PGP-AIML

# **Capstone Project Proposal**

# **Project Trilochana**

## **Domain & Context:**

File Forgery Detection (FFD) is a serious problem 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 anyone who wants to hide any kind of information in plain sight without being perceived to use steganography. Steganography is the practice of concealing a file, message, image or video within another file, message, image, or video. The word steganography combines the Greek words steganos (στεγανός), meaning “covered” and graphein (γράφειν) meaning “writing”. The most usual cover medium for hiding data are images.

### **Problem Statement:**

As 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 to look for images proving the suspect guilty. However, police suspects that he has managed to change extension and signature of some images, so that they look like pdf files. Additionally, it is highly probable that the suspect has used steganography software to hide messages within some images that could reveal valuable information of his collaborators.

The goals are:

1. Examine if an image has been forged. Perform detection of altered (forged) images (both extension and signature) and predict the actual type of the forged file.
2. Examine if an image could hide a text message. Identify the altered images that hide steganographic content
3. Retrieve the potential message from the forged stego images. Extract the hidden messages (text) from the stego images

### **Dataset:**

The above problem statement is part of the challenge provided by IMAGECLEF Conference (<https://www.imageclef.org/2019>). The IMAGECLEF 2019 is an evaluation campaign that is being organized as part of the CLEF(http://www.clef-initiative.eu) initiative labs. The campaign offers several research tasks that welcome participation from teams around the world.

The dataset is provided by the ImageCLEF and consists of the following data:

1. Training set for forged file discovery (i.e.task 1) consists of 2400 files. 1200 of them are true pdf files and the rest seem to be pdf files but actually they are images (400 of each image type i.e. jpg , png and gif).
2. Training set for stego image discovery (i.e. task 2 ) consists of 1000 images of jpg format. 500 of these images are clean while the rest are stego.
3. Training set for secret message discovery (i.e. task 3) contains 1000 images of jpg format. 500 of them are clean while the rest contain text messages different for every 100 images.

The overall size of the dataset comes to be around 2.5 Gigabytes.

Kindly note only TrainingSet data is provided. The testset can be created by us either from the existing training set or use our own methods to create test set data.

### **References:**

<https://www.crowdai.org/challenges/imageclef-2019-security-forged-file-discovery>

<https://www.crowdai.org/challenges/imageclef-2019-security-stego-image-discovery>

<https://www.crowdai.org/challenges/imageclef-2019-security-secret-message-discovery>