Extracting Representative Image From Web Page

Najlah Gali, Andrei Tabarcea and Pasi Fränti
Speech and Image Processing Unit, School of Computing, University of Eastern Finland, Joensuu, Finland
{najlaa, tabarcea,franti}@cs.uef.fi

Keywords: Representative image, Image extraction, Web page information extraction, Web mining.

Abstract: A web page typically contains a blend of information. For a particular user, only informative data such as main content and representative images are considered useful, while non-informative data such as advertisements and navigational banners are not. In this work, we focus on selecting a representative image that would best represent the content of a web page. Existing techniques rely on prior knowledge of website specific templates and on text body. We extract all images, analyze and rank them according to their features and functionality in the web page. We select the highest scored image as the representative image. Our method is fully automated, template independent, and not limited to a certain type of web pages.

1 INTRODUCTION

The web today is a world of information, filled with videos, images and interactive content. To deal with these forms of data, different techniques have been developed to deliver the informative parts of a web page to the user (see Figure 1), such as information retrieval (Yu et al., 2003), main content extraction (Kim et al., 2013) and image extraction (Kherfi et al., 2004).

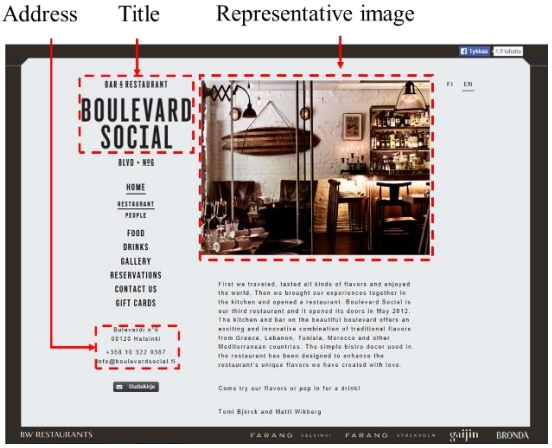

Figure 1: A sample web page and its relevant content to the user: title, image and location.

Images are used in web pages because they can transfer information to the user in a quick and efficient way, they are more informative than text at a glance. Even though a large number of images are embedded into web pages, many of them are less relevant to the content of the web page, such as advertisements, navigational banners, icons and images that serve as section headings (Azad et al., 2014). A solution is needed to ignore the irrelevant images and find a representative image for the web page.

We define the *representative image* of a web page as the image that best represents the content of the page to the user. Representative images are important in many applications, especially in cases when bandwidth limitation restricts the total number of images that can be retrieved or when building a visual category in which a single image must represent an entire category of documents and their associated content (Helfman and Hollan, 2000). Representative image is also important for location based applications such as MOPSI, which is available at cs.uef.fi/mopsi/, where simple thumbnail with title is the minimum information a user needs. It's also used in social applications such as Facebook and Google+ when users share a link of a web page on their wall.

Existing works have been mostly focused on extracting several useful images from a web page (Fauzi et al. 2009) or a collection of web pages (Park et al. 2006) and on selecting an image for a particular category of web pages such as news article. Less attention has been paid to how to select an image that represents the entire web page.

The method in (Joshi and Liu, 2009) focuses on news articles. It assumes that the relevant image is embedded in the article block and has a caption and that non-article images have no caption. It considers

only images with captions as candidate images and may therefore misses potentially useful web images that do not have captions.

The image extraction method in (Adam et al., 2010) focuses on web pages that are written in article-style (title and body). The method locates the border of the article and selects an image from this region based on its size and aspect ratio. It provides image annotation by identifying the captions assigned to them. This method considers only images that accompany the article, which is a section of the web page and may falsely select advertisement images if they have acceptable size and aspect ratio. Our task is wider because we consider all images in the web page and we select an image that represents the entire web page.

Google+ *share preview snippet* (Google+, 2014) summarizes a post made to Google+. It includes a link, a page title, a brief description of the page, and a thumbnail image. The image is selected based on its size and aspect ratio. The image height must be at least 120 pixels, and if the width of the image is less than 100 pixels, then the aspect ratio must be ≤ 3. Although the explicit framework for the snippet has not been published in any scientific forum, the method is described in technical document, and it is used in real application.

The method in (Tsymbalenko and Munson, 2001) focuses on finding relevant images to specific query without downloading or analyzing images. It examines only the text that surrounds the image tag in the source code of the web page and then decides whether the image is relevant or not. However, many web pages do not have text surrounding their useful images which lead to exclude them from being candidate images.

A functional categorization of images is studied in (Hu and Bagga, 2003). The images are classified into categories based on their usage in the web page by defining eight categories: *story*, *preview*, *commercial*, *host*, *heading*, *icons and logos*, *formatting*, *miscellaneous*. These are further grouped in two super-classes: one for useful images (story, preview and host) and one for the images that are not associated with the content (the other categories). We also use image categorization, but we use it directly in the method for helping to choose the best image.

The method in (Gupta et al., 2003) navigates Document Object Model (DOM) tree that is created by parsing the Hypertext Mark Up Language (HTML) code recursively and uses it to extract relevant information, including images. It filters out irrelevant data such as advertisement images by examining the values of the *src* and *href* attributes to determine the servers which the links refer to. If an address matches against a list of common advertisement servers, the node of the DOM tree that contains the link is deleted. We also use the DOM tree of a web page, but we use more image attributes and we define more categories.

The method in (Parmar and Gadge, 2011) removes advertisement images by using a rule-based classifier. Seven rules are defined to decide whether the image is an advertisement or not: *domain name difference*, *dimension*, *well-known advertisement provider*, *advertisement related keywords*, *advertisement by scripting*, *dynamic advertisement*, *flash plug-in removal*. This method eliminates most advertisement images in the web page.

Despite of several researchers have been working in related areas, none of the existing methods is directly applicable to our problem as such. To our knowledge, the only existing methods are the commercial ones implemented in Google+ and Facebook but, according to our experiments, neither of them is working perfectly.

In this paper, we propose a method that parses the source code of a web page, detects all the images and selects one that best represents the content. Instead of analyzing the content of the images or examining the text surrounding them, we rely on the functional purpose of the images within the web page and on the features such as the size, the aspect ratio, the format of the image and the attributes of HTML tags. Similarly to (Hu and Bagga, 2003) we classify the images into categories. We define the following categories based on image functionality: *representative*, *logos*, *banners*, *advertisements*, and *formatting including icons*. We rank the categories in this order, based on how important they are with respect to the content of the web page. The images in each category are ranked based on their features.

The main contribution of our method is that it does not rely on the surrounding text, on certain template or web page categories. Instead, it is targeted to work with all types of web pages. It is therefore general and not limited by the writing style or the layout of the web page. Besides the selection of the threshold values, the method does not require any training data. It is designed to work in real time, without the need to store the results in a database or to query a set of pre downloaded web pages. Since we consider prior classification of images, our method is useful in several applications such as automatic identifying adverts, saving bandwidth by web crawlers by downloading carefully only most relevant media objects, and automatic converting web page for consumption on mobile small screen devices.

The proposed method is implemented in two places in a mobile location-based application called *Mopsi* (Fränti et al., 2011). The first one is to show search results to the mobile user, and the second one is the interactive tool for adding new services to the database using data from web pages.

2 EXTRACTING REPRESENTATIVE IMAGES

The workflow of the algorithm is shown in Figure 2. We start by downloading the source code of the web page and analyzing it using a DOM parser. The DOM representation is a platform- and language-independent interface that allows programs and scripts to dynamically access and update the content, structure and style of documents (www.w3.org/DOM).

We navigate through the DOM tree of the web page to identify links to images by locating <*img*> tags and to Cascading Style Sheets (CSS) files by locating <*link*> tags (*type*=text/css or *rel*=stylesheet), and JavaScript (JS) files by locating <*script*> tags. After analyzing the HTML source code, we use regular expressions to extract the images in CSS and JS files. If the considered web page does not contain any images and is not the root page of the domain, we also analyze the root page in the same way.

We notice that most of the images found in CSS are formatting or background images, although sometimes they have good features. Therefore, we chose to use images from CSS only in the case where no images are detected from HTML, even though best image is sometimes found from CSS (see Figure 3).

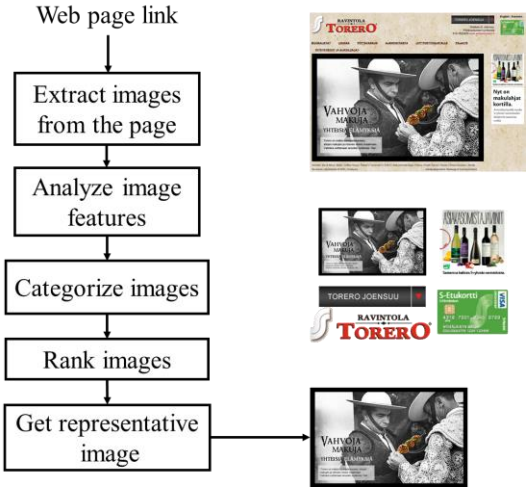

Figure 2: Extracting the representative image.

We then extract a list of features for each image, which are *src*, *alt*, *title*, *from*, *format*, *width*, *height*, *size*, and *aspect ratio* (see Figure 4).

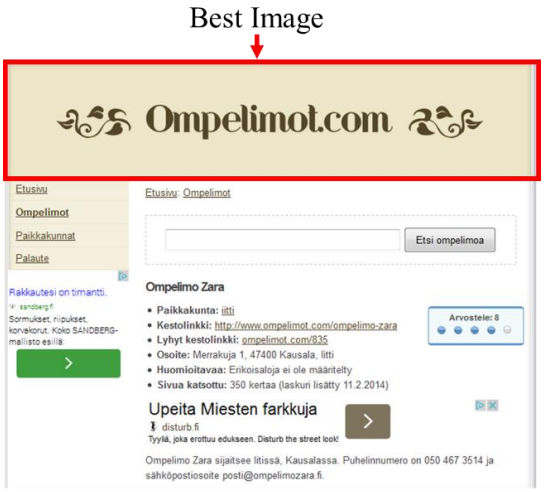

Figure 3: Best image is found from CSS source.

src	http://www.martina.fi/sites/martina.fi/files/styles/fiiliskuva/public/Valitse%20alikansio/Ravintolat/ravintola-martina-paakuva-pasta.jpg?itok=z8DMqAu2
alt	Ravintola Martina Joensuu
title	--
from	html
format	jpg
width	920
height	313
size	287.96 px
Aspect ratio	2.94
Parent tag	
Class	header_fiilis
Class of parent	content clearfix

Figure 4: A sample banner image and its features.

Because we aim at a real-time application, we do not download the images but we calculate width and height either by using the attributes of the <*img*> tag,

retrieving them from CSS or by downloading just the image header (the first kilobytes of the file, which contain the image meta-information).

2.1 Categorization

We define five image categories based on its usage within the web page (see Figure 5) and rank them in the following priority order:

- *Representative:* images that are directly related to the content or the topic of the web page (see Figure 6);
- *Logos:* recognizable images that are used to identify the company or institution that owns the website (see Figure 7);
- *Banners:* images placed on a web page above, below or on the sides of the content. They are generally used for decoration. Headers and footers are classified in this category (see Figure 8);
- *Advertisements:* images that promote products or services that are irrelevant to the topic or the content of the web page (see Figure 9);
- *Formatting and Icons:* images that are used to enhance the visual appearance such as spacers, bullets, borders, backgrounds, or pictures used purely for decoration. We also include the small images which are not classified as logos and serve a functional purpose, such as icons which link to the home page or icons which are used for changing language (see Figure 10).

All images are first assigned into a proper category, and the images in the same category are ranked according to a secondary criteria. The image is chosen from the highest priority category that contains any image.

Figure 5: A sample web page which contains images from all the 5 categories we defined.

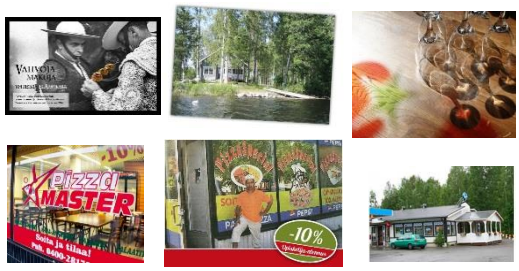

Figure 6: Examples of representative images.

Figure 7: Examples of logos.

Figure 8: Example of banners.

Figure 9: Examples of advertisement images.

Figure 10: Examples of formatting images and icons.

We categorize the images using the rules in Table 1. In all categories, a predefined set of keywords is used. If any of these are found in the image URL, in the class name of the <*img*> tag, or of the parent element, then the image is assigned to that category. Banners and Formatting are also categorized according to image size and aspect ratio.

Table 1: Rules used for image categorization.

Category	**Features**	**Keywords**
Representative	Not in other category	
Logo		logo
Banner	Ratio>1.8	banner, header, footer, button
Advertisement		free, adserver, now, buy, join, click, affiliate, adv, hits, counter
Formatting and Icons	Width<100 px Height<100 px	background, bg, spirit, templates

Note that the categories are overlapping, so the same image may meet the conditions of multiple categories. In this case, we use a decision tree to assign the image to a single category (see Figure 11). We categorize logo images first because their size and aspect ratio might satisfy the conditions of Banner and Formatting categorizes. We categorize advertisement images next because their aspect ratio or their HTML assigned keywords might satisfy the conditions of Formatting or Banner categories. Formatting category is followed because its image aspect ratio might satisfy the condition of Banner category. An image can belong to the class of Representatives only if it does not belong to any other category. The same prioritization is applied for all HTML, CSS and JS, and the images in these file types are considered equal.

The criterion for Logos category is that at least one of the HTML tag attributes (URL, the detected classes, the IDs of the element, or the IDs of the parent element) contains the keyword ”*logo*”. The criterion for Banners category is that at least one of the HTML tag attributes contains any of the keywords: ”*banner*”, ”*header*”, ”*footer*”, ”*button*”, or that the aspect ratio of the image is higher than a threshold 1.8, which was experimentally obtained using small training set of 50 web pages.

The criterion for Advertisements category is that at least one of the HTML tag attributes contains any of the advertising keywords: ”*free*”, ”*now*”, ”*buy*”, ”*join*”, ”*adserver*”, ”*click*”, ”*affiliate*”, ”*adv*”, ”*hits*”, ”*counter*”.

Advertisement images can be hosted either on the same domain as the web page or on a separate server. It is also common that useful images are stored on a different domain, as more and more websites are using a separate server to store images on a cloud storage server. Therefore, the domain where the image is hosted is not a consistent rule that could be used to determine if the image is an advertisement.

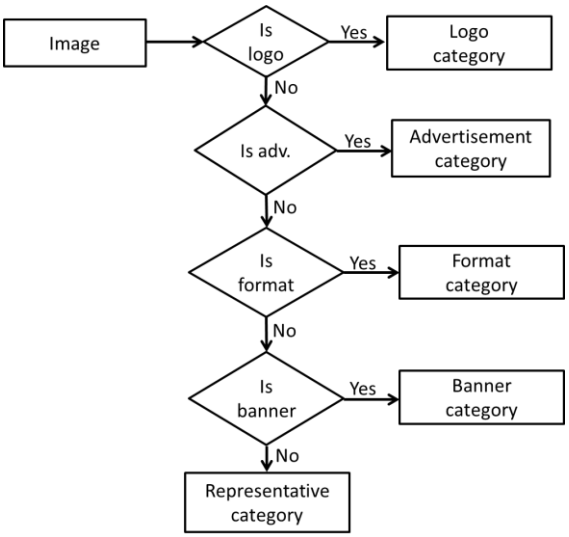

Figure 11: Decision tree used in assigning image categories.

The criterion for Formatting and icons category is that at least one of the HTML tag attributes contains any of the keywords: “*background*”, “*bg*”, “*sprite*”, “*templates*” or if the height or width are smaller than an experimentally selected threshold of 100 pixels.

2.2 Scoring

We analyze the features of the image according to a set of rules as shown in Table 2. We score the image based on the following criteria:

- *Image size*: we consider the image has a good size if it meets the following condition:

$$Size = width \times height \geq 10.000\ px \qquad (1)$$

- *Aspect ratio*: we consider the image that has aspect ratio ≤1.8 is more important than other images, which tend to be either wide and short, or narrow and long, which are usually features of banners, formatting or advertisements. We calculate the aspect ratio as follows:

$$Ratio = \frac{\max(width, height)}{\min(width, height)} \leq 1.8 \qquad (2)$$

- *Image alt and title*: the alt and title attributes describe the content of the image. Images that have alt or title attributes are more important than other images in the web page. Therefore, we extract the keywords of image alt and title and compare them with the keywords of the web page <*title*> and <*h1*> contents. We firstly extract the content of the web page <*title*> and <*h1*> by xpath. Secondly, we use a predefined set of patterns which consists of space and delimiters such as ',', ';', '/', '|', '>', '¦', '«', '–', '.', ':', '?', '::' to separate the words and the phrases of the web page <*title*> and <*h1*> tags. Thirdly, we remove any special characters such as '[', '{', '?', '!' from the text. Finally, we use string comparison to match the keywords of image *alt* and *title* with the keywords of the web page *title* and *h1s*;
- *Image path and URL*: we parse the image path, extract its keywords, and match the keywords with the web page <*title*> and <*h1*> keywords. If a match is found then we consider the image is more related to the content of the web page;
- *In the sub-tree of <h1> or <h2> tags*: we consider an image that is a child of <*h1*> or <*h2*> in the DOM tree is more related to the content of the web page because <*h1*> and <*h2*> contain the main topics of the page, therefore, we consider the images located in these sub-trees are important;
- *Image format*: we analyze four types of formats, which are Joint Photographic Experts Group (jpg), Scalable Vector Graphics (svg), Portable Network Graphics (png), and Graphics Interchange Format (gif). We consider jpg format is more important because it is used for photographs. Although png format is suitable and is increasingly used for compressing photos, most of the web pages use it just for logos and icons. Therefore, we consider it less important than jpg. Less importance is also given to image of format svg and gif because these types of formats are used for graphics and they are more often used for formatting images.

All rules are considered equally important and therefore assigned the same weight of 1 except for some types of image formats as mentioned earlier.

The scores are calculated only for images in the highest priority category, in our case Representative category. If no image is assigned to this category then the scores will be calculated for the images of next highest priority category and so on.

The scores are summed up and the images in the category are ranked based on their scores, except for logo category where the images are sorted based on size, because we consider the web page logo has biggest size among other logos that exist in the page.

Table 2: Rules used for image scoring.

Rule	**Score**
Image size ≥ 10.000 px	1
Aspect ratio ≤ 1.8	1
Image alt or image title has a value	1
Keywords of alt or title are in web page <*title*>	1
Keywords of alt or title are in web page <*h1*>	1
Keywords of path are in web page <*title*> or <*h1*>	1
the image is in the sub-tree of <*h1*> or <*h2*> tags	1
Format : jpg	1
Format: svg	0.5
Format: png	0.5
Format: gif	0.5

3 EXPERIMENTS

To collect a reasonable size of ground truth database we asked volunteers to select at most three images from websites of their own choice, or Mopsi search result. The interface of the data collection can be found here: cs.uef.fi/mopsi/img/. It works as follows:

Search: The user can copy/paste the link of any website he/she visits often, like, or at least knows about. Website selection is therefore not limited to a country, category, specific domain or size of website. Alternatively, the user should give a keyword such as a favorite hobby and a city in Finland. Mopsi search will then provide resulting pages for him/her to evaluate. If the search results returned are service directory, the content is unclear, or the user just cannot decide, then he/she can skip the page and try another keyword/city combination. The user should select maximum three images that best describe the web page.

Evaluation: We have collected a dataset of 1002 websites and 2363 ground truth images (2.36 images per webpages, on average) by 117 volunteers during September 2014. The number of images in each website varies between 1 to over 154. Although the selected images serve as useful ground truth for our purpose, users' choices can sometime be subjective,

which makes it impossible for any realistic algorithm to get 100% accuracy with this data, even if the algorithm was trained for this particular website and knew the user who made the selection. Nevertheless, the ground truth collection is still useful for evaluating the performance of our method, on average.

We compare the performance of our method (WebIma) with Google+ algorithm because it also looks for an image to represent the entire webpage and it uses the same heuristics as in (Adam et. al., 2010), which are the size and the aspect ratio of the image. In addition, we also compared our method with the method used in Facebook when user shares web link on his/her wall. We evaluated these three algorithms by counting how many times they select one of the ground truth images as the output. The results were obtained by input the web link to the algorithms (ours by script, the others manually one by one) and comparing their first choice against the ground truth. These were done using their public web pages during 17-25 September 2014. The results are summarized in Table 3. It shows that our method finds correct image 642 times out of 1002 (64%), which outperforms the comparative results of Google+ (48%) and Facebook (39%). The results also show that our parser extracted the images from 99% of the tested websites.

To find out why the methods perform differently, we have selected two sets of samples from the collected dataset. The first set contains 30 websites where our algorithm performed 100% accuracy while both Google+ and Facebook failed. The second set contains 30 websites where our algorithm failed.

According to our categorization, for the first set the analysis shows that 63% of the ground truth images are selected from Representative category and the rest 37% of the images are from logos category. Less importance was given to the images of banner category, and no images were selected from advertisements and Formatting categories.

Comparing these results with the selections made by Google+ and Facebook, we can observe that both Google+ and Facebook preferred banner images because of their big size. About (57-60) % of images selected by them belong to banner category, which reduces the performance of both algorithms (see Table 4).

Further analysis of the images features shows that the users did not rely mainly on the images that are big in size, long or wide. About 73% of the ground truth images in set 1 are relatively smaller in size, height and width in comparison with those selected by Google+ and Facebook.

Table 3: Performance results for Mopsi WebIma dataset.

	Accuracy	**Extracted Images**
WebIma	64%	99%
Google+	48%	92%
Facebook	39%	90%

The aspect ratio of the images recorded in the first set lies between 1 and 6.2 and about 50% of images selected by our algorithm have aspect ratio lower than the ratio of the images selected by Google+ and Facebook. This indicates that users prefer more square images than images like banner. The statistics also show that most of the ground truth images in this set are of jpg format.

In the second set, as shown in Table 4, the ground truth images are distributed among three categories, which are Representative, Logos and Banners. Our algorithm did not select the correct images because the websites in this set contain many images that meet the criteria of Representative category, but are not selected by the users. The statistics shows that logo images are important and should be given bigger weight in the scoring.

The analysis of the images features in set 2 shows that both Google+ and Facebook selected more small images compared to those selected by WebIma.
This means that the users do not necessarily prefer biggest images in the web and therefore, bigger size should not be considered as the only threshold to select the representative image. Both Google+ and Facebook ignore the logo images if they do not meet the thresholds of the size and the aspect ratio, which affects their accuracy.

Jpg is still the most popular format for the representative images and the statics of the whole data set of 2363 images shows that 63% of the images are of jpg format. Png and gif are preferred for the logo and formatting images. These results make our early assumption of giving extra score to jpg images, is correct. Added to that, both Google+ and Facebook do not select CSS images even though some websites use only CSS and JS images in their design (see Figure 12).

We conclude that some of image features such as the aspect ratio and the jpg format are more important than other features such as the size of the image and therefore should have been given extra weights. Better optimization of the weights, however, is left for future work. Image categorization is also important because it identifies advertisements and formatting images as being harmful and excluded them from being candidates for image representation.

Table 4: Number of images selected from each category.

	Set 1				Set 2			
Image Category	Ground truth (%)	WebIma (%)	Google+ (%)	Facebook (%)	Ground truth (%)	WebIma (%)	Google+ (%)	Facebook (%)
Representative	63	63	13	20	33	83	27	40
Logo	37	37	13	10	30	7	30	7
Banner	0	0	57	60	37	3	27	33
Advertisement	0	0	0	0	0	0	3	3
Formatting	0	0	17	10	0	7	13	17

Figure 12: A website uses images from CSS source only.

No significant differences were concluded with respect to the other features from the selected sets.

4 CONCLUSIONS

We have introduced a method for extracting representative image from a web page. With the collected dataset, it finds correct image in 64% of the cases, which outperforms the results provided by Google+ and Facebook. Our method is implemented in the framework of MOPSI, which is a research project of location-based services developed at the University of Eastern Finland in two places: Search and Service upgrade.

The method works well especially for business oriented touristic places that have their own web page, whereas smaller enterprises in small towns or rural areas rely more on service directories. The service directories include two challenges: they include content of multiple services and it would be more difficult to detect an image that relates to the service in question. Many commercial service directories have also quite poor content, usually showing only name, contact info and map, followed by the service provider's own information. Image of the services themselves are often missing completely. Some web pages also have rather complicated structure where the visual appearance is not a single image, but more complicated product of several independent components. Future research should focus on these challenges.

Some improvement can also be done to the current method such as training the parameters for better results. This would require a large set of training data, which we are currently lacking. Nevertheless, the data we used has 1002 web sites, which makes the results statistically significant.

ACKNOWLEDGEMENTS

The work described in this paper was supported by MOPIS project, University of Eastern Finland.

REFERENCES

Adam, G., Bouras, C., & Poulopoulos, V., 2010. Image Extraction from Online Text Streams: A Straightforward Template Independent Approach without Training. In *Advanced Information Networking and Applications Workshops (WAINA), 24th International Conference*, pp. 609-614. IEEE.

Azad, H. K., Raj, R., Kumar, R., Ranjan, H., Abhishek, K., & Singh, M. P. 2014. Removal of Noisy Information in Web Pages. *In Proceedings of the 2014 International Conference on Information and Communication Technology for Competitive Strategies.* ACM.

Fauzi, F., Hong, J. L., & Belkhatir, M. 2009. Webpage segmentation for extracting images and their surrounding contextual information. *In Proceedings of the 17th ACM international conference on Multimedia*, pp. 649-652. ACM.

Fränti, P., Chen, J., & Tabarcea, A. 2011. Four Aspects of Relevance in Sharing Location-based Media: Content, Time, Location and Network. In *WEBIST*, pp. 413-417.

Google+ platform, 2014, *https:/developers.google.com/+/web/snippet/?hl=no*

Gupta, S., Kaiser, G., Neistadt, D., & Grimm, P., 2003. DOM-based content extraction of HTML documents. In *Proceedings of the 12th international conference on World Wide Web*, pp. 207-214. ACM.

Helfman, J. I., & Hollan, J. D., 2000. Image representations for accessing and organizing Web information. In *Photonics West 2001-Electronic Imaging*, pp. 91-101. International Society for Optics and Photonics.

Hu, J., & Bagga, A., 2003. Functionality-Based Web Image Categorization. *WWW (Posters)*, *2*(003).

Joshi, P. M., & Liu, S., 2009. Web document text and images extraction using DOM analysis and natural language processing. In *Proceedings of the 9th ACM symposium on Document engineering*, pp. 218-221. ACM.

Kherfi, M. L., Ziou, D., & Bernardi, A., 2004. Image retrieval from the world wide web: Issues, techniques, and systems. *ACM Computing Surveys (CSUR)*, *36*(1), pp. 35-67.

Kim, M., Kim, Y., Song, W., & Khil, A., 2013. Main Content Extraction from Web Documents Using Text

Block Context. In *Database and Expert Systems Applications*, pp. 81-93. Springer Berlin Heidelberg.

Park, G., Baek, Y., & Lee, H. K. 2006. Web image retrieval using majority-based ranking approach. *Multimedia Tools and Applications*, *31(2)*, pp.195-219.

Parmar, H. R., & Gadge, J., 2011. Removal of Image Advertisement from Web Page. *International Journal of Computer Applications*, 27(7).

Tsymbalenko, Y., & Munson, E. V., 2001. Using HTML metadata to find relevant images on the world wide web. *Proceedings of internet computing*, *2*, pp.842-848.

Yu, S., Cai, D., Wen, J. R., & Ma, W. Y., 2003. Improving pseudo-relevance feedback in web information retrieval using web page segmentation. In *Proceedings of the 12th international conference on World Wide Web*, pp. 11-18. ACM.