Web Page Segmentation Evaluation

Explaining Results

Evaluation Overview

The evaluation is based on the geometric comparison of to blocks. Segmentations algoritms are compared to a ground truth predefined. One pair of blocks can have two possibles states:

- Equal
- Contained

The evaluation algorithm construct a bipartite graph, where the first set of nodes represents those blocks of the ground (G) thuth while the other set represent the blocks found by a segmentation algorithm (S).

Algorithm consideration:

- if the coordinates of a block in G are similar to those of a block in S, we say there is a one-to-one match, and viceversa.
- If the coordinates of a block in G resides into the area of a block in S, we say there is a containment, and viceversa.

Evaluation vocabulary

The terms used to understand the evaluation are the following:

- **Correct**, if between two nodes (one in G and other in S) exists a one-to-one match.
- **Oversegmentation**, if one node of G has several blocks of S contained.
- **Undersegmentation**, if one node of S has several blocks of G contained.
- **Missed**, if a block in G has no relationship with any block in S.
- **False alarm**, if a block in S has no relationship with any block in G.

Experiment Overview

We have chosen 400 pages from 16 categories taken from dmoz.org Open Directory. The dmoz directory consists in several levels and subleves of categories and results. One subcategory page can have subcategory list, result list or both. The mechanism taken to select pages was the following:

1 For each category:
  1.1 If there is a result list
    1.1.1 take the first hyperlink found
  1.2 If there are subcategories
    1.2.1 explore each subcategory sorted by the amount of subcategories it has and return to step 1
  1.3 stop if 25 results are taken

Then for for each page the following steps have been done:

1. ground truth segmentation
2. Segmentation using BOM
3. Segmentation using VIPS

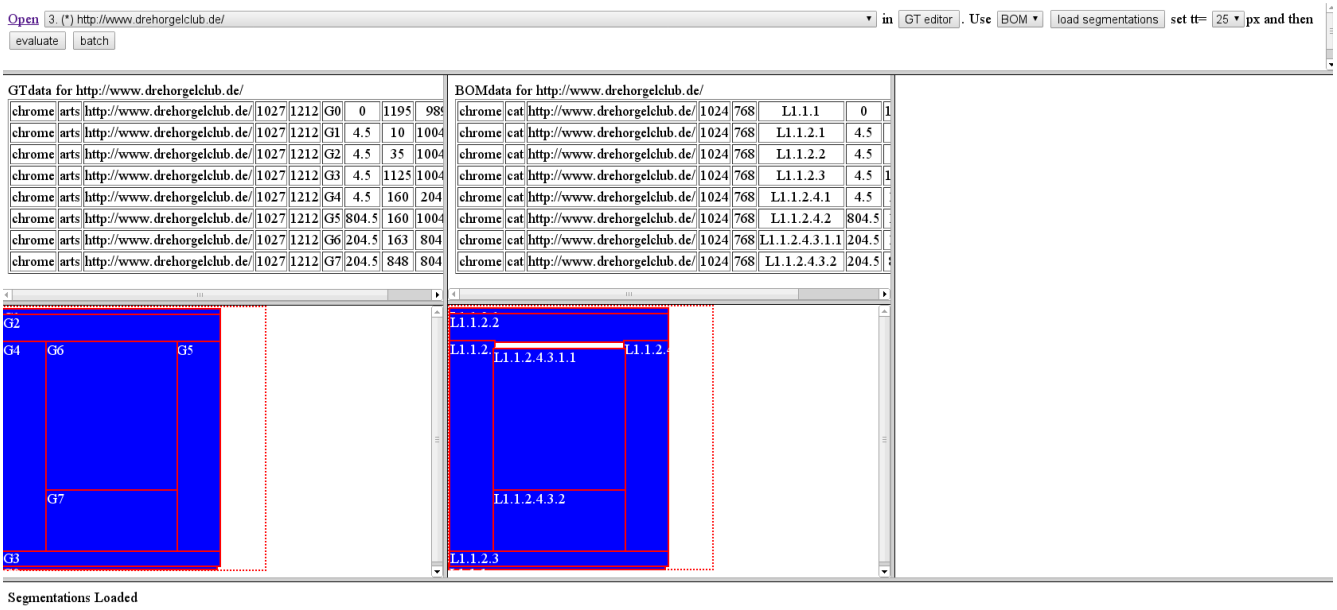
For each block found we register:

Browser	Category	Url	Doc width	Doc height	Block x	Block y	Block w	Block h

For example:

GT	chrome	Arts	http://www.drehorgelclub.de	1009	1632	G0	24.5	0	984.5	152
BOM	chrome	Arts	http://www.drehorgelclub.de	1013	1659	L1.1.1	24.5	0	984.5	152
VIPS	Iexplorer	Arts	http://www.drehorgelclub.de	960	1571	VB1.1	98	0	1058	152

These segmentations have been compared independently to the Ground Truth. In the following figure we can see a screenshot taken form the web page http://www.drehorgelclub.de/ using the BOM algorithm. In the same figure we chose the geometric threshold (tt=25px).



The same procedure has been used to VIPS algorithm:

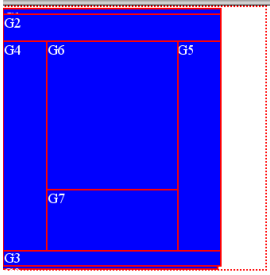
Open 3. (*) http://www.drehorgelclub.de/ in GT editor . Use VIPS load segmentations set tt= 25 px and then
evaluate batch

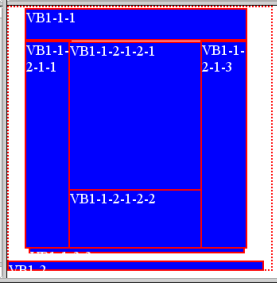
GTdata for http://www.drehorgelclub.de/

chrome	arts	http://www.drehorgelclub.de/	1027	1212	G0	0	1195	989
chrome	arts	http://www.drehorgelclub.de/	1027	1212	G1	4.5	10	1004
chrome	arts	http://www.drehorgelclub.de/	1027	1212	G2	4.5	35	1004
chrome	arts	http://www.drehorgelclub.de/	1027	1212	G3	4.5	1125	1004
chrome	arts	http://www.drehorgelclub.de/	1027	1212	G4	4.5	160	204
chrome	arts	http://www.drehorgelclub.de/	1027	1212	G5	804.5	160	1004
chrome	arts	http://www.drehorgelclub.de/	1027	1212	G6	204.5	163	804
chrome	arts	http://www.drehorgelclub.de/	1027	1212	G7	204.5	848	804

VIPSdata for http://www.drehorgelclub.de/

iexplorer	cat	http://www.drehorgelclub.de/	1156	1194	VB1-1-1	78	10	1(
iexplorer	cat	http://www.drehorgelclub.de/	1156	1194	VB1-1-2-1-1	78	160	2
iexplorer	cat	http://www.drehorgelclub.de/	1156	1194	VB1-1-2-1-2-1	278	163	8
iexplorer	cat	http://www.drehorgelclub.de/	1156	1194	VB1-1-2-	278	835	8





Segmentations Loaded

After all data in the category is processed for all threshold values (0px-50px) the results can be processed in tabular form.

For each evaluation we register the following:

Algorithm	Url	Tc	To	Tu	Co	Cu	Cm	Cf	tt	Gt	St

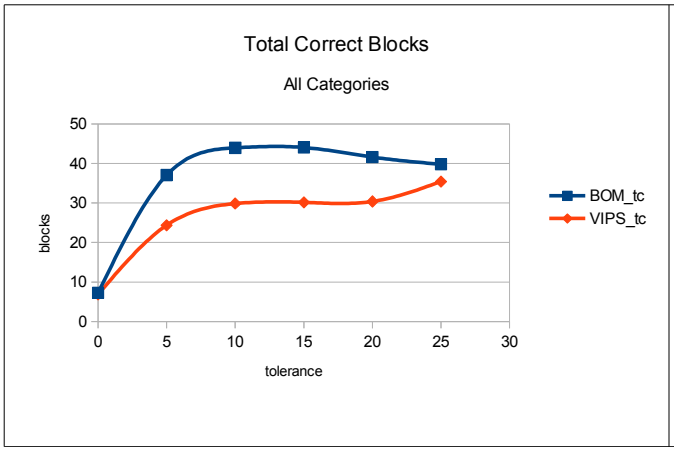
Where:

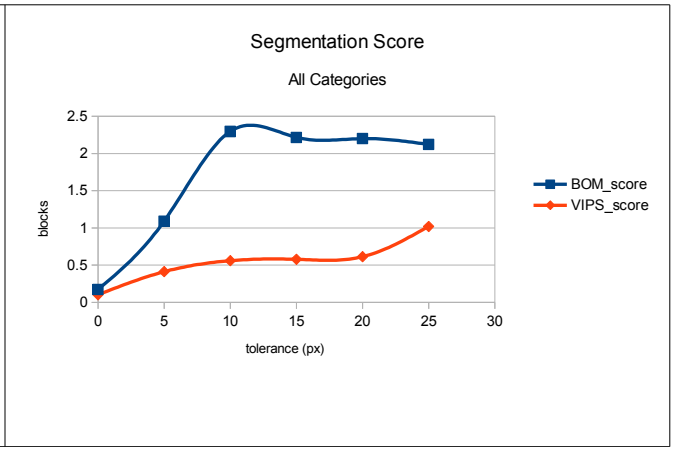
- Algorithm can be BOM or VIPS
- URL is the web page url been evaluated
- Tc = Total Correct blocks
- To = Total Oversegmented blocks (togheter both G and S)
- Tu = Total Undersegmented blocks (togheter both G and S)
- Co = Oversegmented blocks (affects only G)
- Cu = Undersegmented blocks (affects only S)
- tt, tolerance in pixels
- Gt, total blocks in Ground truth
- St, total found by algorithm

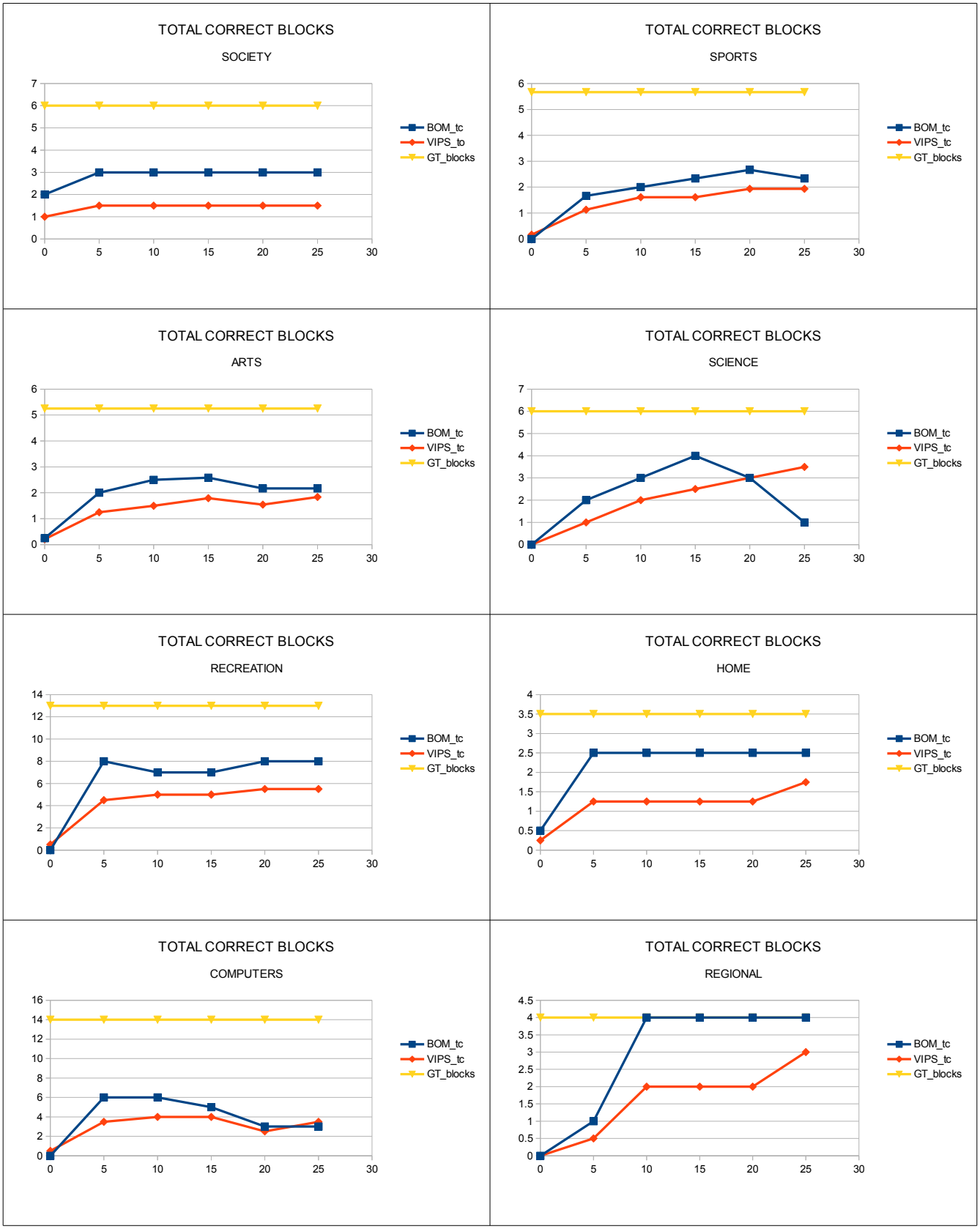
For example,

Alg.	URL	Tc	To	Tu	Co	Cu	Cm	Cf	Tt	Gt	St
BOM	http://www.drehorgelclub.de/	6	2	2	2	2	0	0	25	8	8
VIPS	http://www.drehorgelclub.de/	5	0	0	0	0	2	3	25	8	7

In the following figures we can see an example for the category society:






TOTAL CORRECT BLOCKS
SOCIETY
BOM_tc
VIPS_to
GT_blocks
TOTAL CORRECT BLOCKS
SPORTS
BOM_tc
VIPS_tc
GT_blocks
TOTAL CORRECT BLOCKS
ARTS
BOM_tc
VIPS_tc
GT_blocks
TOTAL CORRECT BLOCKS
SCIENCE
BOM_tc
VIPS_tc
GT_blocks
TOTAL CORRECT BLOCKS
RECREATION
BOM_tc
VIPS_tc
GT_blocks
TOTAL CORRECT BLOCKS
HOME
BOM_tc
VIPS_tc
GT_blocks
TOTAL CORRECT BLOCKS
COMPUTERS
BOM_tc
VIPS_tc
GT_blocks
TOTAL CORRECT BLOCKS
REGIONAL
BOM_tc
VIPS_tc
GT_blocks


TOTAL CORRECT BLOCKS
NEWS
BOM_tc
VIPS_tc
GT_blocks
TOTAL CORRECT BLOCKS
HEALTH
BOM_tc
VIPS_tc
GT_blocks
TOTAL CORRECT BLOCKS
BUSINESS
BOM_tc
VIPS_tc
GT_blocks
TOTAL CORRECT BLOCKS
WORLD
BOM_tc
VIPS_tc
GT_blocks
TOTAL CORRECT BLOCKS
SHOPPING
BOM_tc
VIPS_tc
GT_blocks
TOTAL CORRECT BLOCKS
REFERENCE
BOM_tc
VIPS_tc
GT_blocks
TOTAL CORRECT BLOCKS
KIDS_AND_TEENS
BOM_tc
VIPS_tc
GT_blocks
TOTAL CORRECT BLOCKS
GAMES
BOM_tc
VIPS_tc
GT_blocks