Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı, S. 444-451, Ekim 2019

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue, pp. 444-451, October 2019

Research Article

Local Image Descriptor Based Phishing Web Page Recognition as an Open-Set Problem

Ahmet Selman Bozkir[1*] and Murat Aydos[2]

[1] Hacettepe University, Faculty of Engineering, Department of Computer Engineering, Ankara, Turkey (ORCID: 0000-0003-4305-7800)
[2] Hacettepe University, Faculty of Engineering, Department of Computer Engineering, Ankara, Turkey (ORCID: 0000-0002-7570-9204)

(This publication has been presented orally at HORA congress.)

(First received 1 August 2019 and in final form 25 October 2019)

(**DOI:** 10.31590/ejosat.638404)

ATIF/REFERENCE: Bozkır, A. S. & Aydos, M. (2019). Local Image Descriptor Based Phishing Web Page Recognition as an Open-Set Problem. *European Journal of Science and Technology*, (Special Issue), 444-451.

Abstract

With the advent of e-commerce, digital services and social media, scammers have changed their way to gain illegal benefits in various forms such as capturing the credit card information or exploiting personal cloud accounts which is termed as phishing. For this reason, against this cyber crime, last two decades have witnessed a variety of combatting methodologies like HTML content based similarity analysis, URL based classification and recently visual similarity based matching since phishing web pages visually mimic to their legitimate counterparts in order to create an illusion to deceive innocent users. To this end, in this study, we propose a computer vision and machine learning based approach in order to classify whether a suspicious web page is phishing and further recognize its original brand name. In this regard, we have utilized and investigated two different local image descriptors namely Scale Invariant Feature Transform (SIFT) and DAISY. Apart from their common properties such as scale invariance, the aforementioned descriptors have apparent differences such that in addition to rotational invariance, SIFT employs key-point based sampling whereas DAISY applies dense sampling by default. Therefore, we first aimed to investigate the feasibility of these two local image descriptors in addition to revealing the effects of sampling strategy and rotational invariance in problem domain. Furthermore, in order to create a discriminative representation of a web page, we followed the bag of visual words (BOVW) approach having different vocabulary sizes such as 50, 100, 200 and 400. In order to evaluate the proposed approach, we have utilized a publicly available phishing dataset including snapshots of webpages sampled from both 14 different highly phished brands and ordinary legitimate web pages yielding a challenging open-set problem. The aforementioned dataset involves 1313 training and 1539 testing image samples in total. The visual features extracted via SIFT and DAISY were first transformed to a BOVW histogram and fed to three different machine learning methods such as SVM, Random Forest and XGBoost. According to the conducted experiments, based on a 400-D visual vocabulary, SIFT descriptor along with XGBoost has been found as the best descriptor-learner configuration having reached up to 89.34% validation accuracy with 0.76% false positive rate. Moreover, SIFT has outperformed DAISY descriptor in all settings. As a result, it has been shown that SIFT descriptors equipped with BOVW representation can be effectively used for brand identification of phishing web pages.

Keywords: Phishing, Computer Vision, Machine Learning, SIFT, DAISY.

Bir Açık Küme Problemi Olarak Yerel Görsel Betimleyicilerle Oltalayıcı Web Sayfalarının Tanınması

Öz

E-ticaret, sayısal hizmetler ve sosyal medyadaki gelişmelerle birlikte siber saldırganlar illegal kazanç sağlama adına günümüzde "Oltalama" olarak ifade edilen ve kredi kartı veya kişisel bulut hesaplarına ait hesap bilgilerini ele geçirmek gibi amaçları olan yeni bir saldırı türünü benimsemişlerdir. Bu nedenle bu siber suça karşı son yirmi yılda HTML içerik temelli benzerlik analizi, URL tabanlı sınıflandırma ve masum kullanıcıları yanıltmak için sahte sayfaların özgün sürümlerini andırmasından dolayı son zamanlarda görsel

[*] Corresponding Author: Hacettepe University, Faculty of Engineering, Dept. of Computer Engineering, Ankara, Turkey, ORCID: 0000-0003-4305-7800, selman@cs.hacettepe.edu.tr

benzerlik temelli eşleştirme gibi çeşitli mücadele yöntemleri geliştirilmiştir. Bu çalışmada şüpheli bir web sayfasının oltalayıcı sayfa olup olup olmadığını sınıflandırmak ve orijinal marka adını daha iyi tanımak için bilgisayar görüsü ve makine öğrenmeye dayalı bir yaklaşım önerilmiştir. Bu bağlamda Scale Invariant Feature Transform (SIFT) ve DAISY olmak üzere iki farklı yerel görsel betimleyicisi araştırılmış ve kullanılmıştır. Ölçek duyarsızlığı gibi ortak özelliklerinin yanı sıra, bahsi geçen betimleyicilerin dönme duyarsızlığına ek olarak bazı bariz farklılıkları bulunmaktadır. Örnek olarak SIFT betimleyicileri anahtar nokta temelli örnekleme uygularken, DAISY varsayılan olarak yoğun bir örneklemeyi tercih etmektedir. Bu nedenle, bu calışmada ilk önce örnekleme stratejisi ve dönel değişmezliğin problem uzayındaki sonuçlarından ziyade bu iki yerel görüntü betimleyicisinin uygulanabilirliği araştırılmıştır. Ayrıca, web sayfalarından ayırt edici bir temsil elde etmek için görsel kelime çantası (Bag of Visual Words - BOVW) yaklaşımı benimsenmiş ve 50, 100, 200 ve 400 gibi farklı kelime sayısına sahip temsiller üretilmiştir. Önerilen yaklaşımın değerlendirilmesinde oltalama saldırısına yoğunlukla maruz kalan 14 markaya ve çeşitli özgün web sayfalarına ait sayfa şipşakları içeren zorlayıcı bir veri kümesinden yararlanılmıştır. İlgili veri kümesi makine öğrenimi açısından "açık küme problemi" taşımakta ve bünyesinde toplam 1313 eğitim ve 1539 test görsel örneği ihtiva etmektedir. SIFT ve DAISY betimleyicileri ile çıkarılan görsel özellikler ilk olarak BOVW histogramına dönüştürülmüş, sonrasında SVM, Random Forest ve XGBoost gibi üç farklı makine öğrenme yöntemleri kullanılarak eğitilmiştir. Yapılan deneylere göre 400 görsel kelime dağarcığı ile yapılandırılan SIFT betimleyicileri, XGBoost ile birlikte %0.76 FPR ve %89.34 geçerleme doğruluğuna ulaşmış ve en iyi betimleyici-makine öğrenimi modeli çifti olarak tespit edilmiştir. Ayrıca, SIFT tüm konfigurasyonlarda DAISY betimleyicisindan daha iyi performans göstermektedir. Sonuç olarak, BOVW temsiline dayalı SIFT betimleyicilerinin oltalayıcı web sayfalarının hangi markaya ait olduğunun tanınmasında etkin bir şekilde kullanılabileceği gösterilmiştir.

Anahtar Kelimeler: Oltalama saldırıları, Bilgisayarlı Görü, Makine Öğrenmesi, SIFT, DAISY.

1. Introduction

With the advent of e-commerce, digital services and social media, scammers have changed their way to gain illegal benefits in various forms such as capturing the credit card information or exploiting personal cloud accounts which is called as phishing. These kind of private information is usually employed for various scamming activities such as credit card frauding and stealing accounts for cloud and streaming services. According to the Anti Phishing Working Group's 2017 4th quarterly report, the number of unique phishing attacks for 2017 was detected higher than 700.000 [3]. Further, as of October 2017, the number of affected companies has increased to 348 [4]. It should also be noted that, payment services, insurance companies and digital cloud based servicing firm come into prominence among the main sectors on which phishing attacks usually have the major impact. In general, life cycle of a phishing attack starts with designing and sending spoofed emails to innocent users over the Internet [1] and making them to believe that they are being received from their legitimate counterparts such as banks or governmental agencies [2].

Note that, not all phishing attacks are relying on transmission via emails. In fact, today, various mediums such as social media, SMS and other platforms are being exploited for transmitting the URLs of real phishing assets – namely *phishing web pages*. Therefore, to date, various phishing combatting methologies have been proposed in literature. According to Rao and Pais [5], these methodologies can be categorized into 4 groups: (a) list based, (b) heuristics based, (c) machine learning based and finally (d) visual similarity based. Most of the studies under these categories attempt to identify phishing instances by employing different source of information that can be extracted from web pages. On the other hand, in order obtain a fast and reliable security, list based approaches such as Google Safe Browsing API, utilize a *black* or *white* list in order to keep phishing or legitimate URLs respectively. However, keeping such kind of a huge list is costly and susceptible to "zero-hour" attacks causing vulnerability to undiscovered and non-reported new phishing web pages. As stated in [2,4], heuristics bases approaches employ several features from various source of information such as image, text, URL and DNS records from both legitimate and phishing web sites. Meanwhile, these features are often fed into machine learning methods in order to create effective classifiers for identifying whether a suspicious web page is phishing. Nevertheless, Varshney et al. [2] adress some drawbacks of this kind of approach in perspective of (1) the required time and computational resources for training, (2) limited ability of usage in browsers and (3) the potential adversarial attacks once the key features are discovered by scammers. In contrast, vision based studies employ pure computer vision methods for decision making processes. Considering the phishing domain, vision based studies analyze either whole screenshot image (i.e. hollistic) or part of the input images to find out discriminative information for recognition purposes.

Apart from one dimensional code, web pages can actually be considered as two dimensional visual stimuli since they are being perceived by users as a graphical entity following to rendering process by browsers [4]. Based on this fact, scammers design and implement phishing web pages being visually similar to their counterparts defined as *visual deception* [6]. Thus, it is possible to detect phishing web pages through visual similarity perspective. This approach has several benefits. First, analyzing visual similarity makes the combatting mechanism invariant to underlying HTML and related source codes. Second, same rendering outcome can be generated even with large DOM (Document Object Model) modifications. At this point, vision based analysis will also be invariant to DOM pertubations and the language of the page. Third, phishers often replace textual contents with other type of elements such as images, Flash objects and Java Applets [7]. Therefore, HTML content based methods may fail against this type of delusions. Fourth, though phishers might slightly alter the visual appearance, these modifications should not exceed a certain threshold; otherwise users can understand the difference and may have suspicion yielding to lower deception. Thus, the distortions would be limited to avoid comprimising the nature of the scam. Nonetheless, a smart vision based recognition scheme must handle slight layout alterations as well as color distortions.

In this paper, we proposed a computer vision based phishing web page recognition system based on bag of visual words pooling scheme employing SIFT (rotational invariant and keypoint based sampled) and DAISY (rotationaly variant and dense sampled)

descriptors seperately in order to either classify the brand name of the suspicious phishing web page or label it as legitimate. In this context, we aimed to examine the effect of sampling strategy and rotational variance in problem domain.Furthermore, we have investigated and compared the performance of these two descriptors in the field of phishing web page identification. Moreover, we have trained and tested the suggested scheme on a verified publicly available dataset named as "Phish-Iris" [4]. The dataset subjected to this study involves 1313 training and 1539 testing samples covering 14 different highly phished companies along with a dominant legitimate set. Therefore, the aim of this dataset is evaluating the performance of the classifier in an *open-set* environment. As is known, apart from the target classes, if the dataset involves the "unknown" or "other" class then the classification problem becomes an "open-set" rather than a "closed-set" where each instance is sampled from a known distribution. To our best knowledge, this study is the first which employes these two descriptors in order to create a histogram based bag-of-visual-words representation which will be fed into various machine learning methods such as Support Vector Machine (SVM), Random Forest (RF) and XGBoost (i.e. regularized decision tree boosting). To this end, we have experimented various codebooks in which their sizes are ranging from 50 to 400. According to the conducted experiments, SIFT based visual words trained by XGBoost learner outperforms the DAISY based scheme achiving 89.34% accuracy rate along with 0.0076 false positive rate. In line with this finding, we have also examined the effect of rotational invariance and different keypoint sampling strategies. The total duration regarding the process of inference including feature extraction, representation and classification was counted as 1.5 seconds in average. Moreover, the low memory footprint was also detected. These findings have also enable the proposed approach to be an effective and resource friendly scheme for several environments such as email server backends and mobile platforms. One another key advantage of our approach is that it can process screenshots without any resize operation since it has no size limit in terms of width and height.

The rest of this paper is organized as follows. In section 2, a brief summary about the vision based phishing detection studies was given. Section 3 explains some key points about the utilized image descriptors and the way we represent them. Section 4 briefly introduces the dataset we have used throughout the study. In section 5, details of the experiments were presented. Next, section 6 reports the results of the experiments along with a short discussion. Finally, section 7 concludes the paper.

2. Related Work

To date, several vision based anti-phishing approaches have been suggested in literature. Rao and Ali [8] have attempted to detect phishing attacks by employing SURF descriptors and pair-wise similarity measurement. Bozkir and Akcapinar Sezer [9] have proposed to employ HOG descriptors to reveal visual cues of the web pages in terms of corner and edges by also preserving spatial information. Concerning the whole page snapshots, similarity exceeding 75% indicates a strong phishing evidence in terms of canonical HOG based feature similarity. Zhang et al. [10], have proposed a scheme to capture phishing samples by considering the similarity of web page layout. For this purpose, they extracted web page blocks and employed graph based similarity metric powered by an r-tree based indexing technique. In another study Hara et al. [11] have extracted the images and URL of the web pages in order to be used as a source of information in order to identify phishing web pages. Moreoever, authors have used the "ImgSeek" tool to retrieve similar images on Internet regarding the image found in suspicious web page. In case of similarity exceeds a threshold value, the domain name of the images are being retrieved and compared with the tested instance. The main shortcoming of the suggested scheme is the neccesity of a third party service. Besides, the retrieval quality of the dependent service plays a key role for the success of detection. Dalgic et al. [4] have utilized MPEG-7 and MPEG-7 like compact descriptors for representing the snapshots of the legitimate and phishing web pages. In their work, they built holistic and multi-level path based representations via 6 different descriptors encoding color and edge based information. These representations were fed into multi class classifiers in order to create a brand recognition focused phishing detection system. Corona et al. [12] have followed a hybrid approach which encounters use of both image and HTML features in order to later fuse for a classification based phishing detection system. Apart from HTML content based analysis, they have leveraged descriptors of HOG and Color Context Histograms for revealing color and edge structures belonging to web page snapshots. Similar to Dalgic et al.'s work [4] they have also applied tile based finer grain analysis in terms of feature representation. Apart from having achieved high classification accuracy their work also exhibits a robust scheme against evasion techniques for phishing instances. The shortcoming of their work is that they built classifiers trained with more than 12000 features. Since the arms race between anti-phishers and scammers never ends up, anti phishing research employ every kind of source information for better generalization and accuracy. Li et al. [13] have presented a stacking model utilizing URL and HTML contents for creating an efficient and effective phishing detection system. Similar to other heuristics based studies, they used HTML related conventional features such as "existence of a login form" and "number of dots in domain name". However, ulike other traditional models, they employed deep learning based techniques such as replacing all textual contents with *embedding* vectors obtained from WordVec and modelling of deep convolutional neural networks for screenshot classification. In their work, they evaluated the performance of their work on a not publicly available dataset covering 50K dataset. According to the promising results, they reported 98.60% on accuracy and 1.54% on false alarm rate.

3. Visual Descriptors and Representation

In this study, we have explored the feasibility of two different yet common techniques involving local image descriptors namely SIFT and DAISY. This section aims to provide readers a simple background for these two descriptors. The main principles of creating bag-of-visual-words representation were also given in this section. Due to the page number limitation, explanations have been provided briefly. Thus, readers who need more details about the methologies given in this section are suggested to refer related papers.

3.1. Scale Invariant Feature Transform (SIFT)

Invented by Lowe [14], the main aim of SIFT descriptor is to produce a rotational and scale invariant description of the keypoints found in an input image. In other words, SIFT encodes the image gradients into a 128 dimensional vector on local patches each centered around the detected keypoints. Sachdeva et al. [15] has pointed out that SIFT features also show robustness against geometric transformations such as affine tranformation. Since its discovery, SIFT has found a widespread usage in numerous works such as image identification [16] and logo detection for vehicle images [17].

Broadly speaking, SIFT feature computation has two main steps: (1) key point detection and (2) key point descriptor generation. Lowe [14] employs *Difference of Gaussians* (DoG) rather than Laplacian of Gaussians (LoG) which presents high computational cost. Note that, there also exists different kinds of keypoint detectors such as Harris corner detector which is not scale invariant. In order to detect important key points Lowe [14] constructs a gaussian pyramid involving several layers which are successively blurred versions of the input image. The formulas of this procedure are given below in (1) and (2). Given a blurred instance of image I that is generated by convolution operation (at scale of σ) with a two dimensional Gaussian function, every layer of DoG pyramid can be calculated via subtraction of sucessive blurred instances of I.

$$L_\sigma(x,y) = I(x,y) * G_\sigma(x,y) \tag{1}$$

$$DoG_{k^{n+1}\sigma}(x,y) = L_{k^{n+1}\sigma}(x,y) - L_{k^{n}\sigma}(x,y) \tag{2}$$

Following to revealing the potential keypoints, descriptor generation mainly includes four sub-procedures which are: (1) scale space detection, (2) keypoint localization, (3) orientation assignment and finally (4) keypoint description. The main steps in keypoint descriptor generation is gradient computation and orientation assignment for patches around the detected keypoints. A sample SIFT keypoint detection along with DoG pyramids has been depicted in the Fig. 1 given below.

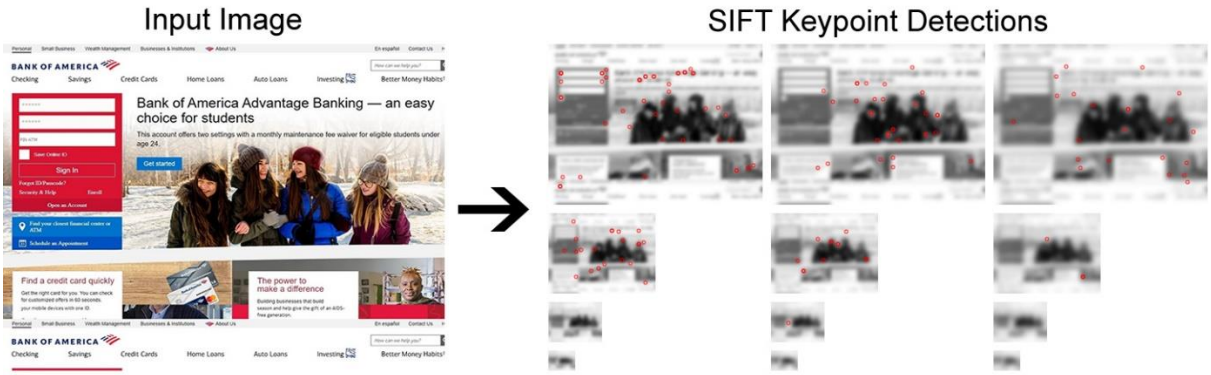

Figure 1. Example of SIFT keypoint detection via DoG pyramids for a phishing web page.

3.2. DAISY Descriptor

Created by Tola et al. [18], the DAISY descriptor has been designed as a local image descriptor aiming fast dense feature extraction. Further, in addition to being robust to photometric transformation such as contrast changes, DAISY descriptor also handles occlusions and perspective distortions [18]. Note that, DAISY's main goal is to be useful for estimating dense depth maps from wide-baseline image pairs. As its name suggested, the name of the descriptor comes from the flower *daisy* whose shape is like region of interest. In general, for feature extraction DAISY involves six hyperparameters: (1) R (radius) – the distance from the center of the inner most circle to the center of the most outer circle, (2) Q – the count of convolved orientations, (3) T – the count of histograms for a single layer, (4) H – number of bins in the histogram, (5) S – the count of histograms employed to generate the descriptor = Q×T +1 and (6) Ds – the size of the final descriptor = S×H [18]. As a result, produced descriptor length is directly dependent on Q, T and H parameters. As can be seen from the Fig. 2, the radius for each circle is proportional to the standard deviation of the Gaussian kernels. Moreover, the marks indicated with '+' highlight the locations from which the resultant descriptor is produced from the convolved orientation maps. In Fig. 2, an enlargement of 3 layers having different radius values and colors can be seen. According to this configuration, eight sampling points appear for each layer and the '+' indicated points are located having a distribution of 45 degree intervals.

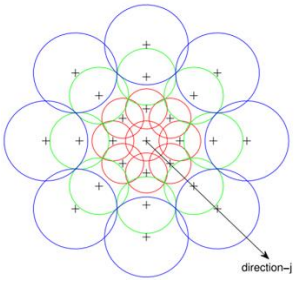

Figure 2. DAISY's central-symmetrical circle feature extraction pattern. Adopted from [18]

3.3. Representation of Bag of Visual Words

The bag of visual words (BOVW) representation is a visual feature oriented version of the conventional well known bag of words (BOW) concept which has been widely used in fields such as text classification and natural language processing. The main goal of BOVW is to represent an image as a set of *pooled* visual features regardless of how they were extracted. In practice, generation of BOVW representation covers following stages: (a) extracting *n* number of feature vectors from the input documents (i.e. image), (b) clustering the accumulated features with a certain cluster count (i.e. visual words) and determining the cluster centroids , (c) for the image *I*, assigning each extracted feature vector to the nearest cluster centroid (i.e. *quantization*) and (d) representing an image with a histogram computed by counting the frequency of each visual word following to cluster assignment. As a result of this pooling scheme, an image involving many feature vectors is being transformed to a histogram involving counts of visual words obtained before. Here, the concept holds for BOW also exists in BOVW based representation that is, rather than of textual words, we use pooled visual features as the "words".

4. The Dataset

In fact, although there exists numerous studies in phising detection literature, the number of vision based approaches is relatively low [21]. Since our aim to create a classifier based on visual features, a labeled dataset is crucial. For this reason, we have searched for a suitable phishing dataset including web page snapshots which were correctly labeled according to the brand name they mimic. We finally found a mid sized yet comprehensive dataset - provided by [21] - namely "Phish-IRIS" located at "https://web.cs.hacettepe.edu.tr/~selman/phish-iris-dataset/" containing totally 2852 snapshots covering 14 different highly phished brands and legitimate web page samples. The sizes of the images are ranging from 700 pixels to 1280 pixels in terms of both width and height. According to the authors, "Phish-IRIS" dataset has been collected between March-May 2018. The sample distribution of the brands in both training and testing categories were listed in Table 1. One of the good properties of the aforementioned dataset is that they supply a large legitimate sample set in order to test the open-set based classification behaviour against to the methodology applied.

Table 1. The snapshot distribution in Phish-IRIS dataset

Brand Name	Training Instances	Testing Instance
Adobe	43	27
Alibaba	50	26
Amazon	18	11
Apple	49	15
Bank of America	81	35
Chase Bank	74	37
Dhl	67	42
Dropbox	75	40
Facebook	87	57
Linkedin	24	14
Microsoft	65	53
Paypal	121	93
Wellsfarno	89	45
Yahoo	70	44
Other (i.e. Legitimate)	400	1000
Total	1313	1539

5. Experimental Study

In this study, as stated before, we have followed bag of visual words approach along with two different local image descriptors. In order to evaluate the effect of strong or weak features in terms of classification performance, we have chosen various number of visual words ranging from 50 to 400 (i.e 50, 100, 200 and 400). Since the increase in word count results sparser visual signatures, we are also enabled to measure the accuracy and false positive rate (FPR) according to the strength of the pooled features. In order to make the experiments we have written a Python 3 based script to accomplish following workflow: (1) K-means based clustering, (2) generation of codebooks (3) final histogram generation and (4) machine learning based classification via SVM [19], Random Forest and XGBoost [20] methods. In this context, we have utilized various Python libraries such as "Sklearn", "Numpy" and "XGB". The workflow of the proposed scheme has been depicted in Fig 3.

In order to extract SIFT features, we have utilized OpenCV's SIFT module built for Python. For obtaining DAISY features we have employed "SkImage" library. During the SIFT based extraction we have not changed any parameters. However, for DAISY based feature extraction, we have set the radius as 16, the number of rings as 3, the number of histograms as 6 and the number of orientations as 8. We have also experimented different hyper-paramaeters such as reducing the number of rings. Nonetheless, reduction of ring count has yielded decrease in performance. So, we have decided to keep it as 3. Meanwhile, throughout the feature extraction, regarding to whole training set we had extensive amount of feature vectors revealed by both of the descriptors separately. This has caused out of memory problems in conventional K-means clustering module. Therefore, in order to fix this issue, we have applied "Mini-Batch" K-Means algorithm with the batch size of 100. Furthermore, we employed initialization of "k-means++" method for determining the initial cluster centroids. In this way, we accelerated the convergence speed.

Throughout the training stage, we have utilized SVM, RF and XGBoost methods. For Support Vector Machine based learning, we have kept values of all parameters as default except the kernel choice. According to our initial experiments, radial basis function (i.e. RBF) achieves better results. Thus we have preferred the RBF kernel during SVM based training. The same regime (keeping the default values) has also been embraced for the training with other classification methods. At this point, it should be noted that, XGBoost classifier that we have used has the support of CUDA based GPU computation. Therefore, we have leveraged this performance improvement in order to reduce the amount of time during training.The whole workflow including the training stages were performed on a Ubuntu installed computer equipped with an Intel® Core ™ i7 8750 processor and 24 GB of memory.

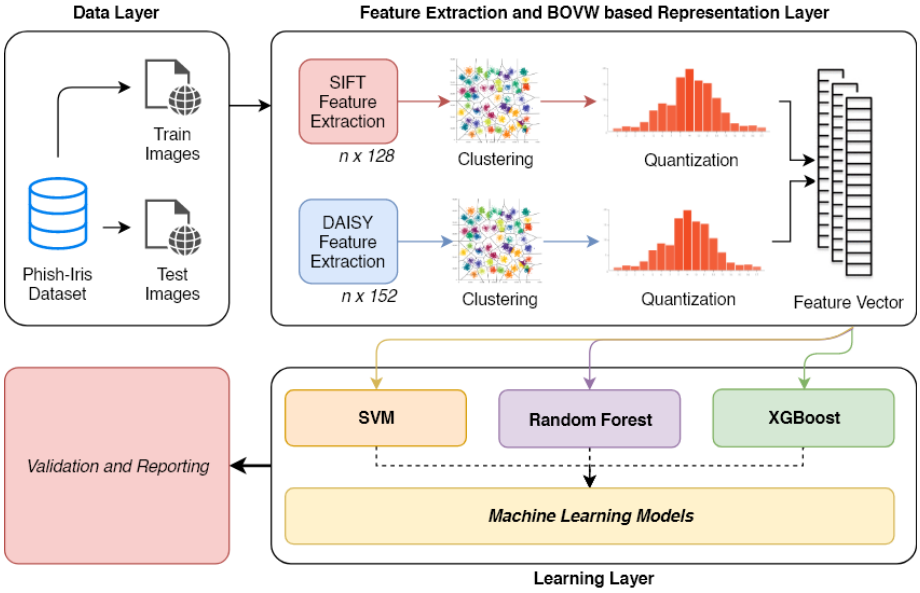

Figure 3. The workflow of the suggested scheme

6. Results and Discussion

This section presents the experimental results of the classification algorithms along with the used descriptor and the preferred codebook size. During the assesment, we have both evaluated the accuracy rate of training and testing samples. Moreover, the evaluation has been carried out by involving the metrics of true positive rate (TPR), false positive rate (FPR) and F-1 measure.According to the SIFT based results given in Table 2, we can infer the findings listed below:

- The highest accuracy has been achieved via XGBoost learner with 400-D codebook size.
- Larger the codebook size, higher the accuracy we have achieved. This implies that, our problem domain requires disciminative visual words even they are rarely found. Thus, increasing the visual word count has a positive effect on SIFT based classification.
- Among the others, for larger codebooks, XGBoost has provided the best results in terms of accuracy and TPR. However it can also be said that RF and XGBoost's classification performance are competitive.

Table 2. Results of the SIFT based classification

Descriptor	Learner	# of Visual Words	Train Acc.	Test Acc.	TPR	FPR	Fl
SIFT	SVM	50	0.611	0.773	0.773	0.0160	0.77
SIFT	XGB	50	0.725	0.818	0.818	0.0120	0.82
SIFT	RF	50	0.729	0.842	0.842	0.1120	0.84
SIFT	SVM	100	0.803	0.803	0.803	0.0140	0.80
SIFT	XGB	100	0.762	0.846	0.846	0.0100	0.85
SIFT	RF	100	0.749	0.860	0.860	0.0099	0.86
SIFT	SVM	200	0.747	0.837	0.837	0.0110	0.84
SIFT	XGB	200	0.799	0.858	0.858	0.0100	0.86
SIFT	RF	200	0.774	0.882	0.882	0.0084	0.88
SIFT	SVM	400	0.821	0.875	0.875	0.0080	0.88
SIFT	XGB	400	0.827	**0.893**	0.893	**0.0076**	**0.89**
SIFT	RF	400	0.801	0.887	0.887	0.0080	0.89

Table 3. Results of the DAISY based classification

Descriptor	Learner	# of Visual Words	Train Acc.	Test Acc.	TPR	FPR	Fl
DAISY	SVM	50	0.648	0.7465	0.746	0.018	0.74
DAISY	XGB	50	0.678	0.7849	0.784	0.015	0.78
DAISY	RF	50	0.699	0.8160	0.816	0.013	0.80
DAISY	SVM	100	0.709	0.7758	0.775	0.016	0.77
DAISY	XGB	100	0.709	0.7953	0.795	0.014	0.79
DAISY	RF	100	0.715	0.8226	0.822	0.012	0.81
DAISY	SVM	200	0.725	0.7901	0.790	0.014	0.79
DAISY	XGB	200	0.722	0.8174	0.817	0.013	0.81
DAISY	RF	200	0.719	0.8310	0.831	0.012	0.82
DAISY	SVM	400	0.725	0.8180	0.818	0.013	0.81
DAISY	XGB	400	0.725	0.8122	0.812	0.013	0.80
DAISY	RF	400	0.716	**0.8356**	0.835	**0.011**	**0.82**

The results belonging to analyses carried out with DAISY decriptor has been presented in Table 3. Regarding to the results reported in Table 3, following deductions have been made:

- Experiments with all codebook sizes show that Random Forest classifer has achieved the best results on test dataset in terms of accuracy, FPR and F1. However, SIFT based analysis has outperformed the classification models created with DAISY descriptors.
- Similar to SIFT, using larger codebook sizes provides better performance. Nonetheless, the improvement rate has not been found as high as SIFT based schemes. Instead, doubling the codebook size has increased the classification results less than 1%.

An honest comparison covering SIFT and DAISY based scores points out that, performance of SIFT based models surpass the DAISY based ones. We believe that, this finding is related to two distinct properties of SIFT namely (1) rotational invariance and (2) key point sampling tecnique. As stated before, SIFT detects keypoints by leveraging DoG technique yielding scale and rotational invariant key points whereas DAISY just samples the points located on a virtually created grid having cell sizes specified by a hyper-parameter. This situation leads the DAISY having potential for sampling redundant and irrelevant keypoints . As a result, SIFT can be thougt as a better and suitabe feature extractor for selecting more important and relevant keypoints in our problem domain. This was not a surprise for us due to DAISY's inherent aim which is focusing on dense stereo image matching for depth map estimation. Apart from this finding, another observation is that the phishing web pages do not show much rotational variance. Therefore, we believe that the performance difference between these two descriptors is highly related to difference of keypoint sampling techniques both of these exhibit.

We also observed that, both of the schemes we explored have achieved considerably good performance in classification of "legitimate" samples in test set. This finding shows that, the brands in phishing web page domain have their own visual characteristics to an extent. It is also possible to infer that, legitimate web page instances do not share much common visual properties with phishing ones. This observation can enable the problem to be treated as an "open-set" classification problem. The results listed in Table 2 and Table 3 show that the hypothesis of "open-set recognition of phishing web pages" is valid.

Regarding the SIFT based run-time speed, we have detected that the total duration of the whole inference including feature extraction, representation and classification took 1.5 seconds in average. In contrast, the best DAISY based inference has took 2.18 seconds in average since the RF algorithm runs on CPU instead of GPU in which the XGBoost runs on. This finding shows that SIFT based detection can be used in scenarios which requires real-time suggestions such as browser based detection.

7. Conclusion

In this study, we have experimented two different local key point based descriptors in order to classify phishing web pages following to extracting visual cues belonging to the web pages. Moreover, we have employed bag of visual words scheme in order to pool and create visual words sourcing from the training set of a publicly available "Phish-Iris" dataset. According to the results, 400 dimensional SIFT based BOVW representation has achieved the best recognition performance among the other configurations. After performing detailed experiments, we have shown that SIFT is a better local descriptor compared to DAISY regarding to phishing web page recognition. These findings highlighted the importance of key point sampling strategy. Moreover, we have also shown that the brand recognition of phishing web pages along with phish/non phish decision is possible through open-set recognition scheme. It is important to remember that, the results we achieved have been obtained by excluding the information of color. Thus, incorporating color based features may improve the recognition performance since color is an important feature for human visual perception. For this purpose, we are planning to include color as a source of information as a future work. Fusion of color and layout based features through autoencoders will be tested in our next work. Besides, as a future work, we also plan to employ deep convolutional neural networks for

improving the generalization abilities of the created models since they are able to combine information of layout and color in a single shot.

References

[1] Drake, C.E., Oliver, J.J. & Koontz, E.J., (2014) Anatomy of a phishing email, In CEAS 2014.
[2] Varshney, G., Misra, M. & Atrey, P.K., (2016) A survey and classification of web phishing detection schemes, *Security and Communication Networks*, *8*, 6266-6284.
[3] APWG, *Phishing Attack Trends Report*. Retrieved from http://docs.apwg.org/reports/apwg_trends_report_q4_2017.pdf), on (02.6.2019).
[4] Dalgic, F. C., Bozkir, A. S., Aydos, M. (2018). Phish-IRIS: A New Approach for Vision Based Brand Prediction of Phishing Web Pages via Compact Visual Descriptors. In 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (pp. 1-8). IEEE.
[5] Rao, R.S. & Pais, A.R., (2018) Detection of phishing web sites using an efficient feature-based machine learning framework, *Neural Computing and Applications*, 1-23.
[6] Lam, I.F., Xiao, W.C., Wang, S.C. and Chen, K.T., (2009) Counteracting Phishing Page Polymorphism: an Image Layout Analysis Approach, LNCS (pp. 270-279).
[7] Chen, K. T., Chen, J. Y., Huang, C. R., Chen, C. S. (2009). Fighting Phishing with Discriminative Keypoint Features. *IEEE Internet Computing*, *13*(3), 56-63.
[8] Rao, R. S. & Ali, S. T. (2015). A Computer Vision Technique To detect Phishing Attacks. In 2015 Fifth International Conference on Communication Systems and Network Technologies (pp. 596-601). IEEE.
[9] Bozkir, A.S. & Akcapinar Sezer, E. (2016). Use of HOG Descriptors in Phishing Detection, In 4th International Symposium on Digital Forensic and Security (ISDFS).
[10] Zhang, W., Lu, H., Xu, B., Yang, H. (2013). Web phishing detection based on page spatial layout similarity. *Informatica*, 37(3).
[11] Hara, M., Yamada., A., Miyake, Y. (2009). Visual similarity-based phishing detection without victim site information. In IEEE Symposium on Computational Intelligence in Cyber Security (pp. 30-36). IEEE.
[12] Corona I, Biggio, B., Contini, M., Piras, L., Corda, R., Mereu, M., Mureddu, G., Ariu, D., Roli, F., (2017). Delta-Phish: Detecting Phishing Webpages in Compromised Websites, In ESORICS 2017.
[13] Li, Y., Yang, Z., Chen, X., Yuan, H., Liu, W. (2019) A Stacking Model using URL and HTML Features for Phishing Webpage Detection, *Future Generation Computer Systems, 94*, 27-39
[14] Lowe,D.G. (2004). Distinctive image features from scale invariant keypoints, *International Journal of Computer Vision 60*
[15] Sachdeva, V. D. et al. (2017) Performance Evaluation of SIFT and Convolutional Neural Network for Image Retrieval, *International Journal of Advanced Computer Science and Applications, 8*
[16] Karami, E., Shehata, M., Smith, A. (2015) Image Identification Using SIFT Algorithm: Performance Analysis against Different Image Deformations, In Newfoundland Electrical and Computer Engineering Conference
[17] Keser, Reyhan. K., Ergun, E., Töreyin, B. U. (2017) Vehicle Logo Recognition with Reduced-Dimension SIFT Vectors Using Autoencoders, In International Workshop on Computational Intelligence for Multimedia Understanding
[18] Tola, E., Vincent L., Pascal F., (2010) DAISY: An Efficient Dense Descriptor Applied to Wide-Baseling Stereo, *IEEE Transactions on Pattern Analysis and Machine Learning, 32*
[19] Cortes, C. & Vapnik, V. (1995) "Support-vector networks", *Machine Learning*, *20*, 273-297
[20] Tianqi C. & Guestrin C., (2016) Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining