Implementation of the Likelihood Ratio framework for camera identification based on sensor noise patterns

Wiger van Houten[1§], Ivo Alberink[†], Zeno Geradts[†]

[§] Criminal Investigation Unit North, Digital Evidence Department, 9728 NP, Groningen, The Netherlands.

[†] Netherlands Forensic Institute, Digital Technology Department, P.O. Box 24044, 2490 AA, The Hague, The Netherlands.

Abstract

It is investigated how implementing the Likelihood Ratio (LR) framework works out in the case of camera identification based on image sensor specific noise patterns. Two typical case scenarios are considered, one with images of low quality, the other with images of high quality. In both cases, it is possible to obtain statistical distributions having a good fit with the reference data both for 'matching' and for 'non-matching' comparisons, and LRs are determined. It turns out that if the reference data is well separated, in the case of 'matching' images/cameras, the statistical fit of the distribution for 'non-matches' is constantly evaluated in a range where there is a lack of reference data. Because of this extrapolation issue, the LRs that emerge are not reliable. This is not a problem that is unique to camera identification: if the informative value of any forensic comparison is high the problem emerges. An alternative approach is presented which consists of choosing a threshold value separating 'matches' from 'non-matches' and

[1] Corresponding author. Tel.: +31 50 587 4742; fax: +31 505875229. *E-mail address*: wawvanhouten@gmail.com.

quantifying the strength of evidence of being larger/smaller than this value. If sample sizes of reference data increase LR results will increase as well, and it is shown that this approach is stable.

Keywords: PRNU, Likelihood Ratio, sensor noise, Bayes' rule

Introduction

In digital forensics the question may arise which particular camera was used to make a certain photograph, e.g. in child pornography casework where an accused is suspected of producing photographs in addition to possessing them. Although photographs often contain EXIF (EXchangeable Image Format) metadata, these often do not list identifying camera characteristics such as serial numbers. Instead, often only classifying characteristics such as brand and model name of the camera are available. An additional problem is that the metadata may easily be removed or changed, either knowingly or unknowingly. Instead of looking at the metadata, one can also look at identifying characteristics present directly in the image due to small deviations in the image sensor itself. These small deviations in the image sensor mostly arise from the pixels in the image sensor having non-uniform sizes. That is, some pixels have slightly larger or smaller areas. These pixels capture more or less light, even when all pixels have the same illumination. This phenomenon is called Photo Response Non-Uniformity (PRNU) [2], and is present in all image sensors. This noise-like pattern is, as far as presently known, stable in time, and can be used as an identifying characteristic. Hence, camera identification

[2] Lukas, J., Fridrich, J., Goljan, M. (2005) Determining Digital Image Origin Using Sensor Imperfections. *Proceedings of SPIE Electronic Imaging San Jose*, CA, January 16-20, 249-260.

comes down to verifying whether the PRNU pattern from a questioned image corresponds to the PRNU pattern from reference images from a camera. Extraction of PRNU patterns can be done effectively and efficiently with state of the art methods. The topic of the current paper will be the assessment of the value of the evidence of eventual similarity of PRNU patterns.

Assessing the similarity between two sources, i.e. individualizing the sources, is classically approached by using a verbal scale (e.g. 'strong support' for a certain hypothesis). This scale may be based on estimations of probabilities or on thresholds set by the expert. It is clear that both approaches are to a certain extent subjective: it likely depends on the amount of experience of the investigator, and may vary from investigator to investigator. In forensics, a framework gaining popularity to assess the value of the evidence is the Likelihood Ratio framework under a Bayesian reasoning approach, from here on: 'LR framework'. The goal of the LR framework is to accurately assess the value of evidence in the light of clearly defined opposing hypotheses, and not to comment on the probability of traces being from a common source, which is considered principally impossible. Furthermore, it should harmonize the value of the evidence and ease the interpretation of the evidence in different disciplines. The LR framework is successfully used for the interpretation of matching DNA profiles, and there is a growing number of publications on the implementation of the framework in fields where comparison of traces takes place, e.g. for comparison of fingerprints, [3,4,5] glass particles,[6] body heights,[7]

[3] Champod, C., Evett, I.W. (2001) A probabilistic approach to fingerprint evidence, *Journal of Forensic Identification*, 51, 101-122.
[4] Egli, N.M., Champod, C., Margot, P. (2007) Evidence evaluation in fingerprint comparison and automated fingerprint identification systems – modeling within finger variability, Forensic Science International, 167(2-3), 189-195.

speech fragments,[8] etcetera. In[9] the same is done for PRNU based camera identification, where the focus is on the measurement uncertainty of the strength of evidence. The current paper focuses on the same subject, but instead of the latter, we investigate the general problems that are encountered when interpretation of results is performed in the LR framework.

The paper starts with a description of the LR framework, the PRNU method, the material used and the way in which the data analysis takes place. In two (fictive) case examples, namely for a mobile phone camera and a good quality camera, it is described what the results of the LR approach are. The paper ends with a discussion of the results.

Methods and materials

The LR framework

In the LR framework[10,11] a distinction is made between competing hypotheses H_p and H_d, evidence E, and background information I in order to make an analysis of the evidential value of E for either of the hypotheses (given I). Instead of looking for the probability that either of two (or more) hypotheses hold given the evidence, the probability of observing this evidence is evaluated given the hypotheses. It is considered the task of the

---

[5] Neumann, C., Champod, C., Puch-Solis, R., Egli, N., Anthonioz, A., Bromage-Griffithis, A. (2007) Computation of Likelihood Ratios in Fingerprint Identification for Configurations of Any Number of Minutiae, Journal of Forensic Sciences, 52(1), 54-64.

[6] Zadora, G. , Neocleous, T. (2010) Evidential value of physicochemical data-comparison of methods of glass database creation. *Journal of Chemometrics*, 24 (7-8), 367-378.

[7] Alberink, I., Bolck, A. (2008) Obtaining confidence intervals and Likelihood Ratios for body height estimations in images, Forensic Science International, 177(2-3), 228-37.

[8] Gonzales-Rodriguez, J., Fierrez-Aguilar, J., Ramos-Castro, D., Ortega-Garcia, J. (2005) Bayesian analysis of fingerprint, face and signature evidences with automatic biometric systems, Forensic Science International, 155, 126–40.

[9] Nordgaard A., Hoglund T. (2011) The use of likelihood ratios in digital camera identification, *Journal of Forensic Sciences*, to appear.

[10] Aitken, C.G.G., Taroni, F. (2004) *Statistics and the Evaluation of evidence for Forensic Scientists*, ed. 2. John Wiley and Sons: Chichester, UK.

[11] Lucy, D. (2006) *Introduction to statistics for forensic scientists*. Wiley, Chichester.

expert to evaluate the ratio of the two, which is called Likelihood Ratio (LR). It is left to the juror/judge to use the LR to update the prior odds of the hypotheses. In a formula this is presented as the so-called Bayes rule:

$$\frac{\Pr(H_p|E,I)}{\Pr(H_d|E,I)} = \frac{\Pr(H_p|I)}{\Pr(H_d|I)} \times LR \qquad (*),$$

with

$$LR = \frac{\Pr(E|H_p,I)}{\Pr(E|H_d,I)}.$$

In the formula (*), $Pr(A)$ stands for the probability of any event A, whereas $Pr(A|B)$ expresses the probability of any event A given any event B. If the evidence consists of continuous readings, the probabilities in the LR formula are substituted by probability densities.

The formula (*) expresses that the posterior odds (left side of the formula) is a product of the prior odds (first term on the right side) and the LR (second term on the right side). The LR conveys the relative support of the evidence for H_p versus H_d. In the LR framework, estimation of the prior odds is considered to be the task of the judge, whereas calculating the LR is the task of the forensic expert. The prior odds are the odds in favor of H_p, without taking the evidence into account. When the evidence is taken into account, the LR is obtained. By multiplying the prior odds with the LR, the posterior odds are obtained. The judge considers the posterior odds when making his decision. Combination of evidence may be achieved by multiplying LRs for separate pieces of evidence, given that they are (conditionally) independent.

The PRNU method

We describe the PRNU method and how the LR framework is applied for comparison based on PRNU patterns.

As put forth in the Introduction, the PRNU pattern is a pattern present in all image sensors mainly as a result of non-uniform sizes of individual pixels. When the image sensor is illuminated uniformly, some pixels will systematically capture slightly more (less) light, resulting in higher (lower) outputs. The word 'systematically' is important here. A pixel in a perfect image sensor (without any non-uniformity) will still exhibit fluctuations (shot noise) in its output when the exact same scene is photographed, due to the statistical nature of light[12]. However, averaging multiple images will remove these fluctuations. Opposite to this, systematic deviations (such as originating from the PRNU) will remain, even after averaging multiple images. Hence, the output of a non-ideal image sensor contains a systematic deviation (PRNU component) and a random component (photon shot noise).

In order to use the PRNU pattern as a means of identifying the source camera, we first need to extract this pattern *P* from the image. In essence, this is accomplished by subtracting a denoised version *F(I)* of the input image *I* from the input image: *P=I-F(I)*. The filtered image *F(I)* is a de-noised version of the original image, an approximation of how the image *I* would have looked if the image sensor would have been perfect, and no noise was present.

Several different filters have been proposed in the literature. In [13], the image is transferred to the wavelet domain and the resulting wavelet-coefficients are denoised with a Wiener filter, giving a filtered image *F(I)*. After transferring the coefficients back

[12] Loudon, R. (2000) *The quantum theory of light*, ed. 3. Oxford University Press, Oxford.
[13] Lukas, J., Fridrich, J., Goljan, M. (2006) Digital Camera Identification from Sensor Noise, *IEEE Transactions on Information Security and Forensics*, 1(2), 205-214.

to the spatial domain, the filtered image is subtracted from the original input image. In [14], the filtered image is obtained by a series of simple convolutions. Put simply, for each pixel, neighbouring data are used as an estimate for the pixel under consideration, often a very good indicator (Lambertian surfaces). Again, by subtracting the filtered image from the original image, the pattern *P* is obtained. For details, see the original publications. By subtracting the denoised image from the original image, we see which pixels have a higher (lower) output. This indicates which pixels are more (less) active.

In practice, we see other characteristics in the PRNU pattern as well. For example, as most digital cameras compress photographs for space-efficient storage, we see periodic patterns along the boundaries of blocks of pixels, as JPEG compression mostly works with small coding units of 8x8 pixels. This compression results in visual artifacts in the image that are inherited in the pattern *P*. Another source for periodic patterns is the following. Image sensors are monochrome devices, which means they are not able to sense colors. In order to give a photo its colors, a color filter array (CFA) is placed on top of the image sensor. A typical color filter consists of repeating patterns of red, green and blue pigments. As a result, each pixel only receives light that is transmitted by its overlaying pigment. In this way, the image to be created forms a mosaic where each pixel only contains information about its own overlaying color. To construct a full color image, where each pixel has a value for red, green and blue, the image needs to be demosaiced. For example, when a pixel has recorded the light intensity of the red color, the intensity for the green and blue color is calculated from its neighbours. Calculating the intensity in this way results in a periodic signal, as the color data is interpolated the same way for

[14] Van Houten, W., Geradts, Z. (2012) Using Anisotropic Diffusion for Efficient Extraction of Sensor Noise in Camera Identification, *Journal of Forensic Sciences*, to appear.

each pixel. This results in visual artifacts in the PRNU pattern. These artifacts are not identifying characteristics, but instead belong to a certain class of cameras. For this reason elevated correlation values are observed between patterns extracted from unrelated images that originate from the same make and model of camera. The degree at which this occurs depends on the relative size of these artifacts with respect to the size of the PRNU.

At our institute, if the question arises whether a certain camera was used to make a certain image, and both the image and camera are available, the PRNU pattern P_q from the questioned image is extracted and compared with the PRNU pattern P_r extracted from a set of reference images from the camera. These reference images are in practice images of a grey surface in which each color channel is approximately uniformly illuminated, so that the PRNU pattern can be reliably obtained. The PRNU patterns are in fact three-dimensional matrices, the first two coordinates fixing the location of the pixel and the third the color (red, green or blue). The similarity between patterns P_q and P_r, denoted as $\rho(P_q, P_r)$, is calculated using a correlation measure, namely Pearson's correlation coefficient:

$$\rho(P_q, P_r) = \frac{(P_q - \overline{P}_q) \cdot (P_r - \overline{P}_r)}{\| P_q - \overline{P}_q \| \cdot \| P_r - \overline{P}_r \|}.$$

Here $\overline{P}_q$ and $\overline{P}_r$ denote the average values of P_q and P_r respectively, whereas

$$a \cdot b = \sum_{i,j,k} a_{i,j,k} b_{i,j,k} \text{ and } \| a \| = \sqrt{a \cdot a}$$

are the usual Euclidian inner product (point-wise multiplication) and norm on the space of three-dimensional matrices. The higher the correlation value is, the stronger the support that there is a relationship between both patterns. However, because of the aforementioned reasons of class-characteristic periodicities, there is no universal

threshold that can be used to conclude whether or not a certain camera is in fact the source of an image.

For this reason, a set of independent reference cameras is used with no relation to the questioned image. Ideally, the number of these independent reference cameras is large, but for practical reasons (they are costly, and taking reference images is time-consuming) it will be limited. When the correlation value between a questioned image q and a questioned camera C is significantly higher than for other cameras, there is a strong indication about the questioned camera being the source camera, and we may try to express this indication in LR form.

We formalize the above in the LR framework as follows. We start by defining two mutually exclusive hypotheses, H_p and H_d:

H_p: the questioned image q was taken with camera C,

H_d: the questioned image q was taken with a camera of the same make and model, but other than C.

We could have taken H_d to imply that any other camera may have been used instead of C. As explained, cameras of the same type will tend to have a higher correlation values with the image than random cameras, so this procedure will tend to come up with conservative results. As a result, a subset of cameras of the same make and model as the suspect camera is used in order to produce the right reference data. To see how well this works, data was obtained or two different types of cameras.

Data used

As stated, two different types of cameras were used: a set of 10 low resolution Motorola V360 mobile phone cameras and a set of 9 Sony DSC-S500 cameras.

For each camera, from the set of reference cameras, a number of reference ('flatfield') images were taken, along with a number of 'natural' images, similar to the questioned image. For the mobile phone camera, we used N=106 natural images and 100 reference images. For each Sony DSC-S500 camera, we used N=100 natural images and 100 reference images. Ideally, the contents of the natural images are the same as of the questioned image, as each pixel has its own characteristic individual response, but this will generally not be feasible in practice. For each camera, the PRNU patterns extracted from the reference images were averaged to obtain a single reference PRNU pattern R_i, and for all individual natural images the PRNU pattern was extracted. For each of the cameras, the correlation of the reference pattern with the PRNU pattern of the natural images of the same and of different cameras (same brand and type) was determined, the correlation being denoted by v. The distributions of outcomes under both hypotheses can be plotted in histograms, and the probability density of both distributions statistically modeled, e.g. by a Generalized Gaussian or a logistic distribution. By dividing the probability density function of H_p at v by the probability density function of H_d at v will yield the LR.

For each of the two sets of cameras (of size n=10 and n=9), each of the cameras was treated in turn as a suspect camera, and LRs were determined for comparisons of patterns under H_p (a total of $N \times n$ values) and under H_d (a total of $N \times n \times (n-1)$ values). Here n is the total number of cameras (either 10 or 9) and N the total number of natural images (either 106 or 100). The results are described below.

Statistical testing methods used

Statistical models for the data derived further on were determined using the Matlab distribution fitting tool 'dfittool'.

In order to statistically test whether a set of observed scores has a good fit with any proposed statistical model, we will use the Kolmogorov-Smirnov and the Lilliefors test statistic, cf. [15]. Results for tests are given by means of *p*-values. If a *p*-value is small this indicates that under the proposed model the findings were unlikely to appear. In this paper the model is rejected if $p<0.05$.

To graphically illustrate the data, we will use kernel density estimations, cf. [15] as well, with normal kernels and standard band widths as calculated by Matlab.

Results

We have a look at the results for the two cases described. We start by looking at the results for the mobile phone camera.

Case 1: the mobile phone cameras

In Figure 1 we plot the results for the mobile phone camera. In the first plot, an illustration is given of the observed correlations in the case of H_p and H_d. We will refer to outcomes under H_p as 'matches' and outcomes under H_d as 'non-matches'. As one can see, there is a fair amount of overlap. The statistical models that were fitted are a generalized extreme value distribution for the 'matching' scores (the Kolmogorov-

[15] Hollander, M., Wolfe, D.A. (1999) *Nonparametric statistical methods*, ed. 2. Wiley, New York.

Smirnov test statistic gives a *p*-value of p=0.68), and a *t* location scale distribution for the 'non-matching' scores (the Kolmogorov-Smirnov test statistic leads to p=0.06). Based on this, in the second plot, the LR is described as a function of the correlation. In the third plot, kernel density estimations are given for the observed LRs under both H_p and H_d. For ease of illustration, instead of LRs, $\log_{10}$(LR)s are given.

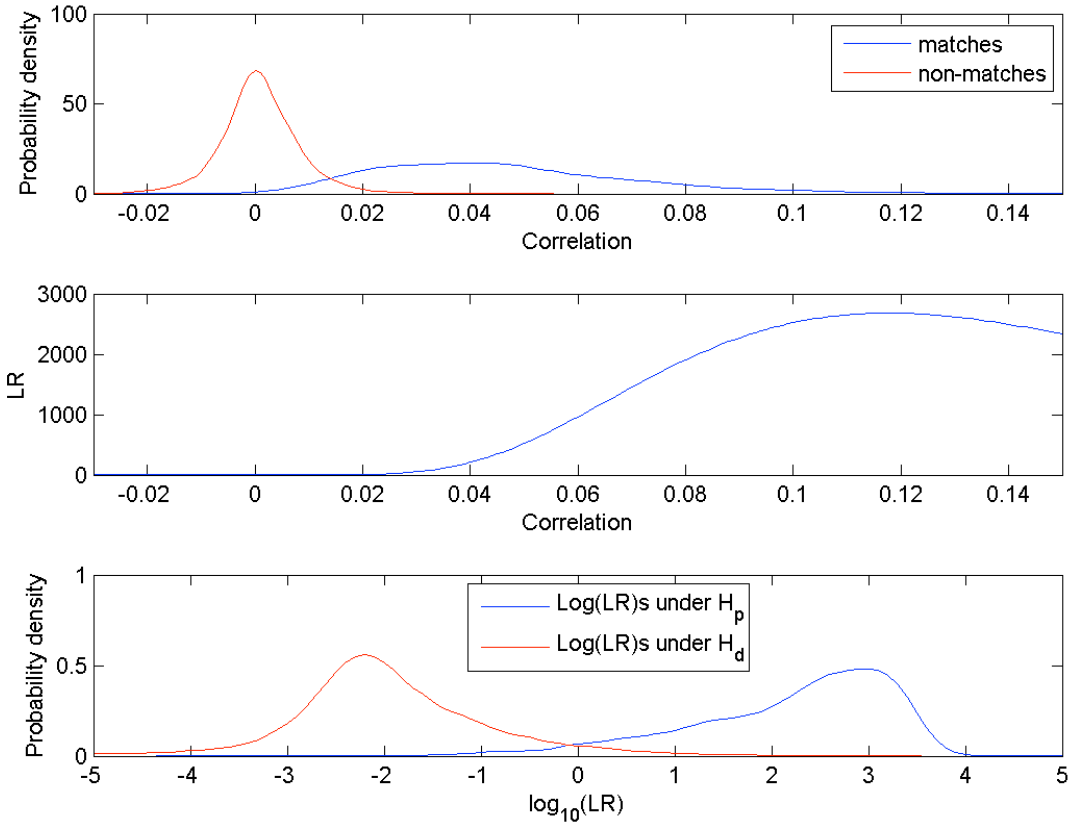

Figure 1. Results for PRNU comparison for mobile phone cameras. In the first plot, kernel density estimations are depicted for 'matching' and 'non-matching' patterns. In the second plot, the LR is described as a function of the correlation. In the third plot,

kernel density estimations are given for the observed LRs under both H_p and H_d. For ease of illustration, instead of LRs, $\log_{10}$(LR)s are given.

In the second plot we see that up to a value 0.12, as the correlation value increases, it becomes more and more likely that the supposed camera was in fact used to make the questioned image and the LR is increasing. This makes sense. The decrease of the function after this value does not. This is a consequence of the statistical model we have adopted. Indeed, only 2% of the observed score for 'matching' patterns are over 0.12, and 0% of the 'non-matches', so in this region it makes no sense to use our statistical models for the data anymore.

Case 2: the Sony DSC-S500 cameras

In Figures 2 and 3 we plot the results for the Sony DSC-S500 cameras. In subplot 1 of Figure 2, an illustration is given of the histograms of outcomes for the observed correlations in the case of H_p and H_d. One can see that there is much less overlap than for the mobile phone cameras. The statistical models that were fitted are a generalized extreme value distribution for the 'matching' scores (the Kolmogorov-Smirnov test statistic leads to p=0.70), and a normal distribution for the 'non-matching' scores (the Lilliefors test leads to p=0.28). Based on this, in the second subplot, the LR is described as a function of the correlation, on a $\log_{10}$ scale.

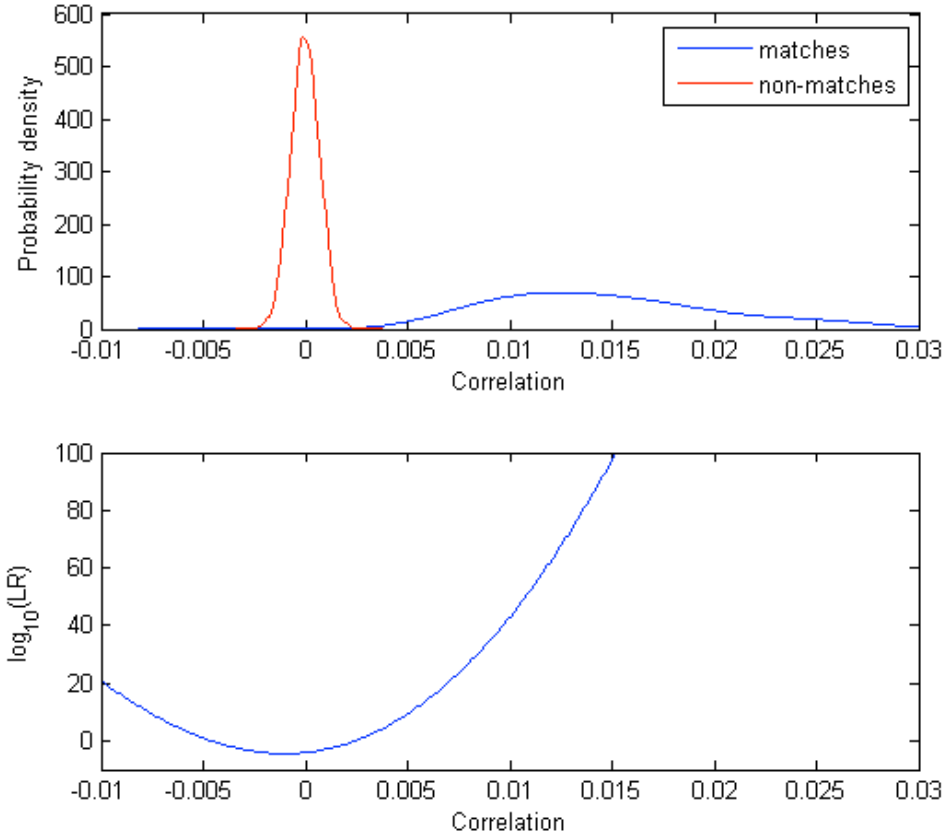

Figure 2. Results for PRNU comparison for the Sony DSC-S500 cameras. In subplot 1, kernel density estimations are given for 'matching' and 'non-matching' patterns. In subplot 2, the LR is described as a function of the correlation. For ease of illustration, instead of LRs, log_{10}(LR)s are given.

In Figure 3, kernel density estimations are given for the observed $\log_{10}$(LR)s under both H_p and H_d.

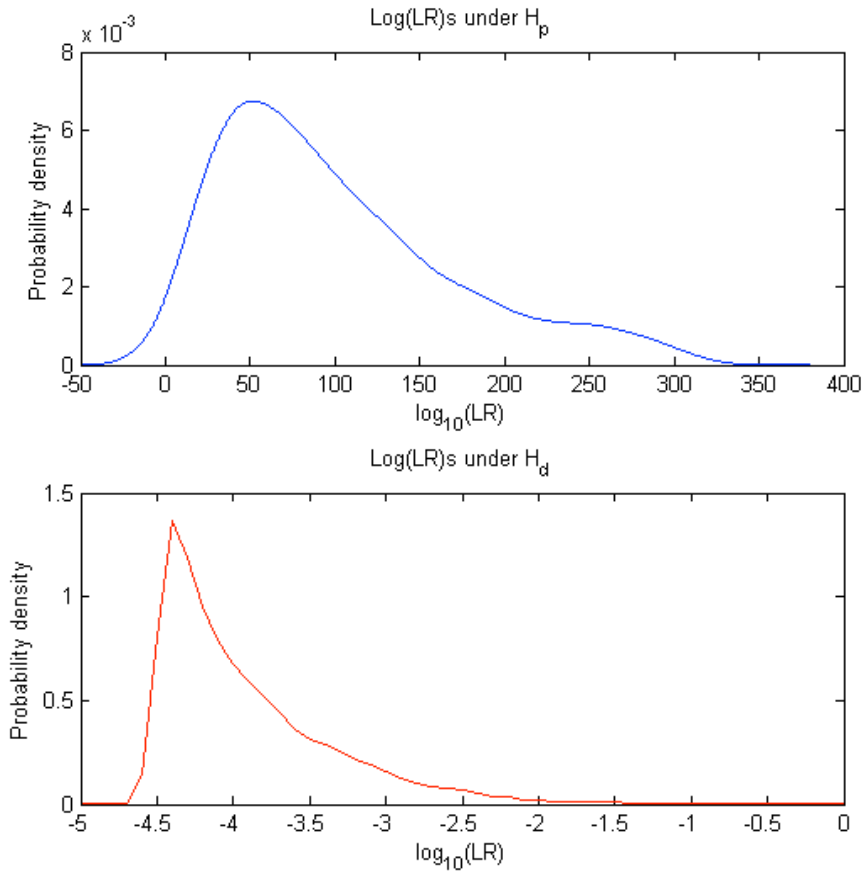

Figure 3. Kernel density estimations of log_{10}(LR)s obtained for PRNU comparison for the Sony DSC-S500 cameras, both under H_p (subplot 1) and H_d (subplot 2).

Note that in Figure 2, subplot 2, up to 0 the LR function is decreasing, which makes no sense, again as a consequence of the statistical models used. The reason for this is that the tail of the normal distribution decreases much more quickly than that of the generalized extreme value distribution. As a second remark, see Figure 3, subplot 1: note the high values of LR under H_p. These outcomes are caused by the model, not the data. Indeed what happens is that for almost every 'matching' score observed, we cannot safely

estimate the probability density for ‘non-matching’ scores at that point, since the reference data does not reach it. The question is: which LRs are reliable here?

A possible way out is to use conservative estimates for the distribution of scores, e.g. stop at some quantile that is still considered dependable and make use of the density value over there for all comparison scores that are higher. In the current set-up this does not work well: if we stop at the 95% quantile, for the Sony cameras we obtain LRs that are <1 under H_p in 100% of the cases, instead of the values like 10^{50} that resulted from the calculations. Given the fact that the size of the reference sample is 7,200, higher quantiles might be dependable which would lift the LRs somewhat.

Alternative approach

A possibility which is more stable is the following. If discrimination of values is very good, we may split the reference data into two parts. The first part, consisting of say 100 scores for ‘matching’ comparisons and 100 for ‘non-matching’, is used to come up with a threshold value for which we expect it separates the reference data of ‘matches’ and ‘non-matches’ optimally. In Figure 2, a plot is presented of the distribution of correlation values for the set of 9 Sony DSC-S500 cameras. In this case, as can be seen in the upper plot of Figure 2, a threshold of (say) 0.003 visually separates the ‘matching’ and ‘non-matching’ correlation values. The remaining reference data, 800 ‘matching’ and 7,100 ‘non-matching’ scores is used to estimate the probability of a comparison score to be over or under the threshold, given both hypotheses. If a new comparison is performed, we just evaluate whether the LR is over 0.003 or not. If the numbers of reference data are high

this may work well. (Note that the reference data needs to be split up because the threshold value may not depend on the data used to determine the strength of evidence.)

We look at a numerical example: suppose we obtain a comparison score of 0.005 and look for the LR of the fact that this outcome is >0.003. Indeed, then

$$LR = \frac{P(v>0.003|H_p)}{P(v>0.003|H_d)} = \frac{p_1}{p_2},$$

where p_1 is the probability for 'matching' scores to be >0.003, and p_2 that for non-'matching' scores. In the reference data, all 'matching' scores were above 0.003, and all 'non-matching' scores below. Now we concentrate on the 'non-matching' scores: here the sample of 7,100 comparisons has a total number of 'successes' (values >0.003) b=0. A priori, this number had a binomial distribution with parameters p_2 and n=7,100, which is approximated well by a normal distribution with expected value np_2 and standard deviation $\sqrt{(np_2(1-p_2))}$. As a consequence, a 95% confidence interval for p_2 is given through:

$$|b-np_2| \leq 1.96 \times \sqrt{(np_2(1-p_2))}.$$

This leads to a quadratic equation in terms of p_2, which leads to:

$$0 \leq p_2 \leq \frac{1.96^2}{n+1.96^2} \approx 5.4\times10^{-4}.$$

In turn, the lower bound for the LR involved is approximately given by $1/(5.4\times10^{-4}) \approx$ 1,800. (Here p_1 was approximated by 1: in fact, using similar arguments, we could have used $1-4.8\times10^{-3}$ as an upper bound, which gives practically the same results.) Note that there is no good upper bound since we cannot prove that $p_2 \neq 0$.

We study the robustness of the above by applying the following procedure for 1,000 times. From the 900 'matching' and 7,200 'non-matching' scores, we select two subsamples of size 100, and calculate the average of the averages of these subsamples.

We use this number as the threshold value. Then we inspect whether any of the remaining 800 ‘matching’ scores is under, or any of the remaining 7,200 ‘non-matching’ scores is over the threshold. In none of the 1,000 repetitions of this experiment this was the case. Hence we always obtain the same lower bound of 1,800 for the LR for a score over the threshold. This illustrates that in the case of the Sony cameras, the procedure is stable.

Discussion

In the paper, it was investigated how implementation of the LR framework under a Bayesian reasoning approach works out in the case of comparison of images and cameras based on PRNU patterns. We considered two typical case scenarios, the one with images of low quality, the other with images of high-quality. In both cases, it turns out to be well possible to obtain statistical distributions underlying the reference data for both ‘matching’ and ‘non-matching’ comparisons. Based on these, LRs can be calculated. For the mobile phone cameras, the second plot of Figure 1 illustrates that here in the tail of the distributions problems will emerge: the LR decreases as a function of the correlation between PRNU patterns, which is nonsensical. For the Sony cameras, this point is even clearer. Again the LR function is not increasing on the whole range of correlations encountered. Moreover, LRs under H_p are absurdly high. The reason for this is that the statistical fit of the distribution for ‘non-matches’ is constantly evaluated in a range where there is no reference data. Clearly the numbers that are coming out cannot be trusted. The problem is that extrapolation takes place in the tail of the fit for correlation scores under H_d, which is bad statistical procedure. All in all, under these circumstances

it is not possible to come up with reliable LRs, and the reason for this is that the correlation scores under both hypotheses are separated too well.

The issue of widely separated distributions, and the resulting unreliable LRs is not a problem that is unique for PRNU-based comparison: if the informative value of any forensic comparison (be it fingerprints, speech, glass particles, etcetera) is high, the problem emerges. Although this may be considered to be a problem of luxury, the question remains how to deal with it. The alternative of checking whether comparison scores are larger or smaller than some threshold value will yield LRs which will be smaller, but at least reliable.