Overview of the ImageCLEFsecurity 2019: File Forgery Detection Tasks*

Konstantinos Karampidis[1], Nikos Vasillopoulos[1], Carlos Cuevas[2], Carlos Roberto del-Blanco[2], Ergina Kavallieratou[1] and Narciso Garcia[2]

[1] AIlab, Department of Information & Communication Systems Engineering, University of the Aegean, Greece
[2] Grupo de Tratamiento de Imágenes, Universidad Politécnica de Madrid, Spain
Imageclefsecurity@aegean.gr

Abstract. The File Forgery Detection tasks is in its first edition, in 2019. This year, it is composed by three subtasks: a) Forged file discovery, b) Stego image discovery and c) Secret message discovery. The data set contained 6,400 images and pdf files, divided into 3 sets. There were 61 participants and the majority of them participated in all the subtasks. This highlights the major concern the scientific community shows for security issues and the importance of each subtask. Submissions varied from a) 8, b) 31 and c) 14 submissions for each subtask, respectively. Although the datasets were small, most of the participants used deep learning techniques, especially in subtasks 2 & 3. The results obtained in subtask 3 -which was the most difficult one- showed that there is room for improvement, as more advanced techniques are needed to achieve better results. Deep learning techniques adopted by many researchers is a preamble in that direction, and proved that they may provide a promising steganalysis tool to a digital forensics examiner.

Keywords: File Forgery Detection, Digital Forensics, Forged Image, Stego Image.

1 Introduction

The File Forgery Detection tasks described in this paper are part of the ImageCLEF benchmarking campaign [1–4], a framework where researchers can share their expertise and compare their methods based on the exact same data and evaluation methodology in an annual rhythm. ImageCLEF is part of CLEF (Cross Language Evaluation Forum). More details about the 2019 campaign are described in Ionescu et al. [5]. In general, ImageCLEF aims at building tasks that are related to benchmark the challenging task of image annotation for a wide range of source images and annotation objectives, since 2003.

The File Forgery Detection has started in 2019 as a new task. It is an important and serious issue concerning digital forensics examiners. Fraud or counterfeits are common causes for altering files. Another example is a child predator who hides porn images by altering the image extension and in some cases by changing the image signature. Many proposals have been made to solve this problem and the most promising ones concentrate on the image content. It is also common that someone who wants to hide information in plain sight without being perceived might use steganography. Steganography is the practice of concealing a file, message, image, or video within another file, message, image, or video. Among them, images are the most usual cover medium for hiding data. Thus, the File Forgery Detection is composed by three different subtasks, namely:

- Forged File Discovery
- Stego Image Discovery
- Secret Message Discovery

This paper presents an overview of the ImageCLEF2019 File Forgery Detection subtasks: the own subtask descriptions are in Section 2, the dataset in Section 3, and an explanation of the evaluation framework in Section 4. The participant approaches are described in Section 5, followed by a discussion and the conclusions in Sections 6.

2 Subtasks

The specific objective of these tasks are first to examine if an image has been forged, and then, if it could hide a text message. Last objective is to retrieve the potentially hidden message from the forged steganography images. Subtask 1 focuses on file forgery. A file can be considered forged whether it has an altered extension or signature (also known as magic bytes). If a file has an altered extension or signature, it is rather simple to identify it. The problem relies in the case when both a file's extension and signature have been altered at the same time. In this case, even the most used digital forensic software cannot identify a file as forged. Subtask 2 concerns the discovery of stego images. Images are the most widespread cover mediums for steganographic content. Steganography concerns the hiding of information into a cover medium which is in plain sight, while steganalysis (our main objective in this subtask) tries to detect its existence (subtask 2) and ideally retrieve the hidden message (subtask 3) [6].

The participant takes the role of a professional digital forensic examiner collaborating with the police, who suspects that there is an ongoing fraud in the Central Bank. After obtaining a court order, police gain access to a suspect's computer in the bank with the purpose of looking for images proving the suspect guilty. However, police suspects that the suspect managed to change file extensions and signatures of some images, so that they look like PDF (Portable Document Format) files or other types. It is probable that the suspect has used steganography software to hide messages within the forged images that can reveal valuable information. The considered subtasks are defined as follows:

- Subtask 1: perform detection of altered (forged) images (both extension and signature) and predict the actual type of the forged file.

- Subtask 2: identify the altered images that hide steganographic content.
- Subtask 3: retrieve the hidden messages (text) from the forged steganographic images.

3 Dataset

The data set consists of 6,400 forged images and pdfs, divided into 3 groups as shown in Table 1. Every group of images was used for a specific task.

Table 1. Number of files per subtask in the data set

	Subtask 1	Subtask 2	Subtask 3
Training Set	2400	1000	1000
Test Set	1000	500	500

All participants had access to the training data sets along with their respective ground truth. The test sets were distributed without the ground truth.

Training set for forged file discovery (i.e. subtask 1) consisted of 2400 files: 1200 of them were true pdf files, whilethe rest seem to be pdf files, but they actually were images (equally distributed among jpg, png, and gif image types). Conversion to pdf files was made by changing their extension to pdf and their signature (the first four bytes) to 25 50 44 46. Training set for stego image discovery (i.e. task 2) consisted of 1000 images of jpg format: 500 of these images were clean, while the rest were stego (Figures 1,2).

Fig. 1. A clean image

Fig. 2. A stego image

Training set for secret message discovery (i.e. task 3) contained 1000 images of jpg format: 500 of them were clean, while the rest contained different text messages (although, the same one for every 100 images). A Least Significant Bit (LSB) insertion technique was used to insert text messages, concerning the presumed dialogue the suspect had with his abettor.

4 Evaluation Framework

For assessing the performance, classic metrics were used:

a) Precision, Recall, and F-measure for Task 1 and Task 2.
b) Edit distance for Task 3.

In pattern recognition, information retrieval, and binary classification, Precision is the fraction of relevant instances among the retrieved instances. For the task 1, Precision could be defined as the fraction of actual detected altered images among all the images detected as altered:

$$Precision = \frac{\text{nº of actual detected altered images}}{\text{Total detections of altered images}}$$

For the task 2, Precision could be defined as the fraction of actual detected images with hidden messages among all the detected images with hidden a message:

$$Precision = \frac{nº\ of\ actual\ detected\ images\ with\ hidden\ messages}{Total\ detections\ of\ altered\ images\ with\ hidden\ messages}$$

Recall is the fraction of relevant instances that have been retrieved over the total amount of relevant instances.
For the task 1, Recall could be defined as the fraction of actual detected altered images among all the altered images:

$$Recall = \frac{nº\ of\ actual\ detected\ altered\ images}{Total\ altered\ images}$$

For the task 2, Recall could be defined as the fraction of actual detected images with hidden messages among all the images with hidden a message:

$$Recall = \frac{nº\ of\ actual\ detected\ images\ with\ hidden\ messages}{Total\ altered\ images\ with\ hidden\ messages}$$

F-measure is the harmonic mean of Precision and Recall, mathematically expressed as

$$F = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

For the task3, the edit distance is adopted, which is defined as follows. Given two strings, a and b, on an alphabet Σ (e.g. the set of ASCII characters), the edit distance $d(a,b)$ is the minimum-weight series of edit operations (Insertion, Deletion, Substitution) that transforms a into b.

5 Challenge Submissions

This section shows the results achieved by the participants in the three subtasks. Table 1 contains the results of subtask 1, Table 2 contains the results of subtask 2, and Table 3 contains the results of subtask 3.

5.1 Results for subtask 1

Six runs were submitted by four groups to this subtask. Table 1 shows the details of the results, while Figure 1 summarizes the F-measure, Precision and Recall per run. The correspondences between run IDs and participant names are given in Table 1.

Table 1: Runs summary table for Subtask 1.

Rank	runID	Participant	F-measure	Precision	Recall
1	26850	UA.PT_Bioinformatics	1.000	1.000	1.000
2	26738	nattochaduke	1.000	1.000	1.000
3	26737	nattochaduke	1.000	1.000	1.000
4	26735	agentili	1.000	1.000	1.000
5	26994	abcrowdai	0.748	0.798	0.703
6	26954	abcrowdai	0.538	0.756	0.417

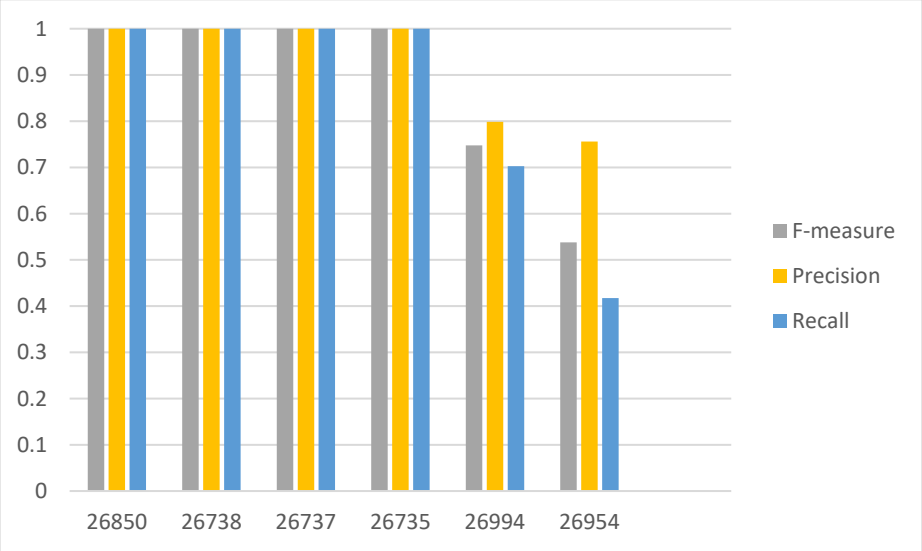

Figure 1. F-measure, Precision and Recall per submitted runID for Task 1.

5.2 Results for subtask 2

Twenty six runs were submitted by six groups to this subtask. Table 2 shows the details of the results, while Figure 2 summarizes the F-measure, Precision and Recall per run. The correspondences between run IDs and participant names are given in Table 2.

Table 2: Runs summary table for Subtask 2.

Rank	runID	Participant	F-measure	Precision	Recall
1	26934	UA.PT_Bioinformatics	1.000	1.000	1.000
2	26929	UA.PT_Bioinformatics	0.986	1.000	0.972
3	26932	UA.PT_Bioinformatics	0.980	0.980	0.980
4	26930	UA.PT_Bioinformatics	0.965	0.939	0.992
5	26867	UA.PT_Bioinformatics	0.945	0.996	0.900
6	26871	UA.PT_Bioinformatics	0.933	0.891	0.980
7	26864	UA.PT_Bioinformatics	0.933	0.874	1.000
8	26868	UA.PT_Bioinformatics	0.932	1.000	0.872
9	26816	agentili	0.888	0.908	0.868
10	26830	nattochaduke	0.660	0.508	0.944
11	26844	Yasser	0.626	0.524	0.776
12	26876	Yasser	0.625	0.537	0.748
13	26825	Yasser	0.614	0.529	0.732
14	26842	Yasser	0.613	0.518	0.752
15	26817	nattochaduke	0.613	0.473	0.872
16	26771	nattochaduke	0.613	0.479	0.852
17	26951	Yasser	0.599	0.542	0.668
18	26950	Yasser	0.599	0.542	0.668
19	26948	Yasser	0.587	0.538	0.644
20	26949	Yasser	0.585	0.525	0.660
21	26885	Yasser	0.576	0.506	0.668
22	26952	Yasser	0.574	0.508	0.660
23	26787	nattochaduke	0.529	0.542	0.516
24	26910	Abcrowdai	0.525	0.467	0.600
25	27454	cen_amrita	0.438	0.422	0.456
26	26770	Nattochaduke	0.243	0.673	0.148

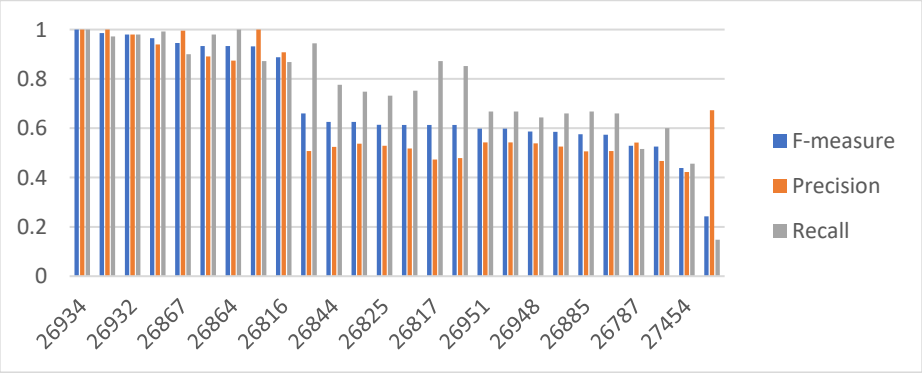

Figure 2. F-measure, precision and recall per submitted runID for Task 2.

5.3 Results for subtask 3

Eleven runs were submitted by two groups to this subtask. Table 3 shows the details of the results, while Figure 3 summarizes the edit (Levenshtein) distance per run. The correspondences between run IDs and participant names are given in Table 3.

Table 3: Runs summary table for Subtask 3.

Rank	runID	Participant	Edit distance
1	27447	UA.PT_Bioinformatics	0.59782861
2	26933	UA.PT_Bioinformatics	0.59558861
3	27162	UA.PT_Bioinformatics	0.588343826
4	27438	UA.PT_Bioinformatics	0.587247762
5	26904	UA.PT_Bioinformatics	0.586426775
6	26898	UA.PT_Bioinformatics	0.571236169
7	26896	JoÃ£o Rafael Almeida	0.563379028
8	26899	UA.PT_Bioinformatics	0.529075304
9	27446	UA.PT_Bioinformatics	0.293547989
10	27445	UA.PT_Bioinformatics	0.27119247
11	26869	JoÃ£o Rafael Almeida	0.083585804

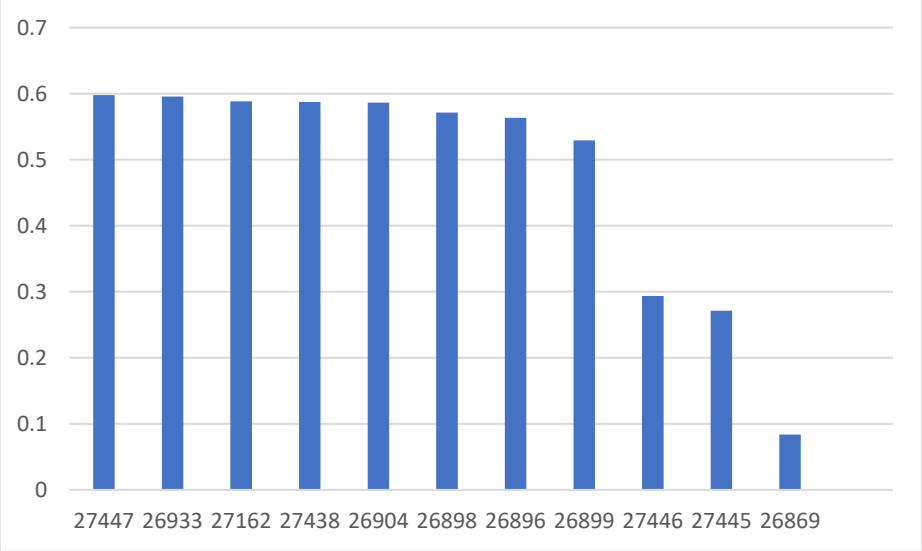

Figure 3. Edit distance per submitted runID for Task 3.

6 Discussion and Conclusions

The security task was introduced in ImageCLEF 2019. The number of the registered teams/individuals and the submitted runs showed that the security challenges receive a significant attention and that they are interesting and challenging. Most participants signed to all three tasks, although this was not mandatory. This fact highlights the importance of each task. The majority of the approaches exploited and combined deep learning techniques, achieving very good results. The third task has been the most challenging one, in which the participants had to retrieve hidden messages from the images. The third task results have also shown that there is room for improvement, as more advanced techniques need to be used for better results. The analysis of the specific task results indicates that the training set was small for the specific problem, i.e., the extraction of the hidden messages. To leverage the power of advanced deep learning algorithms towards improving the state-of-the-art in steganalysis, we plan to increase the data set. We also plan to narrow down the application of the challenges, e.g., focus in steganalysis, probably in another domain.

References

1. Ionescu, B., Muller, H., Peteri, R., Cid, Y.D., Liauchuk, V., Kovalev, V., Klimuk, D., Tarasau, A., Abacha, A.B., Hasan, S.A., Datla, V., Liu, J., Demner-Fushman, D., Dang-Nguyen, D.T., Piras, L., Riegler, M., Tran, M.T., Lux, M., Gurrin, C., Pelka, O., Friedrich, C.M., de

Herrera, A.G.S., Garcia, N., Kavallieratou, E., del Blanco, C.R., Rodrıguez, C.C., Vasillopoulos, N., Karampidis, K., Chamberlain, J., Clark, A., Campello, A.: ImageCLEF 2019: Multimedia retrieval in medicine, lifelogging, security and nature. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 10th International Conference of the CLEF Association (CLEF 2019), LNCS Lecture Notes in Computer Science, Springer, Lugano, Switzerland (September 9-12 2019).

2. Kalpathy-Cramer, J., Garc´ıa Seco de Herrera, A., Demner-Fushman, D., Antani, S., Berick, S., Muller, H.: Evaluating performance of biomedical image retrieval systems: Overview of the medical image retrieval task at ImageCLEF 2004–2014. Computerized Medical Imaging and Graphics 39(0) 55 – 61(2015).
3. Clough, P., Muller, H., Sanderson, M.: The CLEF 2004 cross–language image retrieval track. In Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B., eds.: Multilingual Information Access for Text, Speech and Images: Result of the fifth CLEF evaluation campaign. Volume 3491 of Lecture Notes in Computer Science (LNCS)., Bath, UK, Springer 597–613 (2005).
4. Caputo, B., Muller, H., Thomee, B., Villegas, M., Paredes, R., Zellhofer, D., Goeau, H., Joly, A., Bonnet, P., Gomez, J.M., et al.: Imageclef 2013: the vision, the data and the open challenges. In: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer 250–268 (2013).
5. B. Ionescu, H. Müller, R. Péteri, D.T. Dang-Nguyen, L. Piras, M. Riegler, M.T. Tran, M. Lux, C. Gurrin, Y.D. Cid, V. Liauchuk, V. Kovalev, A. Ben Abacha, S.A. Hasan, V. Datla, J. Liu, D. Demner-Fushman, O. Pelka, C.M. Friedrich, J. Chamberlain, A. Clark, A. García, N. García, E. Kavallieratou, C.R. del Blanco, C. Cuevas, N. Vasillopoulos, K. Karampidis, "ImageCLEF 2019: Multimedia Retrieval in Lifelogging, Medical, Nature, and Security Applications", 41st Eur. Conf. on IR Research, ECIR 2019, Cologne (Germany), pp 301-308, 14-18 Apr. (2019).
6. K. Karampidis, E. Kavallieratou, and G. Papadourakis, ''A review of image steganalysis techniques for digital forensics,'' J. Inf. Secur. Appl., vol. 40, pp. 217–235, Jun. 2018.