q_0(R)\}$. Taking into account all of these considerations, we use a six dimensional feature vector to represent an image:

$$[s_{11}, s_{22}, s_{12}, s_{21}, s_3, s_0] \tag{9}$$

where s_{ij} $(i, j \in \{1, 2\})$ is the fitting score of $\{R_i, Q_i(R)\}$ to CRF (α_j, β_j), given by the root mean square error (RMSE)

$$s_{ij} = \sqrt{\frac{1}{N_i}\sum_{n=1}^{N_i}[Q_i(R)_n - \frac{(\beta_j r_n \ln(r_n) + \beta_j r_n + \alpha_j)^2}{\alpha_j - \beta_j r_n}]^2} \tag{10}$$

N_i is the total number of extracted points in region i.

Similarly, s_k, $k \in \{0, 3\}$ can be computed as

$$s_k = \sqrt{\frac{1}{N_k}\sum_{n=1}^{N_k}[Q_k(R)_n - \frac{(\beta_k r_n \ln(r_n) + \beta_k r_n + \alpha_k)^2}{\alpha_k - \beta_k r_n}]^2} \tag{11}$$

3.4. SVM Classification

The six dimensional feature vectors in Eq. (9) are then fed into SVM to classify authentic and spliced images. Both linear and RBF kernel were experimented, along with a five-fold cross validation in search of the best parameters.

4. EXPERIMENT

4.1. Data Set

There are a total of 363 images in our dataset. 183 of them are authentic images, and 180 are spliced ones. The authentic images are taken with our four digital cameras: Canon G3, Nikon D70, Canon EOS 350D Rebel XT, and Kodak DCS330. The images are all in uncompressed RAW or BMP formats with dimensions ranging from 757x568 to 1152x768. These images mainly contain indoor scenes, eg. desks, computers, or corridors. About 27 images, or a percentage of 15% are taken outdoors on a cloudy day.

We created the spliced images from the authentic image set using Adobe Photoshop. In order to focus only on the effects of splicing, no post processing was performed. Each spliced image contains contents from exactly two cameras. To even out the contribution from each camera, we assign an equal number of images for each camera pair. With four cameras, we have a total of six possible camera pairs, so we create 30 images per pair.

As shown in Fig. 1, each image is manually labelled into four regions: region from camera 1 far from splicing boundary (dark red, equivalent to region 1 in Sec. 3.3), region from camera 1 near splicing boundary (bright red), region from camera 2 far from splicing boundary (dark green, equivalent to region 2 in Sec. 3.3), region from camera 2 near splicing boundary (bright green). Regions labelled with bright red and bright green are then combined as the 'spliced region' (region 3 in Sec. 3.3).

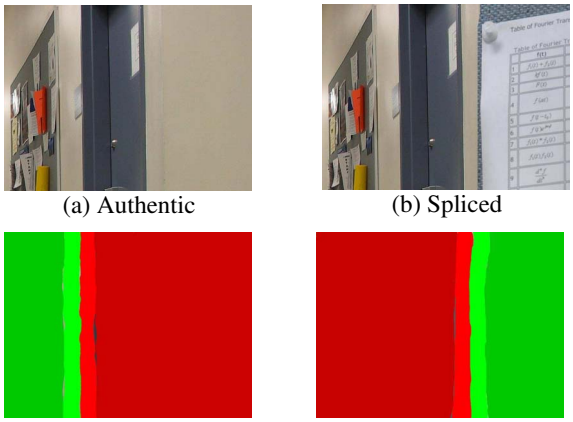
With authentic images, we choose a visually significant area in the image and treat it as the candidate region to be verified. (eg. in Fig. 1(a) the region to the left of the door is treated as the spliced region).

4.2. SVM Classification

We use 11 penalty factors C and 10 Radial Basis Function (RBF) widths γ and use cross validation to find the best set of parameters among them. For each set of (C, γ), we divide the training set into a training subset and a validation subset. We train an SVM on the training subset, test it on the validation subset, and record the testing accuracy. The cross validation is repeated five times for each (C, γ) and the performance is measured by the average accuracy across the five runs. At the end, we choose the (C, γ) with the highest average accuracy and test the classifier on our test set.

5. RESULTS

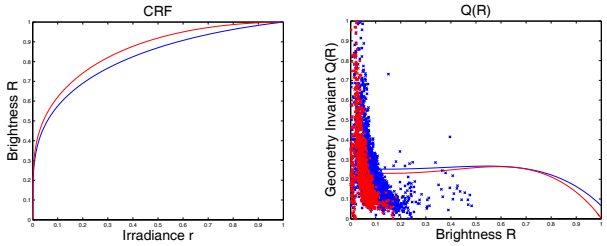
We performed six runs of both linear and RBF kernel SVM with cross validation searching for the best parameters and get average classification rates of 66.54% and 86.42%, respectively. The standard deviations among these six runs are 1.65% and 0.71%, showing that the performance of each SVM is rather insensitive to different runs. The highest RBF kernel SVM classification accuracy is 87.55%, with the spliced image detection rate as high as 90.74%. The confusion matrix of RBF kernel SVM with the highest accuracy is shown in Table 1.



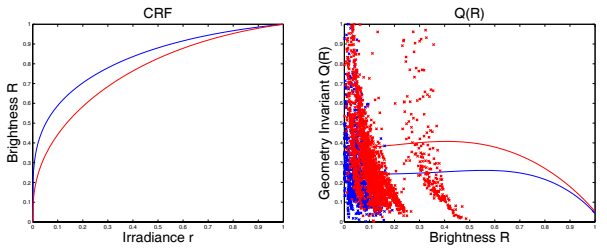
(a) Authentic (b) Spliced

(c) Region labelling of (a) (d) Region labelling of (b)

Fig. 1. Examples of images in our dataset



(a) Two CRF's, authentic (b) Q-R scatterplot, authentic



(c) Two CRF's, spliced (d) Q-R scatterplot, spliced

Fig. 2. CRF's and Q-R's from authentic and spliced images

Fig. 2 shows the estimated CRF's and the fitted Q-R's from an authentic image and a spliced image. In Fig. 2(a)(c), CRF's from two cameras are plotted with different colors. It can be seen that within an authentic image the two CRF's are closer to each other than those within a spliced image. This is consistent with our intuition since authentic images should have its regions coming from the same camera, hence predicting more similar CRF's.

Fig. 2(b)(d) show the scatterplots of $\{R, Q(R)\}$'s extracted from regions 1 (blue) and 2 (red). The population of the $\{R, Q(R)\}$ pool is typically around 2000. A Q-R curve is fitted to each pool of $\{R, Q(R)\}$ to obtain (α, β) in an iterative manner as in [8]. With (α, β) we can construct the CRF, so the irradiance values r's can be related to R's through the CRF. And by using Eq. (8), we can plot the fitted relationship between Q(R) and R, shown in blue/red curves in Fig.

Table 1. Confusion Matrix of RBF Kernal SVM

		Detected As	
		Authentic	Spliced
Actual Category	Authentic	84.42%	15.58%
	Spliced	9.26%	90.74%

2(b)(d).

Both the CRF's and Q-R relationships within an authentic image are indeed more similar to each other than those within a spliced image. Nevertheless, comparing Fig. 2(c) and Fig. 2(d), it is clear that the Q-R curve is more differentiating than the CRF, which justifies the use of $Q(R)$ rather than the CRF itself in cross-fitting.

One would question if $Q(R)$ is an appropriate model. To answer this question we need to look at how this quantity distinguishes cameras and whether it carries physical intuition. Starting from the CRF, there are several possible choices to represent a camera: CRF, $Q(R)$, and $A(R)$. As shown in Fig. 2(a)(c), the CRF does not reflect very well the differences between two cameras, therefore it is not a good model. $Q(R)$ is a good choice since it distinguishes cameras better than the CRF itself, as shown in Fig. 2(b)(d). Lastly, $A(R)$ is also a natural choice. In fact it is perfectly sensible to use $A(R)$, except that an additional dependency on R will be introduced if we plug in Eq. (8) into Eq. (7), which might make the mathematical form of $A(R)$ intractable. Therefore we stay with $Q(R)$ and use it as our cross-fitting model.

$Q(R)$ is also physically meaningful since it is exactly equal to the gamma parameter of the CRF. Note in Eq. (8) if β is zero, then $Q(R)$ reduces to α as in Eq. (7). Therefore, $Q(R)$ is not only physically related to the CRF, but also brings extra advantage when it comes to distinguishing cameras.

6. DISCUSSION

Manual labelling of image regions makes our approach a semi-automatic one. This is not entirely impractical though. One possible scenario would be a publishing agency that handles celebrity photographs. They usually have specific suspicious regions: the contour around human figures. The labelling rule becomes quite clear: label the pixels near the boundary of human figures as 'spliced region'. In fact, [3] uses a similar semi-automatic scheme which allows users to choose suspicious patches to detect CRF abnormality.

If the image has ambiguous splicing boundaries or if the number of candidate images gets large, the manual labelling scheme would become infeasible. In such cases, incorporating image segmentation techniques could be an aide to the current semi-automatic scenario.

Our work is the first of spliced image detection using inconsistency in natural imaging quality. The proposed detection method based on CRF consistency and model cross-fitting is general. The CRF estimation operates on a single image and does not require any calibration. Furthermore, it can be applied to any single-channel, i.e., greyscale image, rather than restricted to multi-channel images.

We also provided a new authentic vs. spliced image dataset, which can serve as a benchmark dataset for spliced images.

Computationally, it takes about 11 minutes to construct a feature vector from one image, including extracting qualifying points, computing $Q(R)$'s, estimating (α, β), and cross-fitting. These operations are all done on a 2.40GHz dual processor PC with 2GB RAM. The time consuming part, however, is SVM training with cross validation. It took us a total of five hours to obtain the best SVM parameters. As time consuming as it is, the SVM training can always be done offline. SVM classification on test data is done in real time.

7. CONCLUSION

We proposed a spliced image detection method in this paper. The detection was based on geometry invariants that relate directly to the CRF, fundamental characteristics of cameras. We used cross-fitting to determine if an image consists of regions from more than one camera. RBF kernel SVM classification results showed that this approach is indeed effective in detecting spliced images. We also discussed the issues of fitting models and the application of our semi-automatic scheme.

8. ACKNOWLEDGEMENT

This work has been supported by NSF Cyber Trust grant IIS-04-30258. The authors also thank Tian-Tsong Ng for sharing the codes and providing helpful discussions.

9. REFERENCES

[1] T.-T. Ng, S.-F. Chang, J. Hsu, L. Xie, and M.-P. Tsui, "Physics-motivated features for distinguishing photographic images and computer graphics," in *ACM Multimedia*, 2005.

[2] M.K. Johnson and H. Farid, "Exposing digital forgeries by detecting inconsistencies in lighting," in *ACM Multimedia and Security Workshop*, 2005.

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

[4] M. Kharrazi, H. T. Sencar, and N. D. Memon, "Blind source camera identification.," in *ICIP*, 2004, pp. 709–712.

[5] J. Lukáš, J. Fridrich, and M. Goljan, "Determining digital image origin using sensor imperfections," in *SPIE*, 2005, vol. 5685, pp. 249–260.

[6] T.-T. Ng, S.-F. Chang, and Q. Sun, "Blind detection of photomontage using higher order statistics," in *ISCAS*, 2004.

[7] H. Farid and S. Lyu, "Higher-order wavelet statistics and their application to digital forensics," in *IEEE Workshop on Statistical Analysis in Computer Vision*, 2003.

[8] T.-T. Ng and S.-F. Chang, "Camera response function estimation from a single grayscale image using differential invariants," Tech. Rep., ADVENT, Columbia University, 2006.

[9] S. Renals and D. Ellis, "Audio information access from meeting rooms," in *ICASSP*, 2003.