Context-aware forgery localization in social-media images: a feature-based approach evaluation

Cédric Maigrot, Ewa Kijak, Vincent Claveau

CONTEXT-AWARE FORGERY LOCALIZATION IN SOCIAL-MEDIA IMAGES: A FEATURE-BASED APPROACH EVALUATION

Cédric Maigrot *Ewa Kijak* *Vincent Claveau*

Univ Rennes, Inria, CNRS, IRISA, Rennes, France

ABSTRACT

In this paper, we study context-aware methods to localize tamperings in images from social media. The problem is defined as a comparison between image pairs: an near-duplicate image retrieved from the network and a tampered version. We propose a method based on local features matching, followed by a kernel density estimation, that we compare to recent similar approaches. The proposed approaches are evaluated on two dedicated datasets containing a variety of representative tamperings in images from social media, with difficult examples. Context-aware methods are proven to be better than blind image forensics approach. However, the evaluation allows to analyze the strengths and weaknesses of the contextual-based methods on realistic datasets.

Index Terms— Forgery localization, Image forensics, Splicing, Context-aware digital forensics, Image tampering

1. INTRODUCTION

Nowadays images are widely spread through social networks and the dissemination of false information is a scourge. In particular, modifying image content is an increasingly popular practice thanks to the widely available, user-friendly image editing software. Therefore, there is a real need for lay or professional users to have tools facilitating the detection of fraudulent images. To address the task of image content verification, any piece of information is valuable, such as geolocation, camera model, contextual information, applied filters, and explicit tampering detection. Recently, a few frameworks or platforms tackle this problem [1, 2, 3].

Various types of modification can occur. One can distinguish *malicious* from *non-malicious tamperings*. The first ones try to change the way humans interpret the content of the images: some visual elements are added, removed, or changed, such that the original semantic of the image is altered. This is typically performed by a copy-move operation (a part of the image is replicated at another location in the same image), or by a copy-paste operation (or *splicing*; the inserted object comes from another image). In contrast, *non-malicious tamperings* are more concerned with the aesthetics of the images Filtering, cropping, rescaling or histogram adjustments are examples of broad-scope image operation, which are generally considered as non-malicious tamperings in the context of images circulating on the Web.

Tampering detection and localization in images has drawn a lot of attention. Passive, blind image forensics are based on single image analysis. Many efforts have been devoted to copy-move detection, where the general principle consists in finding similar blocks, or regions, in the image. Such approaches rely on blocks or keypoints matching to find the highest similarities in the image. Then, a post-processing step groups the matched pairs into coherent candidate regions. The estimation of the parameters of possible affine transforms is used to ensure the robustness of the approach to transformations [4, 5, 6, 7]. Unlike copy-move, splicing detection relies on the assumption that the spliced area significantly differs from the rest of the image. Identifying local discrepancies can rely on traces left by JPEG compression [8, 9, 10, 11], Color Filter Array (CFA) interpolation [12], or noise patterns [13, 14]. Tampering localization methods generally produce a Tampering Heat Map (THM) that indicates regions of the image where tampering is likely to occur. In this work, we are interested in the problem of splicing localization on images found on the Web and social media environments. Such images have particular characteristics. They are frequently subjected to transformations, such as rescaling, filtering, cropping, format conversion, recompression. Users often modify an image before re-posting it, and some image publishing platforms, *e.g.* Facebook or Twitter, automatically operate such transformations, *e.g.* scaling, changing format to JPEG, varying quality, and erasing metadata. These images, posted again and again by users are challenging traditional signal-based method for tampering detection, which are known to perform poorly in this context [15, 16]. While such methods could help in some cases, they certainly can not cope with all the range of user and system visual edits, suggesting that complementary mechanisms should be used to uncover more forgeries.

An emerging paradigm in verifying information from the social media is that several versions of an image are likely to be found on the Web. Finding these images facilitates the tampering detection by using comparisons. Reverse image search can often retrieve near-duplicate versions of the image to be analyzed and then serve as a basis for further manual inspection [1], or as an input for automatic tampering detection based on comparison between pairs of images [17, 18, 16, 19, 20]. The focus of this paper is on this latter case. The paper proposes an approach based on local features in order to localize tampering in images from social media, under the assumption that we are able to crawl the web for similar images and that the source image can be retrieved. In Section 2, we propose a splicing localization method based on matching of local features. Section 3 presents the experimental setup and introduces a realistic dataset we collected, containing characteristic difficult examples and available online for free academic usage and accessible reproducible research. We also use another recent dataset to evaluate the proposed method as well as methods from the state of the art in Section 4, to analyze the strengths and weaknesses of the contextual-based methods on realistic datasets.

2. FORGERY LOCALIZATION METHOD

Localizing regions in an image where tamperings are likely is based on detecting the inconsistencies resulting from a detailed comparison of two versions of the same image: the probe image $\mathcal{F}$ and a retrieved near-duplicate $\mathcal{C}$. In the datasets used for evaluation (see section 3), $\mathcal{F}$ is the forged image while $\mathcal{C}$ is considered as the original one. The tampered areas in images from the web and social medias have most likely undergone transformations such as rotations, scaling, cropping, or affine transformations. The proposed method hence relies on the matching of local features that are robust to such transformations. Matches are then used to both estimate the homography H between the two images, and identify outliers as matches that do not fit H.

1. Local Features Matching based approach (LFM) In our implementation, we use SURF as local features. SURF features are matched from $\mathcal{F}$ to $\mathcal{C}$ according to Lowe's criterion as introduced in [21] that allows to discard most of the false matches. Based on these matches, the homography H between the two images is estimated using RANSAC, which is robust to outliers. We evaluate both dense and detected keypoints extraction. Keypoint detection often results in a higher rate of matches fitting the homography, but provides fewer points, and consequently less matches, which causes troubles in many images. Dense keypoint extraction is retained (stride of 10 pixels) and descriptors are computed over 4 different scales, as in other recognition problems [22].

Once H estimated, outliers are the matches H does not explain. The best candidate match for each keypoint of $\mathcal{F}$ is its nearest neighbor in the keypoints from $\mathcal{C}$. Dense sampling ensures dense outliers detection, and therefore facilitates their localization. A keypoint of $\mathcal{F}$ is considered as an outlier if the distance between its re-projection in $\mathcal{C}$ according to H and the matched keypoint for all scales is above $0.1 \times \text{diag}$, where diag is the diagonal size of the image. A keypoint is an inlier if at least one matched keypoint at one scale is correctly located according to H (Fig. 1d).

The THM is finally obtained by estimating the density of outliers, using a Kernel Density Estimation with a gaussian kernel (*gKDE*), the bandwidth being selected by Scott's Rule of thumb. The objective of this last step is to identify the spatially close outliers and remove the isolated ones. Examples are given in Figs. 1e and 1f. We compare this approach to a simpler one, called *morpho*, consisting in applying morphological operators (dilation, open-close filtering, hole filling) to the binary map given by the outlier points. This can be very roughly seen as an approximation of a kernel density estimation with a uniform (rectangular) kernel (Fig. 1g).

In some particular cases, H is wrongly estimated: when $\mathcal{F}$ has been flipped, or when the retrieved original image $\mathcal{C}$ is a cropped version of $\mathcal{F}$: the area in $\mathcal{F}$ that does not exist in $\mathcal{C}$ will be detected as a tampering while it's not a splicing. These cases are easily detectable after the estimation of H, and prevented by flipping or cropping $\mathcal{F}$ according to H.

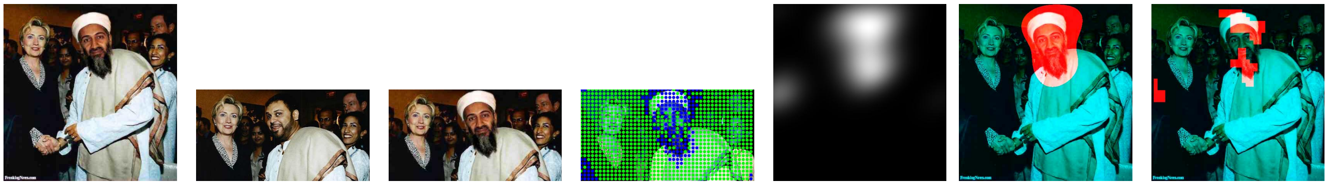
2. Pixel-wise comparison An image comparison framework was also proposed in [16]. We compare the LFM approach to the two best performing methods in [16], namely `IRPSNR` (based on Peak Signal to Noise Ratio), and `SSIM` (based on Structural Similarity Index Measure). We use the code and parameters provided by the authors. For `IRPSNR`, the THM is computed as the pixel-wise PSNR between the Gaussian blurred versions of the two images (with standard deviation $\sigma = 4$). The `SSIM` THM is the pixel-wise SSIM between the two images, using a neighborhood radius of 32 pixels. In both methods, low values of the THM indicate probable tampered areas. As these methods are based on pixel-wise computations, a warp operation is required to transform one of the images. In [16], $\mathcal{C}$ is warped according to the homography H' that map points of $\mathcal{C}$ to the coordinate system of $\mathcal{F}$. In our experiments, as for the LFM approach, $\mathcal{F}$ can be warped when $\mathcal{C}$ is detected as a cropped version of $\mathcal{F}$.

3. EXPERIMENTAL SETUP

1. Image datasets. We use two datasets dedicated to splicing localization. 1) The *WildWeb* dataset (**Wi**), proposed by [3, 15], contains real-world forgeries that have circulated on the web in the last years, accompanied by their original, untampered sources, when those were found by reverse-image search engines. We filtered out cases for which the original, untampered sources are not available and end up with 82 cases of forgeries (65 unique cases) along with their unspliced sources and ground-truth binary masks. In case there are multiple ground truths for a forged image (corresponding to different splices, possibly committed at different times), we merge them. 2) The *Reddit* dataset (**Re**), that we built and made publicly available[1], is composed of 107 photoshopped images and their original version from the Photoshop challenges on the *Reddit* website[2], on which we

[1] http://hoaxdetector.irisa.fr/

[2] http://www.reddit.com



(a) forged $\mathcal{F}$ (b) original $\mathcal{C}$ (c) cropped $\mathcal{F}$ (d) outliers (e) THM by *gKDE* (f) bin. *gKDE* (g) bin. map *morpho*

Fig. 1. Example of results of the tampering detection method. In (d): outliers key points are in blue, inliers in green.

manually annotated the tampered regions. The tampering operations are mainly splicing of various size, from very small tampered areas to very large ones, when for example the whole background of the original image has been replaced. Following the policy used on download media platforms such as Twitter or Facebook, large images are scaled down such that one of the dimension is at most 900 pixels. Fig. 2 shows examples of the 189 image pairs obtained.

2. Performance of the localization. Evaluation of the performance consists in comparing a ground-truth mask with a binarized THM, and is thus sensitive to both the binarization threshold and the quality of the ground-truth mask. Different metrics are used, each evaluating different properties. For each metric, the THMs were binarized using multiple threshold values covering the entire range of possible map values.

Pixel-wise score: One common evaluation method for localization is based on pixel-wise comparison between binarized output and the ground-truth mask. The tampering localization is evaluated by the False Positive Rate (FPR) and the False Negative Rate (FNR), which should both be minimized. False positives pixels may either come from an unfavorable binarization threshold giving regions too large compared to the ground-truth, or from a wrongly detected region. Thus, the same error rate does not have the same meaning in both cases. We can make the same observation for false negatives. Another inherent difficulty is that such evaluation does not take into account the connected components and the ability to separate different (close) objects.

Connected component (CC)-wise score: For these reasons, we also propose a metric based on connected component (CC) that penalizes over-detections. A one-to-one assignment is made between the detected output CCs and the ground-truth CCs. A detected CC is considered as a correct detection if its overlap with a ground-truth CC, defined by the intersection-over-union (IoU) criterion, is greater than 0.5. Hence each detected CC is either a true positive or a false positive, and each ground-truth CC is either a true positive or a false negative. In CC-wise comparison, we compute the precision, the recall and the F-measure which should all be maximized.

WW score: For the sake of comparison, we also use the evaluation methodology presented in [3]. The similarity between a binarized THM and the ground-thruth mask is also a per-pixel performance E (see [3]). Any binarized THM achieving an E larger than a given threshold is considered as an accurate detection. Algorithm performance is given by the number of cases considered as correctly detected.

4. RESULTS AND ANALYSIS

Contextual based methods efficiency. We evaluate the three presented contextual methods on the $\mathbf{Wi}$ dataset, using the same evaluation protocol as in [3]. Table 1 presents the results in terms of the number of detected cases w.r.t. the total number of considered cases, along with the three best performances reported in [3]. This confirms the importance of using contextual clues when they are available.

Table 1. Tampering localization on $\mathbf{Wi}$: ratio between the number of detected cases and the number of cases, using WW score with $E > 0.45$

`LFM`	`IRPSNR`	`SSIM`	GHO [11]	NOI1 [14]	ADQ1 [8]
0.60	0.72	0.79	0.35	0.18	0.16

Tamperings localization. Evaluation results at pixel level are given in Fig. 3 for the two datasets. The binarization threshold for the THMs is controlled by a parameter σ, ranging from 0 to 100, which represents the percentage of the maximum possible value THM_{max}. For `LFM`, the larger σ, the smaller the detected area, while the opposite applies for `IRPSNR` and `SSIM`. `LFM` performs slightly better than `SSIM` on Reddit and `IRPSNR` is always worse. On the other hand, the CC-based evaluation, reported in Tab. 2, gives another perspective.

Table 2. F1-measure on $\mathbf{Wi}$ and $\mathbf{Re}$, with CC-wise score based on the best possible binary THM per image.

Dataset	`LFM`	`IRPSNR`	`SSIM`
$\mathbf{Re}$	**0.84**	0.62	0.69
$\mathbf{Wi}$	0.21	**0.40**	0.21
All images	**0.63**	0.53	0.48

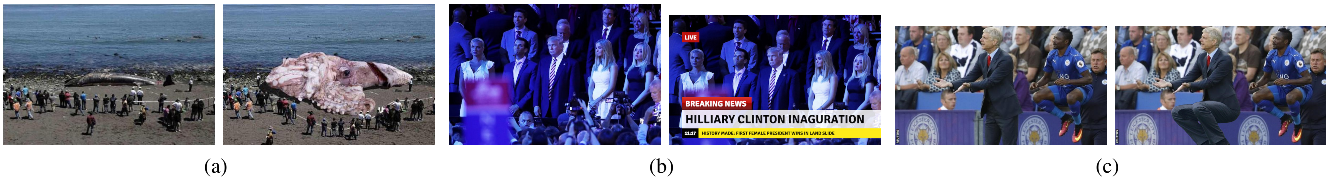


Fig. 2. Examples of 3 pairs of images: Example (a) corresponds to images from $\mathbf{Wi}$, examples (b) and (c) correspond to images from $\mathbf{Re}$. For each pair, the original image is on the left side, and the tampered one is on the right.

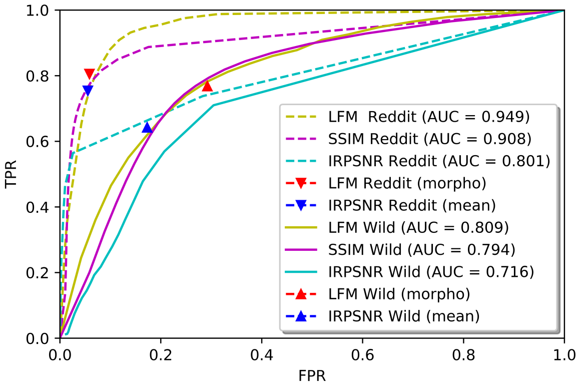
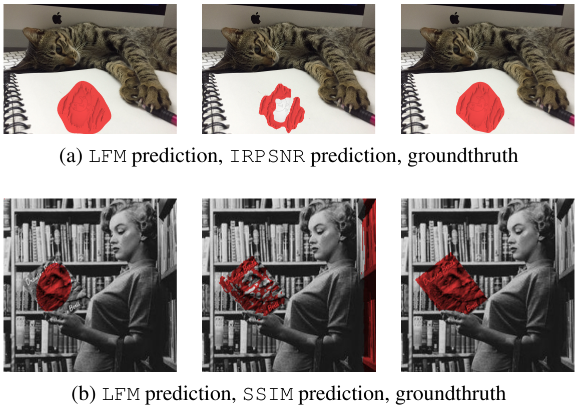


Fig. 3. ROC curve of generated binary THMs, with Pixel-wise score on $\mathbf{Wi}$ and $\mathbf{Re}$ datasets. Red and blue triangles are performances of LFM morpho, and IRPSNR binarized with a fixed threshold set to the mean THM value.



(a) LFM prediction, IRPSNR prediction, groundthruth

(b) LFM prediction, SSIM prediction, groundthruth

Fig. 4. Two examples of errors of methods IRPSNR and SSIM compared with the LFM binarized prediction. Red areas are the predictions of methods and groundthruth.

The gaussian KDE approach used by LFM to cluster outliers tends to produce large and homogeneous detected areas. In contrast, IRPSNR and SSIM tend to over-segment, which is penalized by the CC-wise score. This behaviour is illustrated in Fig. 4. Even if the binary maps can be improved with morphological operations, this post-processing step is sensitive to the structuring elements used and induces new wrong localizations while correcting some defects.

Sensitivity to matching. The studied approaches perform relatively well, strongly depend on the quality of the estimated homography H between $\mathcal{F}$ and $\mathcal{C}$; For IRPSNR and SSIM because they rely on pixel-wise comparison and thus on the quality of the warping, and for LFM because the THM is based on outliers w.r.t. H. IRPSNR and SSIM have to be particularly careful to the registration step, and simply using a warp from $\mathcal{F}$ to $\mathcal{C}$ as in [16] is often insufficient, when dealing with images cropped in different ways.

Results are very different between $\mathbf{Wi}$ and $\mathbf{Re}$. We note two main causes of failure on $\mathbf{Wi}$. The first is when the proposed near-duplicate image $\mathcal{C}$ is quite different from the forged one $\mathcal{F}$ (*e.g.* same location but different point of view, or different period). The matching, and therefore the homography estimation, usually fail. The second is when the near-duplicate image $\mathcal{C}$ is the donor i.e. the image containing the (usually small) element used for the splicing (alien), and not the host image (i.e. which received the alien, producing the forgery). When it occurs the comparable part between the images $\mathcal{F}$ and $\mathcal{C}$ corresponds to the alien, and the forgery will be detected as all the rest of the image.

5. CONCLUSION

In this paper, we propose a context-aware method to localize tamperings in images from social media. The tackled problem is placed in the context of comparison of images, assuming to have an original version of the image to be analyzed, retrieved on the Net. The method is based on local features (e.g. SURF) matching. KDE or morphological post-processings are proposed and compared for producing THMs. This approach is compared to state-of-the-art methods. We evaluate the approaches on two datasets containing real complex examples covering a large variety of tampering in images from social media. Our study allows us to observe the general strengths and limitations of local feature-based approaches, and underline the sensitivity of the results to the evaluation protocol. The proposed task and datasets are very challenging and should grow in interest in the community as there is room for improvements.

6. REFERENCES

[1] C. Pasquini, C. Brunetta, A. F Vinci, V. Conotter, and G. Boato, "Towards the verification of image integrity in online news," in *IEEE Int. Conf. on Multimedia & Expo Workshops*, 2015.

[2] M. Zampoglou, S. Papadopoulos, Y. Kompatsiaris, R. Bouwmeester, and J. Spangenberg, "Web and social media image forensics for news professionals," in *Social Media in the Newsroom, (ICWSM) Workshop*, 2016.

[3] M. Zampoglou, S. Papadopoulos, and Y. Kompatsiaris, "Detecting image splicing in the wild (WEB)," in *IEEE Int. Conf. on Multimedia & Expo Workshops*, 2015.

[4] V. Christlein, C. Riess, J. Jordan, C. Riess, and E. Angelopoulou, "An evaluation of popular copy-move forgery detection approaches," *IEEE Trans. on Information Forensics and Security*, vol. 7, no. 6, pp. 1841–1854, 2012.

[5] C. Pun, X. Yuan, and X. Bi, "Image forgery detection using adaptive oversegmentation and feature point matching," *IEEE Trans. on Information Forensics and Security*, vol. 10, no. 8, pp. 1705–1716, 2015.

[6] D. Cozzolino, G. Poggi, and L. Verdoliva, "Efficient dense-field copy-move forgery detection," *IEEE Trans. on Information Forensics and Security*, vol. 10, no. 11, pp. 2284–2297, 2015.

[7] I. Amerini, L. Ballan, R. Caldelli, A. Del Bimbo, and G. Serra, "A sift-based forensic method for copy–move attack detection and transformation recovery," *IEEE Trans. on Information Forensics and Security*, vol. 6, no. 3, pp. 1099–1110, 2011.

[8] Z. Lin, J. He, X. Tang, and C. Tang, "Fast, automatic and fine-grained tampered jpeg image detection via dct coefficient analysis," *Pattern Recognition*, vol. 42, no. 11, 2009.

[9] T. Bianchi and A. Piva, "Image forgery localization via block-grained analysis of jpeg artifacts," *IEEE Trans. on Information Forensics and Security*, vol. 7, no. 3, pp. 1003–1017, 2012.

[10] I. Amerini, R. Becarelli, R. Caldelli, and A. Del Mastio, "Splicing forgeries localization through the use of first digit features," in *IEEE Int. Workshop on Information Forensics and Security (WIFS)*, 2014, pp. 143–148.

[11] H. Farid, "Exposing digital forgeries from jpeg ghosts," *IEEE Trans. on Information Forensics and Security*, vol. 4, no. 1, pp. 154–160, 2009.

[12] P. Ferrara, T. Bianchi, A. De Rosa, and A. Piva, "Image forgery localization via fine-grained analysis of cfa artifacts," *IEEE Trans. on Information Forensics and Security*, vol. 7, no. 5, pp. 1566–1577, 2012.

[13] D. Cozzolino, G. Poggi, and L. Verdoliva, "Splicebuster: A new blind image splicing detector," in *IEEE Int. Workshop on Information Forensics and Security (WIFS)*, 2015.

[14] B. Mahdian and S. Saic, "Using noise inconsistencies for blind image forensics," *Image and Vision Computing*, vol. 27, no. 10, pp. 1497 – 1503, 2009.

[15] M. Zampoglou, S. Papadopoulos, and Y. Kompatsiaris, "Large-scale evaluation of splicing localization algorithms for web images," *Multimedia Tools and Applications*, vol. 76, no. 4, pp. 4801–4834, 2017.

[16] J. Brogan, P. Bestagini, A. Bharati, A. Pinto, D. Moreira, K. Bowyer, P. Flynn, A. Rocha, and W. Scheirer, "Spotting the difference: Context retrieval and analysis for improved forgery detection and localization," *IEEE Int. Conf. on Image Processing*, 2017.

[17] I. Amerini, R. Becarelli, R. Caldelli, and M. Casini, "A feature-based forensic procedure for splicing forgeries detection," *Mathematical Problems in Engineering*, vol. 2015, 2015.

[18] C. Maigrot, E. Kijak, R. Sicre, and V. Claveau, "Tampering detection and localization in images from social networks: A cbir approach," in *Int. Conf. on Image Analysis and Processing*, 2017, pp. 750–761.

[19] L. Gaborini, P. Bestagini, S. Milani, M. Tagliasacchi, and S. Tubaro, "Multi-clue image tampering localization," in *2014 IEEE Int. Workshop on Information Forensics and Security (WIFS)*, 2014, pp. 125–130.

[20] P. Bestagini, M. Tagliasacchi, and S. Tubaro, "Image phylogeny tree reconstruction based on region selection," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 2059–2063.

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

[22] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods.," in *BMVC*, 2011, vol. 2, p. 8.