DETECTION OF RESAMPLED IMAGES: PERFORMANCE ANALYSIS AND PRACTICAL CHALLENGES

F. Uccheddu [+], *A. De Rosa* [+], *A. Piva* [+], *M. Barni* [*]

[+] Dept. of Electronics and Telecommunications, University of Florence, ITALY
{francesca.uccheddu, alessia.derosa, alessandro.piva}@unifi.it
[*] Dept. of Information Engineering, University of Siena, ITALY
barni@dii.unisi.it

ABSTRACT

The assessment of the practical performance of forensic methods for the detection of resampling operations on digital images is a difficult task requiring a careful experimental analysis. Unfortunately, the experimental analysis reported in multimedia forensics papers is often statistically insufficient, or at least not usable to predict the performance of the proposed systems at scales needed for practical deployment. This paper attempts to move a first step to fill this gap in a twofold way. First of all a proper experimental framework is proposed to analyze the performance of resampling detectors. Then the suggested methodology is applied to evaluate the performance of two of the most significant resampling detectors proposed so far, providing a deeper-than-usual analysis of their behavior under different working conditions.

1. INTRODUCTION

In the last years, in the signal processing research community it has been recognized the need for the definition of standard methodologies for a correct evaluation and comparison of techniques and procedures [2, 11]. As a matter of fact, in the available literature very often proposed algorithms are described without implementation details, thus limiting an easy reproducibility of the schemes; moreover the experimental analysis is usually statistically insufficient, or at least it can not be used to validate the performance of the methods at scales needed for practical scenarios. This trend is currently valid also in the image forensic research field: as a matter of fact, most of the published works lack a clear explanation of the conditions under which a comparison between proposed methods and competing techniques is carried out; often the experimental results are given on very small data sets; the information about the experimental setups (like publicly available data sets, implementation of the proposed algorithm and chosen parameters, steps followed in the experiments) is usually not enough detailed to warrant an easy reproducibility of the presented results. All these limitations do not allow a fair and easy validation and comparison with newly proposed methods hindering the fulfilment also in this field of the so-called Reproducible Signal Processing paradigm [2, 11]. This work aims at starting to fill this gap by studying a methodology for the performance analysis of image forensic techniques; in particular we will focus our attention on the class of forensic algorithms designed for the detection of resampling operations applied to the digital images.

In the next Section we describe the scenario where the considered kind of tampering is used and we analyze the information about the experimental setups and tests provided in the current literature on forensic schemes for resampling detection. In Section 3 we propose an appropriate experimental framework. The proposed framework is used in Section 4 to thoroughly evaluate the performance of two of the most popular resampling detection schemes proposed so far.

2. EXPERIMENTAL TESTS FOR RESAMPLING DETECTORS

Several image forensic techniques have been designed to detect tampering operations on a digital image in absence of any prior knowledge of the original content. In this framework, the class of tools for the detection of interpolation processing has become an important research line.

This interest can be explained by taking into account that in order to obtain convincing forged images, it is often necessary to apply a geometrical transformation to some portions of the manipulated images, thus requiring the application of a resampling step. Although resampling processes do not typically leave perceivable artifacts, they introduce specific periodic correlations between image pixels; these periodic interpolation artifacts present in pixel amplitudes or similar related image statistics are the features that all resampling detectors look for in order to decide if an image, or a subpart of it, has undergone a geometrical transformation.

In general, such schemes work as follows: given a possibly manipulated image, a set of features is extracted from the whole image or from a selected region of interest (usually a central square area of size $N \times N$). Some of the existing methods use a parameter computed from the extracted features as a test function to be compared against a detection threshold in order to decide: if the test function is above the detection threshold the detector decides for the interpolation presence, otherwise a negative answer is given. Other schemes are based on a classifier to decide whether an image is interpolated or not: in such a case the set of extracted features is input to a properly trained classifier that decides whether the image or region under testing belongs to the class of resampled contents or not.

The detection accuracy is assessed mainly by evaluating the missing detection and false alarm probabilities; in the case of classifier-based detectors, the detection accuracy is assessed by separating the considered image set in a training set and a test set, where the missing detection and the false alarms are directly evaluated. Detector robustness is sometimes tested in presence of post processing such as JPEG compression or noise addition. Of course, these measures are highly influenced by the chosen set up: most of the experimental tests are performed on a certain set of unaltered and interpolated images. The chosen images, are of course not the same across different papers: the unaltered datasets differ by the represented contents, numbers of images, image size by the image format (i.e. JPEG/RAW/..); the interpolated images also differ because of the specific geometric transformations applied to the images (resize, rotation,..) as well as for the used interpolation kernel (i.e. bilinear, cubic or nearest-neighbor).

In the following, we briefly review the test settings that are commonly used in the scientific literature to verify the effectiveness of their proposed resampling detectors. No need saying that a common approach is needed both to give more trustworthy accuracy values and to compare different methods under similar conditions.

Concerning the characteristics of the data set used to evaluate the performance of the algorithm, a high variability on the number and kind of images can be found. In [10] only one image is used to test the proposed algorithm and no information about its size or

 ISSN 2076-1465

format is given. In [4] the tests are performed on images coming directly from a digital camera, 13 of them are non interpolated but compressed with the maximum quality, and 101 images are the result of using the in-camera digital zoom. The tested database in [7] consists of 40 uncompressed TIFF images used also for creating the corresponding interpolated images. Authors in [9, 5] use 200 uncompressed TIFF images as unaltered, but while [5] use all of them for generating the corresponding interpolated versions, [9] uses only 50 of them to do so. Finally, [6] considers as unaltered a set of more than 5000 high quality JPEG compressed images from the online Columbia dataset [8].

The validity of the resampling detection algorithms is usually assessed by analyzing supervised resized images (both down-scaled and up-scaled), that use some specific interpolation methods (i.e. bilinear, cubic or nearest-neighbor) though often the method is not specified [10, 4]. In [7] authors test the system also with rotated and skewed images. The robustness to further attacks is tested against JPEG post compression in [9, 5, 7, 10] , against noise addition in [9, 7] and against gamma correction in [9].

The investigated region is 512x512 pixels for [9, 5, 7], the whole image for [4] while no information is reported for [10, 6].

Authors in [9, 5, 7, 4] use a threshold detector to evaluate the detection accuracy, while authors in [10] give just a visual idea of the detector performance, and the authors in [6] use a Support Vector Machine (SVM) classifier to infer if an image is interpolated or not, estimating the accuracy by averaging the results of the tests repeated 30 times by scrambling the training and the test dataset.

The detection accuracy is usually presented by means of graphs that, for a fixed threshold value, show the correct detection rate with the change of scaling factor o rotation degrees.

3. A PROPOSAL FOR AN EXPERIMENTAL METHODOLOGY

By the light of the analysis given in the previous Section, it is evident that the approaches usually adopted to evaluate the performance of resampling detectors are rather naif. To help overcoming this problem, we propose some guidelines that should be followed for testing the reliability of multimedia forensic algorithms, in particular those focusing on resampling detection. We will try to move the analysis away from the laboratory conditions, by taking into account how resampling detection algorithms should by applied in real scenarios.

First of all, the typology of images encountered in real scenarios include not only high resolution, high dimension, and uncompressed images directly coming from digital cameras, but also images with different dimensions, smaller resolution, suffering some level of compression, and last but not least images coming from different sources, like scanners and computer graphic tools.

The number of images used for the tests should be sufficient for a correct statistical analysis, i.e. for measuring a given error probability (1 error on 10^k) the number of analyzed data should be at least $10 \cdot 10^k$.

Focusing on the performance analysis of resampling detectors, the kind of processing applied to the tested images should include at least scaling operations, i.e. different scaling factors for both up-scaling and down-scaling, rotations with different rotation degrees, and skewing applied to y and x directions. Furthermore, each of the previous geometrical operations requires an interpolation process that can be implemented with different interpolators (i.e. nearest neighbor, bilinear, bicubic, and even higher order interpolation polynomials). In many cases the presence of anti-aliasing filtering should be taken into account when an image is down-sampled.

As previously described, it is common that before splicing together parts originating form different images, some geometrical operations are applied to the pasted region. However, to make the tampering more convincing, some post-processing will likely be applied after the pasting and resampling step. For instance it is possible that the tampered image is filtered, gamma-corrected, or it is contrast enhanced. Finally it is very likely that the tampered image is JPEG-compressed. It is necessary then that the robustness of any resampling detection scheme against post processing is measured. On this regard it is worth highlighting a quite common mistake made in the preparation of the dataset used for the experiments. It is quite common in fact that the post-processing attack is applied to the resampled images only. Suppose for instance that we are interested in analyzing the robustness of a certain resampling detector against JPEG compression. It is not rare to find papers in which the test dataset consists of the original images with no JPEG compression and the resampled images to whom a final JPEG compression step is applied. If this is the case, the test results risk to be useless since by relying on them we can not distinguish whether the detector actually recognizes the resampling artifacts or the artifacts introduced by JPEG compression. The reader could even think that the application of a strong JPEG compression helps the detector to identify the resampling artifacts, while the true explanation is that the resampling detector *learned* to distinguish non-compressed from compressed images (an example of this problem will be given in Section 4).

Regarding the dimension of the analyzed region or subregion, i.e. the input for the resampling detector, two different analysis are possible. From one hand we can assume that the geometric operation is applied to the whole original image to obtain a geometrically transformed image: in this case, the tested region should coincide with the tested image, or have dimensions similar to those of the tested image. On the other hand, it may be the case that only the to be pasted region (usually smaller than the final fake image) suffered some geometrical operation: in this case, we can assume that the region to be analyzed by the resampling detector should be no larger than a quarter of the overall analyzed image. In a real scenario the latter situation is the most likely. Finally, since we usually handle RGB images, the resampling detection algorithms could be applied to each RGB channel separately, as well as to the corresponding gray-level image obtained by extracting the luminance only.

In addition to the definition of the dataset to be used in the experiments, particular attention should be given to the selection of a proper performance metric. By looking at resampling detection as a two-class hypothesis testing problem in which the analyzed image is classified either as original or resampled, the receiver operating characteristic (ROC curve) - depicting the tradeoff between correct detection rates and false alarm rates - represents an appropriate means for measuring the performances of the resampling detector [3].

Briefly, the ROC curve plots the true positive rate, also called correct detection probability P_d (i.e. the number of images correctly detected as resampled, on the total number of analyzed resampled images) versus the false positive rate, also called false alarm probability P_{fa} (i.e. the number of images erroneously detected as resampled, on the total number of analyzed no-resampled images), as the decision threshold of the detector is varied.

The ROC curve is a two-dimensional depiction of the properties of the resampling detector: in order to summarize the performance with a unique scalar value representing the general behaviour of the detector, a common method is to calculate the area under the ROC curve (AUC), which should assume values between 0.5 and 1 for realistic and effective (i.e. no random) detectors.

Of course, the resulting ROC curve depends on the images used for computing the false positive rate (the no-resampled image dataset) and the true positive rate (the resampled image dataset): hence, for a given type of image dataset, resampling procedure, dimension of the analyzed region, etc., a different ROC curve and its corresponding AUC should be given. As previously described, very often in the forensic literature of resampling detectors the experimental results are presented as plots of the true positive rate P_d for a fixed false positive rate P_{fa}, by varying the geometrical operations (i.e. scaling factors or rotation degrees). P_{fa} is usually fixed to low values, but no motivations are provided to justify such a choice. On the contrary, for measuring low error probabilities (e.g. a false positive rate of 10^{-2}) a high number of images should be tested (e.g. at least 10^3 images) and this is almost never the case in the state-of-

the-art algorithms. Furthermore, in forensic applications, it is not so clear how worrying are the false positive errors, since this likely depends on the particular application scenario. Consequently, we believe that it is necessary to know the behaviour of the resampling detector for different false alarm probabilities, hence leading us to prefer the AUC value as a synthetic scalar measure of the detector performances.

4. PERFORMANCE OF TWO POPULAR RESAMPLING DETECTORS

In this Section we apply the guidelines given in the previous Section to two of the most significant resampling detection algorithms available in the scientific literature. In this way we reach a twofold goal: first of all we show in practice how the guidelines we gave should be applied, secondly we give some insights about the current state of the art in resampling detection and the practical challenges that still need to be faced before such techniques are applied in a realistic scenario. Specifically, we carried out some experiments to evaluate the performance of the detectors proposed by Kirchner & Gloe [5], (that inherently extends and improves the detector of Popescu and Farid [9]) and Mahdian & Saic [7] (from here on referred as KG and MS detectors). The two algorithms were chosen among others, because of the immediacy of their reproducibility: a clear and complete description of the whole algorithm is given in [5], and the implementation is available online for [7].

The first, non-trivial step consisted in fact in the implementation of the to-be tested algorithms according to the information available in the papers. This is in general a difficult task since the details provided in the papers may not be sufficient at all, and a *personal* approximated re-implementation of the system, is the best that one can do. In addition, if the test conditions differ very much from those adopted in the original papers, it may be necessary to tune the detector parameters to the new conditions thus deviating significantly from the set up described in the original paper. Stated in another way, we have to face a scalability problem, since no cue is given about how the detectors should be tuned to work under different experimental conditions (e.g. images with different size, or images produced by different sources).

In the case of [5], we achieved a good reproduction of the algorithm and its performance compared to those indicated in the paper, also having the possibility to test our implementation of the algorithm on the same dataset of images used by the authors [1]. Concerning the system described in [7], we used the software made publicly available by the authors for feature extraction [1] and we completed the algorithm with the test function evaluation, in order to implement the detection as a two-class hypothesis testing problem. The test function is evaluated as the maximum of the ratio between the feature vector and its averaged version obtained with a window of size 9 samples, skipping the previous and the following samples in the mean computation.

Following the experimental methodology proposed in the previous Section, we tested both the algorithms under different conditions. We used 5 databases, each containing 200 images, namely:

- [natural] natural images coming from a Nikon D70 and converted from raw to uncompressed TIFF images using Nikon View NX;
- [scanned] scanned images coming from a Canon CanoScan 5600F scanner;
- [CG] computer generated images found on the Web, images generated by computer graphic tools;
- [JPEG90] JPEG compressed images: the natural image dataset was compressed with a JPEG quality factor of 90 by using the *imwrite* function of Matlab ®
- [JPEG75] JPEG compressed images: the natural image dataset was compressed with a JPEG quality factor of 75 by using the *imwrite* function of Matlab ®.

[1]We would like to thank Matthias Kirchner for helping us in the reproducibility of his results

All these images are supposed to be unaltered, that is a true assumption for all the databases except for the computer generated images that were found on the Web and for which we do not know whether some particular processing were used after image generation. For each image in the 5 databases, we considered only the luminance channel.

For the particular case of the images coming from digital cameras, we know that the CFA interpolation inside the camera could disturb the resampling detectors: in order to remove the periodic patterns introduced by the demosaicking step, images used for testing could be down-sampled by a factor two; since such a solution could be used only in a supervised laboratory conditions, we did not consider this possibility. We tested the comparing system, by analyzing the detectors performances on scaled interpolated images, but it is worth noting that the same analysis could be performed on rotated images as well as for other tampering operations implying interpolation, or a combination of such manipulations. Starting from these "original" databases, we obtained the corresponding resampled images by applying a scaling operation for different up and down scaling factors, namely $\{0.6, 0.75, 0.9, 0.95, 1.0, 1.05, 1.10, 1.20, 1.50, 1.80\}$ by using the *imresize* function of Matlab ®. In particular, we considered 4 different situations, corresponding to the use of different interpolation procedures - i.e. bilinear (BL) and bicubic (BC) interpolation - and to the possible presence of an anti-aliasing filtering (AA). The results we obtained for Kirchner & Gloe's detector are given in Figure 1 while those produced by Mahdian & Saic's algorithm are given in Figure 2. Both figures refer to the natural database.

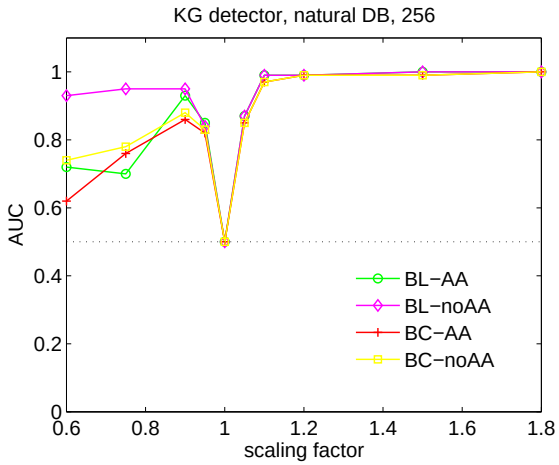

Figure 1: AUC values for the KG's detector on the natural database analyzed on the central 256x256 image region.

Upon inspection of the results several important conclusions can be drawn. First of all both schemes perform better for upscaling rather than for down-scaling. In the latter case, the presence of an antialiasing filter causes a degradation of the performance. Kirchner & Gloe's method tends to perform better than Mahdian & Saic's detector, especially for down-scaling. On this regard, an interesting behavior is observed for the Mahdian & Saic's for down scaling: from a certain point the AUC is lower than 0.5. At a first sight this is a surprising result since AUC = 0.5 corresponds to a random decision, hence values lower than 0.5 should never occur. Indeed, by simply exchanging the decision of the detector better results could be obtained since the detector tends to deterministically exchange original and resampled images. A closer investigation of the experiments yielding a value of AUC lower than 0.5 reveals that the detector statistic used by Mahdian and Saic is higher for the original images than for the resampled images. In other words, the decision statistic permits to discriminate between original and resampled images, however the sign of the decision should

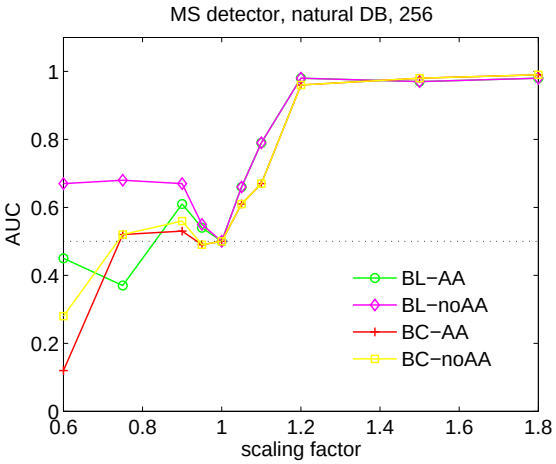

Figure 2: AUC values for the MS's detector on the natural database analyzed on the central 256x256 image region.

be changed, if the decision statistic is lower than the threshold a positive answer should be given.

Later on we will see that this phenomenon is by far stronger when a JPEG compression is applied both before resampling (see Figure 6 for the curves corresponding to JPEG75 and JPEG90 databases) and after resampling (see Figure 8 for the curves corresponding to compressed, 0.90 and compressed, 1.20). When we do not have any a priori information about the kind of processing that have been applied (as in real scenario), the problem of the sign of the decision becomes critical. Very likely a two-sided hypothesis testing problem will have to be used whereby both unusually high or low values of the detector statistic are taken as evidence of resampling.

Regarding the size of the image (or image subpart) under test, we compared, analyzing the natural database and considering bilinear interpolation with an anti-aliasing filtering, three different cases, by giving as input to the detectors a central portion of the image of dimensions 64×64, 128×128 and 256×256 respectively, obtaining the results shown in Figures 3 and 4. As expected when the

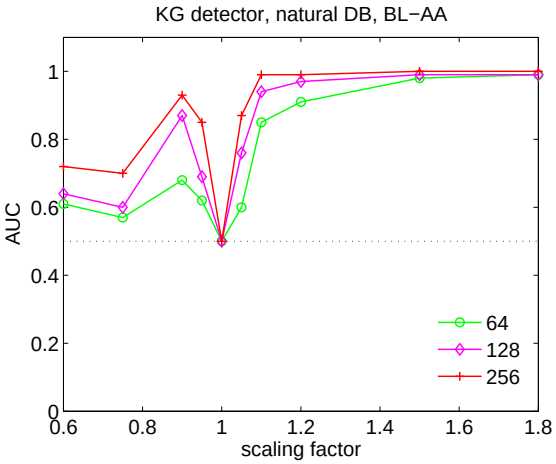

Figure 3: KG detector performance for different size of the investigated region, on the natural image database (BL, AA).

image size decreases the performance get worse, with both methods working better for larger-than-one scaling factors. As before KG method tends to perform better, especially in the presence of moderate down-scaling factors.

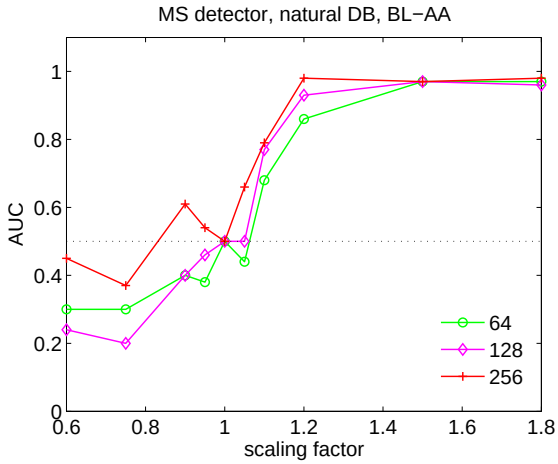

Figure 4: MS detector performance for different size of the investigated region, on the natural image database (BL, AA).

In order to analyze the behavior with respect to the 5 different image databases, we fixed a particular set of conditions: bilinear interpolator, size of analyzed region (256×256), application of the anti-aliasing filter. Figures 5 and 6 show the results we obtained. It is interesting to observe that while the MS detector strongly depends on the analyzed images, the performance of KG detector - at least for up-scaling - is invariant to the image datasets, thus it could be successfully used in real scenarios where there is no a-priori knowledge on the origin of the images under test.

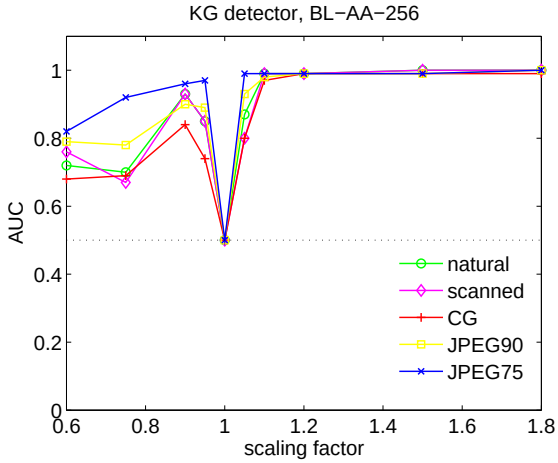

Figure 5: KG detector performance on all the available databases, (256x256 central region, BL, AA).

As last analysis we measured the robustness of KG and MS resampling detectors against JPEG post-compression. We considered as "original" dataset the collection of images belonging to the natural database to whom a JPEG compression was applied with quality factors set to $\{50, 75, 90, 95, 98, 100\}$; as "manipulated" dataset the collection of images belonging to the natural database, resampled with scaling factors fixed to 0.90 or 1.20 and finally JPEG compressed as before. Corresponding results are given for the KG detector in Figure 7, where as expected best performances are achieved for higher JPEG quality factors and higher scaling factors. The behaviour of the MS detector is described in Figure 8 (see the curves corresponding to compressed, 0.90 and compressed, 1.20), where the inversion of the detector sign is evident. In the same figure, the results obtained using as "original" dataset the natural database without any compression, are also plotted (see the

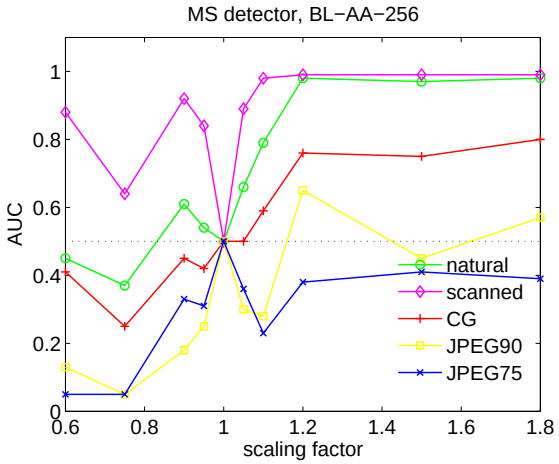

Figure 6: MS detector performance on all the available databases, (256x256 central region, BL, AA).

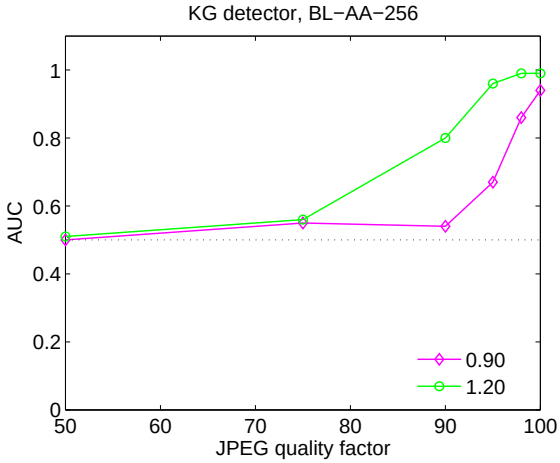

Figure 7: KG detector for JPEG post-compressed images (natural dataset, 256x256 central region, BL, AA, scaled with factors 0.90 and 1.20).

curves corresponding to natural, 0.90 and natural, 1.20). Indeed, as anticipated in Section 3, with this "original" dataset we achieve much better performances, since it is evident that MS detector is identifying the artifacts introduced by JPEG compression, as better as the JPEG quality factor decreases.

5. CONCLUSIONS

In this paper we proposed a proper experimental methodology to make a deeper-than-usual analysis of the performance of resampling detectors. The suggested framework has been applied to two of the most popular resampling detectors proposed so far. Results coming from the applied methodology, provided an interesting analysis of the behaviour of these detectors under different working conditions, thus giving an indication of how such algorithms work in unsupervised scenarios where there is no a-priori knowledge on the origin of the images under test and on the possible manipulation suffered by them.

6. ACKNOWLEDGEMENTS

This work was sponsored by the European Commission under the Project LivingKnowledge (IST-FP7-231126) and by MIUR under project n. 2007JXH7ET.

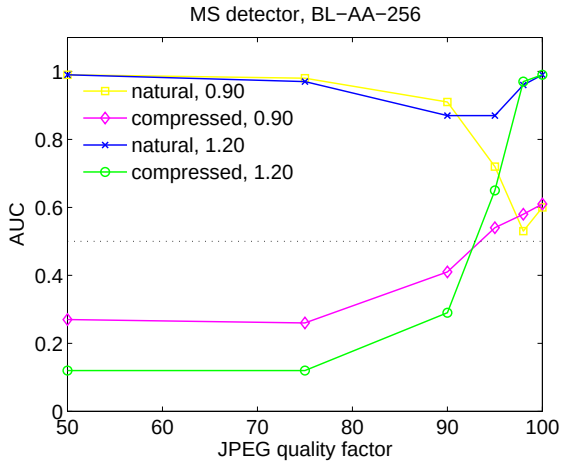

Figure 8: MS detector for JPEG post-compressed images (natural dataset, 256x256 central region, BL, AA, scaled with factors 0.90 and 1.20) using the two different "original" datasets (with and without compression).

REFERENCES

[1] http://zoi.utia.cas.cz/files/rsmp_core.txt.

[2] M. Barni and F. Perez-Gonzalez. Pushing science into signal processing. *IEEE Signal Processing Magazine*, 22(4):119–120, July 2005.

[3] T. Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[4] A. C. Gallagher. Detection of linear and cubic interpolation in JPEG compressed images. *Computer and Robot Vision, Canadian Conference*, 0:65–72, 2005.

[5] M. Kirchner and T. Gloe. On resampling detection in recompressed images. In *Proceedings of the 2009 First IEEE International Workshop on Information Forensics and Security*, pages 21–25, London, UK, 2009.

[6] Q. Liu and A. H. Sung. A new approach for jpeg resize and image splicing detection. In *MiFor '09: Proceedings of the First ACM workshop on Multimedia in forensics*, pages 43–48, New York, NY, USA, 2009. ACM.

[7] B. Mahdian and S. Saic. Blind authentication using periodic properties of interpolation. *IEEE Transactions on Information Forensics and Security*, 3(3):529–538, 2008.

[8] T. Ng and S. Chang. A data set of authentic and spliced image blocks. Technical Report 203- 2004-3, Columbia University, June 2004.

[9] A. C. Popescu and H. Farid. Exposing digital forgeries by detecting traces of re-sampling. *IEEE Transactions on Signal Processing*, 53(2):758–767, 2005.

[10] S. Prasad and K. R. Ramakrishnan. On resampling detection and its application to detect image tampering. In *Proceedings of the 2006 IEEE International Conference on Multimedia and EXPO (ICME 2006)*, pages 1325–1328, 2006.

[11] P. Vandevalle, J. Kovacevic, and M. Vetterli. Reproducible research in signal processing. *IEEE Signal Processing Magazine*, 26(3):37–47, May 2009.