Image Copy Detection Based on Convolutional Neural Networks

Jing Zhang[1], Wenting Zhu[2], Bing Li[2], Weiming Hu[2], and Jinfeng Yang[1]

[1] College of Electronic Information and Automation,
Civil Aviation University of China, Tianjin, 300300, China
[2] CAS Center for Excellence in Brain Science and Intelligence Technology,
National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, No. 95, Zhongguancun East Road,
Beijing, 100190, China
zhangjingfighting@gmail.com

Abstract. In this paper, we present a model that automatically differentiates copied versions of original images. Unlike traditional image copy detection schemes, our system is a Convolutional Neural Networks (CNN) based model which means that it does not need any manually-designed features. In addition, a convolutional network is more applicable to image copy detection whose architecture is designed for robustness to geometric distortions. Our model uses fully connected layers to compute a similarity between the CNN features, which are extracted from image pairs by a deep convolutional network. This method is very efficient and scalable to large databases. In order to see the comparison visually, a variety of models are explored. Experimental results demonstrate that our model presents surprising performance on various data sets.

Keywords: image copy detection, feature extraction, CNN

1 Introduction

Due to the rapid development of technologies like communication, computer network and multimedia, the online multimedia resources are growing exponentially. These advanced technologies are like a double-edged sword, bringing a variety of convenience as well as much challenges to the information security. The convenience of digital image acquisition makes more and more unauthorized copies on the internet. In order to effectively protect the copyright of legitimate users, copyright piracy needs to be monitored. This involves the primary technology, detecting the copies of images.

There are usually two kinds of methods of copyright protection, digital watermarking and copy detection. The digital watermarking technology is to embed watermark information into the image before the image published. Images with no embedded watermark information can not be detected, but the image copy detection only works on the image itself. Because of legal and other reasons, it can not modify image content. Image copy detection becomes a reasonable complement of digital watermarking technology.

In recent years, a technique of mimicking human brain learning—deep learning has made a big leap. It evolved from the initial neural network and achieved great success in a multitude of pattern recognition problems. Take CNN for example, incredible strides had been made on image representation [14]. Because its invariance for translation, scaling, tilting, and other forms of distortion, we build our method surrounding this powerful tool to automatically learn copy image detection.

The contributions of this paper are as follows:

1. The convolutional neural networks are used to directly learn the copy detection task, which does not need any manually-designed features.
2. Different neural network models are proposed and formed a clear contrast, which pave the way for further studies.
3. We apply our models to several data sets, showing tremendous successes in accuracy. It proves that using CNN architecture is perfectly competent for image copy detection.

2 Related Work

2.1 Traditional Image Copy Detection

A traditional image copy detection, includes three main technology parts: feature extraction, feature based index construction and feature comparison. In this section, we briefly review previous works on feature extraction and index construction methods.

The extraction of feature information has two forms, global feature extraction and local feature extraction. The global information can be obtained (like texture) on the whole image. Just like Li in [16], he proposed Gabor texture descriptors using the adjustability of Gabor on the direction and scale. The descriptor is invariant to rotation and scaling invariance. But there is a high false alarm rate for large angle rotation. Copy detection algorithm based on global features is simple in calculation and has high efficiency. But the poor resistance to geometric attacks, especially cropping and rotation, makes scholars prefer detecting image on local features. Berrani [2] computed local differential descriptors for each image which corresponded to the local regions of interest in the image. As [18] puts it, the best performance among all the local descriptors is the Scale Invariant Feature Transform (SIFT) descriptor. This inspires many scholars to do more in-depth research in this field. For example, a method of VLAD (vector of locally aggregated descriptors) based on a compact representation of SIFT was proposed in [11]. Similarly, Cao et al. [3] introduced a new tilt parameters in the affine SIFT transform, using geometric consistency constraints in similarity detection.

In the second part, Bags-of-visual-Words (BOW) and hash algorithms are the most representative methods in the feature based index construction. The BOW [20] extracts image features as visual words. Some others improved it by changing the grouped visual words into visual phrases [28] and visual sentences [23].

H. Ling [17] abandoned previous methods and propose a local binary fingerprint as visual word. This method is available on large data sets and more efficient. But the visual vocabulary discards spatial information of local feature, which will influence the matching work. For indexing algorithm using hashing function, Locality-Sensitive Hashing (LSH) [9] is a popular algorithm in multimedia applications. Recently, many new kinds of hashing schemes have been proposed such as Spectral hashing [15] and Self-taught hashing [27]. These schemes appear a significant improvement over other methods. However, it brings a high computational cost based on local embedding techniques. A robust image hashing based on Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT) for copy detection was presented by Tang [22]. The algorithm for conventional digital operation has good robustness and uniqueness. Subsequently, he put forward a hash based on fan-beam transform algorithm [21] which could resist rotation attacks at any angle.

2.2 Research Progress of CNN

In contrast to the above approaches, we learn features end-to-end directly from input pixels. So a simple but efficient deep learning method is proposed for image copy detection, and it achieves more favorable results on publicly available data sets. Deep learning is a successful model for learning useful representation in research field of images, owing to its strong computing power especially in a large dataset [14,5]. In the task of copy detection, we conduct experiment on two images, and let the model make a decision whether the detected image is a copy of original image. Therefore, in so many models of deep learning, the siamese CNN model is more suitable for our study.

More recently, researchers in the siamese network have presented different approaches. In[26], researchers compared image patches via a similar architecture. The results showed an good performance with a high computational cost. Because they focused on the center of images too much. In older to learn a high precision arithmetic, Han et al. [7] used a multi-layer network followed by a fully connected network. In [25], the best results were obtained on the KITTI benchmark using CNN to compute the stereo matching cost. The success of these methods means that, for the study of the two inputs, just like our work, siamese CNN model has a broad research space. Inspired by these methods, our network structure is similar to them. But there is a notable difference that we show how to build small but powerful model using discriminative strategies. Contrast with[25] we use pooling layers to increase the robustness for the different variations of copy images. And softmax-with-loss is used to make a simple classification for input images compared to [26]. Our models still have other differences in architecture, like an additional dropout layer and Local Response Normalization (LRN) layer. The proposed method will be described in next section.

3 The Proposed Method

3.1 Overview

Our goal is to learn a image copy detection model. When two images C (copy image) and O (original image) are input to the model, the model give the corresponding output 0 or 1 in which the 1 represents a copy relationship between C and O, and the 0 opposite.

The early Siamese networks [6] used contrastive loss functions to train a similarity metric from real data, which is able to place similar images nearby and keep dissimilar images separated. But in our work we want to identify a copy relationship between the pair of images instead of a metric. Here we put our work as a binary classification task. Input images are divided into two categories, copy pair (1) or not (0).

The models we used in this paper include multiple convolutional and spatial pooling layers to obtain feature vectors, followed by fully connected layer to compare features of input images. This has some similarities with [7,26]. But in order to adapt to our tasks, on the top of models, softmax-with-loss instead of cross-entropy loss or hinge-based loss leads to the learning objective function. After the fundamental models, we offered an additional method which also achieved a good result.

3.2 Network Architectures

Fundamental models. As shown in Fig. 1, depending on whether or not the weights of the two branches are shared, the two branches models have two forms, siamese and pseudo-siamese [26]. Following Simonyans [19] advice, 3×3 kernels are adopted to make decision function more discriminative. Rectified Linear Units (ReLU) [14] as a non-linearity for the convolution layers can be used. At the same time, we applyed dropout [8] with probability 0.5 between the fully connected layers to avoid overfitting. For the spatial pooling layers, we use max-pooling layers to deal with scale changes. In addition, the proposed 2-channel thought in [26] is also used, which is deemed that two pictures as two channels of a picture is directly fed to a single branch network. The parameters setting in this model is similar to CaffeNet model [12] which is a slightly modified of [14]. But we used fewer layers to construct our model and replaced the final fully connected layer with a two neuron layer.

Hybrid 2-channel siamese model. According to the results of experiment about 2-channel and two branches models (siamese and pseudo-siamese), jointing two images then extracting features has better performance than extracting features separately before concatenating. So a hybrid 2-channel siamese model (see Fig. 2) is constructed. On the one hand, we convert color images to gray scale images just like data preprocessing of 2-channel method (see Fig. 3). On the other hand, the two color images are spliced together from top to bottom. So we build a deep CNN to learn feature representations from color and gray

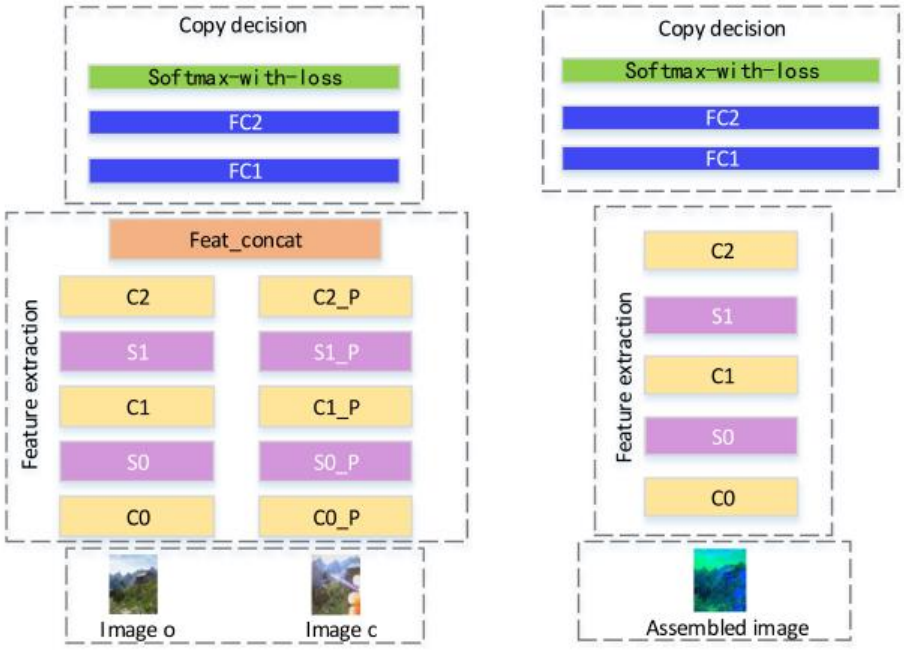

Fig. 1. The fundamental models: On the right side is two branches model. It represents siamese (parameter sharing) network and pseudo-siamese (without parameter sharing) network. On the left side is 2-channel model.

scale images separately, which are then connected and fed into the final fully connected layer. As illustrated in Fig. 2 , the convolution property of CNN is to remain convolution order of each layer unchanged. This ensures that the features extracted on up and down spliced image are still the features of two original images. So even though we only use a pseudo-siamese model to extract the characteristics of the input, it can also be regarded as a hybrid of 2-channel and siamese model. The system not only use gray scale images to prevent the impact of brightness but also discover the most discriminative features using RGB color space.

4 Learning

Training. A strongly supervised manner is applied to all models for training. The output $y_i \in (0, 1)$ is the corresponding label (denoting a non-copy and a copy pair, respectively). We use softmax to compute the loss function and to initialize the backpropagation. Stochastic gradient descent with momentum 0.9 and weight decay $\lambda = 0.0005$ are applied to all architectures.

Different learning rates have been tried on the proposed models. We find that setting the initial learning rate 0.00001 then decreasing it every 100,000 iterations produces better results than using larger learning rates. Depending on the network architecture, it takes about 2 to 3 days to train the full network.

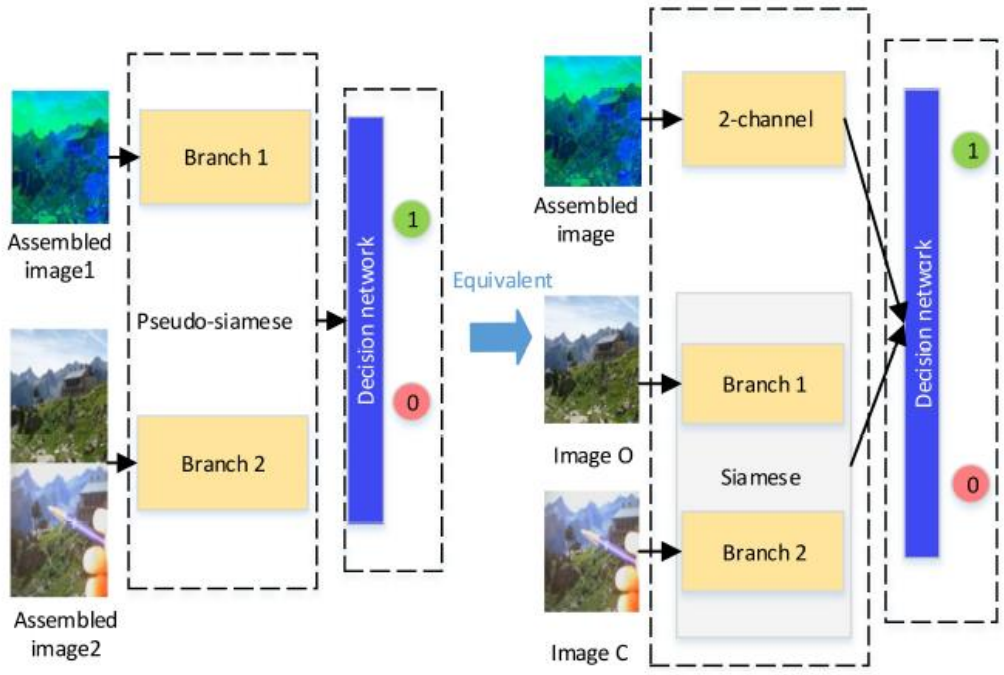

Fig. 2. Hybrid 2-channel siamese model:A simple image preprocessing and a pseudo-siamese model (left) achieves a combination of 2-channel and siamese network (right), but the model parameters have not increased with the change.

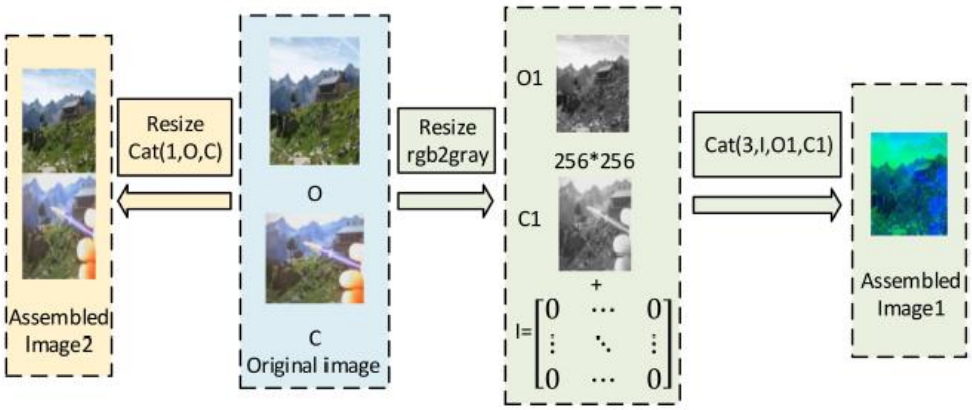

Fig. 3. Image preprocessing: For image pair, O and C, when they are processed into two channels of a picture, they will be used in the 2-channel model. Besides above process, in the hybrid 2-channel siamese model, jointing the two picture up and down is also needed.

The weights are initialized from a zero-mean Gaussian distribution with standard deviation 0.01. Finally, all models are trained from scratch.

Data preprocessing. For 2-channel model and hybrid 2-channel siamese model, a process of image preprocessing is needed. It is showed in Fig. 3.

5 Experiments

5.1 Dataset

For deep learning networks, the training data is critical for final results. To collect large data sets, four different data sets were put together to train models, which ultimately constituted more than 52,000 image pairs. We set aside seventy percent of the total data for training. The data sets are widely applied in image copy detection, and can be downloaded conveniently. They are INRIA Copydays dataset [10], CoMoFoD database [24], Image Manipulation Dataset [4], and MICC-F220, MICC-F2000 dataset from [1]. A brief summary is showed in Table 1. In addition, the INRIA Copydays dataset on the copy image has great transformation called strong copy picture. In order to expand data set and increase the difficulty of experiment, a copy relationship is also thought to exist between the strong copy pictures and any other copy pictures. We chose several combinations and finally generated approximately 10500 image pairs. We also had 25800 image pairs as negative samples, which were crawled from "Web Queries" dataset [13].

Table 1. Data set summary and operation. Table shows corresponding numbers. Each of the original picture and its copy form a pair of input images, then go to test set or train set. Copy image contains a variety of transformation, such as, rotation, scaling, distortion, noise adding and image blurring.

	Dataset name	original image	copy transformation	generate pair	train	test
1	INRIA Copydays	157	>19	10490	7746	3744
2	CoMoFoD	200	50	10000	10000	0
3	Image Manipulation	48	83	3984	0	3984
4	MICC-F220	121	10	110	0	110
5	MICC-F2000	50	14	700	700	0

5.2 Result

In our experiment, we have first trained our proposed models in a small database about 4000 image pairs, which only chose INRIA Copydays dataset as positive samples. It turns out that 2-channel architecture exhibit best performance in all networks. In the two branches networks, pseudo-siamese model is better than

siamese model. This result is in line with [26] inference. However, as we expanded our data set to a new training, a little change happened. It is presented in Table 2. The expansion of the data set lead to an significant advancement of overall performance, and result of siamese network also catch up with and surpass the other models.

Table 2. Test accuracies (%) of models on small and large dataset

Networks	siamese	pseudo-siamese	2-channel	hybrid 2-channel siamese
small dataset	84	86.5	94.5	92
large dataset	99.07	99.04	98.6	97.93

The results of experiment show that our models have desired robustness against JPEG, added noise and filtering attacks. We find that each model has advantages and disadvantages of its own, although the results of models are approximate. The judgment errors of two branches model (siamese and pseudo-siamese) is closely related to the colors of original images. If the color of original picture is too simple, the network will not be able to correctly identify the copies of it (image 2 in Fig. 4, it is a black and white picture). But these models are robust to image cropping. Conversely, for the 2-channel, its false results mainly come from cropping. But for a picture similar to the image 2 in Fig. 4, there is a good ability to identify, which is related to gray scale transformation in preprocessing.

The hybrid 2-channel siamese model is more like a combination of two branches and 2-channel method. Although its results are not outstanding, but it can effectively make up for the defects of two branches and 2-channel method. It can recognize a simple color picture and improve the capability against cropping attack to some extent. We analyzed the results of this model and discovered that most of the errors are due to the strong copy pictures (image 4 in Fig. 4). And such pictures accounts for large percentage in test set, which may be the reason for the low accuracy rate. In addition, we choose uniform parameters for each model in the experiment, but in fact if the best outcomes need to be presented, the parameters of each model are different.

In experiment, we chose a unified batch-size 100 for training, 60 for testing, and 300 test-iteration. We also tried to increase feature dimension, or add fully connected layer in two branches respectively, but the effect was not significant.

In order to compare with the traditional methods, we simply designed and extracted SIFT feature, and then utilized the feature matching to obtain matching point set. Lastly we got the final result through setting a proper threshold. SIFT algorithm has good robustness against the changes of the object shape, translation and rotation, but the detection of the feature points is too small for fuzzy image and edge smooth image. If we depend on traditional method, it is time-consuming to get ideal result while being strenuous.

Image O	Image C		siamese	pseudo-siamese	2-channel	hybrid 2-channel siamese
 1		0	0.00019	0.000077	0	0
		1	0.99981	0.999923	1	1
		0	0	0	0.0004	0.009533
		1	1	1	0.9996	0.990467
 2		0	0.956165	0.833159	0.0806	0.152655
		1	0.043835	0.166841	0.9194	0.847345
 3		0	0.0256	0.024573	0.9982	0.401183
		1	0.9744	0.975427	0.0018	0.598817
 4		0	0	0	0.0002	0.563827
		1	1	1	0.9998	0.436173

Fig. 4. Performance comparison of siamese, pseudo-siamese, 2-channel and hybrid 2-channel siamese: Results for the classification are presented. Image 1 is a positive example and three of the other images are negative examples. The red line box shows the error result of classification.

6 Conclusions

In this paper we presented an effective and efficient image copy detection method based on CNN, which learned a comparing function directly from raw image pixels. Several architectures were studied and each of them displayed extremely good performance. These results indicate that CNN based methods are specifically suited to copy detection task. We compared these models, and summarized their advantages and disadvantages. In the following work, the advantages are to be inherited, the disadvantage help us target our improvement efforts.

Finally, we have to say, such a good result, has a great relationship with the obviously discriminated database. In the next step, we will increase the complexity of the database, but a larger training set and a deeper network can still improve the overall performance. (because our training set in the present experiments is considered smaller than todays standards).

Acknowledgement. This work is partly supported by the 973 basic research program of China (Grant No. 2014CB349303), the Natural Science Foundation of China (Grant No. 61472421), the National Nature Science Foundation of China (No. 61370038) and the Strategic Priority Research Program of the CAS (Grant No. XDB02070003).

References

1. Amerini, I., Ballan, L., Caldelli, R., Del Bimbo, A., Serra, G.: A sift-based forensic method for copycmove attack detection and transformation recovery. IEEE Transactions on Information Forensics & Security 6(3), 1099–1110 (2011)
2. Berrani, S.A., Amsaleg, L., Gros, P.: Robust content-based image searches for copyright protection. In: ACM International Workshop on Multimedia Databases, Acm-Mmdb 2003, New Orleans, Louisiana, Usa, November. pp. 70–77 (2003)
3. Cao, Y., Zhang, H., Gao, Y., Guo, J.: An efficient duplicate image detection method based on affine-sift feature. In: Broadband Network and Multimedia Technology (IC-BNMT), 2010 3rd IEEE International Conference on. pp. 794–797 (Oct 2010)
4. Christlein, V., Riess, C., Jordan, J., Riess, C., Angelopoulou, E.: An evaluation of popular copy-move forgery detection approaches. IEEE Transactions on Information Forensics and Security 7(6), 1841–1854 (Dec 2012)
5. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 580–587 (June 2014)
6. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). vol. 2, pp. 1735–1742 (2006)
7. Han, X., Leung, T., Jia, Y., Sukthankar, R., Berg, A.C.: Matchnet: Unifying feature and metric learning for patch-based matching. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3279–3286 (June 2015)
8. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. Computer Science 3(4), 212–223 (2012)
9. Indyk, P., Motwani, R.: Approximate nearest neighbors: Towards removing the curse of dimensionality. In: Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing. STOC '98, ACM, New York, NY, USA (1998), http://doi.acm.org/10.1145/276698.276876
10. Jégou, H., Douze, M., Schmid, C.: Hamming embedding and weak geometric consistency for large scale image search. In: Proceedings of the 10th European Conference on Computer Vision. pp. 1.1–1.1 (October 2008)
11. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. pp. 3304–3311 (June 2010)
12. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. Eprint Arxiv pp. 675–678 (2014)
13. Krapac, J., Allan, M., Verbeek, J., Juried, F.: Improving web image search results using query-relative classifiers. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. pp. 1094–1101 (June 2010)

14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems 25(2), 2012 (2012)
15. Li, P., Wang, M., Cheng, J., Xu, C., Lu, H.: Spectral hashing with semantically consistent graph for image indexing. IEEE Transactions on Multimedia 15(1), 141–152 (Jan 2013)
16. Li, Z., Liu, G., Jiang, H., Qian, X.: Image copy detection using a robust gabor texture descriptor. In: Proceedings of the First ACM Workshop on Large-scale Multimedia Retrieval and Mining. pp. 65–72. LS-MMRM '09, ACM, New York, NY, USA (2009), http://doi.acm.org/10.1145/1631058.1631072
17. Ling, H., Yan, L., Zou, F., Liu, C., Feng, H.: Fast image copy detection approach based on local fingerprint defined visual words. Signal Processing 93(8), 2328–2338 (2013)
18. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. Pattern Analysis & Machine Intelligence IEEE Transactions on 27(10), 1615–30 (2005)
19. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Eprint Arxiv (2014)
20. Sivic, J., Zisserman, A.: Video google: a text retrieval approach to object matching in videos. In: Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on. pp. 1470–1477 vol.2 (Oct 2003)
21. Tang, Z., Huang, L., Yang, F., Zhang, X.: Robust image hashing based on fan-beam transform. Icic Express Letters 8(8), 2365–2372 (2014)
22. Tang, Z., Yang, F., Huang, L., Wei, M.: DCT and DWT based image hashing for copy detection. Icic Express Letters 7(11), 2961–2967 (2013)
23. Tirilly, P., Claveau, V., Gros, P.: Language modeling for bag-of-visual words image categorization. In: Proceedings of the 2008 International Conference on Content-based Image and Video Retrieval. pp. 249–258. CIVR '08, ACM, New York, NY, USA (2008), http://doi.acm.org/10.1145/1386352.1386388
24. Tralic, D., Zupancic, I., Grgic, S., Grgic, M.: CoMoFoD - new database for copy-move forgery detection. In: 55th International Symposium ELMAR-2013. pp. 49–54 (September 2013)
25. Žbontar, J., LeCun, Y.: Computing the stereo matching cost with a convolutional neural network. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1592–1599 (June 2015)
26. Zagoruyko, S., Komodakis, N.: Learning to compare image patches via convolutional neural networks. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4353–4361 (June 2015)
27. Zhang, D., Wang, J., Cai, D., Lu, J.: Self-taught hashing for fast similarity search. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 18–25. SIGIR '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1835449.1835455
28. Zheng, Q.F., Wang, W.Q., Gao, W.: Effective and efficient object-based image retrieval using visual phrases. In: Proceedings of the 14th ACM International Conference on Multimedia. pp. 77–80. MM '06, ACM, New York, NY, USA (2006), http://doi.acm.org/10.1145/1180639.1180664