Multiple Classifier Systems for Image Forgery Detection

Davide Cozzolino, Francesco Gargiulo, Carlo Sansone, and Luisa Verdoliva

DIETI, University of Naples Federico II
{davide.cozzolino,francesco.grg,carlosan,verdoliv}@unina.it

Abstract. A large number of techniques have been proposed recently for forgery detection, based on widely different principles and processing tools. As a result, each technique performs well with some types of forgery, and under given hypotheses, and much worse in other situations. To improve robustness, one can merge the output of different techniques but it is not obvious how to balance the different sources of information. In this paper we consider and test several combining rules, working both at the abstract level and at measurement level, and providing information on both presence and location of suspect tampered regions. Experimental results on a suitable dataset of forged images show that a careful fusion of detector's output largely outperforms individual detectors, and that measurement-level fusion methods are more effective than abstract-level ones.

Keywords: Forgery detection, digital forensics, image tampering.

1 Introduction

Thanks to the diffusion of simple and powerful software tools for digital source editing, it is extremely simple to modify the content of an image. This has motivated, in these last years, an intense research of algorithms, to be used in the forensics field, which help deciding about the integrity of digital images. Much attention have drawn passive techniques since they require no collaboration on the part of the user through some types of watermarks or signatures. These techniques in fact are based on the observation that each step of the digital image life cycle (from the acquisition process in the camera to its recording in a compressed format and its successive editing) leaves a trace in the image, that can be extracted by the algorithm in order to reveal the tampering.

There are a large variety of approaches proposed in the literature [15], which are tipically based on some hypotheses made on the forgery. Some techniques, for example, are able to detect copy-and-move forgeries, others are designed to look for physical inconsistencies in the image, others exploit the usual adoption of some lossy compression scheme, like JPEG or take advantage of specific features of any different camera models. It is clear that no ultimate solution exists to the image forgery detection problem [4]. Each technique is based on some important hypotheses which limit its applicability, and therefore it is always possible to

A. Petrosino (Ed.): ICIAP 2013, Part II, LNCS 8157, pp. 259–268, 2013.

find cases where it fails. On the other hand, one should take for granted that a malicious tamperer, aware of the principles on which each technique works, will be able to trick it, given enough time and resources. Having a multiplicity of different tools is therefore essential to guarantee a high probability of detecting forgeries. In fact, given different approaches one can try to make a decision properly merging the results from each of them.

The first work proposing to improve forgery detection robustness through fusion applies Discriminative Random Field based methods to incorporate both local-block authenticity and inter-block inconsistency measures [9]. More recently the problem has been faced in [6][1], where a decision fusion strategy based on the Dempster-Shafer approach and the fuzzy theory has been used, respectively. These last approaches, however, work only at decision level and then are not able to locate the forgery within the image under test.

Starting form these proposals, in this work we perform an experimental performance comparison of different combination methods in order to explore how the use of assessed fusion approaches can improve the reliability of a single forensic tool. In particular, we selected five different forensic tools, which exploit quite different approaches for forgery detection, and tested several combining rules, both trainable and non-trainable ones. Moreover, we implemented the chosen rules so as to provide results at pixel level, and therefore to be able to correct locate forgeries.

The paper is organized as follows. In the next section we describe briefly the forgery detection algorithms under comparison, while in section 3 the combination methods are presented. Finally, in section 4 the experimental setting is presented and results are commented.

2 Forensic Tools

In order to evaluate the different combination techniques, a set of five forgery detection tools have been selected, that can cope with *copy and paste* forgeries, but are based on different approaches. In this way the combination process can take advantage of the specificity of each method, described briefly in the following and whose main characteristics are summarized in table 1.

Since most images are JPEG compressed, it is possible to detect traces left during the coding process. In particular, one can exploit the blocking effect introduced by JPEG, which gives rise to the so-called Block Artifact Grid (BAG). In fact, manipulating images in this format causes an alteration of these artifacts, since the BAG of the original image and that of the copied region very likely mismatch. In [12] a simple method is proposed to identify this type of forgery, named here Li-2009, the name of the first Author followed by the year of publication, a convention followed for the other techniques as well. The basic idea is to extract the horizontal and vertical edges due to JPEG blocking effect: if the image has been subject to a *copy and paste* processing a BAG mismatching can be detected when no edges are present in correspondence of the JPEG grid.

Table 1. Synthetic description of the forensic tools

	To detect	False alarms
Li-2009	inconsistency of BAG	non uniform areas
Farid-2009	traces of double compression	uniform areas
Bianchi-2011	non-aligned double JPEG artifacts	uniform and light areas
Mahdian-2008	resampling operations	regular patterns areas
Lukás-2006	inconsistency of PRNU	saturated, dark or highly textured areas

Rather than considering the blocking artifacts, in [5] (Farid-2009) the double-quantization effect is used to detect forgeries. In fact, when creating tampered images through a splicing process, images which originally were coded at different compression ratios are combined together, hence the composite image may contain a trace of the original compression quality. In particular, it is possible to reveal this double quantization by comparing the forged image with different qualities JPEG compressed versions of the image itself. The difference between the image and its JPEG-compressed counterpart presents localized local minima (JPEG ghosts), that reveal the presence of a forgery. Note that the methods described above give good results when considering JPEG low-quality images. The principle of double compression can be applied also to DCT coefficients [2], in particular Bianchi-2011 proposes new probability models for the DCT coefficients of singly and doubly compressed regions and a reliable technique for estimating them.

A different approach is followed in [14] (Mahdian-2008) and relies on finding traces of resampling in the image. The idea is based on a previous work [7], where the authors observed that when a signal is subjected to resampling, the variance shows a periodic behaviour whose period is a function of the rescaling factor. Mahdian has extended this approach to detect also rotations and skewing through the Radon Transform applied to the derivative of the involved image and followed by a periodicity search in the Fourier domain.

The last examined method relies on artifacts introduced by the digital camera, and in particular the photo-response non uniformity (PRNU) which can be considered as a sort of intrinsic fingerprint of a specific digital camera. The PRNU arises from differences and imperfections in the silicon wafer used to manufacture the imaging sensor: these physical differences provide a unique sensor fingerprint which can be used for forgery detection. It was originally proposed in [13] (Lukás-2006) and requires a large number of images taken by the digital camera itself, in order to estimate the camera PRNU. Then, to detect the tampering, the PRNU of the image under investigation is estimated and compared with the reference. This step is quite challenging, since this fingerprint is much weaker than the image, therefore a denoising step is used. In [3] we replaced the original denoising algorithm with state-of-the-art nonlocal filtering, obtaining a significant performance improvement.

Table 2. The considered Combiners

Combiner	Trainable	Required Output Type of the Base Classifiers
WMV	Yes	Abstract
BKS	Yes	Abstract
NB	Yes	Abstract
DS	Yes	Abstract or Measurement
PROD	No	Measurement
SUM	No	Measurement

3 Combination Methods

The forensic tools presented in the previous Section can be seen as classifiers that provide the most likely class (e.g. forged/non-forged), or even a class probability, for each pixel of the image under test. By considering a pool of forensic tools, a multiple classifer approach can be used in order to improve the detection performance, especially if the selected tools are based on quite different approaches. The rationale of a multiple classifier system, in fact, lies in the assumption that, by suitably combining the results of a set of base classifiers (i.e., our forensic tools), the obtained performance is better than that of any base classifier. Generally speaking, the implementation of a multiple classifier system implies the definition of a *combiner* [11] for determining the most likely class a sample should be attributed to, considering the answers of the base classifiers.

Different combiners have been proposed in the literature [11]. In the following we will give a short description of those used in this work.

We considered combiners applicable to *Type 1* classifier (i.e., a classifier that outputs just the most likely class, so working at *Abstract level* [17]), as well as combination schemes that require class probability outputs (i.e., the so-called *Type 3* classifiers that works at *Measurement level* [17][1]). As regards combination rules working at Abstract level, we considered: Weighted Majority Voting (WMV), Behaviour Knowledge Space (BKS) and Naïve Bayes Combiner (NB). Among the combiners that can exploit the measurement values provided by *Type 3* classifiers, we took into account the Product (PROD), the Sum (SUM) and the Dempster-Shafer (DS) rules.

It is also worth noting that, among the above cited combiners, SUM and PROD do not need a training phase before their use (they are called *nontrainable* combiners in [11]), whereas the other ones (*trainable* combiners) need to be trained on a suitable set of data. The main characteristics of the chosen combiners are summarized in Table 2.

Before entering in details, it is worth recalling that some *trainable* combiners make use of the so-called *confusion matrix* [17] for combining *Type 1* classifiers.

[1] For the sake of completeness let us recall that *Type 2* classifiers operate at *Rank level*, providing as output (a subset of) all the possible classes, with the alternatives ranked in order of plausibility of being the correct class.

The classification confusion matrix E^k is such that the generic element e^k_{ij} ($1 \leq i, j \leq m$, where m is the number of the classes) represents in our case the percentage of pixels belonging to the i-th class that the k-th tool assigns to the j-th class. Therefore, the value e^k_{ii} represents the percentage of pixels belonging to the i-th class which are correctly classified by the k-th tool. The values of the elements of E^k have to be computed on a suitable set of images.

In case of majority voting the guess class is the one voted by the majority of the classifier. In general, if more classes obtain the same number of votes, the values e^k_{ii} are used for tie breaking, i.e. the vote of each classifier is weighted by the number representing the confidence degree of that classifier when it assigns a pixel to the class it is voting for. The confidence degree evaluated by means of the confusion matrices was used by **Weighted Majority Voting** (*WMV*) for weighting the votes given by each classifier. The combiner assigns each pixel to the class C such that: $C = \arg\max_i \sum_k e^k_{ii} \cdot V^k_i$, where V^k_i is 1 if the guess class of the k-th classifier is i and 0 otherwise.

A **Behavior-Knowledge Space** (*BKS*) is a N-dimensional space where each dimension corresponds to the decision of a classifier. Given a pixel to be assigned to one of the 2 possible classes, the ensemble of the classifiers can in theory provide 2^N different decisions. Each one of these decisions constitutes one *unit* of the BKS. In the training phase each BKS *unit* can record 2 different values c_i, one for each class. Given a suitably chosen data set, each pixel x of this set is classified by all the classifiers and the *unit* that corresponds to the particular classifiers' decision (called *focal unit*) is activated. It records the actual class of x, say j, by adding one to the value of c_j. At the end of this phase, each *unit* can calculate the best representative class associated to it, defined as the class that exhibits the highest value of c_i. It corresponds to the most likely class, given a classifiers' decision that activates that *unit*. In the operating mode, the BKS combiner acts as a look-up table. For each pixel x to be classified, the N decisions of the classifiers are collected and the corresponding *focal unit* is selected. Then x is assigned to the best representative class associated to its *focal unit*. Note that the BKS combiner sometimes suffers from the overtraining problem [11].

For the **Naïve Bayes** (*NB*) combiner the guess class is instead the one which maximizes the *a posteriori* probability. Applying the Bayes' formula and standing the assumption of the independence of the classifiers, it can be simply shown, starting from the results presented in [11], that the class C which maximizes the *a posteriori* probability is: $C = \arg\max_i M_i \cdot \prod_{k=1}^{N} e^k_{ij}$, where M_i is the number of samples belonging to the i-th class, N is the number of classifiers and j is the guess class provided by the k-th classifier.

While *NB* uses the values of the confusion matrix in order to estimate the *a posteriori* probability for each class, the **Sum** (*SUM*) and the **Product** (*PROD*) Rules directly use the continuous outputs provided by the tools for making such an estimate. In particular, SUM (respectively, PROD) calculates for each class the sum (the product) of the class probabilities provided by the different tools, and assign the pixel under test to the class which maximizes this sum (product).

Fig. 1. Some examples of tampered images

In [10] a formal derivation of these rules from the Bayesian framework has been presented. It is worth noting that PROD is sometimes oversensitive to estimates close to zero. On the other hand, SUM is derived under strongest assumptions, which could not be verified in practical cases.

The **Dempster-Shafer** combiner (*DS*) is based on the Dempster-Shafer theory [8]. It has been frequently applied to deal with uncertainty management and incomplete reasoning. Differently from the classical Bayesian theory, the DS theory can explicitly model the absence of information, whereas Bayesian approaches assign the same probability to all the possible events in case of absence of information. According to the DS theory, a basic probability assignment (bpa) can be associated to each base classifier, which describes the subjective degree of confidence attributed to it. What is modeled, then, is not the analyzed phenomenon, but the belief in how good the base classifiers are at reporting about it. The typical formulation of the DS combiner [17] can be used for *Type I* classifiers; in order to exploit the fact that the considered tools can also output a probability value for each class, we implemented a modified version, as proposed by Fontani et al. (but without exploiting the so-called *tool compatibility*) [6].

4 Experimental Results and Discussion

In this section we report and discuss the performance of the fusion process. Experiments have been conducted on 200 photos (1024×1024 pixels) subject to copy-paste forgeries cropped from the images of the Uncompressed Colour Image Database (UCID) [16]. We considered various scenarios: compressed forgeries at different level of quality (low, medium and high) and uncompressed forgeries subject to resampling operations, such as scaling and rotation. In Fig. 1 we show some examples of tampered photos from our dataset.

First, we present the results obtained by running each tool individually. The output of each technique has been converted into a probability index map, with continuous values in the range [0,1]. After a thresholding operation and a morphological processing (all regions smaller than 0.2% of the whole image are attributed to random errors and removed) we obtain a binary map that, eventually, allows us not only to detect the presence of forgeries but also locate them in the image. For each tool, we compute on the entire database several performance measures (Table 3). In particular, results are given in terms of sensitivity, specificity, harmonic mean of sensitivity and specificity and accuracy, defined as:

Table 3. Results (in percent) obtained by each tool individually

Methods	sensitivity	specificity	harmonic mean	accuracy
Li-2009	**91.59**	45.24	60.56	47.21
Farid-2009	37.70	90.02	53.14	87.80
Bianchi-2011	59.29	**95.17**	73.07	**93.65**
Mahdian-2008	37.84	82.09	51.80	80.21
Lukás-2006	66.93	92.93	**77.81**	91.82

Table 4. Results obtained by combining all the forensic tools

Combiner	sensitivity	specificity	harmonic mean	accuracy
WMV	65.79	98.05	78.75	96.68
NB	**83.60**	92.41	87.78	92.03
BKS	47.86	**99.57**	64.64	**97.37**
DS	77.69	96.27	85.99	95.48
PROD	**83.39**	95.83	**89.18**	95.30
SUM	82.48	96.37	**88.88**	95.77

$$\text{sensitivity} = \frac{TP}{TP+FN} \qquad \text{specificity} = \frac{TN}{TN+FP}$$

$$\text{harmonic mean} = 2\frac{\text{sensitivity} \cdot \text{specificity}}{\text{sensitivity} + \text{specificity}}$$

$$\text{accuracy} = \frac{TP+TN}{TN+FP+TP+FN}$$

with $TP[FP]$ the true[false] positive rate, and $TN[FN]$ true[false] negative rate.

The performance is not always satisfactory. Only Li-2009 provides a large sensitivity to forgeries, but at the price of unacceptably low specificity, under 50%. Bianchi-2011 and Lukás-2006 guarantee a more reasonable compromise, with very good specificity and acceptable sensitivity, as synthesized by the pretty large harmonic mean. and available only in the presence of the camera or a large collection of photos taken by it, a requirement not always met in practical applications. For this reason, when considering the fusion process applied to these forensic tools, we will present results both with and without the PRNU-based algorithm to account for the different situations.

In table 4 we present results obtained through the combination of all techniques, including Lukás-2006, using the various fusion approaches described before. It is immediately clear that the fusion process, in general, grants a significant performance improvement. All individual detectors, even the best ones, are dominated by several combiners, with the only exception of sensitivity for Li-2009. Among the abstract-level combiners (top part of the table) the Naïve

Table 5. Results obtained by the combiners when excluding the PRNU-based tool

Combiner	sensitivity	specificity	harmonic mean	accuracy
WMV	43.58	98.17	60.36	95.85
NB	68.46	95.31	79.68	94.17
BKS	22.12	**99.93**	36.23	**96.62**
DS	58.41	97.36	73.02	95.71
PROD	**73.99**	95.58	**83.41**	94.66
SUM	68.14	96.81	79.98	95.59

Bayes is clearly superior to the others, gaining almost 10 percent points on the harmonic mean w.r.t. the second best. Some measurement-level combiners, however, in particular the PROD and SUM, provide a further small gain, especially in terms of specificity. By excluding the PRNU-based tool, one of the most reliable when applicable, the performance of all combiners drops significantly, as shown in table 5. However, several of them, notably the abstract-level Naïve Bayes, and the measurement-level PROD and SUM, keep providing very good results, superior to those of individual tools, including the PRNU-based one.

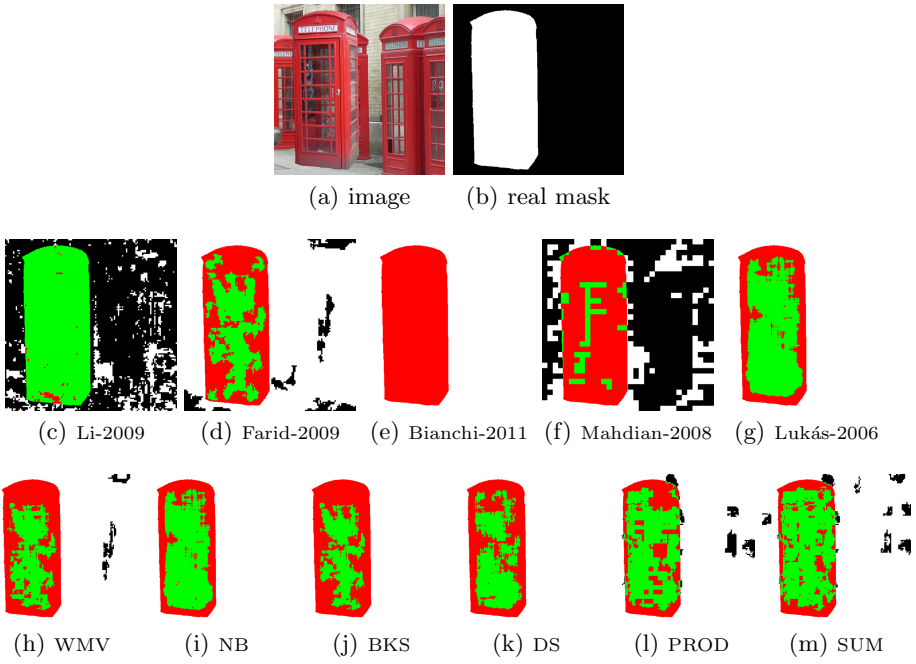

Fig. 2. Example 1: masks obtained by the single tools and the combination approaches. Green: TP, White: TN, Red: FN, Black: FP.

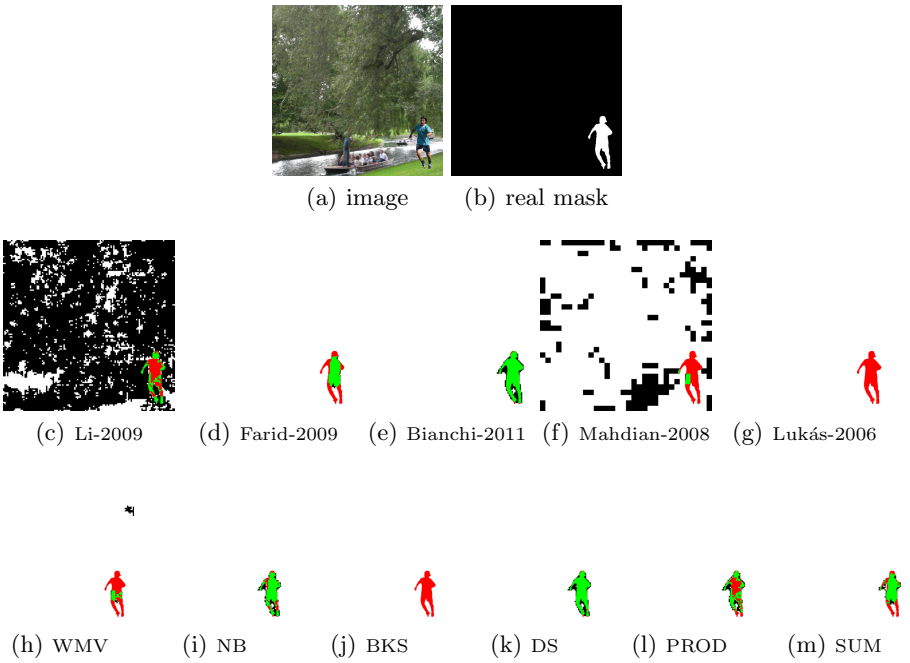

Fig. 3. Example 2: masks obtained by the single tools and the combination approaches. Green: TP, White: TN, Red: FN, Black: FP.

Figures 2 and 3 provide some more insight into the performance of the various individual tools and combiners. In particular, the information fusion increases robustness against unfavorable conditions that might totally foul some techniques. In Fig.2, for example, Bianchi-2011, one of the best tools on the average, misses the forgery altogether, because both the forgery and the image are uncompressed, making the detection tool completely ineffective. Other tools, however, make up for this failure, and most combiners detect the forgery with a high degree of accuracy. The situation is reversed in Fig.3. Lukás-2006 now misses the forgery, too small to trigger the PRNU detector but, again, several combiners provide the correct detection map.

Both the tables and the examples clearly show the advantage of merging the output of multiple detectors, based on alternative approaches and sensitive to different properties of images and forgeries. In particular, working at the measurement level grants a claer advantage, as one can take into account information about the reliability of the each individual classification act to take a decision. In addition, applying morphological operators after the fusion is certainly a good idea. This work establishes the ground for future and more stimulating research showing that there is still space for improvement. Besides including more and more diverse individual tools, it will be fundamental to define a merging rule that takes into full account the specificities of the forgery detection problem.

References

1. Barni, M., Costanzo, A.: A fuzzy approach to deal with uncertainty in image forensics. Signal Processing: Image Communication 27(9), 998–1010 (2012)
2. Bianchi, T., De Rosa, A., Piva, A.: Improved dct coefficient analysis for forgery localization in jpeg images. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2444–2447 (May 2011)
3. Chierchia, G., Parrilli, S., Poggi, G., Sansone, C., Verdoliva, L.: On the influence of denoising in PRNU based forgery detection. In: ACM Workshop on Multimedia in Forensics, Security and Intelligence, Firenze, Italy, pp. 117–122 (2010)
4. Cozzolino, D., Poggi, G., Sansone, C., Verdoliva, L.: A comparative analysis of forgery detection algorithms. In: Gimel'farb, G., Hancock, E., Imiya, A., Kuijper, A., Kudo, M., Omachi, S., Windeatt, T., Yamada, K. (eds.) SSPR&SPR 2012. LNCS, vol. 7626, pp. 693–700. Springer, Heidelberg (2012)
5. Farid, H.: Exposing Digital Forgeries From JPEG Ghosts. IEEE Transactions on Information Forensics and Security 4(1), 154–160 (2009)
6. Fontani, M., Bianchi, T., De Rosa, A., Piva, A., Barni, M.: A dempster-shafer framework for decision fusion in image forensics. In: IEEE International Workshop on Information Forensics and Security (WIFS), November 29-December 2, pp. 1–6 (2011)
7. Gallagher, A.: Detection of linear and cubic interpolation in jpeg compressed images. In: The 2nd Canadian Conference on Computer and Robot Vision, pp. 65–72 (May 2005)
8. Gordon, J., Shortliffe, E.: The dempster-shafer theory of evidence. In: Buchanan, B.G., Shortliffe, E. (eds.) Rule-Based Expert Systems, pp. 272–292 (1984)
9. Hsu, Y.F., Chang, S.F.: Statistical fusion of multiple cues for image tampering detection. In: Asilomar Conference on Signals, Systems, and Computers (2008)
10. Kittler, J., Hatef, M., Duin, R., Matas, J.: On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell. 20(3), 226–239 (1998)
11. Kuncheva, L.: Combining Pattern Classifiers: Methods and Algorithms. Wiley-Interscience (2004)
12. Li, W., Yuan, Y., Yu, N.: Passive detection of doctored JPEG image via block artifact grid extraction. Signal Processing 89(9), 1821–1829 (2009)
13. Lukas, J., Fridrich, J., Goljan, M.: Detecting digital image forgeries using sensor pattern noise. In: Proceedings of the SPIE, vol. 6072, SPIE (2006)
14. Mahdian, B., Saic, S.: Blind Authentication Using Periodic Properties of Interpolation. IEEE Transactions on Information Forensics and Security 3(3), 529–538 (2008)
15. Piva, A.: An Overview on Image Forensics. ISRN Signal Processing pp. 1–22 (2012)
16. Schaefer, J., Stich, M.: Ucid - an uncompressed colour image database. In: Proc. SPIE, Storage and Retrieval Methods and Applications for Multimedia, pp. 472–480 (2004)
17. Xu, L., Krzyzak, A., Suen, C.: Methods of combining multiple classifiers and their applications to handwriting recognition. IEEE Transactions on Systems, Man, and Cybernetics 22(3), 418–435 (1992)