Available online at www.sciencedirect.com

ScienceDirect

AASRI Procedia 9 (2014) 31 – 36

AASRI Procedia

www.elsevier.com/locate/procedia

2014 AASRI Conference on Circuits and Signal Processing (CSP 2014)

# A Novel Imaging Approach of Web Documents Based on Semantic Inclusion of Textual and Non – Textual Information

Martina Zachariasova*, Patrik Kamencay, Robert Hudec, Miroslav Benco, Slavomir Matuska

*Department of Telecommunications and Multimedia, University of Zilina, Zilina, Slovakia*

**Abstract**

This paper deals with research in the area of a novel imaging approach of web documents based on semantic inclusion of textual and non-textual informations. The main idea was to create a robust method for relevant display results into search engine based on search by keywords or images. Thus, we proposed method called Semantic Inclusion of Images and Textual (SIIT) segments. The output SIIT method is short web document. It contains image and textual segments, which are semantic linked. Creation of short web document to possible three steps was divided. Firstly, the all images and textual segments from main content web document were extracted. Secondly, extraction images were analyzed in order to obtain of semantic description objects into image. Finally, linked images and textual segments using linguistic analysis.


Peer-review under responsibility of Scientific Committee of American Applied Science Research Institute

*Keywords:* digital images; image classification; support vector machine; descriptors

## 1. Introduction

In the last year's, there are several mechanisms for semantic inclusion of web objects using text analysis. Sh. Behnami [1] described design of Filimage system, which is intended for the images automatic extraction

* Corresponding author. Tel.: +421 41 513 2239.
*E-mail address: martina.zachariasova@fel.uniza.sk.*


Peer-review under responsibility of Scientific Committee of American Applied Science Research Institute
doi:10.1016/j.aasri.2014.09.007

and their textual comments available on the web. Mulendra Parag Joshi and Sam Liu in [2] described a technique for extracting text and images from a web document using the Document Object Model analysis and natural language processing. In the first phase, the article body from a Web document is extracted. Then, the semantic similarity based on NLP (Natural Language Processing) is used to find relevant images corresponding to the article. J. Pasternack and D. Roth [3] introduce maximum subsequence segmentation, method of global optimization over token-level local classifiers, and applied it to the domain of news websites. L. P. Florence [4] described an image and text mining tool named TNT. This tool is based on Contextual Exploration and work on different points of view. This tool offers a reorganization of the text guided by the images and annotated segments that are associated.

## 2. Proposed Method

Recently, the university research in the area of processing web pages for creation short web documents focuses on the analysis of text segments around images. Research in image processing is currently at a high level, would have been wrong not to use this fact to creation short web document. Our proposed method is based on existing systems, which are described in previous section. In this system we put the emphasis on the minimization or elimination of potential deficiencies (see Tab. 1). The proposed method allows an automatic processing of various components of web pages. The system architecture called Semantic Inclusion of Image using Textual segments is shown in Fig. 1. Automatic extraction takes place along two axes, one oriented towards the text and the other oriented image.

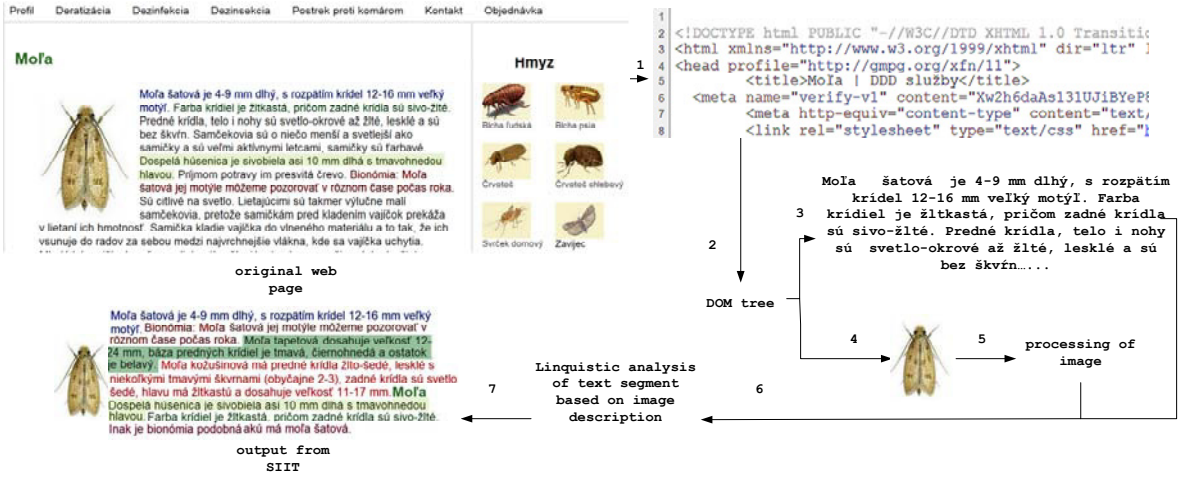


Fig.1. Sample web page processing of our proposed method

The important steps for automatic extraction:

- Loading web page
- Reading source code of web page and creating Document Object Model tree
- Identification and extracting of textual segments from the main content of web page
- Extracting images from around textual segments
- Processing of images using extraction features and its classification
- The textual segments are coming from semantic analysis according to semantic description
- A connection between the two modules allows automatic matching and associating operation