Volume 5, Issue 3, March 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Image Based Website Summarization Model

Apurva Dafe, Mithila Mondkar, Anshul Gupta, Nikheel Bhangale, Prof. Zaheed Shaikh

Computer Engineering Department
K J Somaiya College of Engineering
Maharashtra India

Abstract*— The concept of image-based summarization is used for improving the quality of Web site summaries and as a tool for more effective Web browsing and retrieval. Image-based summarization of a Web site is the process of extracting the most characteristic images from it. The criteria for measuring the importance of images in Web sites are based on their frequency of occurrence, characteristics of their content and Web link information. This project focuses on logo and trademark images. Most of the corporate Web sites are characterized by their logos and trade mark. Our system first retrieves the images of a keyword query, clusters the results based on extracted image features and then the most important logos and trademarks are finally selected to form the image-based summary of a Web site that is inferred to be the most relevant to the search query.

The proposed system is used to support fast and accurate responses to queries addressing text and images in Web pages by incorporating state of-the-art text and Web link information indexing and retrieval methods in conjunction with efficient ranking of Web pages and images by importance.

***Keywords*— Extraction, image processing, histogram, features, ranking, summarization**

I. INTRODUCTION

Image based Website Summarization is the process of extracting the most important characteristic image from it. It is used for improving the quality of website summaries and a tool for effective web browsing and retrieval.

Current image search engines on the web rely purely on the keywords around the images and the filenames, which produces a lot of unwanted output in the search results. Web-based image search engines do not cater to the actual content of images due to which the result of user queries is often cluttered with irrelevant data. This leads to finding out more effective search methods for retrieving information from the web with less amount of irrelevant data in the output of user searches. To support fast and accurate responses to user queries which can focus more on the image content, Image based summarization can be used. The goal of this system is to produce summaries that are as good as human authored summaries. This leads to better understanding of the contents of a Website without first browsing through its content. Image summarization requires image retrieval on the web accompanied by image annotations and image content. In this paper we present a system in which a user need only provide a keyword query, as is the case with standard web-based image search engines. Our system then extracts the most characteristic logo and trademark of a Website as a case study for the evaluation of the proposed method. As shown in Fig.1, images are extracted according to the user query keyword. These all images are then processed for removal of non-logo images. Logo images are further processed to generate image summary for the user.

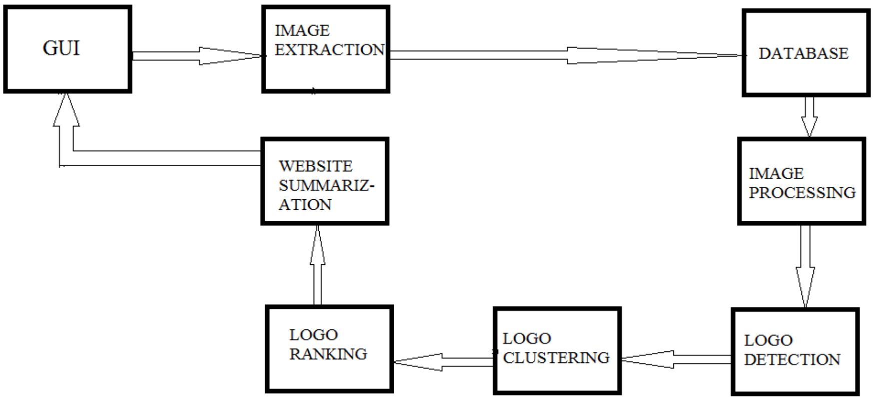
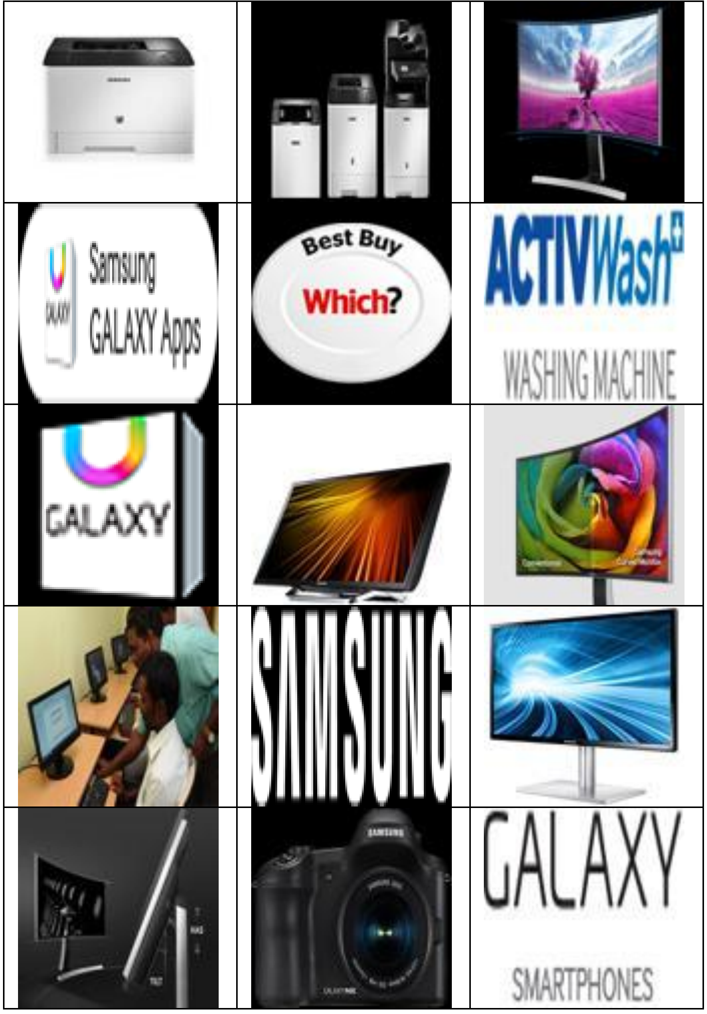


Fig. 1 System Architecture

II. DYNAMIC IMAGE EXTRACTION

Images from any particular website are retrieved from the web. Images are described by text surrounding them in web pages. Image filename, page title and image caption features are extracted along with the image. Image filename includes URL entry in the src field of the img formatting instruction. Page title is the title of the web page in which the image is displayed. Image caption is the sentence which describes the image which usually follows or precedes the image when it is displayed on the browser. The url of any particular website is taken as input from the graphical user interface and all images from that website are downloaded for further processing. As a case study, images are being extracted from website www.samsung.com/in/home as shown in Fig.2.



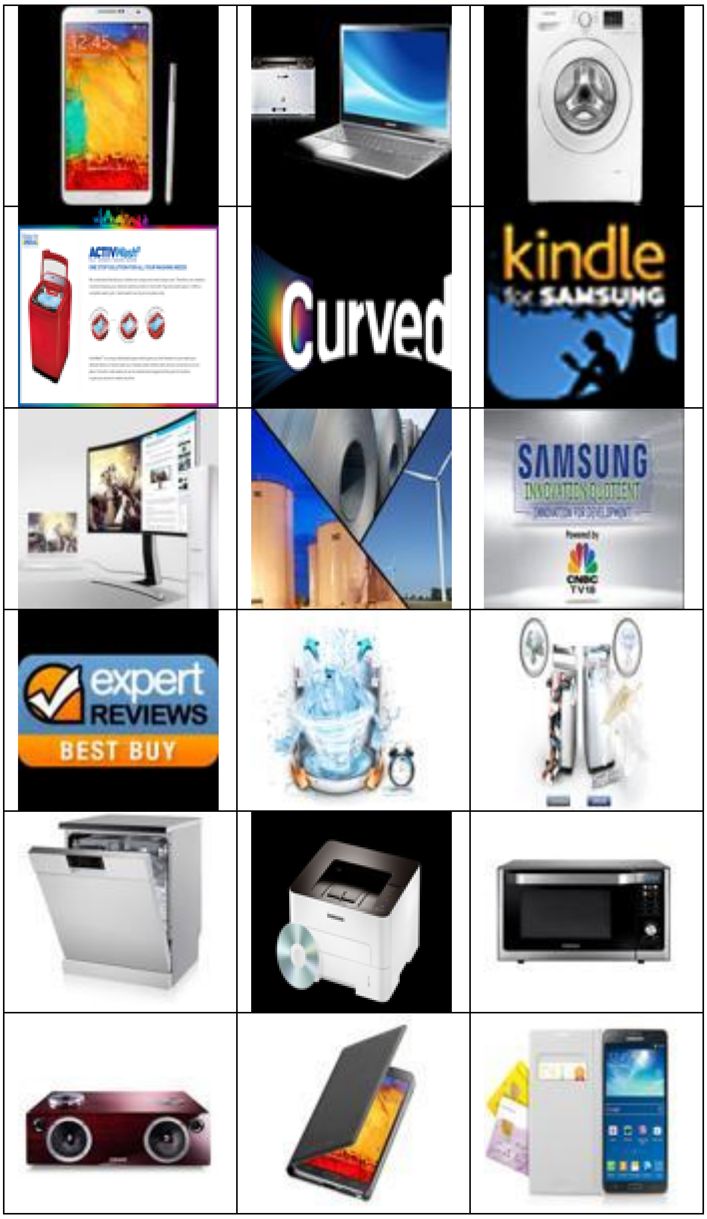


Fig. 2 Image Database

III. IMAGE FEATURE EXTRACTION

Extracted images are then further processed by extracting their features for separation of logo and non-logo images. This information is mostly captured by grey-level intensity information .For this reason, all images are converted to grey scale. Image information is captured by intensity histogram. Considering the grey level histogram [z_i=0, 1,2,…..,L-1], z_i is a random variable indicating intensity ,p(z_i) is histogram of intensity levels in region, L is the number of possible intensity levels and m is the mean average intensity ,as in [2]. These features are described below:-

Mean (m) = $\sum_{i=0}^{L-1} zi\, p(zi)$

Standard deviation=$\sqrt{\mu_z(z)} = \sqrt{\sigma^2}$

Variance$(\sigma) = \frac{1}{mn-1}\sum_{x=0}^{m-1} \sum_{y=0}^{n-1}(\mathrm{I}(x,y) - \mathrm{Mean})^2$

Where, Mean= $\frac{1}{mn-1}\sum_{x=0}^{m-1}\left(a_n \cos\frac{n\pi x}{L} + b_n \sin\frac{n\pi x}{L}\right)$

Entropy= $\sum_{i=0}^{L-1} \mathrm{p(zi)} \log_z z_i$.

IV. LOGO AND TRADEMARK DETECTION

For each image, an estimate of its probability of being logo or trademark is computed. Based on this probability logos and non-logos are separated and further processing is carried out on logo images. This stage will be totally based on machine learning as it will detect the differences in images based on the properties defined above.

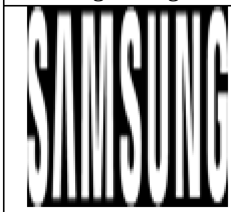
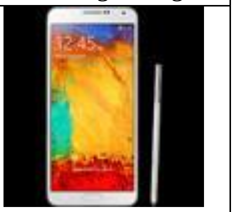
Logo Image	Non-Logo Image
	

Fig. 3 Logo and Non-Logo image

A. Duplicate Logo and Trade Mark Detections

Because the same logo or trademark can appear in the website multiple times in different forms it is important to find the logo or trademark with most originality. Clusters will be formed depending on the image and the most important image from the cluster will be selected.

B. Logo and Trade Mark Ranking

In this stage the most important image from the website will be selected based on the following criteria:

1) *Probability: The* higher the probability of being logo or trademark, the more important the image is. It corresponds to the classification accuracy of the decision tree measured for the image.

2) *Instance:* The more the instances of an image in the Web site hierarchy, the more important it is. It takes values in by normalizing by the total number of logo-trademark images in the Web site.

3) *Depth:* The higher an image is in the Web site hierarchy, the more important it is.

The following formula combines the above idea and computes the importance of an image as:-

Image Importance = Probability ・Depth ・Instances

Based on the importance scores, one of these images will be selected for summary.

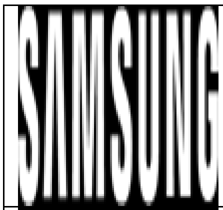
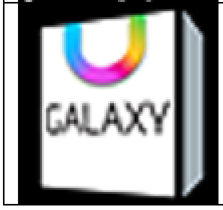
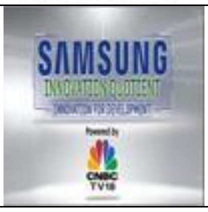
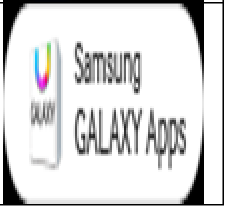
		
		

Fig. 4 Sample Logo Database

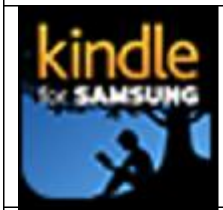
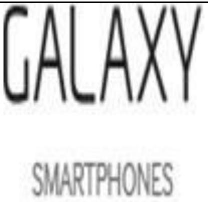
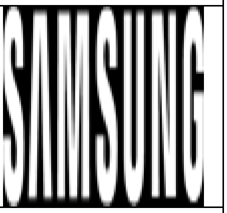
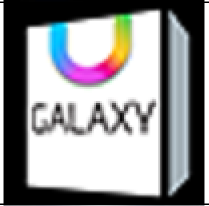
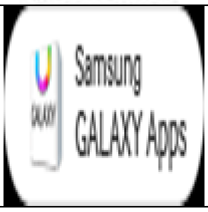
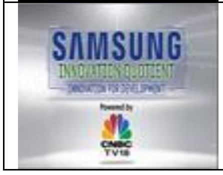
Cluster1	Cluster2	Cluster3
		
		
		

Fig. 5 Clustering of Logo Images

V. IMAGE BASED SUMMARY GENERATION

The images within a cluster represent the same logo or trademark and a cluster may contain many images so only the most characteristic image from each cluster is represented in the summary. However, the number of clusters can be very large, and it becomes meaningful to rank the clusters themselves by importance so that only the clusters ranked higher are represented in the summary. This number of clusters is user defined. The importance of a cluster depends on the importance of the image and is computed as:

Cluster Importance=∑Image Importance image i€ cluster

The more the images in a cluster and the more important these images are the more important is the cluster.

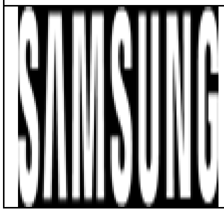
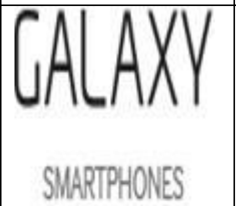
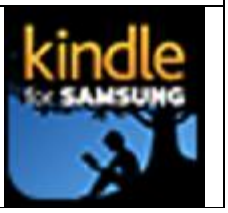
Rank1	Rank2	Rank3
		

Fig. 5 Ranked Clusters with most important Logo Image

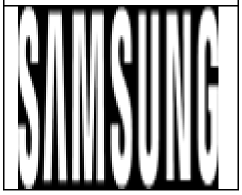
Rank1


Fig. 7 Image Summary

VI. RESULTS

Summarizing the above method, its working is as follows.

- Based on user query,particular webpages are crawled and images are extracted dynamically.
- All logo and trademark images are detected based on their features.
- Similar logo images are grouped into clusters and from each cluster one image is selected to represent the cluster in the summary.
- Clusters which have a higher ranking are represented in the summary to the user.
- Image along with text information is presented to the user in the form of website summary.

VII. CONCLUSIONS

The above Image based summarization model is a case study which can be used to provide more accurate user query search results for more effective web browsing. As a case study, the above model works on logo and trademarks of websites to provide appropriate summaries to the users.

ACKNOWLEDGMENT

The authors would like to thank all the earlier work regarding content based and image based retrieval techniques that contribute the work made in this paper.

REFERENCES

[1] Evdoxios Baratis ,Euripides G.M .Petrakis, and Evangelos Milios, "Automatic Website Summarization based on Image content: A case study on Logo and Trade Image," IEEE ,ISSN 1041-4347 Vol:20,Issue:9(2008)

[2] U.Akilandeswari ,R. Nithya and B.Santhi ,Sastra University Tamil Nadu ,"Review on Feature Extraction Methods in Pattern Classification ," European Journal of Scientific Research ,ISSN 1450—16 Vol.71 No.2(2012)

[3] Dr.Murugappan ,Abirami S ,Mizpha Poorana Selvi S ,"Automatic Web Image Categorization by Image Content:A case study with Web Document Images ," IJSCE Vol. 02, No. 02 (2010)

[4] Epimenides Voutsakis , Euripides G.M. Petrakis ,and Evangelos Milios, "IntelliSearch :Intelligent Search for Images and Text on the Web," ICIAR'06 Proceedings of the Third international conference on Image Analysis and Recognition - Volume Part I (2006)

[5] Raviraj Kasture and Dr.A.M .Dixit,"Internet Image Search based on User Intention, "IJARCSMS, Volume 2, Issue 6, June 2014.

[6] Ms.Sayali.S.Pawar and Prof.R.S.Chaure ,"A New Trend Content –Based Image Retrieval Technique used in Real time application,"IJARCSSE , Volume 4, Issue 6, June 2014