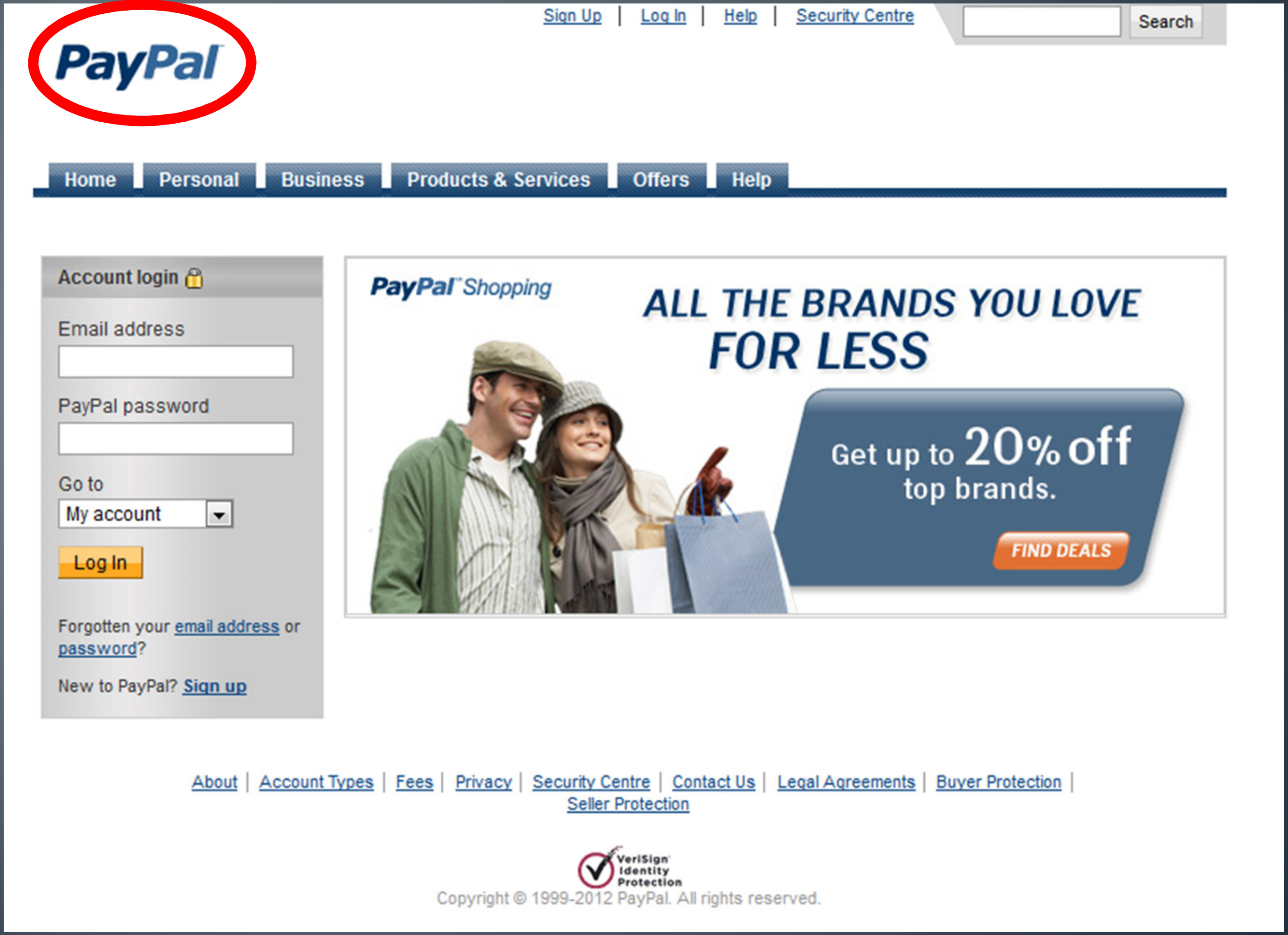
Image Matching For Branding Phishing Kit Images

Chengucui Zhang, Rajan Kumar Kharel, Song Gao, and Jason Britt

University of Alabama at Birmingham

Department of Computer and Information Science

Phish, Phish Kits, and Drop Mails


Sign Up | Log In | Help | Security Centre
Search
PayPal
Home
Personal
Business
Products & Services
Offers
Help
PayPal Shopping
ALL THE BRANDS YOU LOVE FOR LESS
Get up to 20% off top brands.
FIND DEALS
VeriSign Identity Protection

Account login

Email address

PayPal password

Go to

My account

Log In

Forgotten your email address or password?

New to PayPal? Sign up

About | Account Types | Fees | Privacy | Security Centre | Contact Us | Legal Agreements | Buyer Protection | Seller Protection

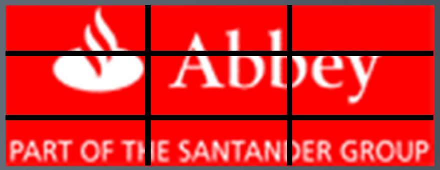
Copyright © 1999-2012 PayPal. All rights reserved.

Color Histogram

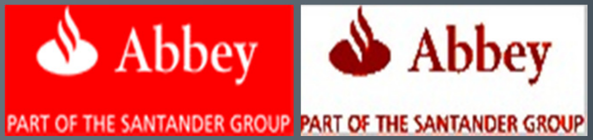
- Standard image analysis technique
- Represents the color distribution in an image.
- Algorithm:
  - Extract two most significant bits of the 8 bit representation from each R, G, and B channel.
  - Forms 64 bins for each image
  - Bin percentage is percentage of the image with a particular color code.
  - Calculate histogram dissimilarity using histogram intersection

Algorithms Evaluated

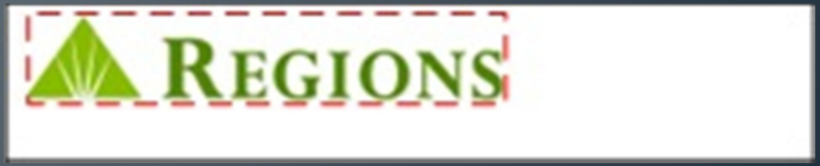
- GCH: Global Color Histogram
- LCH: Local Color Histogram



- LCH+: Local Color Histogram with dimensional constraint and background removal



- LCH++: Dimensional constraint, minimum bounding box (MBB), and background removal



Data Set

- Data source, UAB Kit Data Mine
- Collection of 56,926 kits
- 10,130 unique images with 215 brand images and 9,915 general images
- Manually viewed images to determine ground truth
- 42 different brands covered
- Training Set: 109 training brand images
- Testing Set: 106 testing brand images and 9,915 general images.

Image Matching

- Determined optimal minimum distance threshold
- 106 brand and 9,915 general images used for testing
- Found most similar training image for each testing image
- Calculated true positive (TP), true negative (TN), false positive (FP), false negative (FN) for the four algorithms

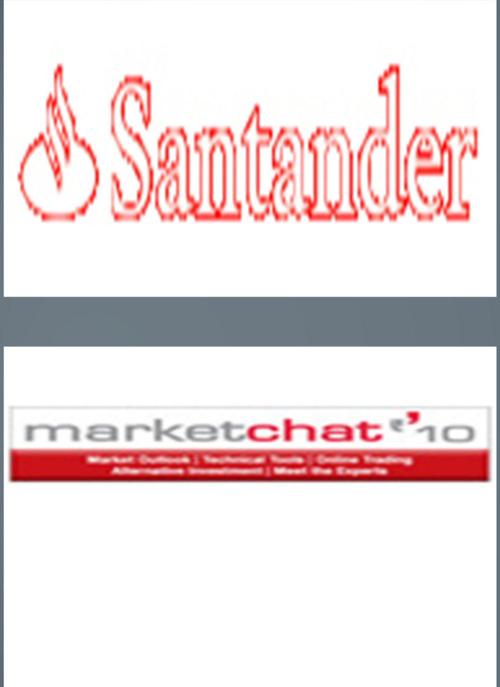
Results

Algorithm	GCH	LCH	LCH+	LCH++
TP	67	62	71	65
TN	8,945	9,046	9,822	9,864
FP	971	870	94	52
FN	38	43	34	40
Accuracy (%)	89.93	90.88	98.72	99.08

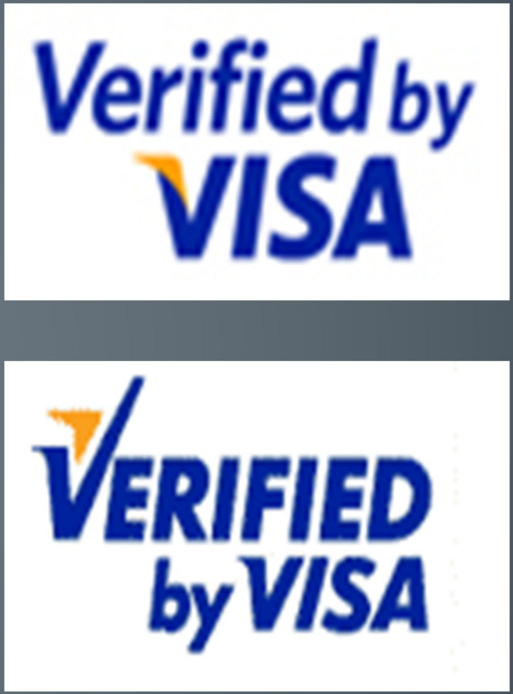
Evaluation of result(Total # of test images: 10,021)

Accuracy = (TP+TN)/(Total # of testing images)

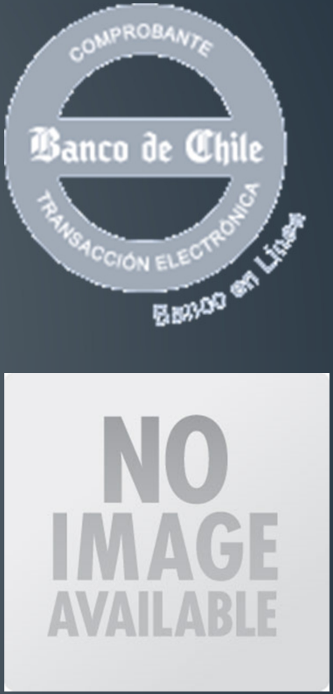
Discussion



GCH False Positive



LCH False Negative



GCH, LCH, LCH+, and LCH++ False Positive

Conclusion

- LCH is generally better than GCH.
- Dimensional constraints, background removal, and minimum bounding box (foreground extraction) improve accuracy.
- Sufficiently accurate for kit retrieval

Future Work

- Explore other visual features
- Use multi-dimensional index to decrease run time
- Develop phish kit branding strategies using image brands
- Brand phish using screen shots
- Spam campaign identification using images

Questions