J. Vis. Commun. Image R. 42 (2017) 65–77

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Reliability assessment of principal point estimates for forensic applications

Massimo Iuliani [a,c], Marco Fanfani [b], Carlo Colombo [b], Alessandro Piva [b,c,*]

[a] Dept. of Mathematics and Computer Science, University of Florence, Firenze, Italy
[b] Dept. of Information Engineering, University of Florence, Firenze, Italy
[c] FORLAB Multimedia Forensics Laboratory, University of Florence, Prato, Italy

ARTICLE INFO

Article history:
Received 7 July 2016
Revised 4 November 2016
Accepted 16 November 2016
Available online 18 November 2016

Keywords:
Image Forensics
Scene level analysis
Geometric constraints
Minimum Vanishing Angle
Cropping detection
Splicing detection

ABSTRACT

Although quite recent as a forensic research domain, computer vision analysis of scenes is likely to become more and more important in the near future, thanks to its robustness to image alterations at the signal level, such as image compression and filtering. However, the experimental assessment of vision-based forensic algorithms is a particularly critical task, since they cannot be tested on massive amounts of data, and their performance can heavily depend on user skill. In this paper we investigate on the accuracy and reliability of a vision-based, user-supervised method for the estimation of the camera principal point, to be used in cropping and splicing detection. Results of an extensive experimental evaluation show how the estimation accuracy depends on perspective conditions as well as on the selected image features. Such evidence led us to define a novel visual feature, referred to as Minimum Vanishing Angle, which can be used to assess the reliability of the method.

1. Introduction

Image Forensics has been proposed as a solution for authenticating the contents of digital images [1–3]. This technology is based on the observation that each phase of the image history – from the acquisition process, through its storage in a compressed format, to any editing operation – leaves distinctive traces on the data, as a sort of digital fingerprint [4]. It is then possible to determine whether a digital image is authentic or modified, by detecting the presence, the absence or the incongruence of such traces, that are intrinsically tied to the digital content itself. Forensic traces can be found both at "signal level" (invisible footprints introduced in the signal statistics, like demosaicing artifacts [5], sensor noise [6], or compression artifacts [7,8]) and at "scene level" (inconsistencies in shadows [9], lighting [10,11], or in perspective and geometry of objects [12,13]). The former are typically detected by automatic methods, but they often exhibit lower effectiveness when the investigated content has been subjected to an unknown chain of processes (e.g., filtering, resizing, compression) that may partially or completely spoil the traces left by previous operations [14]. The latter usually require particular constraints on the scene (e.g. the presence of Lambertian convex surfaces for lighting estimation [15]) but have the advantage of being robust to common image processing operations, thus appearing suitable even for low resolution images, or when the content has undergone multiple compressions. While in the literature a great effort has been devoted to evaluate the performance of signal-based forensic methods in terms of detection accuracy and reliability, a limited analysis has been carried out until now on scene-based techniques. This is mainly due to the fact that such algorithms are usually tested on small datasets only, since they cannot exclude some human intervention, e.g. image feature selection or analysis supervision.

This paper represents – to the best of our knowledge – the first attempt to analytically evaluate the performance of a scene level trace. In particular, we addressed the problem of estimating the camera principal point (PP) (whose position in the image under analysis is usually detected by exploiting vanishing points related to three mutually orthogonal directions [16]); whose application in a forensic scenario has been proposed in some recent works [17–19]. For our evaluation, several tests have been performed, on both synthetic and on real images, by varying both the point of view – so as to obtain different perspective conditions – and the number and position of the extracted features. A critical study of the obtained results has led us to define a novel feature, referred to as *Minimum Vanishing Angle* (MVA), allowing us to measure the reliability of the estimated PP. Using the MVA concept, we have also been able to establish a feature selection criterion. Specifically, one should just

* Corresponding author at: via S. Marta 3, 50139 Firenze Italy.
E-mail address: alessandro.piva@unifi.it (A. Piva).

http://dx.doi.org/10.1016/j.jvcir.2016.11.010
1047-3203/

care about choosing the image lines that provide the widest possible MVA, since the accuracy of PP estimation relies more on MVA amplitude than on the amount of data (i.e. image lines) used.

The paper is organized as follows: in Section 2 the State of the Art is briefly presented, and in Section 3 we briefly review the theory behind the adopted PP estimation method. In Section 4 we introduce the MVA and its relation with the image perspective conditions. Then in Section 5 an in deep analysis of the reliability of the method is given. Section 6 presents two possible forensic applications of the PP: cropping detection – for which we provide a detailed accuracy analysis – and splicing detection. Section 7 concludes the paper and summarizes the contributions in light of the achieved results.

2. State of the art

The estimation of the PP from a single image is a known issue in computer vision and photogrammetry, usually embedded into the camera calibration problem [20, Chapter 2]. In order to calibrate the camera, accurate off-line techniques usually require a known pattern in the scene [21,22]. Other methods use video sequences or multiple images to self-calibrate the camera while solving the Structure from Motion problem [23]. In addition, other scene elements such as coaxial circles, or Manhattan-World structure [24] can be exploited for calibration tasks [25–28].

Reported methods assume to use genuine images only, without any malicious modification. This hypothesis allows the authors to impose constraints on the parameters to ease and improve the estimation (for example, the PP is often initialized in the image center). In a forensic application scenario, however, this assumption doesn't hold; Moreover, we have to typically deal with single images already acquired. So, a calibration approach has to exploit useful characteristics of the scene. Given the abundance of images depicting man-made environments, we focus on techniques based on the Manhattan-World assumption.

Given these difficulties, in the forensic literature only a few methods have been presented that try to exploit the camera PP as a clue for tampering detection. In [17], the authors presented a method based on the estimation of the homography mapping a person's eyes to the image plane. Then, the PP is recovered by homography decomposition (supposing focal length is known) and exploited for splicing detection. A similar approach, that exploits circles in the scene to obtain the PP position, is presented in [18]. In [19], the authors notice that asymmetric cropping of an image introduces a correspondent shift of the principal point. Hence, they suggested that the distance between the estimated PP and the image center can be exploited as evidence of cropping. Slightly different, but still related to this topic, is the approach described in [29] where, instead of estimating the PP, tampering detection is based on the direct observation of the vanishing points of different 3D structures (e.g. buildings).

3. Principal point estimation

The mapping between the 3D world and its 2D images is usually modeled as a central projection of a world point onto the image plane (pinhole model [30], see Fig. 1a). The projection rule can be formally written as $\mathbf{m} = K[I|\mathbf{0}]\mathbf{M}$, where $\mathbf{m} = (x, y, 1)^\top$ and $\mathbf{M} = (X, Y, Z, 1)^\top$ are the homogeneous coordinates of a 2D image point and its corresponding 3D world point respectively, whereas K is the camera matrix, embedding the internal parameters of the acquisition device. I is the identity matrix, and $\mathbf{0}$ a column vector of zeros. Typically, the camera matrix is represented as

$$K = \begin{bmatrix} f & s & p_x \\ 0 & \rho f & p_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{1}$$

where f is the focal length, while the aspect ratio ρ and skew s take into account the actual shape of a pixel. Lastly, (p_x, p_y) are the coordinates of the PP (see again Fig. 1a). Modern cameras have reached a high level of quality, with unity aspect ratio and zero skew. So, without significant loss of accuracy, the K matrix can be modeled with $\rho = 1$ and $s = 0$, passing from 5 to 3 degrees of freedom [31].

To obtain the PP, we can exploit the relation among three vanishing points, related to mutually orthogonal directions in the 3D space [16]. A vanishing point (VP) is the intersection point of all the projected lines that are mutually parallel in the scene (i.e. they share the same 3D direction). Note that, in a practical scenario, if more than two concurrent image lines are available, their intersection will not be unique (see Fig. 1b) – since noise can perturb the image line detection – and the VP has to be estimated with an optimization algorithm. In our experiments we employ the solution reported in [16], where after initializing the VP by solving a linear least square problem, a non-linear optimization is carried out.

Let $\mathbf{v}_1$ and $\mathbf{v}_2$ be two VPs related to 3D orthogonal directions. Then $\mathbf{v}_1^\top \omega \mathbf{v}_2 = 0$, where $\omega = (KK^T)^{-1}$ is the *image of the absolute conic*, depending on the three camera parameters f and (p_x, p_y). Given three vanishing points corresponding to three orthogonal directions, we can thus define three independent linear constraints on ω, and finally estimate ω by solving a linear homogeneous system. Eventually K can be obtained using the Cholesky factorization of ω, from which both focal length and principal point can be estimated [16].

The estimation of the PP on a single image can be summarized in three main steps: (1) selection of three groups of concurrent image lines, corresponding to mutually orthogonal directions in the scene; (2) estimation of vanishing points; and (3) computation of ω and recovery of f and (p_x, p_y).

Note that the first step can be done in a manual or automatic way. In the computer vision field, many works have appeared dealing with the problem of line selection and grouping for VP estimation by using Expectation-Maximization approaches [32], the Hough transform [33], or robust estimators, such as the J-Linkage algorithm [34] and employed in [35]. If the camera calibration is known, mutually orthogonal line clusters can be selected automatically [36–38]. On the other hand, with no a priori information about camera calibration (which is our case), it can be extremely hard to check the vanishing point orthogonality without user intervention or by imposing simple heuristics, such as the selection of the most populated clusters. So, in this work we preferred to use a manual line selection scheme. Moreover, notice that also in [29] parallel lines are validated by the user, while in [19] no specific indication is given about the method used to automatically detect orthogonal vanishing points.

4. Perspective analysis

In this section, we evaluate the performance of the PP estimation algorithm under different perspective conditions, so as to determine if and how its accuracy changes when passing from *weak* to *strong* perspective images. The following two subsections report the results of synthetic and real world tests respectively.

4.1. Synthetic tests

In order to carry out extensive tests, a synthetic dataset featuring 248 representative camera poses was built as follows. A 3D cube with unit length sides was placed in the center of the world