December 28, 2018

Handcrafting visual features of emails or landing pages to detect phishing

Kuntal Sengupta

Vijay Eranti

Handcrafting visual features of emails or landing pages to detect phishing

ABSTRACT

In a phishing attack, a perpetrator attempts to obtain the online credentials of a user by impersonating a trusted entity such as a bank, email service provider, etc. Sophisticated phishers attempt to deceive spam filters by structuring the visual look-and-feel of their fake emails to be nearly but not precisely identical to emails sent by a trusted entity, such that spam filters allow the fake email to reach a user's inbox.

This disclosure describes use of hand-crafted visual features of emails or landing pages, and classification based on earth-mover's distance, to assess the visual similarity of genuine and phished emails. The techniques detect visual near-duplicates of a trusted entity's email and thereby achieve resilience against phishing attacks.

KEYWORDS

- phishing
- spoofing
- fake email
- email security
- visual distance
- earth-mover distance

BACKGROUND

In a phishing attack, a perpetrator attempts to obtain the online credentials of a user by impersonating a trusted entity such as a bank, email service provider, etc. Phishers send emails to users (e.g., bank account-holders, email account holders, etc.) that look nearly identical to those sent by the trusted entity. Email service providers deploy spam/phishing filters to thwart such

attacks, e.g., by identifying such emails and classifying them as spam/suspicious. Currently, such filters use features computed from the text, embedded links, sender domain, etc. of an email to decide if the email is a phishing attempt. For example, an email that is visually indistinguishable from one sent by a trusted entity, but with embedded links that point to online entities other than the trusted entity, is an indicator of a phishing attempt.

DESCRIPTION

The visual look-and-feel of an email, including but not limited to a brand logo included in the email, can provide useful cues as to whether the message is a phishing email or not. Per techniques of this disclosure, visual signals that arise from the email are used to classify emails. The techniques are able to perform the visual classification in a short time, e.g., < 10 milliseconds, and are scalable even for email service providers that serve billions of emails daily.

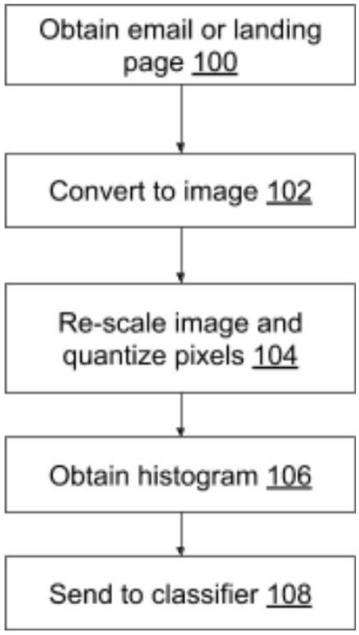

Fig. 1: Obtaining visual features of an email and using such features for classification

Fig. 1 illustrates obtaining visual features of an email per techniques of this disclosure, and using such features to classify the email. An email or a landing page is obtained (100) and converted to an image (102). The image is re-scaled (104) to a predetermined canonical size and form factor. The canonical size and form factor is based on performance, e.g., as measured via a receiver operating curve (ROC), time-to-match, etc. Pixels of the image are quantized, e.g., to three bits. A histogram is created (106) out of the re-scaled and quantized-pixel image, the details of which are explained below. This histogram serves as a visual feature that is sent to a classifier (108).

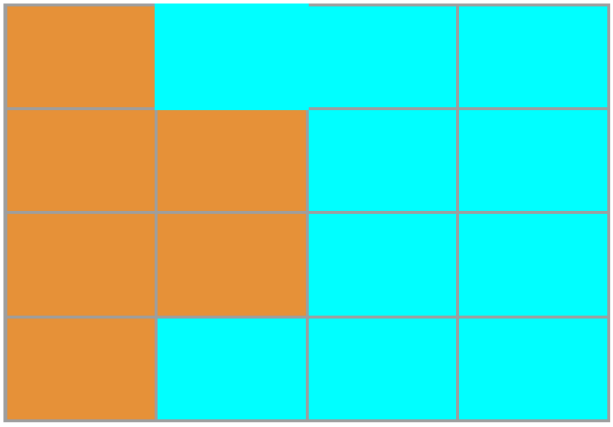

Fig. 2: An example 4x4 image illustrating histogram computation

The histogram computation is illustrated with reference to the example 4x4 image shown in Fig. 2. The color of a pixel is mapped to an integer, `color_value`. The color is an `(r, g, b)` vector, each vector component being three bits wide (due to the aforementioned quantization). Mapping of the color to `color_value` is carried out via the formula

`color_value = 64 × quantized_b + 8 × quantired_g + quantized_r.`

Example: A pixel with `(r, g, b)` given by (7,7,0) maps to a `color_value` 504.

The `(x,y)` coordinates of pixels of a given `color_value` are grouped together to form a component. For example, in Fig. 2, the yellowish pixels are grouped together to form the

component {(0,0), (0,1), (0,2), (0,3), (1,2), (1,3)} where the first number refers to the row and the second number refers to the column. For each component, the centroid is computed by averaging the x- and y-coordinates of the pixels that belong to the component. A bin is created for the `k`th component corresponding to color value `k`, with features such as centroid coordinate values `(x_k, y_k)` and the quantized color coordinates `(b_k, g_k, r_k)`. The entire image is mapped to `M` such bins, with the `k`th bin being described by the feature `[(x_k, y_k) (b_k, g_k, r_k)]`. The population of the bin, denoted `n_k`, is the number of pixels in the image that have the `color_value` of the bin. The population is normalized to the weight value `w_k`, such that sum of the weights over all the bins is unity.

The histogram corresponding to an image is fed to a classifier. The classifier analyzes the histogram to classify images based on a distance measure such as the earth-mover's distance (EMD). For example, two images that have histograms are close to each other in the sense of EMD are deemed to belong to the same class. Conversely, two images that have histograms that are far away from each other in the sense of EMD are deemed to belong to different classes.

The EMD between two histograms is described as the amount of earth (samples) that needs to be moved from one histogram to the other in order to make the two histograms identical. Imagining each histogram as an earthen mound, the match process aims to make the first histogram identical to the second by optimally moving earth from one location to another. The number of samples to be moved and the distance of the move for each sample in feature space are factored into computing the cost of the move. The cost drives the schedule of how many samples from a bin in the first histogram are moved to another bin of the second histogram. Note that the EMD computation seeks to minimize the global, not local, cost of move.

Linear optimization formulations that compute the EMD between two histograms with approximate O(N) complexity are utilized, where N is the number of bins.

The Rubner technique for computing EMD is an established technique to compare intensity or color histograms of two images. This disclosure generalizes the Rubner technique such that histogram features, e.g., component centroids, quantized color value vectors, etc., are used during EMD computation. A pointer is provided to a comparison function between two bins, where the bins belong to the two histograms being compared. For example, if the two histograms have bins respectively described as:

Bin1_k = {(x_k, y_k) (b_k, g_k, r_k)}, and

Bin2_j = {(x_j, y_j) (b_j, g_j, r_j)},

the distance function computes a L2 distance between the centroid location and the color vector respectively, normalizes each of these by the maximum possible value for centroid distance and color distance, and computes a weighted average. Thus, the L2 distance d_{xy} between (x_k, y_k) and (x_j, y_j) is computed. The L2 distance d_{rgb} between (b_k, g_k, r_k) and (b_j, g_j, r_j) is computed. The overall distance is computed as

$$\alpha d_{xy} + \beta d_{rgb},$$

where α and β are weighting parameters. An example selection for α and β is $\alpha=\beta=0.5$.

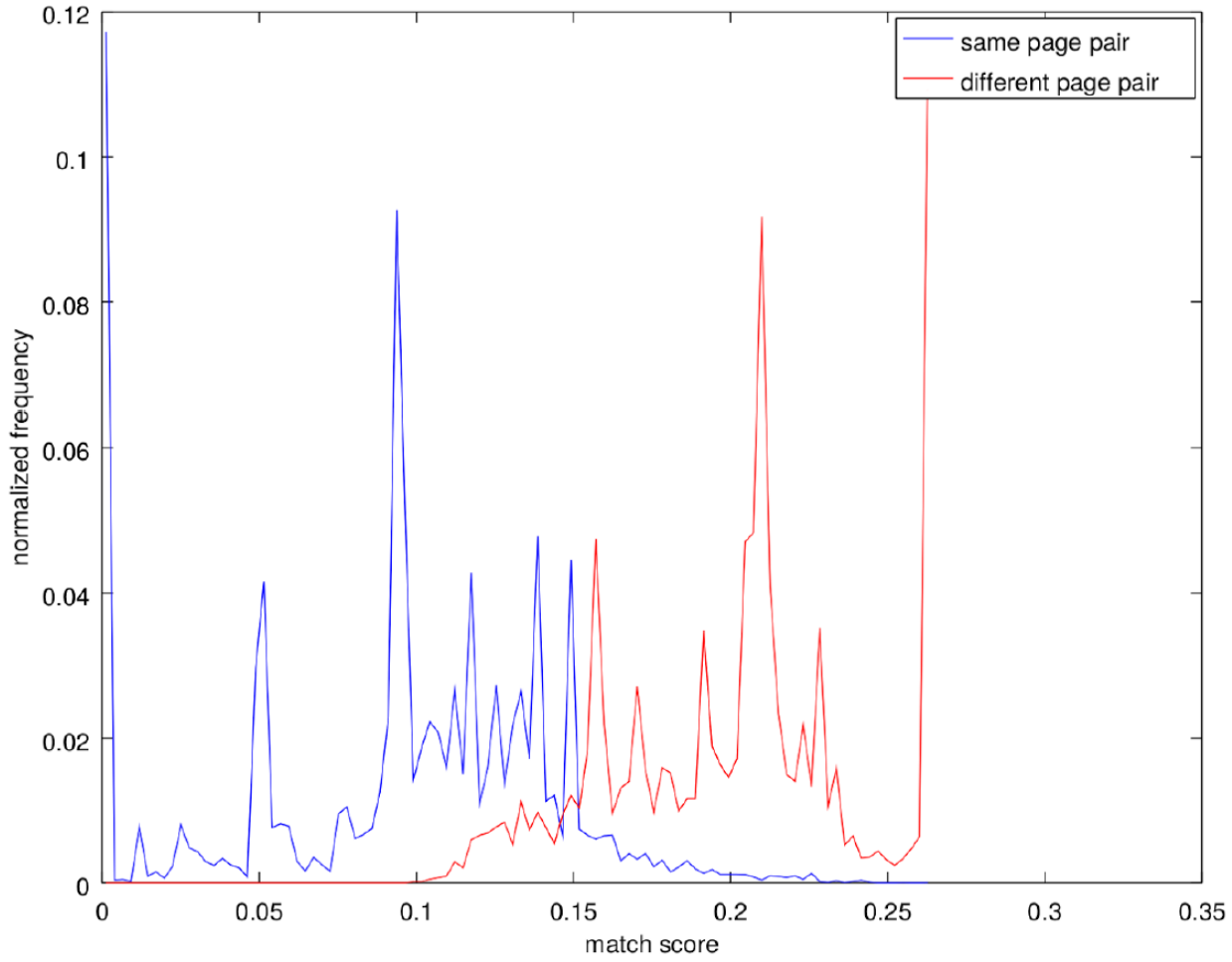

Fig. 3: Histogram of EMD scores for same-brand landing pages (blue) versus different-brand landing pages (red)

Fig. 3 illustrates an example of the performance of the EMD-based classifier described above when applied to same-brand versus different-brand landing pages. The blue curve is a histogram of the match score, e.g., EMD, between landing pages (both genuine and phished) of the same brand. The red curve is a histogram of the EMDs between landing pages (both genuine and phished) between different brands. As expected, the blue curve lies generally to the left, e.g., shows a greater frequency of smaller EMDs, than the red curve. As seen from Fig. 3, if the match-score threshold is set at 0.1, at least 50% of the phished pages can be flagged without incurring any extra false accepts. The problem of clustering landing pages is similar to the

problem of clustering email renderings, although somewhat more challenging, as background images in landing pages are occasionally different even for the same brand.

Alternative to using EMD as a distance metric for the purposes of classification, one may use other distance metrics, e.g., sum of squared differences, sum of absolute differences, etc. Other histogram-based metrics, e.g., scale-invariant feature transform (SIFT), histogram of oriented gradients (HOG), can be used to classify email or landing-page renderings. Other techniques such as near-duplicate detection, deep-learning neural networks, etc. can also be applied for classifying email (or landing page) renderings. The criteria for choice of classifier are false-accept vs. false-reject rates, time-to-match, etc.

CONCLUSION

This disclosure describes use of hand-crafted visual features of emails or landing pages, and classification based on earth-mover's distance, to assess the visual similarity of genuine and phished emails. The techniques detect visual near-duplicates of a trusted entity's email and thereby achieve resilience against phishing attacks.