DETECTING PHOTO MANIPULATION ON SIGNS AND BILLBOARDS

Valentina Conotter, Giulia Boato *

Department of Information Engineering
and Computer Science
University of Trento, Trento (ITALY)

Hany Farid †

Dartmouth College
Department of Computer Science
Hanover NH 03755 (USA)

ABSTRACT

The manipulation of text on a sign or billboard is relatively easy to do in a way that is perceptually convincing. When text is on a planar surface and imaged under perspective projection, the text undergoes a specific distortion. When text is manipulated, it is unlikely to precisely satisfy this geometric mapping. We describe a technique for detecting if text in an image obeys the expected perspective projection, deviations from which are used as evidence of tampering.

Index Terms— Digital Forensics, Digital Tampering

1. INTRODUCTION

In 2004 a controversial photo of a U.S. Marine posing with two Iraqi children while purportedly holding an inappropriate sign was widely circulated on the Internet. The Marine claimed that the image was manipulated, and that the sign originally read "Welcome Marines". The photo created a significant enough controversy that a military inquiry was launched. The investigation, however, was inconclusive and the authenticity of the image was never determined.

The adding or changing of text in an image is relatively easy to do in a way that is perceptually convincing (see examples in Fig. 1). When inserting text into an image, however, it is likely that the precise rules of perspective projection will be violated, and that these violations will not be perceptually obvious [1].

In this paper, we describe a new forensic technique for determining if typed text on a sign or billboard obeys the rules of perspective projection. This method explicitly identifies the projection of text on a planar surface and detects deviations from this model. We consider the case when the font style of the text in question is known and when it is unknown. In the context of geometric-based forensic techniques [3, 4, 5], this is a first approach dealing with detecting manipulated text.

*This work was supported by the European Union within the Seventh Framework project LivingKnowledge (IST-FP7-231126).

†This work was supported by a gift from Adobe Systems, Inc., a gift from Microsoft, Inc. and a grant from the National Science Foundation (CNS-0708209).

Fig. 1. Doctored photos created by manually distorting text onto a planar surface.

2. METHODS

2.1. Planar Homography

The perspective 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 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 geometric transform where all of the world points $\vec{X}$ lie on a single plane and P reduces to a 3×3 planar projective transform H, 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.

We briefly review the estimation of the planar homography H, Eq. (1), from known world and image coordinates [6]. Reformulating Eq. (1) as a cross product yields:

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \times \left[\begin{pmatrix} h_1 & h_2 & h_3 \\ h_4 & h_5 & h_6 \\ h_7 & h_8 & h_9 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \right] = 0. \quad (2)$$

Evaluation of the cross product yields:

$$\begin{pmatrix} x_2(h_7X_1 + h_8X_2 + h_9X_3) - x_3(h_4X_1 + h_5X_2 + h_6X_3) \\ x_3(h_1X_1 + h_2X_2 + h_3X_3) - x_1(h_7X_1 + h_8X_2 + h_9X_3) \\ x_1(h_4X_1 + h_5X_2 + h_6X_3) - x_2(h_1X_1 + h_2X_2 + h_3X_3) \end{pmatrix} = \vec{0}.$$

This constraint is linear in the unknown elements of the homography h_i and may be rewritten as:

$$A\vec{h} = \vec{0}. \quad (3)$$

where A is a 3×9 matrix and $\vec{h}$ is a 9-vector containing the entries of the matrix H. A matched set of points $\vec{x}$ and $\vec{X}$ appear to provide three constraints on the eight unknown elements of $\vec{h}$ (the homography is defined only up to an unknown scale factor, reducing the unknowns from nine to eight). The rows of the matrix A, however, are not linearly independent. As such, this system provides two constraints in eight unknowns. In order to solve for $\vec{h}$, we require four or more points with known image, $\vec{x}$, and (planar) world, $\vec{X}$, coordinates. From four or more points[1], standard least-squares techniques can be applied, as described in [6].

In our case, the required world coordinates are determined by re-creating the text in question with no distortion. We next consider the case when the font style is known and when the font style is unknown. Since the homography is only estimated up to an unknown scale factor, the font size is arbitrary.

2.2. Known Font

Shown in Fig. 2(a) is the text string "ABC" after distortion by a planar homography, as in Eq. (1). Assuming that the font style is known, this string in its world coordinate system can easily be determined, as depicted in Fig. 2(b). From this pair of images, we automatically extract the image and world coordinates required for the planar homography estimation, as described next.

We employ the SIFT operator [7] to extract the coordinates of distinctive image keypoint positions. These keypoints are invariant to certain amounts of image scale, rotation, affine distortion, noise, and illumination differences. Shown in Fig. 2(c) and (d), for example, are a subset of the extracted keypoints (dots) for the images shown in panels (a) and (b).

[1]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 [6].

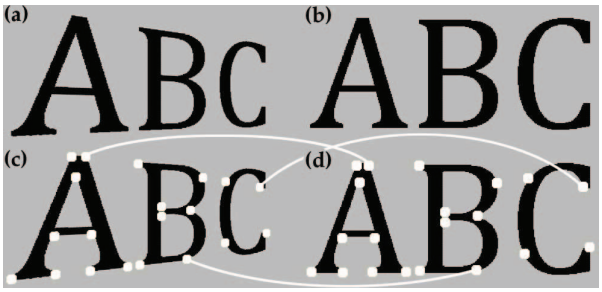

Fig. 2. Shown in panels (a) and (c) is a text string in image coordinates, and shown in panels (b) and (d) is the same string in world coordinates. The dots in panels (c) and (d) correspond to a subset of the extracted coordinates used to estimate the image to world homography.

A feature vector consisting of local gradients is measured at each keypoint position. Keypoints are matched between two images using a variant of nearest neighbor matching on the feature vectors. This association accounts for a geometric transformation between the images by matching keypoints up to a planar homography. The RANSAC algorithm [8] is used to minimize the effect of mis-matched keypoints.

2.3. Unknown Font

When the font of the text in question cannot be easily determined by visual inspection, we adopt the following technique for automatically identifying the font style. We begin by constructing the text in question in undistorted world coordinates with all available font styles. Then, the SIFT operator is applied to each of these images, as described in Section 2.2. The font style that returns the largest number of matched keypoints is taken to be the correct font.

2.4. Photo Composite

Given an image of text that has undergone planar perspective projection (i.e., a homography), we have described how to determine the required image and world coordinates, and how to estimate the world to image homography. Except for degenerate cases, it is always possible to calculate a homography regardless of the authenticity of the underlying text. We will show, however, that when the text is inconsistent with a perspective planar projection, the estimated homography yields a large reconstruction error. Specifically, the inverse homography is applied to the keypoints in image coordinates yielding rectified world coordinates:

$$\vec{X}_r = H^{-1}\vec{x} \quad (4)$$

It is unlikely in an inauthentic image to have the image coordinates precisely satisfy the proper planar perspective distortion. In this case, the rectified image $I_r(x, y)$ is unlikely to

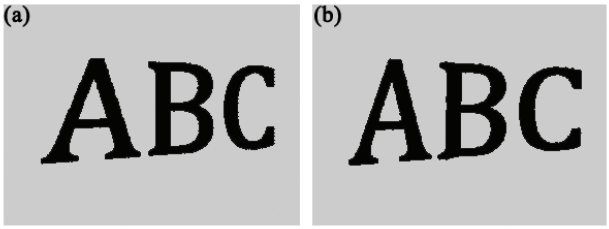

Fig. 3. Shown are (a) authentically projected text and (b) matched inauthentic text generated by isotropically scaling and affine transforming text to best match panel (a).

match the world image $I_w(x, y)$. On the other hand, in the case of an authentic image, the rectified image should be a good approximation of the world image. As such, we use the root mean square (RMS) error between the world and rectified image as a measure of authenticity:

$$e = \frac{1}{\sqrt{n_x n_y}} \|I_w - I_r\|, \quad (5)$$

where n_x and n_y are the image dimensions and $\|\cdot\|$ denotes vector 2-norm. Note that this error is computed on the underlying intensity image, as opposed to the extracted keypoint coordinates.

Because a homography captures a broad range of distortions, we have found it more effective to estimate the homography from several small subsets of the matched keypoints and then compute the RMS error for each estimated homography. The average RMS error is used as a measure of authenticity:

$$E = \frac{1}{N} \sum_{i=1}^{N} e_i, \quad (6)$$

where N is the total number of subsets and e_i is the RMS error, Eq. (5), for the i^{th} subset. Specifically, an error E above a specified threshold is taken to be evidence of tampering.

3. RESULTS

We describe a set of simulations to verify the efficacy of the proposed technique. A set of authentic images were first created by generating images consisting of a text string with six letters in one of 350 font styles. A planar homography, Eq. (1), was then applied by considering physically plausible intrinsic and extrinsic camera parameters. A matched inauthentic image was generated that approximated the appearance of the authentic image, while not precisely satisfying a planar homography, Fig. 3. Specifically, the image in world coordinates was subjected to an anisotropic scaling followed by a six parameter affine transformation constructed to optimally match (in the least-square sense) the authentic image. The result was a perceptually convincing transformation. We

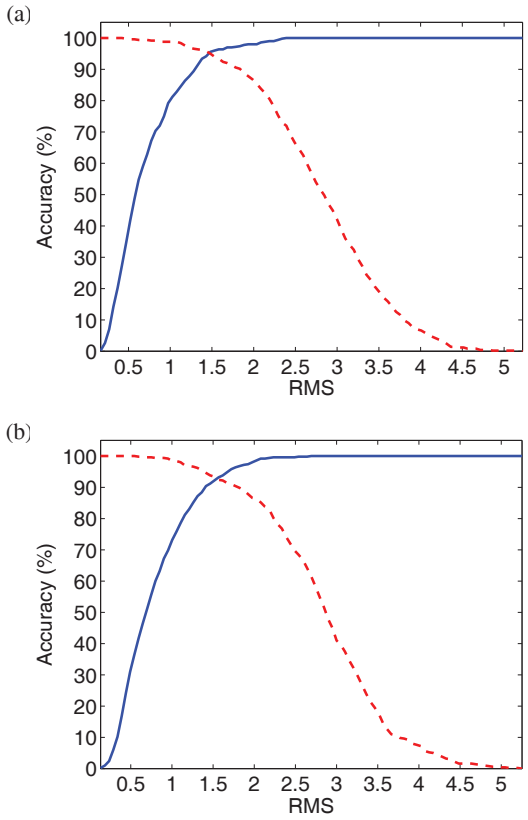

Fig. 4. Shown are ROC curves for classification with (a) known font style and (b) unknown font style. The solid curve corresponds to the authentic images, and the dashed curve corresponds to the inauthentic images.

generated 500 such authentic and inauthentic images. Each image was 1200×900 pixels in size, and rendered as a 1-bit binary image. For each image, we assumed a known font style, automatically extracted the image and world coordinates, estimated the world to image homography, and computed the reconstruction error with $N = 100$, Eq. (6).

Shown in Fig. 4(a) is the resulting ROC curve where the horizontal axis corresponds to the RMS error, and the vertical axis to the classification accuracy. The solid curve corresponds to the authentic images, and the dashed curve corresponds to the inauthentic images. The intersection of these curves corresponds to an overall accuracy of 95%. The detection accuracy and false alarm rate (incorrectly classifying an authentic image as inauthentic) can be controlled by adjusting the RMS threshold. For example, a false alarm rate of 2% yields a detection accuracy of 88% and a false alarm rate of 1% yields a detection accuracy of 82%.

Shown in Fig. 4(b) is the ROC curve for the case when the font style is unknown. In this case, the intersection of the curves corresponds to an overall accuracy of 92%. Note that the overall accuracy is similar, with a slight degradation due to some errors in the font identification stage.

Shown in Fig. 5(a) are two authentic images, and shown in panel (b) are two corresponding visually plausible fakes. For the first image, the reconstruction error, Eq. (6), for the manually extracted strings "Washington" and "Boston" were $E = 0.94$ and $E = 1.9$, respectively. For the second image, the reconstruction error for the strings "Amore" and "Hong Kong" were $E = 1.23$ and $E = 1.94$. These images are correctly classified with a threshold of 1.5 (i.e., an overall accuracy of 92%, the intersection of the ROC curves in Fig. 4(b)). To further validate this method, we tested a total of ten authentic images, obtaining reconstruction errors in the range of 0.93 to 1.35, with a median of 1.05, each of which is below the threshold of 1.5.

4. DISCUSSION

We have presented a new forensic technique for authenticating text in photographs. Because it is relatively easy to digitally insert text into a photo in a visually compelling manner, it can be difficult to manually determine if text is authentic. Our forensic technique explicitly estimates the perspective projection of text onto a planar surface. We have shown that inauthentic text often violates the rules of perspective projection and can therefore be detected. This approach is semi-automatic, requiring only a user to manually select the text in question. In the case when the text font style in question is unknown, this approach requires a sufficiently large database of font styles from which the required world coordinates are extracted.

A determined forger could circumvent this technique by applying the correct homography to the inserted text [9]. This would, of course, require the forger to estimate the correct homography from the image, which is outside of the expertise of the average Photoshop user.

5. REFERENCES

[1] H. Farid and M.J. Bravo, "Image forensic analyses that elude the human visual system," in *SPIE Symposium on Electronic Imaging*, San Jose, CA, 2010.

[2] H. Farid, "A survey of image forgery detection," *IEEE Signal Proc. Magazine*, vol. 2, no. 26, pp. 16–25, 2009.

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

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

[5] 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, pp. 1078–1081.

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

(a)

(b)

Fig. 5. Two (a) authentic and (b) corresponding fake images.

[7] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 2, no. 60, pp. 91–110, 2004.

[8] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[9] H. Ding, R. Bala, Z. Fan, R. Eschbach, C. A. Bouman, and J. P. Allebach, "Semi-automatic object geometry estimation for image personalization," in *SPIE Symposium on Electronic Imaging*, San Jose, CA, 2010, vol. 7533.