Web Image Indexing Using WICE and a Learning-Free Language Model

Nicolas Tsapatsoulis(✉)

Department of Communication and Internet Studies,
Cyprus University of Technology, 30, Arch. Kyprianos Str., 3036 Limassol, Cyprus
nicolas.tsapatsoulis@cut.ac.cy

Abstract. With the advent of Web 2.0 and the rapidly increasing popularity of online social networks that make extended use of visual information, like Facebook and Instagram, web image indexing regained great attention among the researchers in the areas of image indexing and information retrieval. Web image indexing is traditionally approached, by commercial search engines, using text-based information such as image file names, anchor text, web-page keywords and, of course, surrounding text. In the latter case, for effective indexing, two requirements should be met: Correct identification of the related text, known as image context, and extraction of the right terms from this text. Usually, researchers working in the field of web image indexing consider that once the image context is identified extraction of indexing terms is trivial. However, we have shown in our previous work that this is not the rule of thumb.

In this paper we get advantage of Web Image Context Extraction (WICE) using visual web-page parsing and specific distance metrics and following this we locate key terms within this text to index the image using language models. In this way, the proposed method is totally learning free, i.e., no corpus need to be collected to train the keyword extraction component, while the identified indexing terms are more descriptive for the image since they are extracted from a portion of web-page's text. This deviates from the traditional web image indexing approach in which keywords are extracted from all text in the web-page. The evaluation, performed on a dataset of 978 manually annotated web images taken from 243 web pages, shows the effectiveness of the proposed approach both in image context extraction and indexing.

Keywords: Image retrieval · Web image indexing · Web page parsing · Language models

1 Introduction

Since the beginning of the World Wide Web (WWW) and the development of cheap digital recording and storage devices the amount of available on-line digital images, continuously increases. The increasing popularity of online social

L. Iliadis and I. Maglogiannis (Eds.): AIAI 2016, IFIP AICT 475, pp. 131–140, 2016.
DOI: 10.1007/978-3-319-44944-9_12

networks, like Instagram, that are based on visual information push further this tendency. As a result, effective and efficient web image indexing and retrieval schemes are of high importance and a lot of research has been devoted towards this end.

In general, image retrieval research efforts are falling into two broad categories: content-based and text-based. Content-based methods retrieve images by analyzing and comparing the content of a given image example as a starting point. Text-based methods are similar to document retrieval and retrieve images using keywords. The latter is the approach of preference both for ordinary users and search engine engineers. Besides the fact that the majority of users are familiar with text-based queries, content-based image retrieval lacks semantic meaning. Furthermore, image examples that have to be given as a query are rarely available. From the search engine perspective, text-based image retrieval methods get advantage of the well established techniques for document indexing and are integrated into a unified document retrieval framework. However, for text-based image retrieval to be feasible, images must be somehow related with specific keywords or textual description. In contemporary search engines this kind of textual description is, usually, obtained from the web page, or the document, containing the corresponding images and includes HTML alternative text, the file names of the images, captions, metadata tags and surrounding text [1,2]. Text metadata are not always available, and in most cases are not accurate (i.e., they do not fully describe the visual content of the image). In addition, disambiguation of different visual aspects of the same term is very difficult using text metadata without taking into consideration the context.

Surrounding text, is the text that surrounds a Web image inside an HTML document. This text is, indeed, a very important source of semantic information for the image. However, automatic localization of surrounding text is by no means easy mainly due to the modern web-page layout formatting techniques which are based on external files (stylesheets). As a result, visual segmentation (parsing) of the rendered web-page is required in order to identify the surrounding text of an image. The need to automatically extract the semantically related, to an image, textual blocks and assign them to this image led to what we call Web Image Context Extraction (WICE). In that terms, WICE is the process of automatically assigning the textual blocks of a web document to the images of the same document they refer to [3].

In content-based image retrieval features such as color, shape or texture are used for indexing and searching web images. The user provides a target image and the system retrieves the best ranked images based on their similarity from the user's query. Although it has been a long time since the scientists, working on this area, defined the semantic gap [4], *i.e.*, the inability of a system to interpret images based on automatically extracted low-level features, a solution still does not exist. WICE methods may be used as a means of bridging this gap. For instance, in [5] the authors propose an auto-annotation system which combines a content-based search stage to image annotation along with a text-based stage in order to leverage the image dataset in learning from similar annotations.

Despite the fact that image context identification and text-based image indexing is very important per se, the huge amount of images which do no appear in web-pages or they do not have a clearly related context, either as surrounding text or as specific keywords, puts another challenge. Recently, the idea of visual concept modeling [6,7] was proposed as possible solution to this problem. In these approaches keywords are modeled through via low-level features and non-annotated images are passed through these models in order to identify possible matches with the visual representation of the models and assigned the corresponding keywords. The approach is promising but the keyword models require proper training, usually approached as a learning by example procedure, and, thus, they depend heavily on the selected corpus and keyword identification. So far the training set was created using manually annotated data and, for quality assurance on the keyword selection, crowdsourcing methods were adopted [8,9]. Recently, Giannoulakis and Tsapatsoulis [10] investigated the case of Instagram as a source for annotated images concluding that an overall 26 % of hashtags accompanying photos in Instagram are related with photos' visual content. Thus, filtering approaches are still required for a proper training set to be created.

In this paper we investigate the automatic extraction of keywords, from a web-image's context, that can be used either for image indexing or for the automatic creation for training datasets for the visual modeling of keywords mentioned above. Our method deviates from existing approaches in a very important aspect: It does not require any sort of training since it is based on a priori fixed English language model to identify the importance of a keyword in a text fragment. The latter is identified through a computationally efficient visual-based html parsing algorithm. The fact that image context is in most cases concise leads traditional approaches, like probabilistic, *tf-idf* based and clustering based ones, to failure. Thus, image context needs to extended, but, in this case the correlation of the selected text with a specific image in the web page decreases and all images in the web-page tend to share the same indexes (which are also similar to the indexes of the web-page itself). Finally, probabilistic, *tf-idf* and clustering based approaches require training. As a result the problem is recycled: In order to automatically index non-explicitly annotated images you need training examples but in order to automatically create the training examples you need train the indexing extraction algorithms!

2 Related Work

2.1 WICE Methods

In text-based web image retrieval, image file names, anchor texts, surrounding paragraphs or even the whole text of the hosting web page are traditionally used. The user provides keywords or key phrases and text retrieval techniques are used for the retrieval of the best ranked image. Early examples of these methods include [11–13]. However, it turned out very soon that the relevant, to each image, text fragment of the hosting web page must be extracted for better accuracy of retrieval. This is the well-known WICE problem [3], already

mentioned in introduction. The high diversity of designing patterns in web pages, the noisy environment (advertisements, graphics, navigational objects etc.), end the existence of too much textual and visual information in single documents are prohibiting factors a WICE system must overcome.

Several WICE methods have been proposed in the literature. A first category of approaches as [14,15], make use of the DOM tree structure of the hosting web page. In general these methods are not adaptive and they are designed for specific design patterns.

Web page visual segmentation is a second category of approaches to the WICE problem. This kind of approaches was initially proposed in [16], where the authors use Vision based Page Segmentation (VIPS) [17] in order to extract blocks, which contain both image and text, and construct an image graph using link structures. Web page segmentation is indeed a more adequate solution to the problem of text extraction since it is adaptable to different web page styles and depends on the visual cues that form each web page. Most of the proposed algorithms falling within this approach they are computationally heavy [18] and they are not designed specifically for the problem of image indexing [19]; therefore, they often deliver poor results [20]. In addition the creators of VIPS stopped its maintenance as early as in 2011.

In the proposed approach, the WICE problem is tackled through HTML code parsing of the rendered web page. This approach is computationally light and easily applicable and in modern web pages that include several CSS files, for styling purposes, as well as dynamic elements (such as PHP code), is quite effective. It executes html parsing by combining the ideas and the tool presented in [21] with the open source code of Mozilla web browser[1].

2.2 Web Image Indexing from Concise Text Fragments

The text fragments identified by WICE methods are usually very concise; as a result traditional keyword extraction (*tf-idf* like methods) does not apply. Web image retrieval based on clickthrough data is more relevant and effective. Clickthrough data are usually collected from search logs and include text queries and data from relevance feedback [22]. These methods [23–26] are quite effective but they are based on machine learning; thus, they suffer from the scalability problem and they are inappropriate for large scale web image retrieval.

The proposed method uses a learning-free language model for web image indexing using text fragments located by a new WICE method explained next.

3 The Proposed Method

The overall architecture of the proposed method is shown in Fig. 1 along with an illustrative example. It consists of three main steps: (a) html parsing, obtained with the aid of the lxml parser[2], (b) the WICE algorithm, and (c) an English language model accompanied by an keyword extraction algorithm.

[1] http://www.mozilla.org/en-US/.
[2] http://lxml.de/parsing.html.

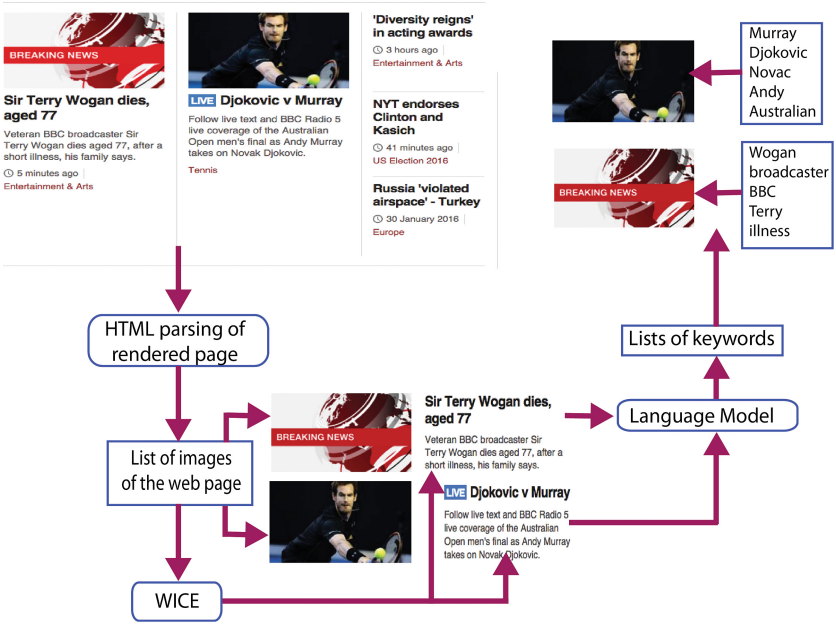

Fig. 1. The architecture of the proposed method through an example

3.1 The WICE Algorithm

The aim of the WICE algorithm is to identify the context (text fragment) of an image given its URL. First the position of the image and its (rendered) dimensions are computed with the aid of the Mozilla open source code. Small images and graphic types (i.e., gif) corresponding to logos and banners are discarded in this stage. Next the nearby sentences (text belonging to the same level as the image in the DOM tree and being within a radius equal to 0.3 of web page's rendered height from image's center) of the image are selected along with the caption, alternative text (if exists) and the hyperlink text. All these text data are merged together to form the text fragment (referred to as image context in the following) related to the given image.

3.2 An English Language Model for Image Retrieval

The relative frequency f_w of appearance of a word 'w' in natural (human) languages follows the well known Zipf law [27]; that is the product of f_w with the ranking R_w of word 'w' is approximately constant:

$$f_w \cdot R_w = c \tag{1}$$

Given that the number of words in web pages is very large and continuously increasing, due to new documents, misspelling, slang, etc., the probability of

Table 1. Examples of the performance of language model. $P(w)$ is the actual probability of word w

Word (w)	R_w	$P(w)$	$\frac{c}{R_w}$	$\frac{c}{R_w+k^{R_w}}$
of	2	0.0280	0.0600	0.0374
to	3	0.0260	0.0400	0.0277
with	14	0.0060	0.0086	0.0067
at	15	0.0056	0.0080	0.0063
his	24	0.0038	0.0050	0.0035
but	25	0.0038	0.0048	0.0034

appearance $P(w)$ of a word 'w' in a web page can be approximated by considering $P(w) = f_w$ as follows:

$$P(w) = \frac{c}{R_w} \tag{2}$$

In text-based retrieval a common keyword identification method involves the well-known *tf-idf* score. Words, or more general, tokens [27] with high *tf-idf* score in a given document or text fragment are considered important (keywords) for its description and can be used as indices. However, while the *tf* (term frequency) value depends only in the specific document the *idf* (inverse document frequency) value is computed based on a relative large number of relevant documents. Thus, in order to compute *tf-idf* we need a training corpus.

In this paper, we argue that web images appear in every type of document in the web, and, as a result it is not necessary to collect a specific domain corpus to compute *idf* and find the keywords of a text fragment. Therefore, we approximate the *idf* value with $\frac{1}{P(w)}$ and we arrive in a learning free indexing method for text fragments related with web images. Equation 2 gives a rough approximation of $P(w)$. After a little experimentation (see Table 1 for some examples) we found that a more accurate language model can be obtained by:

$$P(w) = \frac{c}{R_w + k^{R_w}} \tag{3}$$

where $c = 0.12$ and $k = 1.1$.

In any case, in order for a language model given above to be useful the ranking R_w[3] of every word should be available. In this work we get advantage of the 'Wordcount' project[4] for this purpose.

Keyword extraction is facilitated by the language model and involves a series of steps: (a) text fragment segmentation into sentences, (b) stopwords removal, (c) part of speech (POS) tagging, (e) noun and proper noun selection as candidate indices, (f) ranking of selected terms based on the adopted language model, and (g) final selection of the indexing terms (keywords) based on the $S = \frac{tf}{P(w)}$

[3] http://www.wordcount.org/main.php.
[4] http://www.wordcount.org/about.html.

Table 2. Evaluation results for context extraction. GS denotes gold standard

	# tokens in GS	# tokens found	TP	FP	FN	R	P	F-measure
Context localization	61683	65918	52267	13651	9416	0.847	0.793	0.819

score (terms whose score exceeds an empirically derived threshold T are kept as keywords).

4 Experimental Evaluation

In order to evaluate the proposed method 978 web images taken from 243 web pages were used. Three annotators (students of the Cyprus University of Technology) were asked, independently, to: (a) for each image identify and copy its context (text fragment), and (b) select the keywords from context that best describe the image. The contexts and keywords from the three annotators were merged and used as the gold standard for the evaluation of the proposed method. Indicative examples of images, their context and the keywords chosen by the annotators and found by the algorithm are online available[5].

Table 2 summarizes the results of context extraction. By TP we denote the 'True Positive' rate, that is, the tokens that were in the gold standard and found by the algorithm. Similarly FP denotes 'False Positive' rate, i.e., tokens found by the algorithm but not in the gold standard and FN denotes 'False Negative' rate, that is, tokens in the gold standard that were not found. Recall (R), Precision (P) and F-score (F) values are computed as usual by:

$$R = \frac{TP}{TP + FN} \tag{4}$$

$$P = \frac{TP}{TP + FP} \tag{5}$$

$$F = \frac{2 \cdot P \cdot R}{P + R} \tag{6}$$

We can see in Table 2 that the algorithm tends to include in image context more tokens than those identified by the annotators. This is mainly caused by our decision to include in image context not only the nearby sentences but the tokens in image caption and image's hyperlink. The latter was typically never selected by the annotators although from several studies we know that the information in hyperlinks is of utmost importance in information retrieval. Overall, the results are satisfactory given the simplicity of the proposed method. For comparison see the results reported by Alcic and Conrad [15].

[5] http://cis.cut.ac.cy/~nicolas.tsapatsoulis/ckasapi/showImages.php.

Table 3. Evaluation results for keyword identification

	# keywords in GS	# keywords found	TP	FP	FN	R	P	F-measure
Keyword identification	5966	7237	2836	3130	4401	0.475	0.392	0.430

Table 3 shows the results of keyword identification. We observe that the recall rate is close to 50 % which is very promising compared to similar methods (see for instance [26]). Human annotators tend to use more 'emotional' words to describe images even in cases where the visual content does not correspond clearly to these terms. On the hand the proposed algorithm promotes nouns as keywords (a choice made during the design of the algorithm) and especially named entities (as a result of the use of the proposed language model). Similarly to the context extraction case the algorithm identifies, in general, more keywords than the annotators causing the recall to become higher than precision. Nevertheless, in information retrieval the tendency is to pursuit higher recall than precision (irrelevant results are better than no results).

5 Conclusion and Further Work

In this paper we have presented a method for web page image indexing which is based on language models. The method can be applied to identify keywords for any web image without training on a particular corpus. It is based on raw html parsing of web pages to identify the nearest (to the image) text block and then a metric which combines the frequency of terms in the block with their frequency ranking, in the corresponding language, is used. Preliminary results, on an especially designed and annotated web image database, are promising and show the effectiveness of the proposed method. However, there are also some limitations that need to be surpassed for the method to be widely applied while some improvements are planned for (near) future work. The effectiveness of the proposed system on 'carousel' type web images needs to be tested. Furthermore, the algorithm is currently applied only on English web pages since we get advantage of the ranking of English words to create our language model. However, once such a study for any other natural language exists the extension of the proposed method is straightforward.

An improvement of the proposed method is to consider the structure of the surrounding text to further weight the terms. Thus, terms that appear in headers, subheaders, weblinks, etc., will be given higher importance than the terms in the plain text. Finally, the method will be used in the context of visual concept modeling [8] for automatic creation of image-keywords pairs that are required for training purposes. In this context, the other basic limitation of this work, that is, the inability of applying it to non-web images, will be overcome. For more information on this, please see [6,7].

References

1. Souza Coelho, T.A., Calado, P.P., Souza, L.V., Ribeiro-Neto, B., Munt, R.: Image retrieval using multiple evidence ranking. IEEE Trans. Knowl. Data Eng. **16**(4), 408–417 (2004)
2. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: ideas, influences, and trends of the new age. ACM Comput. Surv. **40**(2), 5:1–5:60 (2008)
3. Tryfou, G., Theodosiou, Z., Tsapatsoulis, N.: Web image context extraction based on semantic representation of web page visual segments. In: Proceedings of International Workshop on Semantic and Social Media Adaptation and Personalization, pp. 63–67. IEEE (2012)
4. Del Bimbo, A.: Visual Information Retrieval. Morgan Kaufmann Publishers Inc., San Francisco (1999)
5. Wang, X.J., Zang, L., Jing, F., Ma, W.Y.: Annosearch: image auto-annotation by search. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 1483–1490. IEEE (2006)
6. Theodosiou, Z., Tsapatsoulis, N.: Image retrieval using keywords: the machine learning perspective. In: Semantic Multimedia Analysis and Processing, pp. 3–30. CRC Press (2014)
7. Xu, G.Q., Mu, Z.C.: Automatic image annotation using modified keywords transfer mechanism base on image-keyword graph. Int. J. Comput. Sci. Issues **10**(2), 267–272 (2013)
8. Theodosiou, Z., Tsapatsoulis, N.: Modelling crowdsourcing originated keywords within the athletics domain. In: Iliadis, L., Maglogiannis, I., Papadopoulos, H. (eds.) Artificial Intelligence Applications and Innovations. IFIP AICT, vol. 381, pp. 404–413. Springer, Heidelberg (2012)
9. Theodosiou, Z., Tsapatsoulis, N.: Crowdsourcing annotation: modelling keywords using low level features. In: Proceedings of the 5th International Conference on Internet Multimedia Systems Architecture and Application, pp.1–4. IEEE (2011)
10. Giannoulakis, S., Tsapatsoulis, N.: Instagram hashtags as image annotation metadata. In: Chbeir, R., Manolopoulos, Y., Alhajj, R. (eds.) AIAI 2015. IFIP AICT, vol. 458, pp. 206–220. Springer, Heidelberg (2015). doi:10.1007/978-3-319-23868-5_15
11. Alexandre, L., Pereira, M., Madeira, S., Cordeiro, J., Dias, G.: Web image indexing: combining image analysis with text processing. In: Proceedings of the 5th International Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS 2004 (2004)
12. Smith, J.R., Chang, S.F.: An image and video search engine for the world-wide web. In: Proceedings of SPIE Storage and Retrieval for Image and Video Databases, pp. 84–95. SPIE (1997)
13. Ortega-Binderberger, M., Mexico, A.: Webmars: a multimedia search engine for the world wide web. University of Illinois at Urbana-Champaign (1999)
14. Fauzi, F., Hong, J.L., Belkhatir, M.: Webpage segmentation for extracting images and their surrounding contextual information. In: Proceedings of the 17th ACM International Conference on Multimedia, pp. 649–652. ACM (2009)
15. Alcic, S., Conrad, S.: A clustering-based approach to web image context extraction. In: Proceedings of the 3rd International Conferences on Advances in Multimedia, pp. 74–79. IARIA (2011)
16. He, X., Cai, D., Wen, J.R., Ma, W.Y., Zhang, H.J.: Clustering and searching www images using link and page layout analysis. ACM Trans. Multimedia Comput. Commun. Appl. **3**(2) (2007)

17. Cai, D., Yu, S., Wen, J.R., Ma, W.Y.: Vips: a vision based page segmentation algorithm. Technical report, Microsoft Research (2003)
18. Tryfou, G., Tsapatsoulis, N.: Using visual cues for the extraction of web image semantic information. In: Zaphiris, P., Buchanan, G., Rasmussen, E., Loizides, F. (eds.) TPDL 2012. LNCS, vol. 7489, pp. 396–401. Springer, Heidelberg (2012)
19. Tryfou, G., Tsapatsoulis, N.: Extraction of web image information: semantic or visual cues? In: Iliadis, L., Maglogiannis, I., Papadopoulos, H. (eds.) Artificial Intelligence Applications and Innovations. IFIP AICT, vol. 381, pp. 368–373. Springer, Heidelberg (2012)
20. Alcic, S., Conrad, S.: Measuring performance of web image context extraction. In: Proceedings of the 10th International Workshop on Multimedia Data Mining, pp. 1–8. ACM (2010)
21. Pappas, N., Katsimpras, G., Stamatatos, E.: Extracting informative textual parts from web pages containing user-generated content. In: Proceedings of 12th International Conference on Knowledge Management and Knowledge Technologies, vol. 4, pp. 1–8. ACM (2012)
22. Park, J.Y., O'Hare, N., Schifanella, R., Jaimes, A., Chung, C.W.: A large-scale study of user image search behavior on the web. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp.985–994. ACM (2015)
23. Tsikrika, T., Diou, C., de Vries, A.P., Delopoulos, A.: Image retrieval using multiple evidence ranking. Multimedia Tools Appl. **55**(1), 27–52 (2011)
24. Fang, Q., Xu, H., Wang, R., Qian, S., Wang, T., Sang, J., Xu, C.: Towards msr-bing challenge: ensemble of diverse models for image retrieval. In: Proceedings of the 2013 MSR-Bing Image Retrieval Challenge, Microsoft (2013)
25. Hua, X.S., Yang, L., Wang, J., Wang, J., Ye, M., Wang, K., Rui, Y., Li, J.: Clickage: towards bridging semantic and intent gaps via mining click logs of search engines. In: Proceedings of the 21st ACM International Conference on Multimedia, pp. 243–252. ACM (2013)
26. Sarafis, I., Diou, C., Tsikrika, T., Delopoulos, A.: Weighted svm from clickthrough data for image retrieval. In: 2014 IEEE International Conference on Image Processing (ICIP), pp. 3013–3017. IEEE (2014)
27. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge (1999)