

1 Preliminary study of Image type recognizing in web pages

The web is full of bad programmed documents that don't follow the defined standards, thus leads to reduced accessibility. The web pages are often limited to the use of a specific browser type and very few are adapted to be used by blind persons. The wrong use of images can lead to reduced accessibility of webpages, as for example mathematical formulas. They should be represented in the format defined by the W3C standard, and not in an image, which is an frequently seen solution.

1.1 Objectives of this work

The goal of this project is to create a crawler and a classifier that downloads the images in web pages and tries to classify the content of each image into different categories. The two main categories are photography (natural picture) or drawing (computer made illustration). The drawing category consists of the subcategories text and no-text, where the text category contains formula and pure text. No-text category consists of button, design element, logo, graph, and other. The whole categorization is indicated in the following diagram.

Button and *logo* is self-evident, but the *design element* requires a bit more explanation. *Design element* is typically an object on a page for only estetic purposes. It is a visual element as e.g. big single colored rectangle in the edge of a frame.

Other will contain the rest that don't fall into any of the other categories. That may be normal computer made drawings.

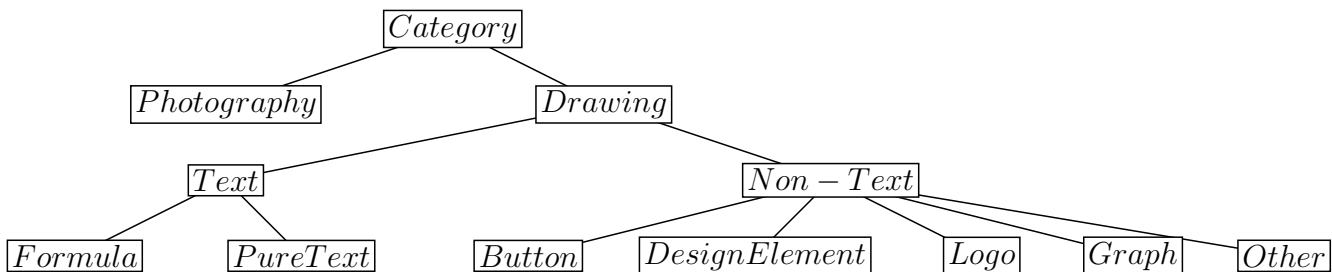


Figure 1: Categorizing tree indicating which categories the images will be categorized in.

1.2 Work plan

The work plan consists of several milestones. The time and amount of problems will decide how much we can complete, but a minimum work plan with milestones is indicated in the following list.

- make the crawler
- categorize a picture into the two first classes, namely photography or drawing
- categorize a drawing into text and non-text

If the time allows it, the rest of the categorization will be completed. First prioritized is to categorize into buttons, design elements and other. Second, the formulas and normal text is categorized. The possibility of how to do it will be researched and a conclusion will be written. Desirable, but way out of this work's scope, is text recognition in such a way that the text in the image automatically can be translated into normal text.

1.3 Planned report contents

The final report will start with an executive summary and an introduction. The introduction contains a description of detection of image usage that reduces the accessibility of a web page. The focus is on use of images to present mathematical formulas or normal text, which is a good example of use of images that reduces the accessibility of a web page. The main content is the description of the categorizing approaches with its tested algorithms, both good working and failures. The final prototype is tested on random web pages, where the image content is checked manually and compared with the results to the prototype. This and its accuracy is presented and commented in the conclusion.