**Embedded Images Source/Abstract**

<http://lbmedia.ece.ucsb.edu/resources/ref/thesis.pdf>

* Embedded-text in images usually carry important messages about the content. In the past, several algorithms have been proposed to detect text boxes in video frames. Previous work often followed a multi-step framework using a combination of image-analysis and machine-learning techniques. In this work, we propose a unified embedded-text detection framework to efficiently and accurately locate text boxes particularly in web and email images. We approach the embeddedtext problem from the angle of object detection. We define position-independent features to capture the essence of characters and a smart-scan algorithm to trace text lines using their spatial and geometrical properties. We also propose a novel anti-spam system which utilizes visual clues, including the embedded-text information. The experimental results demonstrate the effectiveness of the proposed embedded-text detection framework and the anti-spam filtering system.

<https://content.iospress.com/articles/journal-of-computer-security/jcs371>

* Phishing emails usually contain a message from a credible looking source requesting a user to click a link to a website where she/he is asked to enter a password or other confidential information. Most phishing emails aim at withdrawing money from financial institutions or getting access to private information. Phishing has increased enormously over the last years and is a serious threat to global security and economy. There are a number of possible countermeasures to phishing. These range from communication-oriented approaches like authentication protocols over blacklisting to content-based filtering approaches. We argue that the first two approaches are currently not broadly implemented or exhibit deficits. Therefore content-based phishing filters are necessary and widely used to increase communication security. A number of features are extracted capturing the content and structural properties of the email. Subsequently a statistical classifier is trained using these features on a training set of emails labeled as ham (legitimate), spam or phishing. This classifier may then be applied to an email stream to estimate the classes of new incoming emails. In this paper we describe a number of novel features that are particularly well-suited to identify phishing emails. These include statistical models for the low-dimensional descriptions of email topics, sequential analysis of email text and external links, the detection of embedded logos as well as indicators for hidden salting. Hidden salting is the intentional addition or distortion of content not perceivable by the reader. For empirical evaluation we have obtained a large realistic corpus of emails prelabeled as spam, phishing, and ham (legitimate). In experiments our methods outperform other published approaches for classifying phishing emails. We discuss the implications of these results for the practical application of this approach in the workflow of an email provider. Finally we describe a strategy how the filters may be updated and adapted to new types of phishing.

<https://www.researchgate.net/publication/221600234_Efficient_Modeling_of_Spam_Images>

* Image spam has become a real threat to email communication these days, since most prevalent content based spam filters can not efficiently detect them out, even when the latest OCR techniques are employed, spammers could compromise the system easily through text distortion and other obscuring skills. In this paper, we propose a novel and efficient image modeling approach for spam image classification, this content based statistical model does not rely on the availability of text information embedded in the image files, so that it is robust to obfuscations. Experimental results show that the proposed method can perform with good accuracy in practice.

**Metadata/Textual Analysis**

* <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.93.3666&rep=rep1&type=pdf>
  + In this paper, we present the first results of our empirical analysis for the actual use that is made of metadata in Learning Object Repositories, more specifically in the ARIADNE Knowledge Pool System (KPS). We analyze metadata information provided by indexers when they introduce new learning objects into the KPS. This gives a clear indication about their understanding of metadata elements, values, etc. Finally, we study the correlation between metadata elements filled-in by those indexers. These results provide us with empirical guidelines to asses the development of application profiles and metadata toolsets.
* <https://www.deepdyve.com/lp/wiley/a-quantitative-categorical-analysis-of-metadata-elements-in-image-puF04RvbQv>
  + This article reports on a quantitative categorical analysis of metadata elements in the Dublin Core, VRA Core, REACH, and EAD metadata schemas, all of which can be used for organizing and describing images. The study found that each of the examined metadata schemas contains elements that support the *discovery* , *use* , *authentication* , and *administration* of images, and that the number and proportion of elements supporting functions in these classes varies per schema. The study introduces a new schema comparison methodology and explores the development of a class‐oriented functional metadata schema for controlling images across multiple domains.