https://doi.org/10.2352/ISSN.2470-1173.2018.10.IMAWM-445

Multimedia analytics platform for profiling keywords embedded in photo catalogues

Federica Mangiatordi, Andrea Bernardini, Emiliano Pallotti, Licia Capodiferro; Fondazione Ugo Bordoni; Roma, Italy

Abstract

In the recent years, the global penetration of Internet and the rapid spread of mobile devices have led to an exponential rise of trade in counterfeit and pirated goods with a negative impact on the profits of affected firms and consequently damage for employment and economic growth. This peculiar online trade has taken place mostly on deep web, but today it has started to shift to common IM platform and image based social networks. Regard to this context, this work presents a specific multimedia analytics platform, that monitors image catalogues promoting potential counterfeit products on social networks in order to extract useful information (as email, WeChat or WhatsApp, external links to specific online marketplaces) and profile the potential fakers. The preliminary results, derived by considering the image catalogues shared by various sellers on image based social networks, show the effectiveness of the proposed multimedia analytics methodologies.

Introduction

A multi-billion dollar underground economy, based on trade in counterfeit and pirated products, has emerged over the last decades in almost every country in the world. A wide range of products are affected in all market segments, from high-end consumer luxury goods (watches, perfumes or leather goods), to business-to-business products (machines, chemicals or spare parts), to common consumer products (toys, pharmaceuticals, cosmetics and foodstuffs). Counterfeiting and unauthorized use of trademarks represent a major threat to the modern knowledge-based economies having negative impact both on the sustainable intellectual property-based business models and the long term economic growth [1, 2].

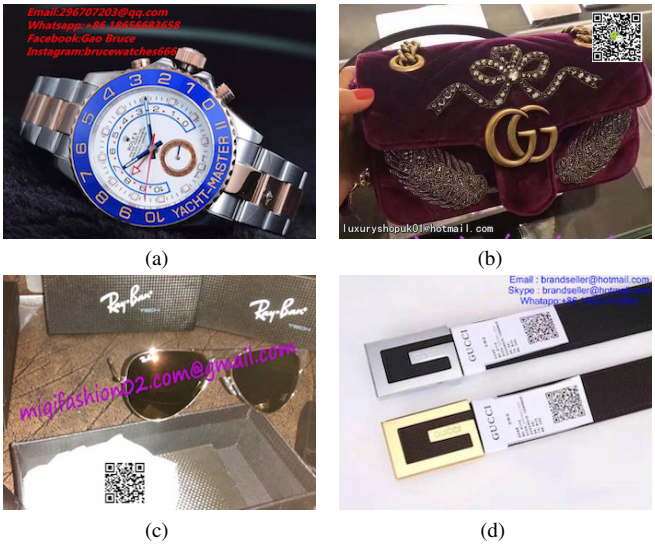
Trade in Counterfeit items not only cause enormous economic loss but can also damage the reputation of a brand. Buyers of fake products often receive low-quality goods and may be exposed to serious safety and health risks [3].

The recent analysis conducted in the European Union jointly by the OECD and the EUIPO (EU Intellectual Property Office) has reported that trade in counterfeit and pirated goods (know as fakes) amounted to up to 2.5 % of world trade in 2013 and to up to 5 % of imports in the EU market [4]. More specifically, the June 2015 study produced by the EUs Observatory on infringements of Intellectual Property (IP) rights has revealed 9.7% of sales lost by the clothing, footwear and accessories sectors with an annual revenue lost of 26.3 billion of euros due to trade of fakes [5].

The global penetration of the Internet and mobile devices has further accelerated these phenomenons pushing counterfeiting and piracy on social media platforms [6]. Social networks have emerged as easy channels for advertising and distributing products that infringe the intellectual property (IP) rights of legitimate holders, formal companies and governments [7]. According to the simple rule of "reaching the largest number of users and leaving as few fingerprints as possible", some unscrupulous sellers have set up a variety of e-commerce activities exploiting popular photo sharing websites and image-based social media to reach a huge audience. The instant messaging and social platforms are used to upload photos of fake products, for which the sellers can be contacted by potential buyers through links embedded in the images itself.

To contrast online counterfeiting and protect the health and safety of consumers, it is mandatory for companies and governments to deploy appropriate technologies to watch over the Internet for IP infringing products [8, 9].
This requires the adoption of specific multimedia analysis techniques to quickly detect and identify peculiar illicit e-commerce activities on image-based social platforms[10]-[13].

This work presents an analytics framework based on HAAR cascading classifiers for monitoring the catalogues posted on photo sharing site [14, 15]. The goal is to detect and analyze the profile of potential counterfeiters that advertise their goods posting images with embedded external links and contacts such as QR code or text. The paper is organized as it follows. The first section describes the proposed web analytics framework. The second section presents the design choices in the implementation of the main modules. The third section discusses the experimental results. Finally, concluding remarks are given in section four.



(a) (b)

(c) (d)

Figure 1: Photos of luxury items posted online with sellers external contact info in overlapping text box and QR-code

Image analytics framework

This paper presents an integrated framework for scraping the web in order to find potential sellers of counterfeit and pirated goods from the study of their post on the image based social platforms.

Crawling the photo catalogues posted online with certain hashtags (i.e. #handbag, #luxurywatches) and analyzing the overlapping box text and QR-code, make possible to extract the external contact info (ie.email, WeChat, Skype) of the sellers and eventual links to their webshop pages. Subsequently, the derived contact information is merged with post descriptions and features of the social accounts (such as nickname, post frequency, number of post, ratio following/followers, number of likes) to study the seller profile.
Fig.1 shows some posted photos extracted from online photo catalogues and HTML documents by our prototype system.

From the highest level of perspective our image analytics architecture is divided into three logical parts, as shown in Fig.2:

- *Scraping Module* is responsible to manage and run the python scripts for crawling the web. It starts from the specific inputs defined by the user (such as particular words in social account nicknames, some hashtags on posted photos or specific URL) and can be adapt to different aspects of image based social media;
- *Image Processing Module* controls the process for the enhancement and segmentation of the images downloaded by the scraper. At this end it processes the single image color components to increase the probability of localizing the overlapping text boxes and QR code patterns. The localization process uses cascade classifiers and provides the positive regions to OCR and QR code reader;
- *Data Visualization and Analysis Module* receives the raw textual data extracted from the photo catalogues and from the Document Object Model (DOM) of the web pages associated with the seller. It transforms these data and summarize them in XML schema. Then XML documents can be used to classify them in visual forms that allow users to fully understand and memorize data insights.

Implementation

The design of the presented data analytics system is focused on the structural analysis of:

- the images extracted from the photo catalogues shared online;
- the textual contents in the web pages linked to the URLs, found in the posted images or in the account description of the social platform user.

The main design choices are briefly described in the subsequent subsections.

Scraping Module

The developed crawler is a automated tool which controls and runs various python scripts in order to browse the web in a methodical manner [17, 18] and in accordance with the inputs specified by the user. Thus it is in charge of the preprocessing activity for the page selection, the parsing of the textual contents, the querying of a certain number of images from a determined URL or photo catalogue with specific hashtags.
The scraper program makes usage of python libraries such as Beatiful Soup [19].

The crawler can be configured to collect specific information such as the prices of particular goods from some web-shops or to query the account description on a specific image based social platforms.

The gathered textual information is transformed in semi-structured data to be saved in the NoSQL databases in Fig.2. For each analyzed accounts of a social platforms, it is stored the user ID, the number of posts, the number of followers, the links to external websites, the nicknames of the follower. Moreover the crawler creates XML documents containing the link to each posted photo, its hashtags, the comments of the followers, its date, the number of likes, its dimensions in pixels and in bytes.

Specifically, the scraping manager follows two loops for exploring all internal and external links (such as those specified in the image contents or HTML documents) and for gathering data from all kind of sources. The goal is to exploit all multimedia data (textual and visual) to understand and analyze the profile of potential seller of counterfeit products.

Image Processing Module

This module manages the visual structural and colour information of the downloaded images to identify embedded textual data in form of QR code or overlapping text box. However the recognition of textual contents overlapped on images can be a challenging task due to the variable background [20, 21].

To maximize the probability of success and reproduce the human detection procedure in text and QR code extraction,it is decided to process each image as follows:

- extraction of grey and colour components;
- adjustment of local contrast of each component through an adaptive filtering;
- computation of low level structural feature;
- analysis of limited portion of the image that will contain text and QRcode pattern by using cascade classifier [22, 23];
- analysis and extraction of textual contents from the selected image region applying specific OCR and QRcode open libraries [24, 25].

The adaptive histogram equalization of colour components improves the local contrast and positively affect the performance of text detector. To make the QR detection more robust to illumination change and noise, it is necessary to process only the luminance of the image. This grey scale component is enhanced applying an adaptive intensity threshold operator. At this end the Bersen method [26] can be considered with a sliding window of NxN pixels over the entire image. In this paper the value N=15 is adopted. The threshold value is given by the equation below:

$$TH_{bersen} = (w_{low} + w_{high})/2 \quad (1)$$

where w_{low} and w_{high} are the minimum and maximum grey level values in the sliding window.

Since the QR code patterns are constituted of white and black squares that define a particular texture it is decided to represent the structural information of the image by its HAAR features.

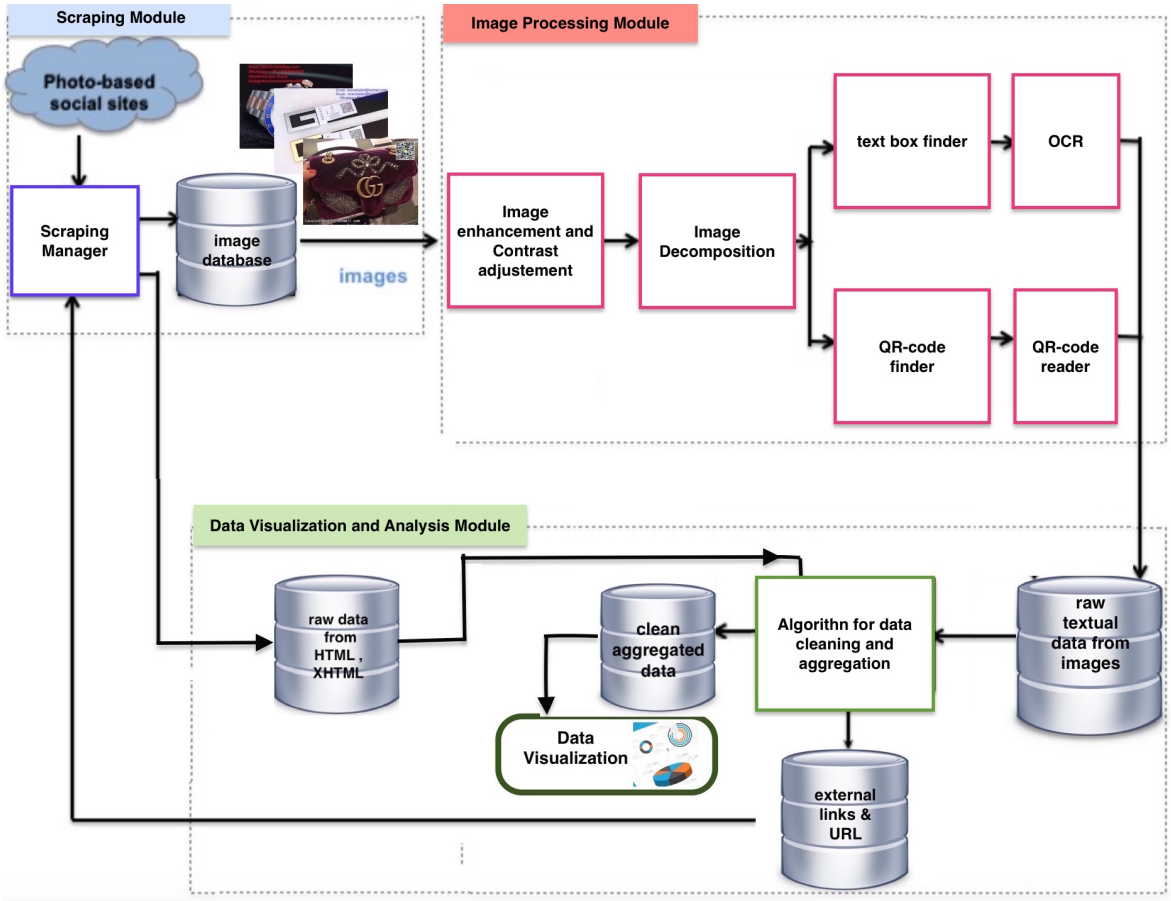


Figure 2: Image analytics framework

These features characterizes equally well the high edge density of text areas.

To localize the full QR code area and identify the minimum text box it is follows a sliding window approach.
Thus each image is scanned from the top left to the bottom, and from right to the left with rectangular sliding windows of different scales.
For each sliding window the corresponding HAAR features are provided to a specific cascade classifier, opportunely trained, in order to decide if a text box or QR pattern is present.

The QR code classifier will label as 'positive samples' the windows that include a full QR-code. Similarly, the text classifier will label as 'positive samples' the windows including text characters. In the case of text box finding, we can have more overlapped areas that are recognized as positive by the cascade classifier. To solve this ambiguity the text finder merges together overlapped text positive areas by the computation of the relative bounding external rectangle.

The Data Analysis and Visualization Module

The module is responsible for the cleaning activity of the raw text from images by parsing and aggregating the recognized text characters. The aim is to recognize the particular contact info (email, WeChat, WhatsApp, etc) used by the seller. This information is combined with the description of the social accounts and provided to traditional text mining algorithm. The main goal is to define a profile of the seller in a XML schema.
Specifically, the text mining python scripts analyzes the textual data in order:

- to find URL to other web pages or other IM accounts;
- to discover useful knowledge about the seller activity, such as profile keywords of potential counterfeiting sellers, number of followers gained, number of posts uploaded, variety of the merchandise associated to the same social accounts.

The crunched data are then summarized in specific semi-structured documents and stored in a specific dB.
This allows to measure the similarity of seller profiles, to aggregate their common features, to compute some significant descriptive statistics.

Experiments

To demonstrate the effectiveness of the proposed multimedia analytics platform we have configured the scripting manager, developed in Python programming language, to download 62250 images from catalogues hosted by Yupoo and Instagram (image sharing sites). To run the crawling activities we have searched for:

- hashtags associated with luxury items, such as #luxurywatches, #handbags, #gucci, #chanel, #shoeswomen, #womenfashion, #fashionglassess, #prada, #rayban, #vuitton, #loubutinshoes;

- words in post descriptions such as 'glam', 'elegance', 'style', 'fashion', 'jewel', 'swissmade', 'sports shoes', 'sneakers', 'slim bag', 'womens wear', 'boutique', 'dream dress', 'outfit', 'men style', 'mens wear', 'leather atelier, 'design', 'luxury life style', 'luxury bag', 'luxury brand', 'quality products'.

The images posted by social user, have different size and resolution, so we have set the script to store in the database only the images with minimum dimension greater than 480 pixels.
To find the full QR code pattern, an HAAR cascade classifier is been trained using OpenCV library with a training time of about 26 hours on an iMac, with 3.4 GHz Intel Core i7 with 16 GB of ram. The default weak classifier parameter values for true positive rate (TPR) and false positive rate (FPR) are respectively 0.995 and 0.5. This implies that the 99.5 % of the positive samples are correctly classified at each stage. The total number of the stage is set at 10. The QR cascade classifier is trained on a db containing 5000 synthetic images of 256 X256 pixels and 2500 cropped real images from those downloaded by scraping module. The positive examples (full QR code pattern) are the 20% of the training dataset. The gained precision is 0.9127 and the F-measure 0.8458.

Similarly, it is trained an HAAR cascade classifier for text box detection. The selected number of stages is 30 and the relative parameters have default values (TPR=0.995, FPR=0.5).

The training set is constituted with 5000 rectangular images of 64 x 16 pixels. These images are obtained segmenting the real images provided by the scraper and choosing randomly the negative examples (absence of text in the image area). The percentage of negative images is the 80% of the training set. The obtained precision is 0.7751 with an F-measure=0.8641.

Finally, we consider the experimental results obtained in the validation of overall image analytics system. We have compared the results between the configurations with and without the QR and text box finders. The test on the 62250 extracted images confirm an improvement of about 28% in text recognition when the OCR is fed with the enhanced image areas identified by text box classifier. Besides, the improvement of the performance of QR code decoder is about 9.2% when it is fed with the scaled and contrast equalized regions identified by QR-code finder.
A closer look at the raw data extracted from the proposed framework shows that:

- it is localized a text or qr-code pattern associated with a contact information (email, phone number, Skype, WeChat, WhatsApp) for the 76.4% of the images;
- the 62,8% of those images has at least two contacts;
- the percentage for the recognized contact info is: 32,37% for WeChat, 56,78% for WhatsApp, 14,8% for email, 46,71% for a phone number, 26,15% for Skype;
- it is localized a QR-code containing contact info or URLs for the 32,46% of images.

The analysis of the web pages linked by the images reveals that:

- the great part of URLs are Chinese or Russian web stores (89.34%);
- these web shops provides a wide range of luxury items associated with a variety of brands. The advertised products are frequently characterized by quite cheap prices.

Conclusions

This work investigates a new image analytics platform to take action against counterfeiting in online sales over image based social networks. The proposed framework takes social intelligence beyond text analytics by exploiting image analysis techniques to extract textual information from images posted by sellers in their ramified photo-catalogues. The goal is to understand the profile of sellers by collecting heterogeneous information from web-page linked by QR-code and/or text advertising in uploaded images. The task of image analytics is especially challenging when the sellers don't provide clean descriptions of their products posting low quality images.

References

[1] OECD/EUIPO,"Trade in Counterfeit and Pirated Goods: Mapping the Economic Impact", OECD Publishing, Paris, pg. 138, (2016)

[2] EPO,OHIM, "Intellectual Property Rights intensive industries: contribution to economic performance and employment in the European Union", pg. 144, (2013), https://oami.europa.eu/ohimportal/en/web/observatory/ip-cotribution

[3] Justin A. Heinonen, Thomas J. Holt, and Jeremy M. Wilson 2012. "Product counterfeits in the online environment: An empirical assessment of victimization and reporting characteristics". International Criminal Justice Review,Vol 22, Issue 4, pp. 353 - 371, (November 2012)

[4] EUROPOL,OHIM, "2015 Situation Report on Counterfeiting in the European Union. A joint project between Europol and the Office for Harmonization in the Internal Market", pg. 70, (April 2015), https://www.europol.europa.eu/publications-documents/2015-situation-report-counterfeiting-in-european-union

[5] OHIM,"The Ecomomic Cost of IPR Infringement in the Clothing, Footwear and Accessories Sector", pg.36, (2015) https://oami.europa.eu/tunnel-web/secure/webdav/guest/document_library/observatory/resources/research-and-studies/ip_infringement/study2/the_economic_cost_of_IPR_infringement_in_the_clothing_footwear_and_accessories_sector_en.pdf.

[6] Fung, I., Blankenship, E., Goff, M., Mullican, L., Chan, K., Saroha, N., Duke C.,Eremeeva M.,Fu K., Tse Z. (2017). Zika-Virus-Related Photo Sharing on Pinterest and Instagram. Disaster Medicine and Public Health Preparedness. Vol. 11, pg. 656-659. (2017)

[7] Cao, Qiang, Michael Sirivianos, Xiaowei Yang, and Tiago Pregueiro. "Aiding the detection of fake accounts in large scale social online services." InProceedings of the 9th USENIX conference on Networked

[8] Rahm, Erhard. "Discovering product counterfeits in online shops: a big data integration challenge." ACM Journal Data and Information Quality, August 2014.

[9] MarkMonitor. "Seven Best Practices for Fighting Counterfeit Sales Online." markmonitor.com. Edited by White Paper. (2012). https://www.markmonitor.com/download/wp/wp-Fighting_Counterfeit_Sales.pdf.

[10] Anitha Kannan, Inmar E. Givoni, Rakesh Agrawal, and Ariel Fuxman, 2011, "Matching unstructured product offers to structured product specifications." In Proceedings of the 17th ACM SIGKDD Conference.

[11] Thomsen, Jakob G., Erik Ernst, Claus Brabrand and Michael I. Schwartzbach. "WebSelF: A Web Scraping Framework." ICWE (2012).

[12] Charu C. Aggarwal , ChengXiang Zhai , Mining Text Data , Springer US (2012)

[13] Li, Zhixu, Xiangliang Zhang, Hai Huang, Qing Xie, Jia Zhu and Xiaofang Zhou. Addressing Instance Ambiguity in Web Harvesting. Proceedings. 18th International Workshop on the Web and Databases, WebDB 2015, pg. 6-12 (2015).
[14] Sander Soo, "Object detection using Haar-cascade Classifier", Institute of Computer Science, University of Tartu (2008).
[15] Xiaobin Zhuang, Wenxiong Kang, Qiuxia Wu,Real-time vehicle detection with foreground based cascade classifier, IET Image Processing, 2015).
[16] Guennouni S., Ahaitouf A., Mansouri A., "A Comparative Study of Multiple Object Detection Using Haar-Like Feature Selection and Local Binary Patterns in Several Platforms" Modelling and Simulation in Engineering, vol. 2015, pg. 8, (2015).
[17] Schrenk, M. Webbots, spiders, and screen scrapers: a guide to developing Internet agents with PHP/CURL. No Starch Press, 2007.
[18] Vargiu, Eloisa and Mirko Urru. "Exploiting web scraping in a collaborative filtering-based approach to web advertising.", Artif. Intell. Research 2 (2013): 44-54.
[19] Beautiful Soup. Python
https://www.crummy.com/software/BeautifulSoup
[20] Khare, Vijeta, Palaiahnakote Shivakumara, Raveendran Paramesran and Michael Blumenstein. Arbitrarily-oriented multi-lingual text detection in video. Multimedia Tools and Applications 76, (2016)
[21] Amandeep Kaur, Renu Dhir, Gurpreet Singh Lehal, "A survey on cameracaptured scene text detection and extraction: towards Gurmukhi script", International Journal of Multimedia Information Retrieval, (2017).
[22] Khare, Vijeta et al. Arbitrarily-oriented multi-lingual text detection in video. Multimedia Tools and Applications 76 (2016).
[23] Bodnar, Peter Nyl, Lszl. "QR Code Localization Using Boosted Cascade of Weak Classifiers". (2014) 22. 338-345. 10.1007/978-3-319-11758-4_37.
[24] https://github.com/zxing/zxing
[25] https://github.com/tesseract-ocr/tesseract
[26] Madhuri Latha, Chakravarthy, "An Improved Bernsen Algorithm Approaches For License Plate Recognition", IOSR Journal of Electronics and Communication Engineering (IOSR-JECE), Vol. 3, Issue 4 (2012).

Author Biography

Federica Mangiatordi received the M.Sc. Degree in Electronic Engineering at University of Rome La Sapienza and the PhD in Electronic Materials, Optoelectronics and Microsystems from the University of Roma TRE. She works at Fondazione Ugo Bordoni from 2007. Her research interest concern multimedia retrieval, image restoration algorithms, novel metrics for full reference and no-reference image objective quality assessment.

Andrea Bernardini received his Dr. Ing. degree in Computer Engineering at the University of Rome "Roma TRE". In 2010, he was Visiting Researcher at the Institute for Computing, Information and Cognitive Systems (ICICS) of the University of British Columbia (UBC). In 2002 he joined The Fondazione Bordoni, where he works as researcher in Information processing and management Department. His research interests include User Experience, Data Mining and User Modeling.

Licia Capodiferro received her Dr. Ing. degree in Electronic Engineering from the University of Rome La Sapienza, Italy. In 1987 she joined the Fondazione Ugo Bordoni where she currently works as head of the Department of Information Processing and Management. Her main research interests are in the field of multimedia processing, with a focus on algorithms that allow the use of images and videos on the different types of terminals.

Emiliano Pallotti received the Laurea Degree in Telecommunications Engineering at the University of Rome La Sapienza, Italy, and PhD in Electronic Materials, Optoelectronics and Microsystems from the University of Roma TRE. In 2007 he joined the Fondazione Ugo Bordoni where his research activities are in the field of on computational algorithms and video processing techniques based on multiresolution image representation in wavelet domain.