

CASIA IMAGE TAMPERING DETECTION EVALUATION DATABASE

Jing Dong, Wei Wang and Tieniu Tan

Institute of Automation, Chinese Academy of Sciences
P.O.Box 2728, Beijing, 10190
E-mail: {jdong,wwang,tnt}@nlpr.ia.ac.cn

ABSTRACT

Image forensics has now raised the anxiety of justice as increasing cases of abusing tampered images in newspapers and court for evidence are reported recently. With the goal of verifying image content authenticity, passive-blind image tampering detection is called for. More realistic open benchmark databases are also needed to assist the techniques. Recently, we collect a natural color image database with realistic tampering operations. The database is made publicly available for researchers to compare and evaluate their proposed tampering detection techniques. We call this database CASIA Image Tampering Detection Evaluation Database. We describe the purpose, the design criterion, the organization and self-evaluation of this database in this paper.

Index Terms— Database, Image Forensics, Tampering Detection, Algorithm Evaluation

1. INTRODUCTION

Seeing is believing? Not really. With easily accessible photomontage software such as Adobe Photoshop, the manipulation of images through forgery can be done easily by even an unprofessional editor. Seeing is no longer believing. Digital image forgery influences the perception of an observer of the depicted scene, potentially resulting in ill consequences if the forgery created with malicious intentions [1]. With the goal of verifying image content authenticity, passive-blind image tampering detection was called for and much work has been done in this area [2][3]. From year 2001 to year 2012, over 350 papers about digital forensics are published according to our records. Fig.1 shows the statistics of publication on digital forensics as well as the number of publications on the detection of image tampering or related issues from 2001 to 2012 according to the information provided in [4].

From Fig.1 we can notice that much work has been done on image tampering (or forgery) detection, e.g.[5, 6, 7, 8, 9, 10, 11, 12, 13]. Image splicing is one of the most popular techniques used in image tampering. It aims to cut and paste image regions from the same or another (other) image(s) to make a "fake" image. Splicing is considered as the core and

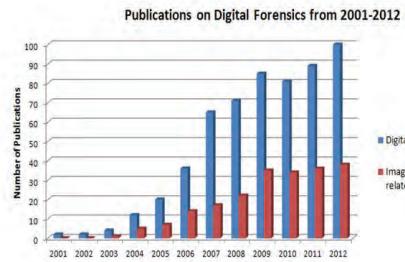


Fig. 1: Statistics of publications on digital forensics from 2001 to 2012.

simplest operation of image tampering since it does not require any pre-processing or post-processing for manipulation of the images. Many researchers have investigated the trace of image splicing to start their study on image tampering detection. In 2004, researchers from Columbia Univ. constructed an image splicing detection evaluation dataset and made it available to the research community [14]. The Columbia image database contains two parts, one part is a gray image dataset named the Columbia Image Splicing Detection Evaluation Dataset which consists of 933 authentic and 912 spliced gray-level image blocks of size 128×128 pixels, extracted from images in CalPhotos image set. There are two main categories of this dataset: authentic category and spliced category. Further, the two main categories are respectively subdivided in five subcategories according to the location of spliced regions. The other part is named as the Columbia Uncompressed Image Splicing Detection Evaluation Dataset, which consists of 183 authentic color image blocks and 180 spliced color image blocks with sizes ranging from 757×568 to 1152×768 and all are uncompressed images, in either TIFF or BMP format. All spliced images in both datasets are created using the authentic images without any post processing. More details about the Columbia image datasets can be found in [14]. This dataset is the only published database in recent years for image tampering detection. However, Columbia database only contains simple tampered image blocks and does not have adequate color image samples. As increasingly tampering detection methods are being developed for natu-

ral images nowadays, the Columbia datasets with simple and limited number of samples can not meet the demand of the state-of-art as an evaluation database.

Without a public standard database, many researchers tested their proposed tampering detection methods on limited examples[5, 6, 7, 8, 9, 12, 13]. For example, in [10], the author proposed a non-intrusive approach for image component forensics and tested their method. The example they used in this paper is only one picture of size 2048×2036 created by combining two different image parts. None detailed information about how to generated the test image. Similarly, in [11], the method they proposed for forgery detection was also tested on their own collected gray image database which is not available to other researchers. In[12],the author tested their proposed algorithm in few examples gathered from downloaded images without further details. In [13], a relatively more larger dataset was used for their experiments since their method only needs to be tested on images with blurring, downsizing and upsizing, and these images could be automatically generated by computer. Considering the increasing demand of public database for image forensics, we collected a more realistic and common platform for researchers to compare and evaluate their passive-blind image tampering detection techniques. In this paper, we introduce our constructed natural color image tampering database with realistic tampering operations. The rest of this paper is organized as follows. Detailed information on our database organization and design criterion are presented in Section II. Section III introduces several evaluation tests about our collected database. Section IV presents some potential uses of this database. Conclusions and database download information are given in Section V.

2. DATABASE CONSTRUCTION

Our collected database is named as CASIA Image Tampering Detection Evaluation Database (CASIA ITDE Database). Our tampered images in this database are all color images generated by using Adobe Photoshop CS3 version 10.0.1 on Windows XP. There are two versions of our image tampering detection evaluation database. Version 1.0 is a smaller set with 1,725 color images and version 2.0 is a larger one with 12,323 color images. The database V1.0 only considers splicing as the manipulation for tampering hence we call the tampered images in this database as spliced images. The image size in database V1.0 is fixed as 384×256 with JPEG format. Compared to database V1.0, tampered examples in database v2.0 are more comprehensive. Besides, various image sizes and format can be found in V2.0. We now describe the construction and organization of the two versions of our database in details in the following paragraphs.



Fig. 2: Two examples of generating the tampered images in CASIA ITDE V1.0



Fig. 3: Example Images in CAISA ITDE v1.0

2.1. CASIA ITDE V1.0

In CASIA ITDE V1.0, we collected an image set containing 1,721 color images of size 384×256 pixels with JPEG format. We divided these images into two subsets: authentic set and tampered set. There are 800 images in the authentic set and 921 images in the tampered set. Images in the authentic set were mostly collected from the Corel image dataset [15] and others are taken by our own cameras. The Corel image database is a well-known used image database and royalty free for many professional applications. It can be downloaded online [16]. Some information about the Corel Image Database can be found in [15]. The Corel Images are with various image contents. Our authentic set can be roughly clustered into 8 categories according to image content (scene, animal, architecture, character, plant, article, nature and texture). The tampered images are generated only by using crop-and-paste operation under Adobe Photoshop on these authentic images, hence we call these tampered image as spliced images. The following criteria are considered when spliced images were generated.

- Spliced images makers are told to randomly use candidate images from authentic set to generate spliced images. The spliced region(s) are either from the same authentic image or from two different authentic images.
- The shape of spliced regions in Photoshop palette can be chosen automatically by customization.
- Cropped image region(s) can be processed with scaling, rotation or other distortion operations before pasting to generate a spliced image. No post processing (like blurring) is utilized after generating a spliced image.
- Different sizes (small, medium and large) of spliced regions are concerned when generating the spliced set.
- There are several texture images in authentic set. We only generate spliced texture images by using one or two texture authentic images since splicing between

Table 1: Some statistical information about the spliced images in CAISA ITDE v1.0.

Category		No. of Images
Source of Tampered Region(s)		921
Manipulation with pre-processing	Same Image	451
	Different Images	470
Manipulation without pre-processing		562
Shape of Tampered Region	Rotation	25
	Resize	206
Distortion		53
Rotation and Resize		45
Resize and Distortion		27
Rotation and Distortion		3
Rotation, Distortion and Resize		0
Circular boundary		114
Rectangular boundary		169
Triangular boundary		102
Arbitrary boundary		536

texture and un-texture image will be more noticeable. Hence, we generate a batch of spliced texture images by randomly cropped a region (in regular or arbitrary shape) of an texture image and paste it to the same or a different texture image.

Fig.3 shows some examples of V1.0. We record the process of each generated spliced image by its filename. Ground truth information of how to generate the spliced image can be read from its filename. Detailed explanation about the filenames can be found in website [15]. Some statistical information about the organization of our generated spliced images in V1.0 is shown in Table.1. Here we also illustrated two examples of the generation of spliced images in our database V1.0 in Fig.2.

2.2. CASIA ITDE V2.0

The structure of CASIA ITDE database V2.0 is similar to database V1.0, but the V2.0 is an extended version. It contains totally 12,323 color images and two image subsets (authentic and tampered). The authentic set contains 7,200 authentic images and the tampered set contains 5,123 tampered images. However, database V2.0 is more challenging and comprehensive compared with database V1.0. Besides splicing, in database V2.0 we introduce blurring when manipulating the tampered image set. Unlike V1.0, the images in V2.0 are with difference sizes, ranging from 320×240 to 800×600 pixels. V1.0 only contains one kind of JPEG images, while V2.0 contains some uncompressed image samples (BMP and TIFF) and also considered JPEG images with different Q factors. The authentic images in V2.0 are collected from the Corel image dataset [16], public websites (with permission) and our own captured images. The image content of authentic set is again roughly clustered into several categories as we did in V1.0, but we collected a bath of "indoor" images for the authentic set to consider illumination variation when generating tampered images. There are 9 categories, which are classified to scene, animal, architecture, character, plant, article, nature, indoor and texture in the authentic image subset. Then we generate the tampered image set. We here consider post processing when we design the tampering criteria. Blur-



Fig. 4: Two examples of generating the tampered images in CASIA ITDE v2.0.



Fig. 5: Example Images in CAISA ITDE v2.0

ring can be used along with the spliced region's edge or any other region of the tampered image. The usage of blurring operation after we generate a spliced image is the most different feature between our database V1.0 and V2.0 tampered sets. But there are several exceptions and more challenging manual tampered images as we want to make this database more comprehensive. The following criteria are considered when generating the tampered images in V2.0.

- Tampered images makers are told to create the tampered images as much as realistic images to human eyes by using those defined manipulations in PhotoShop.
- The tampered image is either created from the same authentic image or from two different authentic images.
- In order to make our tampered images more realistic and challenging, most of the tampered regions are with arbitrary contour defined by the image makers.
- Cropped image region(s) can be processed with scaling, rotation or other distortion operations (defined by PhotoShop users) before pasting to generate a spliced image. Post processing like blurring could be utilized after generating a spliced image. Blurring/filtering can be applied along the tampered region couture or anywhere else in the generated image.
- Different sizes (small, medium and large) of tampered regions are concerned when generating the forgery images.

Fig.5 shows some examples of database V2.0. Similar to V1.0, the ground truth information of how to generate the tampered image can be read from the image filename. Detailed explanation about the filenames can be found in the website[15]. Some statistical information about the organization of generated tampered images in V2.0 are shown in Table.2. Two examples of the generation of the tampered images in database v2.0 are illustrated in Fig.4.

Table 2: Some statistical information about the spliced images in CAISA ITDE v1.0.

<i>Category</i>		<i>No. of Images</i>
	JPEG Format	2064
	TIFF Format	3059
Source of Tampered Region(s)	Same Image	3274
	Different Images	1849
Manipulation with pre-processing	Rotation	568
	Resize	1648
	Distortion	196
	Rotation and Resize	532
	Resize and Distortion	211
	Rotation and Distortion	42
	Rotation, Distortion and Resize	83
Manipulation without pre-processing		1843
Manipulation with post-processing	Blurring along spliced edges	848
	Blurring on other regions	131
Manipulation without post-processing (Blurring)		4144
Size of Tampered Region	Small	3358
	Medium	819
	Large	946

3. DATABASE EVALUATION

Since the Columbia Image Splicing Detection Evaluation Dataset and the Columbia Uncompressed Image Splicing Detection Evaluation Dataset are the first and only published image datasets for image tampering detection in recent years, we here compared the organization of our constructed tampering databases with them. The source, number of images, size, format and tampering operation (method) of these datasets are considered and the comparison can be found in Table 3.

Table 3: The comparison of CAISA databases and Columbia databases.

Database	Size	Components	Format	Operation Method
Columbia	933 authentic and 912 spliced	128 x 128 gray image blocks	BMP	Simple Splicing
Columbia (Uncompressed)	183 authentic and 180 spliced	from 757 x 568 to 1152 x 768 color image blocks	TIFF	Simple Splicing
CASIA V1.0	800 authentic and 921 spliced	374 x 256 color image	JPEG	Splicing using Photoshop with pre-processing
CAISA V2.0	7200 authentic and 5123 tampered	from 320 x 240 to 800 x 600 color image	JPEG, BMP, TIFF	Splicing using Photoshop with pre-processing and/or post-processing

We designed a test among people to evaluate the quality of our generated tampered images. In this test, we asked 30 people to give a judgement of several seeing images. Each person has to label 100 images and give their judgments. The 100 images are consist of 50 authentic and 50 tampered images which randomly selected from our databases (both V1.0 and V2.0). Tester should mark the seeing image if he/she thought it is an tampered one. We collect an average manually tampering detection results which is 58.73% from our tests. The results is barely better than a random thought, which illustrates that our generated tampered images are very realistic to human eyes. Similar tests are done in Columbia Uncompressed Image Splicing Detection Evaluation Dataset. The detection results by human eyes are approaching 100% since the splicing in this dataset is very simple and one can easily notice the spliced region. The test for Columbia Image Splicing Detection Evaluation Dataset is not applied in our

Table 4: Cross validation experimental results of tampering detection by method of [8].

	Columbia	V1.0	V2.0
Columbia	93.1%	43.6%	34.6%
V1.0	55.4%	85.6%	65.7%
V2.0	59.0%	69.1%	96.0%

experiments because this dataset only contains gray images and the image resolution is relatively low for eye detection.

We also designed an cross validation experiment to compare our datasets with Columbia dataset. Usually in machine learning, features can be learned well if the test samples are well discriminative and vice versa. We used the tampering detection features proposed in [8] for our comparison. The detection feature is based on the analysis of image chroma component and its statistics. SVM are used for classification. Detailed information about this method can be found in [8]. Table 4 shows the cross validation experimental results of these three datasets. The first column means training sample source and the first raw indicates testing sample source. The detection rates in the diagonal line of Table 4 show that the training and testing are performed in the same image dataset. Other figures in the table indicate the results by the detection method performed in different training and testing datasets. For example, the detection rate of 43.6% in second raw shows the result of our experiment by using Columbia Uncompressed dataset as the training samples and then tested the trained model on CASIA ITDE V2.0 dataset. From Table 4 we can notice that the testing results trained in CASIA ITDE datasets are better than the results trained in Columbia dataset. These results also indirectly insinuate that our datasets contains more various and challenging tampered image samples.

4. CONCLUSION AND AVAILABILITY

Passive-blind image tampering detection techniques raises researchers' interests and open benchmark databases for these techniques should equipped. In this paper, we have introduced our constructed CASIA Image Tempering Detection Evaluation Database and their motivations, design criterions, structures and self-evaluations. The databases are now available online at <http://forensics.idealtest.org/>. The main purpose for constructing CASIA ITDE Database is to evaluate algorithms. To be sure, our database V1.0 and V2.0 can only reflect a part of real-world image tampering cases. However, the importance of such evaluations as well as the database used for the promotion of this research filed should not be undervalued. Since the release of our database in late 2010, we have received more than 500 times downloaded application over 30 countries. We believe the construction of such image databases and making them available to the public would benefit our community.

5. REFERENCES

- [1] Hany Farid, “Examples of photo tampering throughout history.,” in <http://www.cs.dartmouth.edu/farid/research/digitaltampering>.
- [2] Wei Wang, Jing Dong, and Tieniu Tan, “A survey of passive image tampering detection,” in *8th International Workshop on Digital Watermarking, Springer Verlag*, 2009, pp. 308–322.
- [3] Hany Farid, “A survey of image forgery detection.,” in *IEEE Signal Processing Magazine.*, 2009, vol. 26, pp. 16–25.
- [4] Hany Farid, “Digital forensic database.,” in <http://www.cs.dartmouth.edu/farid/dfd/index.php/publications/showlist/year>.
- [5] Xiaoying Feng and Gwenael Doerr, “Jpeg recompression detection,” in *SPIE Conference on Media Forensics and Security*, 2010.
- [6] Sevinc Bayram, Husrev T Sencar, and Nasir Memon, “An efficient and robust method for detecting copy-move forgery,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2009, pp. 1053–1056.
- [7] Zhenhua Qu, Guoping Qiu, and Jiwu Huang, “Detect digital image splicing with visual cues,” in *11th International Workshop on Information Hiding*, 2009, pp. 247–261.
- [8] Wei Wang, Jing Dong, and Tieniu Tan, “Effective image splicing detection based on image chroma,” in *IEEE International Conference on Image Processing*, 2009.
- [9] Gang Cao, Yao Zhao, and Rongrong Ni, “Edge-based blur metric for tamper detection,” in *Journal of Information Hiding and Multimedia Signal Processing*, vol. 1:1, pp. 20–27.
- [10] Ashwin Swaminathan, Min Wu, and K. J. Ray Liu, “Component forensics of digital cameras: A non-intrusive approach,” in *40th Annual Conference on Information Sciences and Systems*, pp. 1194–1199.
- [11] Mo Chen, Jessica Fridrich, Miroslav Goljan, and Jan Lukas, “Determining image origin and integrity using sensor noise,” in *IEEE Transactions on Information Forensics and Security*, vol. 3, pp. 74–90.
- [12] Husrev T. Sencar Sevinc Bayram and Nasir Memon, “Discrimination of computer synthesized or recaptured images from real images,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing, Taipei, Taiwan*, pp. 1053–1056.
- [13] A. Dirik and N. Memon, “Image tamper detection based on demosaicing artifacts,” in *Proceedings of IEEE International Conference on Image Processing*, p. 1497C1500.
- [14] Tian-Tsong Ng and Shih-Fu Chang, “A data set of authentic and spliced image blocks,” in *ADVENT Technical Report 203-2004-3 Columbia University*, June 2004.
- [15] CASIA ITDE Database, “<http://forensics.idealtest.org/>” .
- [16] Corel Database, “<http://corel.digitalriver.com/>” .