 CMC, vol.57, no.3, pp.533-547, 2018

Phishing Detection with Image Retrieval Based on Improved Texton Correlation Descriptor

Guoyuan Lin[1, 2, *], Bowen Liu[1], Pengcheng Xiao[3], Min Lei[4] and Wei Bi[5, 6]

Abstract: Anti-detection is becoming as an emerging challenge for anti-phishing. This paper solves the threats of anti-detection from the threshold setting condition. Enough webpages are considered to complicate threshold setting condition when the threshold is settled. According to the common visual behavior which is easily attracted by the salient region of webpages, image retrieval methods based on texton correlation descriptor (TCD) are improved to obtain enough webpages which have similarity in the salient region for the images of webpages. There are two steps for improving TCD which has advantage of recognizing the salient region of images: (1) This paper proposed Weighted Euclidean Distance based on neighborhood location (NLW-Euclidean distance) and double cross windows, and combine them to solve the problems in TCD; (2) Space structure is introduced to map the image set to Euclid space so that similarity relation among images can be used to complicate threshold setting conditions. Experimental results show that the proposed method can improve the effectiveness of anti-phishing and make the system more stable, and significantly reduce the possibilities of being hacked to be used as mining systems for blockchain.

Keywords: Anti-phishing, blockchain, texton correlation descriptor, weighted euclidean distance, image retrieval.

1 Introduction

Phishing is the top threat vector for cyberattacks and causes too much economic damage. We focus on detecting phishing attacks through image processing based techniques.

Currently, other approaches have been proposed to anti-phishing, including URLs based

[1] School of Computer Science and Technology, China University of Mining & Technology, Xuzhou, 221116, China.

[2] State Key Laboratory for Novel Software Technology hosted at Nanjing University, Nanjing, 210023, China.

[3] Department of Mathematics, University of Evansville, Indiana, 47722, USA.

[4] Information Security Center, Beijing University of Posts and Telecommunications, Beijing, 100876, China.

[5] SeeleTech Corporation, San Francisco, 94107, USA.

[6] Zabatech Corporation, Beijing, 100088, China.

[*] Corresponding Author: Lin Guoyuan. Email: lingy@cumt.edu.cn.

detection [Zhang, Pan, Wang et al. (2016); Daeef, Ahmad, Yacob et al. (2016); Tahir, Asghar, Zafar et al. (2016)] and webpage anomaly-based detection methods [Moghimi and Varjani (2016)]. But attackers often use images, JavaScript, and so forth, to bypass the anti-phishing system [Aleroud and Zhou (2017)]. Webpage anomaly-based detection methods and URLs based detection fail to detect this type of phishing webpages. Image processing based techniques can detect the embedded objects present in suspicious webpage because these techniques take the snapshot of the webpage and compare it with the corresponding legitimate webpage.

In most of detecting phishing methods, adjusting the appropriate threshold to detect a maximum number of phishing websites is a common way in phishing detection [Haruta, Asahina and Sasase (2017)]. The condition of threshold setting, only two websites that are the detecting website and the corresponding legitimate website which are compared when the threshold is determined, in most methods of detecting phishing is simply and the phishers can easily discover vulnerabilities to realize anti-detection. They can reduce the similarity between the detecting website and the corresponding legitimate website under the assurance of visually to mislead netizens. More websites are used to compare with detecting website to deal with this problem in this paper.

Under the circumstances of using image processing based techniques to detect phishing, image retrieval is considered to get more websites to be compared with detecting website. When a webpage is in the eyes of users, users are attracted to the salient region of the web-page and make a decision whether the webpage is trustworthy [Canfield and Fischhoff (2018)]. According to this, extracting the salient region when the webpage is switched into image can reduce the workload and make phishing detection be in human behavior. Wu et al. [Wu, Liu and Feng (2016)] proposed a method of image retrieval based on TCD. TCD has the advantage of extracting image salient region over other image retrieval methods. So TCD is considered to be applied into phishing detection.

Our proposed method firstly improves TCD to better depict the salient region of images and then makes use of image retrieval based on improved TCD to obtain enough webpages that have similarity in the terms of salient region which then is switched into images. Finally, similarity matrix is introduced to deal with the features produced from improved TCD of images. In the way, our proposed method solves the problem of thresholds mentioned above.

The contributions of this paper are as follows: (1) Our proposed method which is different from other methods obtains enough webpages which have similarity in the aspect of salient region to solve the threats of anti-detection from the view of complicating threshold setting conditions. (2) Two disadvantages of TCD are proposed and solved. Firstly, we propose new strategies to select neighborhoods and compute color difference to solve one disadvantage of TCD which has the lack of coping with the correlation between the number of features and their location. And then the statics on correlation is improved. (3) Secondly similarity matrix is introduced to solve the other one disadvantage of TCD which neglects the similarity relation among images.

The rest of this paper is organized as follows. Section 2 introduces the basic principle of the TCD. In Section 3, the improved TCD is described and applied to detect phishing websites. Experiments and analyzes are shown in Section 4. Conclusion and future

work are in Section 5.

2 Basic principle of TCD

There are three steps followed to generate TCD: 1) Detecting the low-level features (color value and local binary patterns) of pixels to generate color or texture uniform regions which contain discriminative information of images; 2) Color difference feature and texton frequency feature are used to respectively character contrast and spatial structure information in uniform regions; 3) TCD is generated by combining these two features.

The TCD consist of follow two vectors, (1) $\mathbf{H}_C$ which represents color feature generated in color uniform regions, (2) $\mathbf{H}_T$ which represents texture feature generated in texture uniform regions. Due to significance of $\mathbf{H}_C$ and $\mathbf{H}_T$ are different, they should be weighted when combined to represent TCD and is expressed in Eq. (1).

$$\mathbf{H} = (\alpha \cdot \mathbf{H}_C,\ \beta \cdot \mathbf{H}_T) \tag{1}$$

α and β respectively denote weight coefficient of $\mathbf{H}_C$ and $\mathbf{H}_T$.

3 TCSSD and applied to detect phishing

3.1 New strategies to select neighborhoods and compute color difference

In the process of analyzing uniform regions in the TCD, HSV color space is switched to Cartesian coordinates $H^cS^cV^c$ to make statistics on feature of color difference. In the $H^cS^cV^c$, the color difference d_i is computed with Euclidean Distance and expressed in Eq. (2):

$$d_i = \sqrt{\left(\Delta H_i^c\right)^2 + \left(\Delta S_i^c\right)^2 + \left(\Delta V_i^c\right)^2} \tag{2}$$

The 3X3 windows is applied to select neighborhoods g_i of center pixel g_m and thus the value domain of the number *i* is from 1 to 7.

$$DT_C\left(CH_k(g_m)\right) = \sum\sum\left(\sum_{i=0}^{8}\delta_c(g_m, g_i)\cdot d_i\right) \tag{3}$$

The $DT_C\left(CH_k(g_m)\right)$ expressed in Eq. (3) is the sum of color difference between the central pixel g_m and every neighborhood pixel g_i which has the same structure $CH_k(g_m)$ with g_m. $\delta_c(g_m, g_i)$ is expressed in Eq. (4):

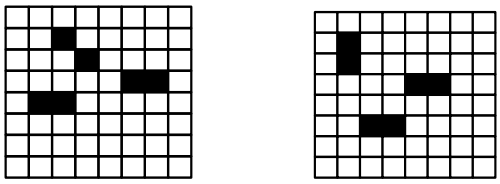
$$\delta_c(g_m, g_i) = \begin{cases} 1, & CH_k(g_m) = CH_k(g_i) \\ 0, & CH_k(g_m) \neq CH_k(g_i) \end{cases} \tag{4}$$

$$\overline{DT_C}\left(CH_k(g_m)\right) = \sum\sum\left(\sum_{i=0}^{8} d_i\right) \tag{5}$$

The $\overline{DT_C}\left(CH_k(g_m)\right)$ expressed in Eq. (5) is the sum of color difference between the central pixel g_m and every neighborhood pixel g_i. And then, the color difference feature S_c^k is expressed in Eq. (6):

$$S_c^k\left(CH_k(g_m)\right) = \frac{DT_C\left(CH_k(g_m)\right)}{sum\left(\overline{DT_C}\left(CH_k(g_m)\right)\right)} + \frac{DT_C\left(CH_k(g_m)\right)}{\overline{DT_C}\left(CH_k(g_m)\right)} \tag{6}$$

The method fails to deal with the correlation between the number of the feature and its location as is described in Fig. 1 bellow:



(a) The first image (b) The second image

Figure 1: Different distribution and same color difference

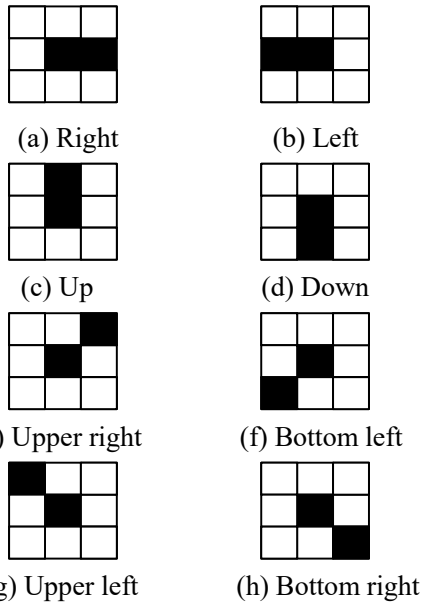
In Fig. 1, assuming that black square is the feature CH_k to be counted. The value of color difference between the same features is d_1 and different features is d_2. According to Eq. (6), $S_c^k(CH_k) = \frac{6d_1}{64} + \frac{6d_1}{6(d_1+7d_2)}$ is obtained in Fig. 1(a) and $S_c^k(CH_k) = \frac{6d_1}{64} + \frac{6d_1}{6(d_1+7d_2)}$ is obtained in Fig. 1(b). The value of two $S_c^k(CH_k)$ is same but they should be different because black squares in Fig. 1(a) and Fig. 1(b) have different distribution. In order to solve this problem, weighted color difference is proposed.

Assuming that color difference, d_i, between the g_m and g_i is the same. In this case, g_0 represents the neighborhood pixel right to the center pixel and the i increases with the clockwise under the 3×3 windows. The weighted color difference dw_i is expressed in Eq. (7).

$$dw_i = (i+1) \cdot d_i \tag{7}$$

In Eq. (7), The range of values for i is 0 to 7.

In the way above, there are a total of 8 kinds of neighborhood location concluded and showed in Fig. 2.



(a) Right (b) Left

(c) Up (d) Down

(e) Upper right (f) Bottom left

(g) Upper left (h) Bottom right

Figure 2: 8 kinds of neighborhood

And then, value of DT_C in Figs. 2(a)-2(h) is computed and showed in Tab. 1. From the Tab. 1, the value of DT_C can be classified into 4 classes and every class has the same value of DT_C. So counting up the value of DT_C can be regarded as counting up the four kinds of relative position.

Table 1: Value of DT_C in Fig. 2

Image number	DT_C	Image number	DT_C
Fig. 2(a)	$10d_i$	Fig. 2(e)	$8d_i$
Fig. 2(b)	$10d_i$	Fig. 2(f)	$8d_i$
Fig. 2(c)	$6d_i$	Fig. 2(g)	$12d_i$
Fig. 2(d)	$6d_i$	Fig. 2(h)	$12d_i$

Let a_i $(i = 1,2,\cdots,\mathrm{n})$ be the value of DT_C correspond to *n* kinds of relative location. A mathematical model following described is proposed to find the regular of a_i. $\forall k_1, k_2, \cdots, k_n,\ m_1, m_2, \cdots, m_n, \exists a_i (i = 1,2,\cdots,\mathrm{n})$, $\sum_{i=1}^{n}(k_i \times a_i) \neq \sum_{i=1}^{n}(m_i \times a_i)$ always stands up and then $\sum_{i=1}^{n} k_i = \sum_{i=1}^{n} m_i$, the value of k_i is different from the value of m_i. The k_i and the m_i is the number of the *n* kinds of relative location. In the terms of *n*=4, the value of a_i computed from the model above is a solution of the problem mentioned under the strategy of 3×3 windows. It is difficult to find a group a_i when the *n* equals 4, but easy to find a relationship between a_1 and a_2, $a_1 \neq a_2$, when the *n* equals 2.

The number of neighborhoods can be inferred as 4 from the situation of *n*=2 and thus double cross windows, cross-window and diagonal cross-window, are proposed. In the 3× 3 windows, 8 neighborhood pixels are selected. In double cross windows, neighborhood pixels, four neighborhood pixels above, below, to the left, and to the right of center pixel, are selected under the cross-window and the rest of the 8 neighborhood pixels are selected under the diagonal cross-window.

Two mappings, presented in Tab. 2, are proposed to make the following work convenient. The first mapping is between angle θ which represents the angle between the neighborhood pixels and center pixel, and the image number in Fig. 2. The second mapping is between the θ and the function $f(\theta)$.

Table 2: Two mappings

Image number	θ	$f(\theta)$	Image number	θ	$f(\theta)$
Fig. 2(a)	0^0	2	Fig. 2(e)	225^0	8
Fig. 2(b)	180^0	4	Fig. 2(f)	45^0	6
Fig. 2(c)	90^0	1	Fig. 2(g)	135^0	5
Fig. 2(d)	270^0	3	Fig. 2(h)	315^0	7

Based on the analysis above, NLW-Euclidean distance is described as follow.

Let $F(x,y)$ be the color image, and then its low-level features image is $CH_k(x,y)$ $(k = C,T)$ which contains the $CH_C(x,y)$, representing the quantized color

image, and the $CH_T(x,y)$, representing the LBP threshold image. On the basis of double cross windows, NLW-Euclidean distance is proposed to compute the color difference $d_{f(\theta)}$ which is between the center pixel, $g_m = (x_m, y_m)$, and the neighborhood pixel, $g_{f(\theta)} = (x_{f(\theta)}, y_{f(\theta)})$, in the Cartesian coordinates $H^cS^cV^c$. And the $g_{f(\theta)}$ has the distance as *D* and angle as θ from the center pixel. The value of *D* is 1 because of the 3X3 windows. Let $CH_k(g_i)$ be the structure of the pixel g_i. The $d_{f(\theta)}$ is expressed in Eq. (8).

$$d_{f(\theta)} = \begin{cases} \sqrt{\left(\Delta H^c_{f(\theta)}\right)^2 + \left(\Delta S^c_{f(\theta)}\right)^2 + \left(\Delta V^c_{f(\theta)}\right)^2} \times num\left(f(\theta)\right) & CH_k(g_m) = CH_k\left(g_{f(\theta)}\right) \\ \sqrt{\left(\Delta H^c_{f(\theta)}\right)^2 + \left(\Delta S^c_{f(\theta)}\right)^2 + \left(\Delta V^c_{f(\theta)}\right)^2} & CH_k(g_m) \neq CH_k\left(g_{f(\theta)}\right) \end{cases} \tag{8}$$

The $num\left(f(\theta)\right)$ in the Eq. (8) is expressed in Tab. 3. On the basis of the mathematical model mentioned above, the value of $num\left(f(\theta)\right)$ when the $f(\theta)$ equals 1 or 2 is different from the value when $f(\theta)$ equals 3 or 4. In the same way, the value of $num\left(f(\theta)\right)$ when $f(\theta)$ equals 5 or 6 is different from the value when $f(\theta)$ equals 7 or 8.

Table 3: The value of $num(f(\theta))$

$num\left(f(\theta)\right)$	Cross-window	Diagonal Cross-window
1	$f(\theta) = 1$	$f(\theta) = 5$
2	$f(\theta) = 2$	$f(\theta) = 6$
4	$f(\theta) = 3$	$f(\theta) = 7$
8	$f(\theta) = 4$	$f(\theta) = 8$

3.2 Improved statics on correlation

In the TCD, the color difference feature, describing the similarity degree of low-level feature between center pixel and neighborhood pixels, and the texton frequency feature, describing the probability of the structure similarity between center pixel and neighborhood pixels, are both generated in color uniform regions or texture uniform regions. Based on method of selecting neighborhoods and computing the color difference mentioned in 3.2, the correlation statistics analyzing uniform regions is improved.

Let $DT_{C,j}\left(CH_k(g_m)\right)$ $(j = 1,2)$ be the sum of color difference between the central pixel g_m and every neighborhood pixel $g_{f(\theta)}$ which has the same structure $CH_k(g_m)$ with g_m and represented in Eq. (9). In the Eq. (10), $\overline{DT_{C,j}}\left(CH_k(g_m)\right)$ is the sum of color difference between the central pixel g_m and every neighborhood pixel $g_{f(\theta)}$ In the Eq. (9) and Eq. (10), the statistics is made under the cross-window when the *j* equals 1 and under the diagonal cross-window when the *j* equals 2.

$$DT_{C,j}\left(CH_k(g_m)\right) = \begin{cases} \sum\sum\left(\sum_{f(\theta)=1}^{4} \delta_c\left(g_m, g_{f(\theta)}\right) \cdot d_{f(\theta)}\right), & j = 1 \\ \sum\sum\left(\sum_{f(\theta)=5}^{8} \delta_c\left(g_m, g_{f(\theta)}\right) \cdot d_{f(\theta)}\right), & j = 2 \end{cases} \tag{9}$$

$$\overline{DT_{C,J}}\left(CH_k(g_m)\right) = \begin{cases} \sum\sum\left(\sum_{f(\theta)=1}^{4} d_{f(\theta)}\right), & j = 1 \\ \sum\sum\left(\sum_{f(\theta)=5}^{8} d_{f(\theta)}\right), & j = 2 \end{cases} \tag{10}$$

And then, $\delta_c\left(g_m, g_{f(\theta)}\right)$ is expressed in Eq. (11).

$$\delta_c\left(g_m, g_{f(\theta)}\right) = \begin{cases} 1, & CH_k(g_m) = CH_k\left(g_{f(\theta)}\right) \\ 0, & CH_k(g_m) \neq CH_k\left(g_{f(\theta)}\right) \end{cases} \tag{11}$$

The description about the color difference feature, the ratios of the $DT_{C,j}\left(CH_k(g_m)\right)$ to $\overline{DT_{C,J}}\left(CH_k(g_m)\right)$, is represented in Eq. (12).

$$CD_j^k\left(CH_k(g_m)\right) = \frac{DT_{C,j}\left(CH_k(g_m)\right)}{\overline{DT_{C,J}}\left(CH_k(g_m)\right)} \tag{12}$$

The $CD_j^k\left(CH_k(g_m)\right)$ only describes the correlation of color difference between pixels, and the color difference histograms, normalized histogram of $\overline{DT_{C,J}}\left(CH_k(g_m)\right)$, is proposed to describe the global probability of color difference to be as supplementary expressed in Eq. (13).

$$CF_j^k\left(CH_k(g_m)\right) = \frac{\overline{DT_{C,J}}\left(CH_k(g_m)\right)}{sum\left(\overline{DT_{C,J}}\left(CH_k(g_m)\right)\right)} \tag{13}$$

A method proposed by Feng et al. [Feng, Wu, Liu et al. (2015)] is adopt to fuse the advantage of $CD_j^k\left(CH_k(g_m)\right)$ and $CF_j^k\left(CH_k(g_m)\right)$, and the fusion is expressed in the Eq. (14).

$$S_{c,j}^k\left(CH_k(g_m)\right) = CD_j^k\left(CH_k(g_m)\right) \times \left(CF_j^k\left(CH_k(g_m)\right) + 1\right) \tag{14}$$

Not only the color difference spatial distribution of the pixels which is adjacent but also the global probability of the color difference is considered in the $S_{c,j}^k\left(CH_k(g_m)\right)$. According to the Eq. (14), the feature description of Fig. 1(a) is that: $S_{c,1}^k(CH_k) = \frac{20d_i}{64} + \frac{20d_i}{32d_i}$, $S_{c,2}^k(CH_k) = \frac{5d_i}{64} + \frac{5d_i}{11d_i}$, and the feature description of Fig. 1(b) is that: $S_{c,1}^k(CH_k) = \frac{25d_i}{64} + \frac{25d_i}{43d_i}$, $S_{c,2}^k(CH_k) = 0$. The $S_{c,1}^k(CH_k)$ of Fig. 1(a) is different from the Fig. 1(b) and the $S_{c,2}^k(CH_k)$ of Fig. 1(a) is different from the Fig. 1(b). This shows that the two images in Fig.1 is distinguished.

The improved texton frequency feature is expressed in the next description. Let the $\overline{DF_{c,J}}$ be as the number of pixels whose structure is $CH_k(g_m)$ in the uniform regions, and the corresponding normalized histogram is defined in the Eq. (15).

$$TF_j^k\left(CH_k(g_c)\right) = \frac{\overline{DF_{c,J}}\left(CH_k(g_m)\right)}{sum\left(\overline{DF_{c,J}}\left(CH_k(g_m)\right)\right)} \tag{15}$$

In the Eq. (15), $sum\left(\overline{DF_{c,J}}\right)$ is used to express the number of all pixels in the uniform regions and $TF_j^k\left(CH_k(g_m)\right)$ is used to express the global distribution character. The relation between the local center pixel and its neighborhood pixel is not considered in the Eq. (15). The strategy of making statistics on texton frequency correlation proposed by Liu et al. [Liu, Li, Zhang et al. (2011)] is introduced to cope with the disadvantage of $TF_j^k\left(CH_k(g_m)\right)$ and is expressed in the Eq. (16) and Eq. (17).

$$DF_{c,j}\left(CH_k(g_m)\right) = \begin{cases} \sum\sum\left(\sum_{f(\theta)=1}^{4}\delta_c\left(g_m, g_{f(\theta)}\right)\right), & j=1 \\ \sum\sum\left(\sum_{f(\theta)=5}^{8}\delta_c\left(g_m, g_{f(\theta)}\right)\right), & j=2 \end{cases} \tag{16}$$

$$TH_j^k(CH_k(g_m)) = \frac{DF_{c,j}(CH_k(g_m))}{4\overline{DF_{c,j}}(CH_k(g_m))} \tag{17}$$

In the Eq. (17), the $DF_{c,j}$ is the probability of neighborhood pixel $g_{f(\theta)}$ which has the same structure $CH_k(g_m)$ with the center pixel g_m. The $4\overline{DF_{c,J}}$ is the frequency of all neighborhood pixels under the cross-window or diagonal cross-window which have the same structure with the center pixel. The frequency feature is expressed in the Eq. (18) via the fusion method proposed by Feng et al. [Feng, Wu, Liu et al. (2015)].

$$L_{f,j}^k\left(CH_k(g_m)\right) = TH_j^k\left(CH_k(g_m)\right) \times \left(TF_j^k\left(CH_k(g_m)\right) + 1\right) \tag{18}$$

The advantages of the histogram and the correlation statics are fused in the Eq. (18). 8 kinds of feature description are obtained from the analysis above and is showed in the Tab. 4.

Table 4: 8 kinds of feature descriptions

Low-level image	Color difference	Texton frequency
$CH_C(x,y)$	$S_{c,1}^C \quad S_{c,2}^C$	$L_{f,1}^C \quad L_{f,2}^C$
$CH_T(x,y)$	$S_{c,1}^T \quad S_{c,2}^T$	$L_{f,1}^T \quad L_{f,2}^T$

3.3 Phishing with the TCSSD

In terms of detecting phishing websites, two kinds of image libraries are established. The first image library is legal website image library which is consist of the screenshots of the homepage of legal websites and the second image library is the phishing websites image library which is consist of the screenshots of the phishing websites detected before.

The website homepage will be identified in detecting phishing websites. It is found that phishers often lure victims in the visual similarity of website homepages. Under the guarantee on accuracy, the workload will be reduced when only homepages are detected. At the time of recognizing sites, identifying screenshot image will be compared with the images from the legal website image library and phishing websites image library.

Table 5: Symbol set

Image number	$S_{c,j}^C$	$S_{c,j}^T$	$L_{f,j}^C$	$L_{f,j}^T$
m_1	$S_{c,j}^C(m_1)$	$S_{c,j}^T(m_1)$	$L_{f,j}^C(m_1)$	$L_{f,j}^T(m_1)$
m_2	$S_{c,j}^C(m_2)$	$S_{c,j}^T(m_2)$	$L_{f,j}^C(m_2)$	$L_{f,j}^T(m_2)$
⋮	⋮	⋮	⋮	⋮
m_n	$S_{c,j}^C(m_n)$	$S_{c,j}^T(m_n)$	$L_{f,j}^C(m_n)$	$L_{f,j}^T(m_n)$

On the basis of this strategy above, according to the method proposed by Qian et al.

[Qian, Li, Liang et al. (2016)] and the method proposed by Guo et al. [Guo, Yang and Liang (2016)], the identifying screenshot images are numbered. Symbol set is established and expressed in the Tab. 5.

The set G =$\{S_{c,1}^C,\ S_{c,2}^C,\ S_{c,1}^T,\ S_{c,2}^T,\ L_{f,1}^C,\ L_{f,2}^C,\ L_{f,1}^T,\ L_{f,2}^T\}$ are the 8 kinds of image features whose weights are the set $\mathrm{W} = \{w_1 = \beta,\ w_2 = \beta_1,\ w_3 = \gamma,\ w_4 = \gamma_1,\ w_5 = \varepsilon,\ w_6 = \varepsilon_1,\ w_7 = \rho,\ w_8 = \rho_1\}$. The set $\mathrm{M} = \{m_1,\ m_2,\ \cdots m_n\}$ are the image numbers.

The 8 kinds of features, $S_{c,1}^C,\ S_{c,2}^C,\ S_{c,1}^T,\ S_{c,2}^T,\ L_{f,1}^C,\ L_{f,2}^C,\ L_{f,1}^T$, and $L_{f,2}^T$, are the feature descriptions of image m_j (j=1,2,…n). The following strategy is proposed to determine the value of the weights, $\beta,\ \beta_1,\ \gamma,\ \gamma_1,\ \varepsilon,\ \varepsilon_1,\ \rho$ and ρ_1.

1) The features obtained in the color uniform regions or the texture uniform regions have different significance. Assuming the μ is the weight of features obtained in the color uniform regions and the ϑ is the weight of features obtained in the texture uniform regions. It is concluded that $\mu + \vartheta = 1$;

2) The color difference features and texton frequency features obtained respectively from the color uniform regions or texture uniform regions under the double cross windows;

3) It is concluded that the weights, $\beta,\ \beta_1,\ \varepsilon$ and ε_1, equal μ and the other weights, $\gamma,\ \gamma_1,\ \rho$ and ρ_1, equal ϑ.

Similarity among images is computed between the image symbol object in pairs and expressed in the Eq. (19).

$$T_{p,q} = \left(\sum_{s=1}^{8} w_s \tau_s(m_p, m_q)\right) \Big/ \sum_{s=1}^{8} w_s \tag{19}$$

In the Eq. (19), $\tau_s(m_p, m_q)$ is the expressed in the Eq. (20).

$$\tau_s(m_p, m_q) = \begin{cases} 1, & S_{c,j}^k(m_p) = S_{c,j}^k(m_q) \text{ or } L_{f,j}^k(m_p) = L_{f,j}^k(m_q) \\ 0, & S_{c,j}^k(m_p) \neq S_{c,j}^k(m_q) \text{ or } L_{f,j}^k(m_p) \neq L_{f,j}^k(m_q) \end{cases} \tag{20}$$

And then, the following similarity matrix **T** of $T_{p,q}$ is obtained;

$$\mathbf{T} = \begin{bmatrix} T_{1,1} & T_{1,2} & \cdots & T_{1,j} & \cdots & T_{1,n} \\ T_{2,1} & T_{2,2} & \cdots & T_{2,j} & \cdots & T_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ T_{i,1} & T_{i,2} & \cdots & T_{i,j} & \cdots & T_{i,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ T_{n,1} & T_{n,2} & \cdots & T_{n,j} & \cdots & T_{n,n} \end{bmatrix}$$

Every column of similar matrix **T** is a new dimension of the identifying website images, and every row of the similar matrix **T**, $\mathbf{T_i} = (T_{i,1}, T_{i,2} \cdots T_{i,n})\ (i \in (1,n))$, is the improved TCD, texton correlation based on space structure descriptor (TCSSD), of every image in the set M. The first row of similar matrix **T**, $\mathbf{T_1} = (T_{1,1}, T_{1,2} \cdots T_{1,n})$, is the TCSSD of identifying website image. And Eq. (21) is proposed to compute the similarity between the identifying website image and the other images in the set M.

$$\cos(\mathbf{T_1}, \mathbf{T_s}) = \frac{\sum_{i=1}^{n} T_{1,i} T_{s,i}}{\sqrt{\sum_{i=1}^{n} {T_{1,i}}^2} \sqrt{\sum_{i=1}^{n} {T_{s,i}}^2}} \quad (s \in (2, n)) \tag{21}$$

In the Eq. (21), the value of $\cos(\mathbf{T_1}, \mathbf{T_s})$ is used to quantize the probability the identifying website is phishing website and larger the value is, the higher the probability is. For the precision of phishing detection, a threshold ω of the $\cos(\mathbf{T_1}, \mathbf{T_s})$ should be set.

4 Experiments

4.1 Determine experiment parameters

The weights, μ and ϑ, and the threshold ω can be determined independently because the precision of the retrieval based on TCSSD has nothing to do with the threshold ω and then μ and ϑ can be determined firstly.

The Corel-1000 image set which is consist of 1000 images which are classified into 10 kinds and every kind has 100 images is used to be as the experimental image set. Every image from the Corel-1000 image set will be as sample to retrieve kindred images and the precision of retrieving is expressed in the Eq. (22).

$$RPre = \frac{\sum_{q=1}^{n} \frac{N_{cq}}{N_{rq}} \times 100\%}{n} \tag{22}$$

In the Eq. (22), the times of retrieving is *n*, the N_{cq} is the number of the kindred images in the *qth* time and the N_{rq} is the number of images retrieved in the time *qth* time. In order to improve the speed of experiment, we use the cloud computing environment mentioned by Lin et al. [Lin, Bie, Lei et al. (2014); Lin, Wang, Bie et al. (2014)].

The final *RPre* is the average of all the *RPre* under the same combination of μ and ϑ. In this way, lots of *RPre* is computed under the different combination of μ and ϑ. The result is showed in the Tab. 6.

Table 6: *RPre* and corresponding combination of μ and ϑ

μ	ϑ	*RPre* (%)	μ	ϑ	*RPre* (%)
0	1	60.13	0.5	0.5	80.16
0.05	0.95	78.36	0.55	0.45	79.94
0.1	0.9	79.65	0.6	0.4	79.53
0.15	0.85	80.29	0.65	0.35	79.13
0.2	0.8	80.46	0.7	0.3	78.76
0.25	0.75	80.86	0.75	0.25	78.24
0.3	0.7	80.59	0.8	0.2	77.92
0.35	0.65	80.43	0.85	0.15	77.48
0.4	0.6	80.38	0.9	0.1	77.06
0.45	0.55	80.26	1	0	76.95

From the Tab. 6, the highest *RPre* is 80.86% and the corresponding combination of μ and ϑ is 0.25 and 0.75. It is concluded that the effect of retrieval is the best when the

μ equals 0.25 and thus 0.25 and 0.75 is selected as the weights μ and ϑ.

The threshold ω can be determined under the μ and ϑ selected above and then there are two steps to be adopt: the first step is that the screenshots of legal website homepages and phishing website homepages are obtained; the second step is that precision of phishing detection, *PRE*, expressed in the Eq. (23) is computed when the threshold is adjusted.

$$PRE = \frac{N_p}{N_w} \times 100\% \tag{23}$$

In the Eq. (23), the N_p is the number of the websites which is detected as phishing website correctly and the N_w is the number of all websites detected. The result of the phishing detection under the different ω is showed in the Fig. 3.

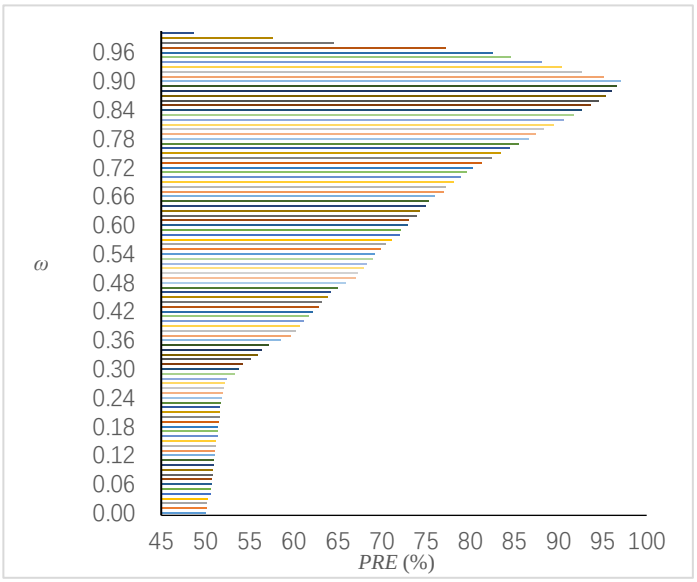


Figure 3: The *PRE* under different ω

When the ω is 0.90, the *PRE* is the highest. The *PRE* would be reduced due to two reasons. On the one hand, it is concluded that the legal websites could be detected as phishing websites when the ω is less than 0.90. On the other hand, it is concluded that the phishing websites could be detected as legal websites when the ω is greater than 0.90. To detect more phishing websites correctly, the ω is selected as 0.90.

4.2 Comparison

The test set is composed of two kinds of websites, one is the 3000 legitimate websites and the other one is 5000 phishing websites, from June 5 to June 20, 2017. The phishing websites which are obtained from PhishTank archive are verified and still online during this period. The legitimate websites are collected from Alexa. We refined Alexa to return only the top 500 sites on the web by category [Chiew, Chang, Sze et al. (2015)]. Categories include but are not limited to arts, business, health, recreation,

shopping, and sports.

To verify the effectiveness of the method of detecting phishing websites based on TCSSD, the rate of correct discrimination, *TF* expressed in the Eq. (24), the miscarriage rate, *EF* expressed in the Eq. (25), and the missing rate, *LF* expressed in the Eq. (26) are proposed to quantify the effectiveness.

$$TF = \frac{F_{p-p}}{s} \times 100\% \tag{24}$$

$$EF = \frac{F_{l-p}}{s} \times 100\% \tag{25}$$

$$LF = \frac{F_{p-l}}{s} \times 100\% \tag{26}$$

The s is the times of phishing detection, F_{p-p} is the times that the phishing websites are detected correctly, F_{l-p} is the times that the legal websites are detected as the phishing websites and F_{p-l} is the times that the phishing websites are detected as legal websites.

The result of comparison on effectiveness between the TCSSD and other methods, a method to detect phishing web pages based on Hungarian matching algorithm [Zhang, Zhou, Xu et al. (2010)], a method of combining local and global features of webpages [Zhou, Zhang, Xiao et al. (2014)] and a method of using HOG descriptors to detect phishing websites [Bozkir and Sezer (2016)] is showed in Tab.7. In order to verify that TCSSD is more suitable than TCD to be applied into phishing detection, TCD is applied to detect phishing and compared with TCSSD.

Table 7: Comparison on effectiveness between the TCSSD and other method

Methods	*TF* (%)	*EF* (%)	*LF* (%)
TCSSD	96.67	1.35	1.98
TCD	95.34	2.47	3.19
Hungarian matching	96.47	1.49	2.04
Local and Global	95.63	2.03	2.34
HOG	95.46	1.79	2.75

In the Tab. 7, the *TF* of TCSSD is 96.67% and is higher than the *TF* of any method including TCD, 95.34%, 96.47%, 95.63% and 95.96%. In the Tab. 7 the *EF* and the *LF* of TCSSD are the lowest value, 1.35% and 1.98%. The reason is that the TCSSD not only has a better precision in retrieving images but also the similarity among images in image set is considered in TCSSD. It is concluded that the TCSSD has more satisfying effectiveness applied to detect phishing and is more suitable than TCD.

SEF is proposed to express the detection speed of those methods and expressed in the Eq. (27).

$$SEF = \frac{Nother_t}{NTCSSD_t} \tag{27}$$

In the Eq. (27), the $Nother_t$ is the number of phishing websites which is detected correctly with methods except the TCSSD within t hours and the $NTCSSD_t$ is the

number of phishing websites which is detected correctly with TCSSD within *t* h.

In experiment, the detection speed is computed in various duration and the result is showed in Tab. 8.

In the Tab. 8, different duration is selected to observe the relationship between the detection speed and the running time of those methods. If the *SEF* of a method is below 1, it will be concluded that TCSSD has a better detection speed, or vice versa. As we can see, the detection speed of the method in Tab. 8 is lower than the TCSSD under the situation that the corresponding *SEF* is below 1. The *SEF* of any method in Tab. 8 is below 1 in any duration and thus the detection speed of TCSSD can be regarded as the fastest. With the increase of duration, the *SEF* of any method decreases and thus TCSSD also has a satisfactory stability.

Table 8: The *SEF* of those methods of phishing

Methods	3 h	9 h	12 h
TCD	0.986	0.983	0.978
Hungarian matching	0.981	0.981	0.980
Local and Global	0.992	0.991	0.989
HOG	0.996	0.995	0.994

Based on the experiments, the TCSSD proposed in this paper has more satisfying stability and effectiveness in phishing detection.

5 Conclusion and expectation

In this paper, a new method of image retrieval is proposed and applied to detect phishing. The proposed approach realizes the combination of image retrieval and phishing detection to solve the problem of anti-detection. Experimental results show that the accuracy of phishing detection and efficiency are improved compared to the other methods used in the experiment. The theoretical analysis and experiments verified that TCSSD has satisfactory stability and high effectiveness in phishing detection.

The proposed method still has some disadvantages. Further studies should be focused on the following aspects: 1) Storage strategy and index method should be studied to cope with the increasing amount of images to be detected; 2) In the regional characteristics of statistical consistency, the algorithm of image rotation invariance needs to be further improved; 3) Efficient methods to deal with the websites also needs to be studied to reduce the workload.

Acknowledgment: The work reported in this paper was supported by the Joint research project of Jiangsu Province under Grant No. BY2016026-04, the Opening Project of State Key Laboratory for Novel Software Technology of Nanjing University under Grant No. KFKT2018B27, the National Natural Science Foundation for Young Scientists of China under Grant No. 61303263, and the Jiangsu Provincial Research Foundation for Basic Research (Natural Science Foundation) under Grant No.

BK20150201.

Reference

Aleroud, A.; Zhou, L. N. (2017): Phishing environments, techniques, and countermeasures: A survey. *Computers & Security*, vol. 68, no. 7, pp. 160-196.

Bozkir, A. S.; Sezer, E. A. (2016): Use of HOG descriptors in phishing detection. *4th International Symposium on Digital Forensic and Security*, pp. 148-153.

Canfield, C. I.; Fischhoff, B. (2018): Setting priorities in behavioral interventions: An application to reducing phishing risk. *Risk Analysis*, vol. 38, no. 4, pp. 826-838.

Chiew, K. L.; Chang, E. H.; Sze, S. N.; Tiong, W. K. (2015): Utilisation of website logo for phishing detection. *Computers & Security*, vol. 54, no. 10, pp. 16-26.

Daeef, A. Y.; Ahmad, R. B.; Yacob, Y.; Phing, N. Y. (2016): Wide scope and fast websites phishing detection using URLs lexical features. *3rd International Conference on Electronic Design*, pp. 410-415.

Feng, L.; Wu, J.; Liu, S.; Zhang, H. (2015): Global correlation descriptor: A novel image representation for image retrieval. *Journal of Visual Communication and Image Representation*, vol. 33, no. 11, pp. 104-114.

Guo, Q.; Yang, H. J.; Liang, X. Y. (2016): Image retrieval method based on new space relationship feature. *Journal of Computer Applications*, vol. 36, no. 7, pp. 1918-1922.

Haruta, H.; Asahina, H.; Sasase, I. (2017): Visual similarity-based phishing detection scheme using image and CSS with target website finder. *IEEE Global Communications Conference*, pp. 1-6.

Liu, G. H.; Li, Z. Y.; Zhang, L.; Xu, Y. (2011): Image retrieval based on micro-structure descriptor. *Pattern Recognition*, vol. 44, no. 9, pp. 2123-2133.

Lin, G. Y.; Bie, Y. Y.; Lei, M.; Zheng, K. F. (2014): ACO-BTM: A behavior trust model in cloud computing environment. *International Journal of Computationl Intelligence Systems*, vol. 7, no. 4, pp. 785-795.

Lin, G. Y.; Wang, D. R.; Bie, Y. Y.; Lei, M. (2014): MTBAC: A mutual trust based access control model in cloud computing. *China Communications*, vol. 11, no. 4, pp. 154-162.

Moghimi, M.; Varjani, A. Y. (2016): New rule-based phishing detection method. *Expert Systems with Applications*, vol. 53, no. 7, pp. 231-242.

Qian, Y. H.; Li, F. J.; Liang, J. Y.; Liu, B.; Dang, C. Y. (2016): Space structure and clustering of categorical data. *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 10, pp. 2047-2059.

Tahir, M. A. U. H.; Asghar, S.; Zafar, A.; Gillani, S. (2016): A hybrid model to detect phishingsites using supervised learning algorithms. *International Conference on Computational Science and Computational Intelligence*, pp. 1126-1133.

Wu, J.; Liu, S. L.; Feng, L. (2016): Image retrieval based on texton correlation descriptor. *Journal of Computer Research and Development.* vol. 52, no. 12, pp. 2824-2835.

Zhang, J. Y.; Pan, Y.; Wang, Z. Q.; Liu, B. (2016): URL based gateway side phishing detection method. *2016 IEEE Trustcom/BigDataSE/ISPA*, pp. 268-275.

Zhang, W. F; Zhou, Y. M.; Xu, Lei; Xu, B. W. (2010): A method of detecting phishing web pages based on hungarian matching algorithm. *Chinese Journal of Computers*, vol. 33, no. 10, pp. 1963-1975.

Zhou, Y.; Zhang, Y. Z.; Xiao, J.; Wang, Y. P.; Lin, W. Y. (2014): Visual similarity based anti-phishing with the combination of local and global features. *IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications*, pp. 189-196.