Tampering detection and localization in images from social networks: A CBIR approach

Cédric Maigrot, Ewa Kijak, Ronan Sicre, Vincent Claveau

Tampering detection and localization in images from social networks: A CBIR approach

Cedric Maigrot[1], Ewa Kijak[1], Ronan Sicre[2], and Vincent Claveau[2]

[1] Univ. Rennes I, UMR 6074 IRISA, Rennes, France
[2] CNRS, UMR 6074 IRISA, Rennes, France
{cedric.maigrot,ewa.kijak,ronan.sicre,vincent.claveau}@irisa.fr

Abstract. Verifying the authenticity of an image on social networks is crucial to limit the dissemination of false information. In this paper, we propose a system that provides information about tampering localization on such images, in order to help either the user or automatic methods to discriminate truth from falsehood. These images may be subjected to a large number of possible forgeries, which calls for the use of generic methods. Image forensics methods based on local features proved to be effective for the specific case of copy-move forgery. By taking advantage of the number of images available on the internet, we propose a generic system based on image retrieval, followed by image comparison based on local features to localize any kind of tampering in images from social networks. We also propose a large and challenging adapted database of real case images for evaluation.

Keywords: Tampering detection and localization, Tweet image analysis, Image forgery, copy-move and splicing detection, matching

1 Introduction

Massive amounts of information are spread over social networks, and among them a large quantity of fake information is conveyed. Messages are often composed of images or videos associated with text. Cases of misinformation take many forms: images can be modified for malicious purpose, or original images can be reused in a wrong context. Detecting such manipulations is now a key issue, and such process usually requires to examine the several modalities to get some contextual information about the transmission channel as well as information from the web. In this work, we focus on the visual aspect of this problem, and we are interested in automatically providing clues about images exchanged on the social networks.

Images may have undergone different types of modifications: some of them are malicious, like duplication of some parts of the image (known as copy-move attack), inserting a region from another image (copy-paste or splicing attack), or deleting some regions (thanks to techniques as inpainting or seam carving); but images posted on social networks can also typically be submitted to editing process, such as combination of several images into one, adding of text or shapes

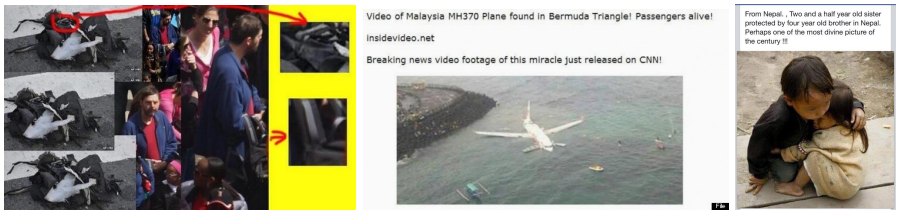


Fig. 1: Examples of images in social networks

(arrows, circles, etc.), aesthetic filters, or simply cropped or re-compressed, see Fig. 1. Rather than only classifying an image as modified or pristine, we are interested in detecting and localizing any type of modifications.

Many studies in the image forensics field tackle the problem of assessing the authenticity of digital images. In the traditional forensics paradigm, no external information but the image is available. This is a difficult task, and forensics methods can usually only cope with copy-move attacks, and are evaluated on clean dedicated databases. We adopt a different paradigm as we rely on the access to external information such as image databases, or Web reverse image search. Indeed, one of the first step in manual checking of image integrity is to search it (or modified versions) on the Web[3], and there's no reason to refuse this information, in particular in the context of social network use. The problem is thus assimilated to a comparison task between pairs of images, which can handle various tampering operations, at a lower cost and faster than tampering detection methods based on a single image. These previous methods can be seen as an alternative approach, when no similar images are retrieved.

Difficulties lie in the wide variety of possible modifications. In this work, we propose a unified framework to detect and localize a large variety of forgeries in an image, by detecting inconsistencies between two images. The image to analyze is compared to the most similar images retrieved by a Content-Based Image Retrieval (CBIR) system. Such a system could be a reverse image search tool, but in our work we query our own database. Thus, we can evaluate the performance of our CBIR system when dealing with the particular class of images considered here, where strong editing process may trouble the recognition. Once similar images are retrieved, a local descriptor based approach is used to identify and localize differences. We also build two datasets containing various types of forgeries to evaluate our system.

In the next section, we discuss related studies on image forensics, image retrieval and social networks analysis. Our approach is described in Section 3, while datasets for evaluation and results are detailed in Section 4. Concluding remarks are presented in Section 5.

[3] http://www.stopfake.org/en/13-online-tools-that-help-to-verify-the-authenticity-of-a-photo/

2 Related Work

Image forensics. The identification of tampered images has been largely studied in the field of image forensics. Various forms of image manipulation exist such as objects deletion, retouching objects, copy-moving parts of an image, or inserting elements taken from a different source, *i.e.* splicing or copy-paste. Such diverse scenarios require specific approaches and techniques. Traditionally in image forensics scenarios, the decision (tampered or not) must be made solely on the basis of the image to be analyzed, without using any external information. Most passive forgery detection techniques aim at revealing alteration of the underlying statistics of the forged image. However, almost all existing forensics methods detect only one type of image processing operations or are based on some assumptions regarding the image format or the camera used. Among these techniques, pixel-based approaches are the most related to our context. Indeed, for images transmitted on social networks, we have neither information about camera (as EXIF informations are erased), nor prior about format.

Pixel-based methods widely address the problem of copy-move forgery detection (CMFD) [24]. These methods, also called Local Descriptor-based forgery detection techniques, are typically based on feature matching. Block-based approaches split the image into overlapping blocks and extract features, such as DCT, DWT, histogram of co-occurrences on the image residual [11], Zernike moments, or Local Binary Pattern (LBP) [9]. Keypoints-based approaches compute features, usually SIFT or SURF [1, 10], on local regions characterized by a high entropy. Features are then matched to detect similar regions, as a cue for copy-move forgery. Generally, it is shown that techniques based on dense fields provide a higher accuracy [7]. Also, some methods propose not only the detection but also the localization of the modified regions.We note that deep Convolutional Neural Networks (CNN) have been recently introduced in image forensics [17, 5]. The general idea is to restrict the first convolutional layer to a set of high-pass filters in order to suppress image content. However, the CNNs are used either only for image binary classification (authentic/forged), without localization [17], or to identify some manipulations such as median filtering or Gaussian blurring, excluding copy-move or splicing attacks [5].

Content-Based Image Retrieval (CBIR). For several years state-of-the-art methods in image retrieval consisted in aggregating local descriptors, such as SIFT, into a global representation. These last years, the use of pre-trained CNN [13] became the new reference for global descriptors. [4] first showed that using fully connected layers of a pre-trained deep network as global descriptors can outperform descriptors based on SIFT features, even without fine-tuning. Similar conclusions were shared by [18] and [23] with region-based descriptors. Also, [3] proposed to aggregate deep local features, while [20] proposed new fusing schemes for compact descriptions.

Social networks information analysis. Analysis of information on social networks raises a growing interest, in particular detecting false information. This is illus-

trated by an increasing number of projects on this topic [4], and the emergence of a task dedicated to tweet classification on true or false at the *Mediaeval* benchmark, named *Verifying Multimedia Use*[5]. Usually the methods are interested in the multimodal nature of the messages to make a decision (text, social networks, image). It was also shown that the use of external knowledge is of great importance in the success of the proposed methods [16, 15].

3 Proposed method

We propose a unified approach to detected a large variety of forgeries, which is composed of two main steps. First, the image to analyze is used to query a database. The system searches for the most similar image. If an image is retrieved, it is then compared to the query image to detect and localize the forged areas; Otherwise, the process ends.

3.1 Content-based image retrieval system

Initial ranking. A CBIR system is used to retrieve candidate images, sufficiently similar to a query (the image to be analyzed), even if the images are different one from another due to tampering operations.

First, images are described using CNN-based representations. Following the recent works of [23, 21], we choose to build descriptors using the seven-th fully connected layer *fc7* of the VGG vd19 CNN [22] trained on ImageNet. Images are first scaled to the standard 224×224 input size. Then, ℓ_2-normalization is performed and we obtain a 4096-dimensional vector.

Once all images descriptors are obtained, cosine similarity is computed between the query and images from the database. The nearest neighbors are retrieved using a KD-Tree to accelerate the search. Only images whose similarity exceeds a given threshold $\mathcal{T}$, which is further evaluated in the experimental section, are considered as relevant. Otherwise no image is considered similar.

Filtering. A geometric verification step, *i.e.* filtering, is then employed to filter the false positives from the short list of top ranked images returned by the CBIR. Filtering is based on the number of inlier matches after estimating the spatial transformation between the query and each candidate images. Finally, only the image with the highest similarity is considered for further processing.

The proposed approach is based on SURF features matching, similarly to several reranking process used in CBIR systems. Specifically, dense SURF features are first extracted in both images and matched [19]. RANSAC algorithm is then applied to estimate the affine transformation H between the two images. To further decrease the number of false matches, only a subset $\mathcal{S}$ of points in

[4] see for example Reveal project (`https://revealproject.eu/`), InVID project (`http://www.invid-project.eu/`), or Pheme project (`https://www.pheme.eu/`)

[5] `http://www.multimediaeval.org/mediaeval2016/verifyingmultimediause`

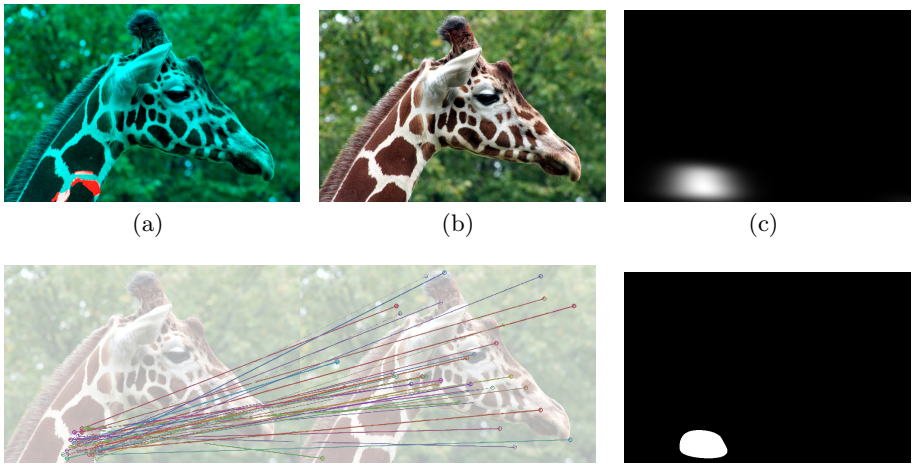


Fig. 2: (a) query image; (b) candidate image returned by the CBIR (d) outliers computed from query to candidate image; (c) density map; (e) binary mask.

the query are kept as candidate matches for the RANSAC algorithm. These are points that match another point with a distance $d \leq 2 \times d_{min}$, where d_{min} is the minimum distance found between 2 descriptors of the pair of images.

After RANSAC estimation, we further apply H to each point of $\mathcal{S}$ and classify them as inlier if the distance d^* between the projected position and its match is lower than $0.15 \times diag$, where $diag$ is the length of the image diagonal in terms of pixels. Images with a majority of outliers in the set $\mathcal{S}$ are discarded as false positives. Among the remaining images, the one with the highest ratio of inliers over outliers of the set $\mathcal{S}$ is selected and given to the following localization part.

3.2 Tampering localization

Once a pair of images is given by the CBIR system, the tampering localization step consists in identifying potential inconsistencies between them. The process should be robust to various transformations, such as rotation, illumination changes, crop, or translation, and is then based on local descriptors. In our case, we are interested in detecting outlier matches spatially close to one another, as a cue of tampering.

Having the homography H computed previously, we apply H to all keypoints of the query to identify inliers and outliers, as detailed in the previous section. Note that the matching criteria considered at this step (1-nn) is weaker than the one used to estimate the homography, in order to enforce a one-to-one matching of keypoints. Since this process is not symmetric, both images are used in turn as query. The image containing the most outliers is selected for the localization step, see Fig. 2(d).

Finally, we identify the areas with high density of outliers and remove the isolated points. These two operations are carried out by a Kernel Density Estimation (KDE) technique. We compute a density map $\mathcal{D}$ on the set of outliers by applying a Gaussian kernel with bandwidth selected by Scott's Rule of thumb, see Fig. 2(c). This density map is then thresholded to obtain a binary mask $\mathcal{B}$ of the suspicious regions. Only points $\mathbf{p}$ of the density map verifying $\mathcal{D}(\mathbf{p}) \geq 1/2\ max_{\mathbf{p}\in\mathcal{D}}(\mathcal{D}(\mathbf{p}))$ are retained in the final segmentation, see Fig. 2(e).

4 Experiments

We evaluate our approach on challenging datasets exhibiting a large variety of modifications. We first give an overview of the datasets involved and describe the different characteristics of the data. The CBIR is further evaluated using all these datasets and the tampering localization is finally tested.

4.1 Datasets

Many datasets of various size and difficulty have been proposed in image forensics to evaluate forgery detection methods. They differ by the realism of their construction (from simple artificial insertion to realistic complex objects with post-processing), by the types of attacks they address, and by the presence of the modification masks allowing the evaluation of the tampering localization.

Most existing datasets focus on copy-move attacks, thus we build two new datasets. *Reddit* is built from real data with every type of forgery especially copy-paste, which are almost not occurring in the other datasets. Similarly, *Synthetic* is artificially built with various and precise forgeries to better understand how our system copes with each type of attack. Also, we are interested in datasets allowing tampering localization and for which the original images are available.

MICC-F600 [1] is a dataset from image forensics. It contains 600 images: 440 original images from the 1,300 images of the MICC-F2000 dataset [2], and 160 forged images from the SATS-130 dataset [6]. Forged images contain realistic and challenging multiple copy-move attacks.

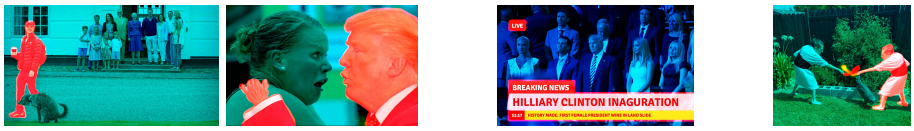
MediaEval (ME) is composed of 316 images associated to the tweets used in the *Verifying Multimedia Use* task of *Mediaeval 2016*. We use 40 images as queries: 17 fake images particularly challenging, which have their original image in the database, and 23 images with typical collage, cropping, or insertion of text and geometrical shapes (see Fig. 3). These last modifications are generally not achieved in a malicious purpose, but are challenging for the CBIR system. The groundtruth maps were manually constructed for these queries.

Reddit is a collection of 129 original images and their photoshopped versions from the Photoshop challenge on the *Reddit* website[6], totalling 383 images. 106 images are used as query and were manually annotated by up to three annotators,

[6] `http://www.reddit.com`

(a) (b) (c) (d)

Fig. 3: Some examples of challenging images from the *ME* dataset.



(a) (b) (c)

Fig. 4: Examples of different kinds of attacks in *Reddit*: (a) copy-paste; (b) text insertion; (c) copy-move. Blue: unmodified regions; Red: tampered regions.

with an inter-annotator agreement of 75.12% in terms of Jaccard's score. The tampering operations are mainly splicing of various size, which is not addressed by *MICC-F600*. Some examples are given in Fig. 4.

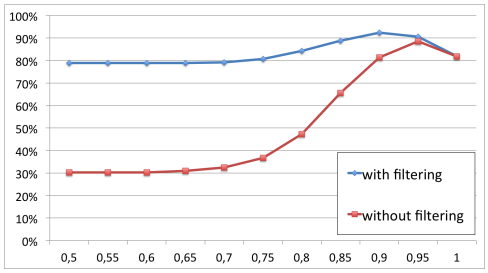
Synthetic is an artificially generated dataset of 3,500 forged images, including both copy-move and copy-paste attacks and different processing of the alien. For each 7 original images, we generate 500 forged versions. Each forged image is created by combining a random selection of different parameters among the number of modifications (between 0 and 3), the size of the alien (10, 20, 30, 40, or 50% of the host image), the rotation applied (0, 45, 90, 135, 180, 225, 270, or 315 degrees), a blurring or not of the alien, and the type of attack, *i.e.* copy-move or copy-paste. Note that we can find both copy-move and copy-paste attacks in a forged image, and that a blur attack can be applied on the whole host image (even without any attack). This dataset is not evaluated with the CBIR.

Distractors Additionally, we collect distractors when evaluating the CBIR system. We use 8,035 images collected from 5 websites dedicated to hoax detection[7]. We further add 82,543 unique images from Twitter, corresponding to the top tweets during January and February 2017, for a total of 170 different topics.

4.2 CBIR system

Most CBIR systems are evaluated on benchmark databases composed of several views of a same object. However, we want to test whether our system is capable

[7] `hoaxbuster.com`, `hoax-busters.org`, `urbanlegends.about.com`, `snopes.com`, and `hoax-slayer.com`



(a) Accuracy of the CBIR system according to the cosine similarity threshold $\mathcal{T}$.

(b) Example of true positive

(c) Example of false positive

Fig. 5: CBIR evaluation results.

of returning a quasi-copy of a query at first rank and none if no copy exists. We further evaluate the behaviour of our system with tampered and noisy images.

The query set is composed of diverse tampered and pristine images and the database contains original images as well as distractors. Specifically, the database to query is composed of 93,121 images: 82,543 images from Twitter, 8,035 images from hoax websites, 316 images from *ME*, 129 original images from *Reddit*, 98 images from the *SATS-130* dataset and 2,000 images from the *MICC-F2000* dataset which contain the original images of *MICC-F600*.

Then, we use a set of 2,151 queries, both positives and negatives (meaning having or not a correspondence in the database): 600 images from *MICC-F600*, 106 photoshopped images from *Reddit*, and 40 tampered images from *ME* are positive examples. Amongst them, 440 images from *MICC-F600* are not tampered. 1,405 images from *Holidays* dataset [12] are used as negative queries.

Results Unlike most CBIR measuring ranking performance in terms of precision (P@k, mAP, etc.), we evaluate our system in terms of mean accuracy, computed over all the queries. Indeed, we wish our CBIR-based system to output either the most similar image or no image, if no quasi-copy is found in the database.

Figure 5(a) shows the accuracy of the CBIR system for various threshold values $\mathcal{T}$. We observe that the best threshold is $\mathcal{T} = 0.9$ with an accuracy of 91.91% with filtering and 81.08% without filtering. The value $\mathcal{T} = 0.9$ is kept for the tampering localization step.

Table 1 shows the performance with respect to each set of queries for given thresholds $\mathcal{T}$. We observe a gain in accuracy for lower $\mathcal{T}$ on *Reddit*, *MICC-F600*, and *ME* (positive queries). However, *Holidays* performs best for a high $\mathcal{T}$, as it only contains negative queries. Indeed, a low threshold allows to list all relevant images, while generating a lot of false positives.

As an insight, we observe that the CBIR mainly fails when the forged area is very large with respect to the image. This is particularly illustrated by poor performances on *ME*. This small set of queries was specially chosen to challenge

$\mathcal{T}$	Reddit	MICC-F600	ME	Holidays
0.75	**73.62%**	**99.83%**	**32.50%**	74.68%
0.80	73.23%	**99.83%**	**32.50%**	80.58%
0.85	71.65%	99.50%	**32.50%**	88.41%
0.90	64.57%	98.50%	20.00%	96.09%
0.95	37.80%	94.00%	15.00%	**100.00%**

Table 1: CBIR accuracy per datasets for different threshold values $\mathcal{T}$

Dataset	*Synthetic*	*Synthetic Unblurred*	*MICC-F600*	*Reddit*	*ME*
F_P	12.41 %	0.93 %	10.82 %	37.11 %	24.37 %
F_N	15.05 %	9.90 %	20.93 %	24.82 %	29.82 %
TPR	100.00 %	100.00 %	95.61 %	100.00 %	100.00 %
FPR	49.64 %	0.00 %	9.10 %	0.00 %	0.00 %

Table 2: Tampering localization results per datasets.

the CBIR system, which is disturbed by overly large insertions (more than 50% of the image size), or border/banners insertions. Fig. 3 shows such queries whose original image has not been retrieved. Examples of successful match despite a quite large forgery and false positive are given in Fig. 5(b) and 5(c).

4.3 Tampering localization

We evaluate the tampering localization on *Synthetic*, and on the pairs of real forged images returned by the CBIR, from MICC-F600 (copy-move attacks), *Reddit* (various attacks, mainly copy-paste), and *ME* (various modifications). For the *Synthetic* dataset, image pairs are directly given.

The performance on patch localization is computed at the pixel level as the percentage of erroneously detected pixels F_P (i.e. false positives) and erroneously missed pixels F_N (i.e. false negatives). To compare with other methods, we also measure the detection performance at the image level in terms of True Positive Rate (TPR) and False Positive Rate (FPR), where TPR is the fraction of tampered images correctly identified, while FPR is the fraction of original images that are not correctly identified.

Results Table 2 shows the localization results per datasets. We observe on *Synthetic* that the localization method is robust to the size, rotation or number of inserted aliens, but unsurprisingly sensitive to blurring of the whole image. The high FPR corresponds to blurred original images classified as forged. Discarding the blurred images (*Synthetic Unblurred*), attacks are precisely detected.

Generally, the pixel-level localization is altered by two factors: (i) our predicted area is often smaller than the alien, which increases F_N. However, we do

Method	Fan2016 [10]	Cozzolino2015 [8]	Li2016 [14]	Ours
TPR	88.13 %	96.25 %	96.25 %	95.00 %
FPR	6.82 %	5.91 %	4.77 %	9.10 %

Table 3: Results on MICC-F600 (best settings for each method, in %)

not focus on having the most accurate localization at the pixel level but rather precisely detecting whether a tampering is detected or not; (ii) when the image is wrongly matched by the CBIR with a false positive, the tampering localization failed, resulting in an increase of F_P. This does not concern *MICC-F600*, which offer cleaner and smaller attacks, and for which the accuracy of the CBIR is the highest, with no false positives.

At the image level, the detection of tampering in *Reddit*, and *ME* offers perfect results. The null FPR is due to the fact all queries are forged for these datasets. When not all queries are forged, as in *MICC-F600*, performance remains very high. In fact, we compute FPR and TPR for the sake of comparison with the state of the art on *MICC-F600*, as most of methods (except [8]) only deal with detection. Comparison with the state of the art is given in Table 3. We note that the CBIR is not applied there (whole images of *MICC-F600* are processed) to allow the comparison. Our system performs on par with recent state of the art methods, with a higher FPR.

Regarding the entire process and all the datasets (including Holidays as negative examples), we measure a TPR of 81.37% and a FPR of 5.14%. Errors are mainly due to the CBIR performance, as false positives at the retrieval step generate false positives for the tampering detection, while false negatives result in missed tampering detections.

5 Conclusion

In this paper, we address the problem of verifying the authenticity of images from social networks. Moreover, we built two complete dataset for the evaluation. We propose a system that detect and localize tampering on such images, based on image retrieval, followed by image comparison based on local features. Unlike methods from the literature, our system is generic and can handle a large variety of modifications. We evaluated our system on diverse datasets, and shown that the proposed method performs on par with the state of the art for copy-move. We also observed that images from social networks are challenging for state of the art CBIR, and there is room for improvement to deal with this particular type of images. Future work will be directed in this direction.

References

1. Amerini, I., Ballan, L., Caldelli, R., Bimbo, A.D., Tongo, L.D., Serra, G.: Copy-move forgery detection and localization by means of robust clustering with j-linkage. Signal Processing: Image Communication (SPIC) (2013)

2. Amerini, I., Ballan, L., Caldelli, R., Del Bimbo, A., Serra, G.: A SIFT-based forensic method for copy-move attack detection and transformation recovery. TIFS (2011)
3. Babenko, A., Lempitsky, V.S.: Aggregating local deep features for image retrieval. In: Int. Conf. on Computer Vision (ICCV) (2015)
4. Babenko, A., Slesarev, A., Chigorin, A., Lempitsky, V.: Neural codes for image retrieval. In: European Conf. on Computer Vision (ECCV) (2014)
5. Bayar, B., Stamm, M.C.: A deep learning approach to universal image manipulation detection using a new convolutional layer. In: Work. on IHMS (2016)
6. Christlein, V., Riess, C., Angelopoulou, E.: On rotation invariance in copy-move forgery detection. In: Workshop on Information Forensics and Security (2010)
7. Christlein, V., Riess, C., Jordan, J., Riess, C., Angelopoulou, E.: An evaluation of popular copy-move forgery detection approaches. Trans. on Information Forensics and Security (TIFS) (2012)
8. Cozzolino, D., Poggi, G.., Verdoliva, L..: Efficient dense-field copy–move forgery detection. Trans. on Information Forensics and Security (TIFS) (2015)
9. Dixit, A., Dixit, R., Gupta, R.K.: Detection of copy-move forgery exploiting lbp features with discrete wavelet transform. Int. Journ. of Computer Application (2016)
10. Fan, Y., Zhu, Y.S., Liu, Z.: An improved sift-based copy-move forgery detection method using t-linkage and multi-scale analysis. IHMSP (2016)
11. Fridrich, J., Kodovsky, J.: Rich models for steganalysis of digital images. Trans. on Information Forensics and Security (TIFS) (2012)
12. Jegou, H., Douze, M., Schmid, C.: Hamming embedding and weak geometric consistency for large scale image search. In: European Conf. on Computer Vision (ECCV) (2008)
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Neural Information Processing Systems 25 (2012)
14. Li, Y., Zhou, J.: Image copy-move forgery detection using hierarchical feature point matching. In: APSIPA Transactions on Signal and Information Processing (2016)
15. Maigrot, C., Claveau, V., Kijak, E., Sicre, R.: Mediaeval 2016: A multimodal system for the verifying multimedia use task. In: MediaEval Workshop (2016)
16. Phan, Q.T., Budroni, A., Pasquini, C., De Natale, F.G.: A hybrid approach for multimedia use verification. In: MediaEval Workshop (2016)
17. Rao, Y., Ni, J.: A deep learning approach to detection of splicing and copy-move forgeries in images. In: Workshop on Information Forensics and Security (2016)
18. Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: An astounding baseline for recognition. In: CVPR Workshops (2014)
19. Sicre, R., Gevers, T.: Dense sampling of features for image retrieval. In: ICIP. pp. 3057–3061. IEEE (2014)
20. Sicre, R., Jégou, H.: Memory vectors for particular object retrieval with multiple queries. In: ICMR. ACM (2015)
21. Sicre, R., Jurie, F.: Discriminative part model for visual recognition. CVIU (2015)
22. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. ICLR (2014)
23. Tolias, G., Sicre, R., Jégou, H.: Particular object retrieval with integral max-pooling of CNN activations. In: ICLR (2016)
24. Warbhe, A.D., Dharaskar, R., Thakare, V.: A survey on keypoint based copy-paste forgery detection techniques. Procedia Computer Science (2016)