### Large-scale Analysis and Comparison of Web Page Segmentation Approaches Defense of Master's Thesis Lars Meyer January 28th, 2020

🔃 Watertown 🗊 Ogdensburg 🕼 Massena 🗊 Lowville 🔃 Carthage 📾 Malone 🔞 Oswego <table-cell> NNY Business 🗋 NNY Living 🖏 NNY Ads 🛛 Fri., Sep. 22 🛄 🔂 MURETISE SUBSCRUBE SIGN IN 🗸

#### Watertown Daily Times

nunities of Jefferson, St. Lawrence and Lewis co In print daily. Online always.

New York

A LOCAL NATIONAL SPORTS OBITUARIES	EATURES OPINION PAGE 2 CALENDAR PHOTOS & VIDEO ARCHIVES CONTACT & HELP NEWSLETTERS	CLASSIFIEDS JOBS AUTO HOMES
Image: Introduction of the state       Description       Belated Stories       OGD NEWS       Image: Introduction of the state       Image: I	<section-header><section-header><section-header><section-header><section-header><section-header></section-header></section-header></section-header></section-header></section-header></section-header>	Connect with Us         Image: Search
	1. Since to the population of the anticoliteter/relation commenters, do not engage in personal attacks or name-calling.     2. Comments that include profanity/obscenties or are libelous in nature will be removed without warning.     Violators' commenting privileges may be revoked indefinitely. By commenting you agree to our full Terms of Use.	
	V Comments     Watercown Dany Times     Login ~     V Recommend     Pt Share     Sort hv Rect ~	
	Start the discussion	
	LOG IN WITH OR SIGN UP WITH DISQUS (?	
	Ivenie	
	Be the first to comment.	

Subscribe D Add Disgus to your site D Privacy

DISQUS

🔢 Watertown 🗊 Ogdensburg 🕼 Massena 🐑 Lowville 🔃 Carthage 🛲 Malone 😻 Oswego 💽 NNY Business 📃 NNY Living 🗒 NNY Ads 🛛 Fri\_, Sep. 22 🖸 ADVERTISE BUBBCRIBE SIGN IN 🗸 Watertown Daily Times Q Search... "Bitten by Magic" by Kelly Ashley published today **Related Stories** GD NEWS North Country Library System launches N 🖨 🖂 🕑 Tweet mobile app Buffalo Soldiers to be Subject of Remington Lecture tten by Magic," the second book in Kelly M. hley's Sisters of Fate series, is out today. St. Lawrence County has second-highest August unemployment rate in the state Ogdensburg man pleads guilty to meth follow up to her fantasy romance novel, "The ur Flies," which was released in May, Mrs. Community Gallery Project on display at hley's new book centers on the passion of exa, a witch, and Wilhelm, a vampire. "Bitten by Magic" Alexa is required to marry lhelm in order to settle a dispute. 'It's a paranormal romance," Mrs. Ashley, who ites under the pen name Kelliea Ashley, said. "If anybody likes the witches and the vampires **Connect with Us** nd the werewolves that's what it is." 🔊 🚹 ⊵ 🗊 ut Mrs. Ashley cautioned, "It's not 'Twilight," referring to the popular book and movie series. My vampires don't sparkle." rn in Gouverneur, Mrs. Ashley currently lives in Ogdensburg and said she draws inspiration om the north country as settings for her novels. "All of my books have a north country feel," she said. "Even with the vampires and werewolves." Wrs. Ashley said she originally intended to write the Sisters of Fate series as a trilogy, but was pushed by her editor at Breathless Press to explore the characters further and extend it. t's open ended," she said, so readers can expect more tales of forbidden love between witches mpires and werewolves. Irs. Ashley said she has been writing since she was 13 but it was her husband, Joshua, who nally convinced her to find a publisher for her novels in 2012. r latest book is available as an e-book at breathlesspress.com, amazon.com, romanceebooks.com, Apple I-Books and bookstrand.com. as long as [people] read them and get away from their everyday life, that's what I'm shooting," Irs. Ashley said. "I'm not looking to get rich; I'm just looking to put my work out there." nmenting rules: Stick to the topic of the article/letter/editorial.
 When responding to issues raised by other commenters, do not engage in personal attacks or name-calling.
 Comments that include profanity/obscrities or are libelous in nature will be removed without warning. tors' commenting privileges may be revoked indefinitely. By commenting you agree to our full Terms of Use 0 Comments Watertown Daily Times 🚺 Login Recommend 🔀 Share Sort by Best Start the discussion .. LOG IN WITH OR SIGN UP WITH DISQUS (?) Name

DISQUS

Be the first to comment.

🖂 Subscribe 🛛 🜔 Add Disqus to your site 🛛 🔒 Privacy

### What is a segment?

"A segment is a part of a web page containing the elements that belong together...

... visually, semantically, and in purpose."

#### accessibility enhancements

• enhanced screen readers

e.g.: Michael Cormier et al., "Towards an improved vision-based web page segmentation algorithm", 2017

#### • adaptation to small screens

e.g.: Shumeet Baluja, "Browsing on small screens: recasting web-page segmentation into an efficient machine learning framework", 2006

#### • ...

### information retrieval

#### content summarization

e.g.: Chitra Pasupathi et al., "Web document segmentation using frequent term sets for summarization", 2012

#### page classification/ranking

e.g.: Lidong Bing et al., "Web page segmentation with structured prediction and its application in web page classification", 2012

### • similarity assessment

e.g.: Marc Teva Law et al., "Structural and visual similarity learning for web page archiving", 2012

- accessibility enhancements
  - enhanced screen readers

e.g.: Michael Cormier et al., "Towards an improved vision-based web page segmentation algorithm", 2017

#### • adaptation to small screens

e.g.: Shumeet Baluja, "Browsing on small screens: recasting web-page segmentation into an efficient machine learning framework", 2006

• ...

#### information retrieval

#### content summarization

e.g.: Chitra Pasupathi et al., "Web document segmentation using frequent term sets for summarization", 2012

### page classification/ranking

e.g.: Lidong Bing et al., "Web page segmentation with structured prediction and its application in web page classification", 2012

### • similarity assessment

e.g.: Marc Teva Law et al., "Structural and visual similarity learning for web page archiving", 2012

- accessibility enhancements
  - enhanced screen readers

e.g.: Michael Cormier et al., "Towards an improved vision-based web page segmentation algorithm", 2017

#### adaptation to small screens

e.g.: Shumeet Baluja, "Browsing on small screens: recasting web-page segmentation into an efficient machine learning framework", 2006

• ...

#### information retrieval

#### content summarization

e.g.: Chitra Pasupathi et al., "Web document segmentation using frequent term sets for summarization", 2012

### page classification/ranking

e.g.: Lidong Bing et al., "Web page segmentation with structured prediction and its application in web page classification", 2012

### • similarity assessment

e.g.: Marc Teva Law et al., "Structural and visual similarity learning for web page archiving", 2012

- accessibility enhancements
  - enhanced screen readers

e.g.: Michael Cormier et al., "Towards an improved vision-based web page segmentation algorithm", 2017

#### adaptation to small screens

e.g.: Shumeet Baluja, "Browsing on small screens: recasting web-page segmentation into an efficient machine learning framework", 2006

• ...

#### information retrieval

#### content summarization

e.g.: Chitra Pasupathi et al., "Web document segmentation using frequent term sets for summarization", 2012

#### • page classification/ranking

e.g.: Lidong Bing et al., "Web page segmentation with structured prediction and its application in web page classification", 2012

#### • similarity assessment

e.g.: Marc Teva Law et al., "Structural and visual similarity learning for web page archiving", 2012

- accessibility enhancements
  - enhanced screen readers

e.g.: Michael Cormier et al., "Towards an improved vision-based web page segmentation algorithm", 2017

#### adaptation to small screens

e.g.: Shumeet Baluja, "Browsing on small screens: recasting web-page segmentation into an efficient machine learning framework", 2006

• ...

#### information retrieval

#### content summarization

e.g.: Chitra Pasupathi et al., "Web document segmentation using frequent term sets for summarization", 2012

#### • page classification/ranking

e.g.: Lidong Bing et al., "Web page segmentation with structured prediction and its application in web page classification", 2012

#### • similarity assessment

e.g.: Marc Teva Law et al., "Structural and visual similarity learning for web page archiving", 2012

- accessibility enhancements
  - enhanced screen readers

e.g.: Michael Cormier et al., "Towards an improved vision-based web page segmentation algorithm", 2017

#### adaptation to small screens

e.g.: Shumeet Baluja, "Browsing on small screens: recasting web-page segmentation into an efficient machine learning framework", 2006

• ...

#### information retrieval

#### content summarization

e.g.: Chitra Pasupathi et al., "Web document segmentation using frequent term sets for summarization", 2012

#### • page classification/ranking

e.g.: Lidong Bing et al., "Web page segmentation with structured prediction and its application in web page classification", 2012

#### • similarity assessment

e.g.: Marc Teva Law et al., "Structural and visual similarity learning for web page archiving", 2012

# Approaches

Category	Name	Document type	Publication
DOM-only	VIPS	Web page	Cai et al., "Extracting Content Structure for Web Pages based on Visual Representation", 2003
	HEPS	Web page	Manabe et al., "Extracting Logical Hierarchical Structure of HTML Documents Based on Headings", 2015
Visual	Cormier et al.	Web page	Cormier et al., "Purely vision-based segmentation of web pages for assistive technology", 2016
	MMDetection	Photo	Chen et al., "MMDetection: Open mmlab detection toolbox and benchmark", 2019
Hybrid	Meier et al.	Newspaper page	Meier et al., "Fully convolutional neural networks for newspaper article segmentation", 2017

# Approaches

Category	Name	Document type	Publication
DOM-only	VIPS	Web page	Cai et al., "Extracting Content Structure for Web Pages based on Visual Representation", 2003
	HEPS	Web page	Manabe et al., "Extracting Logical Hierarchical Structure of HTML Documents Based on Headings", 2015
N/11			
Viewal	Cormier et al.	Web page	Cormier et al., "Purely vision-based segmentation of web pages for assistive technology", 2016
Visual	Cormier et al. MMDetection	Web page Photo	Cormier et al., "Purely vision-based segmentation of web pages for assistive technology", 2016 Chen et al., "MMDetection: Open mmlab detection toolbox and benchmark", 2019

## **Evaluation setup**

 first crowd-sourced dataset for Web Page Segmentation



 first crowd-sourced dataset for Web Page Segmentation



- first crowd-sourced dataset for Web Page Segmentation
- assembled through Amazon Mechanical Turk



- first crowd-sourced dataset for Web Page Segmentation
- assembled through Amazon Mechanical Turk
- 8490 pages, 5 annotators per page  $\rightarrow$  42450 human segmentations



- first crowd-sourced dataset for Web Page Segmentation
- assembled through Amazon Mechanical Turk



- 8490 pages, 5 annotators per page  $\rightarrow$  42450 human segmentations
- Fusion of human segmentations for page based on area agreement  $\rightarrow$  ground truth

Segments are regarded as **clusters of atomic elements**:

Segments are regarded as **clusters of atomic elements**:

• pixels

Segments are regarded as **clusters of atomic elements**:

- pixels
- *edges<sub>fine</sub>* and *edges<sub>coarse</sub>* (Canny edge detection)

Segments are regarded as **clusters of atomic elements**:

- pixels
- *edges<sub>fine</sub>* and *edges<sub>coarse</sub>* (Canny edge detection)
- *nodes* (in the DOM tree)

Segments are regarded as **clusters of atomic elements**:

- pixels
- *edges<sub>fine</sub>* and *edges<sub>coarse</sub>* (Canny edge detection)
- *nodes* (in the DOM tree)
- characters

Segments are regarded as **clusters of atomic elements**:

- pixels
- *edges<sub>fine</sub>* and *edges<sub>coarse</sub>* (Canny edge detection)
- *nodes* (in the DOM tree)
- characters

 $\longrightarrow$  Precision ( $P_{B^3}$ ), Recall ( $R_{B^3}$ ) and F-score ( $F_{B^3}$ ) can be calculated between two segmentations

Segments are regarded as **clusters of atomic elements**:

- pixels
- *edges<sub>fine</sub>* and *edges<sub>coarse</sub>* (Canny edge detection)
- *nodes* (in the DOM tree)
- characters
- $\longrightarrow$  Precision ( $P_{B^3}$ ), Recall ( $R_{B^3}$ ) and F-score ( $F_{B^3}$ ) can be calculated between two segmentations

→ different atomic elements cover variety of algorithm performance aspects Kiesel et al., "Web Page Segmentation from First Principles", 2020

### Terms

**Precision**: how many of the elements in an **algorithm segment** also belong to *one* segment in the ground truth?

**Recall**: how many of the elements in a **ground truth segment** are grouped *together* in *one* algorithm segment?

F-score: harmonic mean of precision and recall

• Webis Web Segments 2020 based on Webis Web Archive 17

- Webis Web Segments 2020 based on Webis Web Archive 17
- high level of completeness and reproduction accuracy within *Webis Web Archiver*

- Webis Web Segments 2020 based on Webis Web Archive 17
- high level of completeness and reproduction accuracy within *Webis Web Archiver*
- archived pages can be reproduced in automated Chromium browser

- Webis Web Segments 2020 based on Webis Web Archive 17
- high level of completeness and reproduction accuracy within
   Webis Web Archiver
- archived pages can be reproduced in automated Chromium browser
- any segmentation algorithm written in JavaScript can be run on the archived web page  $\rightarrow$  highly consistent input

- Webis Web Segments 2020 based on Webis Web Archive 17
- high level of completeness and reproduction accuracy within
   Webis Web Archiver
- archived pages can be reproduced in automated Chromium browser
- any segmentation algorithm written in JavaScript can be run on the archived web page  $\rightarrow$  highly consistent input
- → Contribution: TypeScript/JavaScript port of VIPS



04-<< 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 >>06-

本商會簡介與規章

スポンサーサイト

歌迎新老手加入~此商會是純手動玩家的聖地,因此本商會嚴禁外掛。也嚴禁不做任何努力而四處伸手要東西的行為、對話中總是用注音符號、打劫
」
上記の広告は1ヶ月以上更新のないプログに表示されています。
新しい記事を書く事で広告が預せます。

商會的設施簡介

スポンサー広告 有酒館與銀行及擺攤功能的商館(熱那亞6號會館)、本商會的寶石與貴金屬港(13117,2246)、需格外登錄的建設中分會港(13053,2382)、商會東引 附註:以上任何設施都可自由使用,若是想協助建設開港請與赤斬聯絡討論。

	請追逐自己的夢想記阿茲特克劇情有感
大航海時代相關聯結	2013-05-03
Database Seasons(船隻資料4 巴哈姆特 -攻略百科:索引 博文_夕阳醉_(船隻強化參考)	<sup>參考)</sup> 在漫長的五一假期中,終於可以悠閒的欣賞一下自己期待多時的
管理頁面	這次介紹的是阿茲特克劇情
*可直接打關鍵字搜索文章	
	條件是發現黃金之城"特諾奇帝特闌"
	(単位克魯斯)時代 作(225)700 D ● 参グ回顧
文章類別	
商會公告 (5)	上記広告は1ヶ月以上更新のないプログに表示されています。新しい記事を書くことで広告を消せます。
問言沽動公告(3) 功略:教母:心得(17)	
图系 (5)	
商系 (7)	
軍系 (5)	

CSSBox (VIPS-Java)

	大航海時代-娜斯水狐		[PR] 全広告、完全オフ! 広告オフ機能がグレー 大航海時代-娜斯水狐
<b>04-&lt;&lt;</b> 1 2 <b>3</b> 4 5 6 7 8 9 10 11	1 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 >>06-	<b>04-&lt;&lt;</b> 1 2 <b>3</b> 4 5 6 7 8 9 10	11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 >>06-
本商會簡介與規章	スポンサーサイト	本商會簡介與規章	スポンサーサイト
歡迎新老手加入~此商會是純手動玩家的聖地 商會的設施簡介 有酒館與銀行及擺攤功能的商館(熱那亞6號會	b, 因此本商會嚴禁外掛。也嚴禁不做任何努力而四處伸手要東西的行為、對話中總是用注音符號、打劫」 上記の広告は1ヶ月以上更新のないプログに表示されています。 新しい記事を書く事で広告が用せます。           スポンサー広告           2           (13117,2246)、需格外登録的建設中分會港(13053,2382)、商會東3	歡迎新老手加入~此商會是純手動玩 家的聖地,因此本商會嚴禁外掛。也 嚴禁不做任何努力而四處伸手要東西 的行為、對話中總是用注音符號、打 劫自己商會的人等等行為,以上凡違 反其一來必認知商會	
附註:以上任何設施都可自由使用,若是想做 大航海時代相關聯結	a助建設開港請與赤斬聯絡討論。 請追逐自己的夢想記阿茲特克劇情有感 2013-05-03	商會的設施簡介	
Database Seasons(船隻資料參考) 巴哈姆特 -攻略百科:索引 博文_夕阳醉_(船隻強化參考) 管理頁面	在漫長的五一假期中,終於可以悠閒的欣賞一下自己期待多時的 這次介紹的是阿茲特克劇情	有酒館與銀行及擺攤功能的商館(熱那 亞6號會館)、本商會的寶石與貴金屬 港(13117,2246)、需格外登錄的建 設中分會港(13053,2382)、商會東 亞貿易系統(貿易港為淡水) 附註:以上任何設施都可自由使用。	上記の広告は1ヶ月以上更新のないプログに表示されています。
*可直接打關鍵字搜索文章	條件是發現黃金之城"特諾奇蒂特蘭"	若是想協助建設開港請與赤斬聯絡討 論。	新しい記事を書く事で広告が消せます。
· · · · · · · · · · · · · · · · · · · ·	#世立会費前単冊部出 作27257000 D ● 夢の回聞	大航海時代相關聯結	スポンサー広告
商會公告 (5) 上記広告は1ヶ月 商會活動公告 (3) 攻略·教學·心得 (17)	後1-22mmのよう。 以上更新のないプログに表示されています。新しい記事を書くことで広告を消せます。	Database Seasons(船隻資料參考) 巴哈姆特 -攻略百科:索引 博文_夕阳醉_(船隻強化參考) 管理頁面	請追逐自己的夢想記阿茲特克劇情有感 2013-05-03 在漫長的五一假期中,終於可以悠閒的欣賞一下自己期待多 時的話曲劇標了
商系 (7) 軍系 (5)	and the second se	<u>*可直接打關鍵字搜索文章</u>	<sup>此了1011111111111111111111111111111111111</sup>

#### CSSBox (VIPS-Java)



## Summary: Methodology

# Summary: Methodology



Webis Web Archive 17
# Summary: Methodology



Webis Web Archive 17

# Summary: Methodology

algorithm segmentation



ground truth

# Summary: Methodology

algorithm segmentation



#### ground truth

# Approaches, Evaluations and Results

# Overview

- Evaluation of all algorithms + single-segment baseline against the ground truth
- 2. Parameter analyses: VIPS and Cormier et al.
- 3. Visual/hybrid segmentations fit to DOM nodes
- 4. Cross-evaluation (algorithm similarity)
- 5. *Min-vote* ensemble (combining algorithm segmentations)

# DOM-only approach: VIPS



Deng Cai, Shipeng Yu, Ji-Rong Wen, Wei-Ying Ma, "Extracting Content Structure for Web Pages based on Visual Representation", 2003

# DOM-only approach: VIPS



fixed set of rules down to element level

Deng Cai, Shipeng Yu, Ji-Rong Wen, Wei-Ying Ma, "Extracting Content Structure for Web Pages based on Visual Representation", 2003

# DOM-only approach: VIPS



- fixed set of rules down to element level
- Permitted Degree of Coherence (PDoC) influences granularity

Deng Cai, Shipeng Yu, Ji-Rong Wen, Wei-Ying Ma, "Extracting Content Structure for Web Pages based on Visual Representation", 2003

			pixels			characters	
PDoC	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$
8	80.2	0.46	0.36	0.32	0.93	0.41	0.50
5	13.5	0.35	0.70	0.38	0.74	0.76	0.68
Δ	- 66.7	- 0.11	+ 0.34	+ 0.06	- 0.19	+ 0.35	+ 0.18

• oversegmentation (ground truth: 9.1 segments)

			pixels			characters	
PDoC	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$
8	80.2	0.46	0.36	0.32	0.93	0.41	0.50
5	13.5	0.35	0.70	0.38	0.74	0.76	0.68
Δ	- 66.7	- 0.11	+ 0.34	+ 0.06	- 0.19	+ 0.35	+ 0.18

- oversegmentation (ground truth: 9.1 segments)
- $\rightarrow$  high precision, low recall, low F-score

			pixels			characters	
PDoC	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$
8	80.2	0.46	0.36	0.32	0.93	0.41	0.50
5	13.5	0.35	0.70	0.38	0.74	0.76	0.68
Δ	- 66.7	- 0.11	+ 0.34	+ 0.06	- 0.19	+ 0.35	+ 0.18

### Reason:

• PDoC > 6 applies rules targeting specific element types

			pixels			characters	
PDoC	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$
8	80.2	0.46	0.36	0.32	0.93	0.41	0.50
5	13.5	0.35	0.70	0.38	0.74	0.76	0.68
Δ	- 66.7	- 0.11	+ 0.34	+ 0.06	- 0.19	+ 0.35	+ 0.18

### Reason:

- PDoC > 6 applies rules targeting specific element types
- $\rightarrow$  outdated, detrimental to segmentation quality

			pixels			characters	
PDoC	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$
8	80.2	0.46	0.36	0.32	0.93	0.41	0.50
5	13.5	0.35	0.70	0.38	0.74	0.76	0.68
Δ	- 66.7	- 0.11	+ 0.34	+ 0.06	- 0.19	+ 0.35	+ 0.18

Documentation

php

Example: HTML **<code>** element

Variable and Type Related Extensions     Reflection > Reflection FunctionAbstract	sstatt.getivulliberorkeyülledealalileters »
Change language: English Edit Report a Bug ReflectionFunctionAbstract::getNumberOfParameters	ReflectionFunctionAbstract clone getClosureScopeClass getClosureThis
(PHP 5 >= 5.3.0, PHP 7) ReflectionFunctionAbstract::getNumberOfParameters — Gets number of parameters	getDocComment getEndLine getExtension getExtensionName getFileName
Description	getName getNamespaceName » getNumberOfParameters
<pre>public int ReflectionFunctionAbstract::getNumberOfParameters ( void )</pre>	getNumberOfRequiredParameters getParameters
Get the number of parameters that a function defines, both optional and required.	getReturnType

Search

			pixels			characters	
PDoC	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$
8	80.2	0.46	0.36	0.32	0.93	0.41	0.50
5	13.5	0.35	0.70	0.38	0.74	0.76	0.68
Δ	- 66.7	- 0.11	+ 0.34	+ 0.06	- 0.19	+ 0.35	+ 0.18

Documentation

oho

Example: HTML **<code>** element

P ·· P		
PHP Manual → Function Reference → Variable and Type Related Extensions → Reflection → ReflectionFunctionAbstract	« ReflectionFunctionAbstract::getNamespaceName ReflectionFunctionA	bstract::getNumberOfRequiredParameters »
ReflectionFunctionAbstr	Change language: English Edit Report a Bug act::getNumberOfParameters	ReflectionFunctionAbstract clone getClosureScopeClass getClosureThis
(PHP 5 >= 5.3.0, PHP 7) ReflectionFunctionAbstract::getNumberOfP <b>Description</b>	arameters — Gets number of parameters	getDocComment getEndLine getExtension getExtensionName getFileName getName getName
public int ReflectionFunctionAbstra	ct::getNumberOfParameters ( void )	» getNumberOfParameters getNumberOfRequiredParameters
Get the number of parameters that a funct	ion defines, both optional and required.	getParameters getReturnType

Search

			pixels			characters	
PDoC	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$
8	80.2	0.46	0.36	0.32	0.93	0.41	0.50
5	13.5	0.35	0.70	0.38	0.74	0.76	0.68
Δ	- 66.7	- 0.11	+ 0.34	+ 0.06	- 0.19	+ 0.35	+ 0.18

#### PDoC 5:

• applied rules target only coarse page divisions

			pixels			characters	
PDoC	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$
8	80.2	0.46	0.36	0.32	0.93	0.41	0.50
5	13.5	0.35	0.70	0.38	0.74	0.76	0.68
Δ	- 66.7	- 0.11	+ 0.34	+ 0.06	- 0.19	+ 0.35	+ 0.18

PDoC 5:

• applied rules target only coarse page divisions

 $\rightarrow$  oversegmentation reduced; lower precision, but much higher recall, increased  $F_{B^3}$ 

			pixels			characters	
PDoC	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$
8	80.2	0.46	0.36	0.32	0.93	0.41	0.50
5	13.5	0.35	0.70	0.38	0.74	0.76	0.68
Δ	- 66.7	- 0.11	+ 0.34	+ 0.06	- 0.19	+ 0.35	+ 0.18

#### PDoC 5:

- applied rules target only coarse page divisions
- $\rightarrow$  oversegmentation reduced; lower precision, but much higher recall, increased  $F_{B^3}$
- $\rightarrow$  VIPS (PDoC 5) is best single algorithm

_	Carlotte .		And and a design of the local division of the local division of the local division of the local division of the		
E	<u>j</u>			Siq Start earning r	gn Up Now! noney in less then 10 minutes!
Ρı	ublishers			Full Name	
Maxi	mize your revenue with Po	pCash.l	Net	Email	
HOW	TO START?				
	Learn More about Our Advantages!			_	Create Account
	oull/	7			
Put START Maximi in less All you code oi your vis assure that's le website Are you	Dilishers EARNING EXTRA REVENUE IN LES Ize your revenue with PopCash! You car than 10 minutes. need to do is register, submit your webs in the deaired pages. Once your domain altors will be shown a popunder advertis that your viewer's experience will not b at for you to do is to focus on attracting i a. ready to start? You can <u>register here</u> !	S THAN 1 I start earni ilte and pla has been a ement ever e affected more visito	0 MINUTESI ing extra revenue ce our popunder approved, all of ry 24 hours. This by our ads. All rs towards your		
	Vantages HAGES OF PUBLISHING WITH US Fast Approval Your websile will be approved within 1 hour in	9	Fast Payments Juliy Payments: this is how we make	sure that	\$10 Minimum Withdrawal Ve proces payments through various pays
	Vantages TAGES OF PUBLISHING WITH US Fast Approval Your website with a sproved within 1 hour in our business days and up to 12h on weekends	9	Fast Payments Sally Payments: this is how we make our payments are processed within a ours.	sure that i matter of	HO Minimum Withdrawal Ne poose sayneetis through various pay alarways such a space. Favor, Payza, Netwoore, Sinth Bitcon, Wire Transfer
	Vantages ITAGES OF PUBLISHING WITH US Fast Approval Your vebsile will be approved within 1 hour in our business days and up to 12th on weekends 80% Revenue Share Exclusive advertisers with direct campaigns an high rates.		Fast Payments July Payments: this is how we make our payments are processed within a vorus. Hourly Updated Reports Vawys he up-to-data with your earns' thouly-updated statistics.	sure that matter of gs through	110 Minimum Withdrawal       Ve process payments through various pays patentias such as payout, Pauson, Payou, Alexano, Payou, Alexano, Payou, Alexano, Payou, Alexano, Payou, Sirill, Batcain, Wen Transfer.       First-Class Support Ve offer all of our users live support on e-m payoe.
	Vantages ITAGES OF PUBLISHING WITH US Fast Approval Your website will be approved within 1 hour in our business days and up to 12h on weekends 80% Revenue Share Exclusive advertisers with direct campaigns and high rates. Safe and Clean Ads All ads are continuously moderated by our learn using both internal and third-party tools.		Fast Payments July Payments: this is how we make any payments are processed within a vorum. Hourly Updated Reports Todays hours-applied statistics. Simple Implementation July and Acode and Payland Statistics Simple Implementation Are all code and assign by implement what it. We also have playing available statistics and acode and acode and acode and statistics and acode and acode and acode and statistics and acode and acode and acode and acode and acode and acode	sure that matter of gs through de lot oarry	F10 Minimum Withdrawal         Fig notes apyments through various payr plenays such as Payral, Paura, Payra, Methomory, Starti, Bioton, Vier Tansfer.           First-Class Support Ne offer all of our uses live support on e-m layre.         First-Class Support Not Support.           Soost Your Website's Earnings Dura divertisements have no interference wit urrent barner ads.         Soost Your Support.
	Vantages TRACES OF PUBLISHING WITH US Fast Approval Your vebsile will be approved within 1 hour in our business days and up to 12h on weekends BO% Revenue Share Exclusive advertisers with direct campaigns an high rates. Safe and Clean Ads All dis are continuously moderated by our learn using both internal and third-party tools. 10% Referral Program Refer your finends and sam passive income for ther activity without any limits.		Fast Payments July Payments: this is how we make any payments are processed within a vars. Hourly Updated Reports Namys In up-to-tails with your earnin ar hourly updated statistics. Simple Implementation July and code on aveilable VoodTwas and Biogarcom weblake Understand Biogarcom weblake Jinque Integration Dra da code works semethesity on all and code nones semethesity on all and code nones semethesity on all	sure that matter of gs through ed into any and into any sectors	k10 Minimum Withdrawal         No process payments through various payments through various payments such an Payol, Paxon, Payon, Nethermore, Sarit, Bloton, Wer Tarender.           First-Class Support         Nee diff all do users the support on empression.           Post of users and on users the support on empression.         Society Cour Website's Earnings on a diversion was under the no interference with unrent barrier ads.           Society Court Vebsite's Experience         Mile PayCash was established in 2012, the am has been involved in the online adventainees for more than a decade.
	Vantages ITAGES OF PUBLISHING WITH US Fast Approval Your website with be approved within 1 hour in our business days and up to 12th on weekends 20% Revenue Share Exclusive adventisers with direct campaigns an high nates. Safe and Clean Ads All ads are continuously moderated by our lear using both internal and third-party tools. 10% Referral Program Refer your finends and earn passive income for their activity without any limits.		Fast Payments July Payments: this is how we make any payments are processed within a currs. Many be up-b-date with your earnin we hourly updated Reports We hourly spatiated statistics. Single Implementation Dura at odds on the up bying a sunkil Wordhwes and Biogram cum website Unique Integration Jnique Integration at at odds works summersely on all or date be sailly integrated.	sure Pat matter of gs Brough de for de for de vices	H0 Minimum Withdrawal       Ne process payments through vertices payments through vertices payments through vertices payments and the second seco
	Vantages Intraces or PUBLISHING WITH US Fast Approval Your evaluates with a reproved within 1 hour in our business days and up to 12th on weekends Bight Action and Approximation and the second and and any		Fast Payments Day Payments the is how we make our payments are processed within a source of the second second second second days to the up-to-date with your arms/ to hourly-updated statistics. Simple implementation Data at code an easily be implement weaters. We also have julgers availed Data code an easily be implement of can be easily integrated. Register Now Contact Uss	sure that and and a sure that and and a sure that and a sure that	\$10 Minimum Withdrawal Ne process payments through various pays plannays such a Brook, With Tander, Vehnonzy, Sarth, Bickon, With Tander, Vehnonzy, Sarth, Bickon, With Tander, Part Chass Support Ne offer all of our users live support on e- highpe.       Part Class Support We offer all of our users live support on e- highpe.       Do advisor of the support Support of the support warms barrier ads.       Dar advisor of the support Support of the support warms barrier ads.       Dar advisor of the support of the support warms barrier ads.       Dar advisor of the support warms barrier ads.
	Vantages TRACES OF PUBLISHING WITH US Fast Approval Work rebails will be approved stilling in face in or bailness days and up to 12th on weekends O <mark>% Revenue Share Bigh rates. Safe and Clean Ads All da ere ontinuously moderated by our team uing bath internal and third-party tools.</mark>		Contact US	sure that matter of ge through the form	10 Minimum Withdrawal Re process payments through various pay- pleneopy such as Payol, Pasan, Payo, Veternoory, Skrill, Biccon, With Tarade.       First-Class Support Re offer all of our uses the support on e-m age.       Societ Voor Website's Earnings Dy advertisiment have no interference with Pay advertisiment have no interference with an the been involved in the ordine advert supress for more than a decode.       When Pay-Cash was established in 2012, It- sam has been involved in the ordine advert supress for more than a decode.       anticless and
	Vantages ITRACES OF PUBLISHING WITH US Fast Approval Your extelle with be approved within 1 hour in or business days and up to 12th on exectends 80% Revenue Share Encluse advertisers with direct campaigns an high rates. Safe and Clean Ads All ads are continuously moderated by our tear using both internal and third-party tools. 10% Referral Program Refer your friends and earn passive income for their activity without any limits. Our mains for Our mains for	it is flexible req	Fast Payments Daily Payments its in bow we make our payments are processed within a service of the service of the service <b>Jourdy Updated Reports</b> Waves the uso-breake with your earners <b>Jourdy Updated Reports</b> Control of the service of the service <b>Jourdy Updated Reports</b> Control of the service of the service <b>Jourdy Updated Reports</b> Control of the service of the service <b>Control of the service of the service</b> Service of the service of th	sure Pat matter of gs through de for de for	\$10 Minimum Withdrawal         Ne process payments through various pays         determine starting and the Paysial (Pasum, Paysa, Network, Sark, Bickon, With Tanake, Paysa, Pasum, Paysa, Pasum, Paysa, Pasum, Paysa, Pasum, Paysa, Pasum, Paysa, Pasum, Pasum
	Vantages ITAGES OF PUBLISHING WITH US Fast Approval Wor websiles will be approved within 1 hour in our business days and up to 12th on weekends BOCK Revenue Share Excluse activetisers with direct campaignes are high rates. Safe and Clean Ads All ads are continuously moderated by our lear using both internal and third-parity tools. 10(% Referral Program Refer your foreids and earm passive income for their activity without any limits. Cour main tree Cour main tree	it is flexib req or any buz	Contact Us	sure that matter of gs through de tor to de to any de to	10 Minimum Withdrawal         Re process payments through various pays planeways such as Payol, Pasan, Payon, Veternoory, Saritt, Bicton, Wei Taranter.         First-Class Support We offer all of our uses the support on ent Pays.         Socs Toyor Website's Earnings Day advertisement have no interference with warest barner ads.         Veer 10 Years' Experience Mile PayCath was established in 2012, th ann has been involved in the orien adverta- varies for more than a decade.         Initial State of the support of the orient of the orient adverta- varies for more than a decade.
	Vantages ITRACES OF PUBLISHING WITH US Fast Approval Your existes with exproved within 1 hour in or business days and up to 12th on exertends 80% Revenue Share Enclaive advertisers with direct campaigns an high rates. Safe and Clean Ads All ads are continuously modernated by our therr using both internal and third-party tools. 10% Referral Program Refer your finends and earn passive income for their activity without any limits. Our mains the Our mains the Court mains the same of the same of the same of the Safe and Clean Ads Our mains the Safe and Clean Ads Our mains the Safe and Clean Ads Our mains the Safe and Clean Ads Safe and Clean Ads Safe and Clean Ads Our mains the Safe and Clean Ads Our mains the Safe and Clean Ads Safe and Clean Ads Sa	it is flexib req or any bus		sure that matter of the second	I to Minimum Withdrawal Re process primerits through vertices pays thermorys such as Paysil, Pasam, Peysa, Vectorovy, Gord, Bickon, Vier Tarake. First-Class Support to citerat of our users live support on em appe. Boost Your Website's Earnings Dar advertisements have no interference ai unrent barner ads. Vert 10 Years' Experience man has been involved in the colore advert usizess for more than a decide. antibes and
	Vantages Vantages Vantages Varaces or PUBLISHING WITH US Fast Approval Variances advertisers with direct campaignes an variance advertisers and advertisers and variance advertisers variance vari	a lit is flexible req or any bus source us		sure that and and an and an and an and an	10 Minimum Withdrawal         Ma process payments through various payments through various payments such as Payad, Paware, Payad, Water, Paya
	Vantages ITRAGES OF PUBLISHING WITH US Fast Approval Vor existing with the approved within 1 hour in our business days and up to 12th on exekends S0% Revenue Share Exclusive advertisers with direct campaigns an high rates. Safe and Clean Ads All ads are continuously moderned by our tear using both internal and that-party tools. Our main tra Court marks and campasive hours for grade Party and P	a it is flexible or any bus or an		sure that matter of ge through for de too any for de too any for det or det or d	I to Minimum Withdrawal  Parsons payments through various pay determorey. Skett Broad, Various pay determorey. Skett Broad, Various pay determorey. Skett Broad, Various Payae, Paara, Parst-Class Support  Frist-Class Support  Frist-Class Support  For our varies the support on ent appea.  Social Your Website's Earnings Dur Johnson State Statements  Vert Of Varis' Experience  Vert Of Varis' Experience  Vert Of Varis' Experience  Anna has been involved in the orline advert supress for more than a decade.  Anna has been involved in the orline advert supress for more than a decade.  Anna has been involved in the orline advert supress for more than a decade.  Anna has been involved in the orline advert supress for more than a decade.  Anna has been involved in the orline advert anna been involved in the orline advert advert anna been involved in the orline advert advert advert  Anna has been involved in the orline advert advert advert Anna has been involved in the orline advert advert advert  Anna has been involved in the orline advert advert advert Anna has been involved in the orline advert advert advert Anna has been involved in the orline advert advert advert Anna has been involved in the orline advert advert Anna has been involved in the orline advert advert advert Anna has been involved in the orline advert advert Anna has been involved in the orline advert advert Anna has been involved in the orline advert advert Anna has been involved in the orline advert advert Anna has been involved in the orline advert advert Anna has been involved in the orline advert advert Anna has been involved in the orline advert advert Anna has been involved in the orline advert advert Anna has been involved in the orline advert advert Anna has been involved in the orline advert advert Anna has been involved in the orline advert advert Anna has been involved in the orline advert advert Anna has been involved in the orline advert

PDoC 8



#### Publishers

START EARNING EXTRA REVENUE IN LESS THAN 10 MINUTESI Maximize your revenue with PopCash! You can start earning extra revenue in less than 10 minutes. All you need to do is register, submit your website and place our popunder code on the deaired pages. Once your domain has been approved, all of your visitors will be shown a popunder advertisement every 24 hours. This sauruss that your viewers' experimence will not be affected by our ads. All that's left for you to do is to focus on attracting more visitors towards your website. Are you ready to start? You can register here!



	TAGES OF PUBLISHING WITH US Fast Approval Your website will be approved within 1 hour in our business days and up to 12h on weekends.	Ð	Fast Payments Daily Payments: this is how we make sure that your payments are processed within a matter of hours.		\$10 Minimum Withdrawal We process payments through various payment gateways such as Payte, Paum, Payta, Wethoncow Skill Briston Wen Tander.
6	80% Revenue Share Exclusive advertisers with direct campaigns and high rates.	Ē.	Hourly Updated Reports Always be up-to-date with your earnings through our hourly-updated statistics.		First-Class Support We offer all of our users live support on e-mail or Skype.
	Safe and Clean Ads All ads are continuously moderated by our team, using both internal and third-party tools.		Simple Implementation Our ad code can easily be implemented into any website. We also have plugins available for WordPress and Blogger.com websites.	5	Boost Your Website's Earnings Our advertisements have no interference with your current banner ads.
<u>8</u> 28 88	10% Referral Program Refer your friends and earn passive income from their activity without any limits.	<u>.</u>	Unique Integration Our ad code works seamlessly on all devices and can be easily integrated.		Over 10 Years' Experience While PopCash was established in 2012, the core team has been involved in the online advertising business for more than a decade.
			Register Now		
v ∠≡×	Our main trait i	s flexib	Contact Us ility and adaptability to any project's p uirements, regardless of their scale!	articula	rities and



PDoC 5



	antages					
	Fast Approval four website will be approved within 1 hour in our business days and up to 12h on weekend	6	Fast Payments Daily Payments: this is how we make sure that your payments are processed within a matter of hours.	() () () () () () () () () () () () () (	510 Minimum V Ve process payme lateways such as F Vebmoney, Skrill, E	Vithdrawal nls through various payment 'aypal, Paxum, Payza, titcoin, Wiro Transfer.
6	30% Revenue Share Exclusive advertisers with direct campaigns a high rates.	nd	Hourty Updated Reports Always be up-to-date with your earnings through our hourly-updated statistics.	F v s	First-Class Sup Ve offer all of our u kype.	port sers live support on e-mail or
	Safe and Clean Ads All ads are continuously moderated by our tea using both internal and third-party tools.	m,	Simple Implementation Our ad code can easily be implemented into any website. We also have plugins available for WordPress and Blogger.com websites.		Boost Your We Dur advertisements urrent banner ads.	bsite's Earnings have no interference with your
282 22	10% Referral Program Refer your friends and earn passive income fr heir activity without any limits.	om	Unique Integration Our ad code works seamlessly on all devices and can be easily integrated.	<b>ر التحديث</b> منابع منابع	Over 10 Years' While PopCash was eam has been invo usiness for more t	Experience s established in 2012, the core eved in the online advertising han a decade.
			Register Now			
	Our main tr	ait is flexit req For any bu	Contact Us billy and adaptability to any projects p juirements, regardless of their scale! siness proposal or query, please cont Contact Us	oarticulariti act us!	ies and X V	
Pay	Pal Payza pox	um	WebMoney 🛞 Skrill	<mark>0</mark> 6	oitcoin	
Pop The Popu f in S PopCast	I SA	ABOUT US Home Ferms of use	PRODUCTS JOIN US Advertiser Login Publisher Register Contact u		If want to r please clic	report a malware to us, k the link below. Report Malware
Bucha	rest, Romania		PopCash Network. All rights reserved.		Design by	difiz

#### ground truth



• *probabilistic* algorithm based on edge detection, optimized for *locally significant* edges



- *probabilistic* algorithm based on edge detection, optimized for *locally significant* edges
- designed to detect *extended lines* (visually non-continuous lines that may form segment borders)



- *probabilistic* algorithm based on edge detection, optimized for *locally significant* edges
- designed to detect *extended lines* (visually non-continuous lines that may form segment borders)



- *probabilistic* algorithm based on edge detection, optimized for *locally significant* edges
- designed to detect *extended lines* (visually non-continuous lines that may form segment borders)



- *probabilistic* algorithm based on edge detection, optimized for *locally significant* edges
- designed to detect *extended lines* (visually non-continuous lines that may form segment borders)



- *probabilistic* algorithm based on edge detection, optimized for *locally significant* edges
- designed to detect *extended lines* (visually non-continuous lines that may form segment borders)



- *probabilistic* algorithm based on edge detection, optimized for *locally significant* edges
- designed to detect *extended lines* (visually non-continuous lines that may form segment borders)



Popcash The Popunder Network	Home Publ	there Advertisers EAO About Us Contact Login	Register
		Start earning money in less them to minutes:	
Publishers		Full Name	
Makimize your revenue with PopCas		Email	
HOW TO START?		Create Account	

### Cormier et al. $(s_{min} = 45, t_l = 512)$

				pixels			characters			
	Parameters	# segments	$P_{B^{3}}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$		
(worst)	s <sub>min</sub> = 90 px t <sub>l</sub> = 256 px	18.4	0.29	0.86	0.35	0.60	0.87	0.63		
(best)	$s_{min} = 45 \ px$ $t_l = 512 \ px$	38.0	0.34	0.77	0.36	0.67	0.78	0.62		
	Δ	+ 19.6	+ 0.05	- 0.09	+ 0.01	+ 0.07	- 0.09	- 0.01		
	VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68		

Primary observations:

• Purely visual approach comes close to VIPS' performance

				pixels			characters			
	Parameters	# segments	$P_{B^{3}}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$		
(worst)	s <sub>min</sub> = 90 px t <sub>l</sub> = 256 px	18.4	0.29	0.86	0.35	0.60	0.87	0.63		
(best)	$s_{min} = 45 \ px$ $t_l = 512 \ px$	38.0	0.34	0.77	0.36	0.67	0.78	0.62		
	Δ	+ 19.6	+ 0.05	- 0.09	+ 0.01	+ 0.07	- 0.09	- 0.01		
	VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68		

Primary observations:

- Purely visual approach comes close to VIPS' performance
- but: needs more than **3x** segment count to come close

				pixels			characters			
	Parameters	# segments	$P_{B^{3}}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$		
(worst)	s <sub>min</sub> = 90 px t <sub>l</sub> = 256 px	18.4	0.29	0.86	0.35	0.60	0.87	0.63		
(best)	$s_{min} = 45 \ px$ $t_l = 512 \ px$	38.0	0.34	0.77	0.36	0.67	0.78	0.62		
	Δ	+ 19.6	+ 0.05	- 0.09	+ 0.01	+ 0.07	- 0.09	- 0.01		
	VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68		

Primary observations:

- Purely visual approach comes close to VIPS' performance
- but: needs more than **3x** segment count to come close

→ expresses fundamentally different operation (visual vs. DOM-based)

				pixels		characters			
	Parameters	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	
(worst)	$s_{min} = 90 \ px$ $t_l = 256 \ px$	18.4	0.29	0.86	0.35	0.60	0.87	0.63	
(best)	s <sub>min</sub> = 45 px t <sub>l</sub> = 512 px	38.0	0.34	0.77	0.36	0.67	0.78	0.62	
	Δ	+ 19.6	+ 0.05	- 0.09	+ 0.01	+ 0.07	- 0.09	- 0.01	
	VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68	

Parameters:

- $t_l$  (max. line length for probability estimation)  $\in \{256, 512\}$  px
- increasing t<sub>l</sub> finds extended lines across larger gaps



(a)  $s_{min} = 45, t_l = 256 \text{ px}$ 



				pixels			characters			
	Parameters	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$		
(worst)	s <sub>min</sub> = 90 px t <sub>l</sub> = 256 px	18.4	0.29	0.86	0.35	0.60	0.87	0.63		
(best)	s <sub>min</sub> = 45 px t <sub>l</sub> = 512 px	38.0	0.34	0.77	0.36	0.67	0.78	0.62		
	Δ	+ 19.6	+ 0.05	- 0.09	+ 0.01	+ 0.07	- 0.09	- 0.01		
	VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68		

#### Parameters:

- $t_l$  (max. line length for probability estimation)  $\in \{256, 512\}$  px
- increasing t<sub>l</sub> finds extended lines across larger gaps



(a)  $s_{min} = 45, t_l = 256 \text{ px}$ 



				pixels		characters			
	Parameters	# segments	$P_{B^{3}}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	
(worst)	$s_{min} = 90 \ px$ $t_l = 256 \ px$	18.4	0.29	0.86	0.35	0.60	0.87	0.63	
(best)	s <sub>min</sub> = 45 px t <sub>l</sub> = 512 px	38.0	0.34	0.77	0.36	0.67	0.78	0.62	
	Δ	+ 19.6	+ 0.05	- 0.09	+ 0.01	+ 0.07	- 0.09	- 0.01	
	VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68	

#### Parameters:

- $t_l$  (max. line length for probability estimation)  $\in \{256, 512\}$  px
- increasing t<sub>l</sub> finds extended lines across larger gaps



Year and year

				pixels		characters			
	Parameters	# segments	$P_{B^{3}}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	
(worst)	$s_{min} = 90 \ px$ $t_l = 256 \ px$	18.4	0.29	0.86	0.35	0.60	0.87	0.63	
(best)	$s_{min} = 45 \ px$ $t_l = 512 \ px$	38.0	0.34	0.77	0.36	0.67	0.78	0.62	
	Δ	+ 19.6	+ 0.05	- 0.09	+ 0.01	+ 0.07	- 0.09	- 0.01	
	VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68	

#### Parameters:

•  $s_{min}$  (minimum segment border length)  $\in \{45, 90\}$  px



				pixels			characters		
	Parameters	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	
(worst)	s <sub>min</sub> = 90 px t <sub>l</sub> = 256 px	18.4	0.29	0.86	0.35	0.60	0.87	0.63	
(best)	s <sub>min</sub> = 45 px t <sub>l</sub> = 512 px	38.0	0.34	0.77	0.36	0.67	0.78	0.62	
	Δ	+ 19.6	+ 0.05	- 0.09	+ 0.01	+ 0.07	- 0.09	- 0.01	
	VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68	

#### Parameters:

- $s_{min}$  (minimum segment border length)  $\in \{45, 90\}$  px
- influences segmentation granularity directly



				pixels		characters			
	Parameters	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	
(worst)	s <sub>min</sub> = 90 px t <sub>l</sub> = 256 px	18.4	0.29	0.86	0.35	0.60	0.87	0.63	
(best)	s <sub>min</sub> = 45 px t <sub>l</sub> = 512 px	38.0	0.34	0.77	0.36	0.67	0.78	0.62	
	Δ	+ 19.6	+ 0.05	- 0.09	+ 0.01	+ 0.07	- 0.09	- 0.01	
	VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68	

#### Parameters:

- $s_{min}$  (minimum segment border length)  $\in \{45, 90\}$  px
- influences segmentation granularity directly
- *interaction* with  $t_l$


# Results: Cormier et al.

				pixels		characters		
	Parameters	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$
(worst)	s <sub>min</sub> = 90 px t <sub>l</sub> = 256 px	18.4	0.29	0.86	0.35	0.60	0.87	0.63
(best)	s <sub>min</sub> = 45 px t <sub>l</sub> = 512 px	38.0	0.34	0.77	0.36	0.67	0.78	0.62
	Δ	+ 19.6	+ 0.05	- 0.09	+ 0.01	+ 0.07	- 0.09	- 0.01
	VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68

#### Parameters:

- $s_{min}$  (minimum segment border length)  $\in \{45, 90\}$  px
- influences segmentation granularity directly
- *interaction* with  $t_l$



(b)  $s_{min} = 45, t_l = 256 \text{ px}$ 

			pixels			characters		
Variant	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	
best	38.0	0.34	0.77	0.36	0.67	0.78	0.62	
best, fitted	16.8	0.42	0.77	0.38	0.68	0.81	0.65	
Δ	- 21.2	+ 0.08	_	+ 0.02	+ 0.01	+ 0.03	+ 0.03	
VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68	

			pixels	pixels			characters		
Variant	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$		
best	38.0	0.34	0.77	0.36	0.67	0.78	0.62		
best, <b>fitted</b>	16.8	0.42	0.77	0.38	0.68	0.81	0.65		
Δ	- 21.2	+ 0.08	_	+ 0.02	+ 0.01	+ 0.03	+ 0.03		
VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68		

• *pixels*: consistently lower  $P_{B^3}$  and therefore  $F_{B^3}$  than for all other atomic elements

			pixels			characters	
Variant	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$
best	38.0	0.34	0.77	0.36	0.67	0.78	0.62
best, <b>fitted</b>	16.8	0.42	0.77	0.38	0.68	0.81	0.65
Δ	- 21.2	+ 0.08	_	+ 0.02	+ 0.01	+ 0.03	+ 0.03
VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68

- *pixels*: consistently lower  $P_{B^3}$  and therefore  $F_{B^3}$  than for all other atomic elements
- primary culprit: blank space

			pixels			characters	
Variant	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$
best	38.0	0.34	0.77	0.36	0.67	0.78	0.62
best, <b>fitted</b>	16.8	0.42	0.77	0.38	0.68	0.81	0.65
Δ	- 21.2	+ 0.08	_	+ 0.02	+ 0.01	+ 0.03	+ 0.03
VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68

- *pixels*: consistently lower  $P_{B^3}$  and therefore  $F_{B^3}$  than for all other atomic elements
- primary culprit: blank space
- correct segmentation of blank space important for some uses (e.g. design mining), irrelevant for others (e.g. text extraction tasks)

			pixels			characters	
Variant	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$
best	38.0	0.34	0.77	0.36	0.67	0.78	0.62
best, <b>fitted</b>	16.8	0.42	0.77	0.38	0.68	0.81	0.65
Δ	- 21.2	+ 0.08	_	+ 0.02	+ 0.01	+ 0.03	+ 0.03
VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68

- *pixels*: consistently lower  $P_{B^3}$  and therefore  $F_{B^3}$  than for all other atomic elements
- primary culprit: blank space
- correct segmentation of blank space important for some uses (e.g. design mining), irrelevant for others (e.g. text extraction tasks)
- can be somewhat mitigated by *fitting to DOM nodes*

			pixels			characters	
Variant	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$
best	38.0	0.34	0.77	0.36	0.67	0.78	0.62
best, <b>fitted</b>	16.8	0.42	0.77	0.38	0.68	0.81	0.65
Δ	- 21.2	+ 0.08	_	+ 0.02	+ 0.01	+ 0.03	+ 0.03
VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68

• human segmentations are fit to DOM nodes (containment threshold  $\theta_c$  = 0.75)

			pixels		characters		
Variant	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$
best	38.0	0.34	0.77	0.36	0.67	0.78	0.62
best, <b>fitted</b>	16.8	0.42	0.77	0.38	0.68	0.81	0.65
Δ	- 21.2	+ 0.08	_	+ 0.02	+ 0.01	+ 0.03	+ 0.03
VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68

- human segmentations are fit to DOM nodes (containment threshold  $\theta_c$  = 0.75)
- → fair treatment: fit visual/hybrid algorithms to DOM nodes, too

# DOM fitting: example



original



fitted

			pixels			characters			
Variant	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^{3}}$	$R_{B^3}$	$F_{B^3}$		
best	38.0	0.34	0.77	0.36	0.67	0.78	0.62		
best, <b>fitted</b>	16.8	0.42	0.77	0.38	0.68	0.81	0.65		
Δ	- 21.2	+ 0.08	-	+ 0.02	+ 0.01	+ 0.03	+ 0.03		
VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68		

• Reduced oversegmentation **and** increased precision (*pixels*) and recall (*characters*)

			pixels			characters	
Variant	# segments	$P_{B^{3}}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$
best	38.0	0.34	0.77	0.36	0.67	0.78	0.62
best, fitted	16.8	0.42	0.77	0.38	0.68	0.81	0.65
Δ	- 21.2	+ 0.08	_	+ 0.02	+ 0.01	+ 0.03	+ 0.03
VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68

• Reduced oversegmentation **and** increased precision (*pixels*) and recall (*characters*)

 $\longrightarrow$   $F_{B^3}$  matches VIPS for *pixels* and comes closer for *characters* 

• **NOT** a Web Page Segmentation algorithm!

- **NOT** a Web Page Segmentation algorithm!
- … in fact, not an algorithm by itself
  → machine learning toolkit

- **NOT** a Web Page Segmentation algorithm!
- … in fact, not an algorithm by itself
  → machine learning toolkit
- designed for real-world object detection and *instance segmentation* (i.e. segmenting real-world images)

- **NOT** a Web Page Segmentation algorithm!
- … in fact, not an algorithm by itself
  → machine learning toolkit
- designed for real-world object detection and *instance segmentation* (i.e. segmenting real-world images)



- **NOT** a Web Page Segmentation algorithm!
- … in fact, not an algorithm by itself
  → machine learning toolkit
- designed for real-world object detection and *instance segmentation* (i.e. segmenting real-world images)
- offers high-performance, pre-trained, state-of-the-art neural network models



- **NOT** a Web Page Segmentation algorithm!
- … in fact, not an algorithm by itself
  → machine learning toolkit
- designed for real-world object detection and *instance segmentation* (i.e. segmenting real-world images)
- offers high-performance, pre-trained, state-of-the-art neural network models
- currently leads Microsoft COCO challenge in instance segmentation



- **NOT** a Web Page Segmentation algorithm!
- … in fact, not an algorithm by itself
  → machine learning toolkit
- designed for real-world object detection and *instance segmentation* (i.e. segmenting real-world images)
- offers high-performance, pre-trained, state-of-the-art neural network models
- currently leads Microsoft COCO challenge in instance segmentation
- → transfer to Web Page Segmentation possible?



			pixels		characters				
Variant	# segments	$P_{B^{3}}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$		
original	252.2	0.47	0.41	0.33	0.80	0.44	0.48		
fitted	14.7	0.67	0.38	0.35	0.80	0.54	0.56		
Δ	- 237.5	+ 0.20	- 0.03	+ 0.02	_	+ 0.10	+ 0.08		
VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68		

• Real-world image segmentation does not directly transfer well

			pixels		characters				
Variant	# segments	$P_{B^{3}}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$		
original	252.2	0.47	0.41	0.33	0.80	0.44	0.48		
fitted	14.7	0.67	0.38	0.35	0.80	0.54	0.56		
Δ	- 237.5	+ 0.20	- 0.03	+ 0.02	_	+ 0.10	+ 0.08		
VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68		

- Real-world image segmentation does not directly transfer well
- *massive* oversegmentation

			pixels		characters		
Variant	# segments	$P_{B^{3}}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$
original	252.2	0.47	0.41	0.33	0.80	0.44	0.48
fitted	14.7	0.67	0.38	0.35	0.80	0.54	0.56
Δ	- 237.5	+ 0.20	- 0.03	+ 0.02	_	+ 0.10	+ 0.08
VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68

#### Reasons:

			pixels		characters		
Variant	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$
original	252.2	0.47	0.41	0.33	0.80	0.44	0.48
fitted	14.7	0.67	0.38	0.35	0.80	0.54	0.56
Δ	- 237.5	+ 0.20	- 0.03	+ 0.02	-	+ 0.10	+ 0.08
VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68

#### Reasons:

• segmenting real-world objects found in images on web pages

			pixels		characters			
Variant	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	
original	252.2	0.47	0.41	0.33	0.80	0.44	0.48	
fitted	14.7	0.67	0.38	0.35	0.80	0.54	0.56	
Δ	- 237.5	+ 0.20	- 0.03	+ 0.02	_	+ 0.10	+ 0.08	
VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68	

#### Reasons:

- segmenting real-world objects found in images on web pages
- neural network model not trained on web pages

			pixels		characters		
Variant	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$
original	252.2	0.47	0.41	0.33	0.80	0.44	0.48
fitted	14.7	0.67	0.38	0.35	0.80	0.54	0.56
Δ	- 237.5	+ 0.20	- 0.03	+ 0.02	-	+ 0.10	+ 0.08
VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68

			pixels		characters			
Variant	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	
original	252.2	0.47	0.41	0.33	0.80	0.44	0.48	
fitted	14.7	0.67	0.38	0.35	0.80	0.54	0.56	
Δ	- 237.5	+ 0.20	- 0.03	+ 0.02	-	+ 0.10	+ 0.08	
VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68	

• 94.2% reduction in segment count

			pixels		characters			
Variant	# segments	$P_{B^{3}}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	
original	252.2	0.47	0.41	0.33	0.80	0.44	0.48	
fitted	14.7	0.67	0.38	0.35	0.80	0.54	0.56	
Δ	- 237.5	+ 0.20	- 0.03	+ 0.02	-	+ 0.10	+ 0.08	
VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68	

- 94.2% reduction in segment count
- **best precision** for *pixels* across all single algorithms

			pixels		characters		
Variant	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$
original	252.2	0.47	0.41	0.33	0.80	0.44	0.48
fitted	14.7	0.67	0.38	0.35	0.80	0.54	0.56
Δ	- 237.5	+ 0.20	- 0.03	+ 0.02	-	+ 0.10	+ 0.08
VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68

- 94.2% reduction in segment count
- **best precision** for *pixels* across all single algorithms
- $F_{B^3}$  approaches VIPS for *pixels*

# Further experiments

• *F*<sub>B<sup>3</sup></sub> expresses segmentation *similarity* (interpreted as quality when comparing to ground truth)

•  $F_{B^3}$  expresses segmentation *similarity* (interpreted as quality when comparing to ground truth)

 $\rightarrow$  possibility of expressing similarity between algorithms

•  $F_{B^3}$  expresses segmentation *similarity* (interpreted as quality when comparing to ground truth)

→ possibility of expressing similarity between algorithms

$F_{B^3}$	S				${F_{B^3}}$				
$S^*$	VIPS	HEPS	Cormie	r MMDet.	$S^*$	VIPS	HEPS	Cormier	MMDet.
VIPS	1.00	0.41	0.51	0.31	VIPS	1.00	0.48	0.60	0.41
HEPS	0.41	1.00	0.50	0.31	HEPS	0.48	1.00	0.43	0.36
Cormier	0.51	0.50	1.00	0.37	Cormier	0.60	0.43	1.00	0.40
MMDet.	0.31	0.31	0.37	1.00	MMDet.	0.41	0.36	0.40	1.00

pixels

characters

## *Min-vote* ensemble

• Applying segmentation fusion (used in ground truth creation) to algorithm segmentations

## *Min-vote* ensemble

- Applying segmentation fusion (used in ground truth creation) to algorithm segmentations
- *Min-vote@n*,  $n \in [1, 2, 3, 4]$  = number of algorithm segmentations that put a given element in a segment

## *Min-vote* ensemble

- Applying segmentation fusion (used in ground truth creation) to algorithm segmentations
- *Min-vote@n*,  $n \in [1, 2, 3, 4]$  = number of algorithm segmentations that put a given element in a segment
- Initially evaluated for paper with unoptimized parameters
#### *Min-vote* ensemble

- Applying segmentation fusion (used in ground truth creation) to algorithm segmentations
- *Min-vote@n*,  $n \in [1, 2, 3, 4]$  = number of algorithm segmentations that put a given element in a segment
- Initially evaluated for paper with unoptimized parameters
- now: optimized parameters, fitted segmentations → what improvements do we see?



*n* = 1 *n* = 2 *n* = 3 *n* = 4



*n* = 2

ground truth

#### Results: Min-vote ensemble

			pixels		characters		
Variant	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$
unoptimized n = 2	32.9	0.39	0.64	0.38	0.76	0.68	0.65
optimized n = 2	16.0	0.37	0.77	0.40	0.71	0.80	0.69
Δ	- 16.9	- 0.02	+ 0.13	+ 0.02	- 0.05	+ 0.12	+ 0.04
VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68

• influence of optimized parameters and **fitting** 

#### Results: Min-vote ensemble

		pixels			characters		
Variant	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$
unoptimized n = 2	32.9	0.39	0.64	0.38	0.76	0.68	0.65
optimized n = 2	16.0	0.37	0.77	0.40	0.71	0.80	0.69
Δ	- 16.9	- 0.02	+ 0.13	+ 0.02	- 0.05	+ 0.12	+ 0.04
VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68

- influence of optimized parameters and **fitting**
- segment count cut in half, only minor losses in precision

#### Results: Min-vote ensemble

		pixels			characters		
Variant	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$
unoptimized n = 2	32.9	0.39	0.64	0.38	0.76	0.68	0.65
optimized n = 2	16.0	0.37	0.77	0.40	0.71	0.80	0.69
Δ	- 16.9	- 0.02	+ 0.13	+ 0.02	- 0.05	+ 0.12	+ 0.04
VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68

- influence of optimized parameters and **fitting**
- segment count cut in half, only minor losses in precision
- Min-vote@2 beats VIPS, provides best overall results

characters

Approach / Variant	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$
Min-vote@2 (optimized)	16.0	0.37	0.77	0.40	0.71	0.80	0.69
VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68
Cormier et al. (s <sub>min</sub> = 45 px, t <sub>l</sub> = 512 px, fitted)	16.8	0.42	0.77	0.38	0.68	0.81	0.65
MMDetection (fitted)	14.7	0.67	0.38	0.35	0.80	0.54	0.56
HEPS	35.8	0.39	0.54	0.32	0.72	0.50	0.50
Meier et al. ( <b>fitted</b> )	7.0	0.56	0.39	0.26	0.66	0.48	0.42

characters

Approach / Variant	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$
Min-vote@2 (optimized)	16.0	0.37	0.77	0.40	0.71	0.80	0.69
VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68
Cormier et al. (s <sub>min</sub> = 45 px, t <sub>l</sub> = 512 px, fitted)	16.8	0.42	0.77	0.38	0.68	0.81	0.65
MMDetection (fitted)	14.7	0.67	0.38	0.35	0.80	0.54	0.56
HEPS	35.8	0.39	0.54	0.32	0.72	0.50	0.50
Meier et al. ( <b>fitted</b> )	7.0	0.56	0.39	0.26	0.66	0.48	0.42

characters

Approach / Variant	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$
Min-vote@2 (optimized)	16.0	0.37	0.77	0.40	0.71	0.80	0.69
VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68
Cormier et al. (s <sub>min</sub> = 45 px, t <sub>l</sub> = 512 px, fitted)	16.8	0.42	0.77	0.38	0.68	0.81	0.65
MMDetection (fitted)	14.7	0.67	0.38	0.35	0.80	0.54	0.56
HEPS	35.8	0.39	0.54	0.32	0.72	0.50	0.50
Meier et al. ( <b>fitted</b> )	7.0	0.56	0.39	0.26	0.66	0.48	0.42

characters

Approach / Variant	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$
Min-vote@2 (optimized)	16.0	0.37	0.77	0.40	0.71	0.80	0.69
VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68
Cormier et al. (s <sub>min</sub> = 45 px, t <sub>l</sub> = 512 px, fitted)	16.8	0.42	0.77	0.38	0.68	0.81	0.65
MMDetection (fitted)	14.7	0.67	0.38	0.35	0.80	0.54	0.56
HEPS	35.8	0.39	0.54	0.32	0.72	0.50	0.50
Meier et al. ( <b>fitted</b> )	7.0	0.56	0.39	0.26	0.66	0.48	0.42

characters

Approach / Variant	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$
Min-vote@2 (optimized)	16.0	0.37	0.77	0.40	0.71	0.80	0.69
VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68
Cormier et al. (s <sub>min</sub> = 45 px, t <sub>l</sub> = 512 px, fitted)	16.8	0.42	0.77	0.38	0.68	0.81	0.65
MMDetection (fitted)	14.7	0.67	0.38	0.35	0.80	0.54	0.56
HEPS	35.8	0.39	0.54	0.32	0.72	0.50	0.50
Meier et al. ( <b>fitted</b> )	7.0	0.56	0.39	0.26	0.66	0.48	0.42

characters

Approach / Variant	# segments	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$	$P_{B^3}$	$R_{B^3}$	$F_{B^3}$
Min-vote@2 (optimized)	16.0	0.37	0.77	0.40	0.71	0.80	0.69
VIPS (PDoC 5)	13.5	0.35	0.70	0.38	0.74	0.76	0.68
Cormier et al. $(s_{min} = 45 px,$ $t_l = 512 px,$ fitted)	16.8	0.42	0.77	0.38	0.68	0.81	0.65
MMDetection (fitted)	14.7	0.67	0.38	0.35	0.80	0.54	0.56
HEPS	35.8	0.39	0.54	0.32	0.72	0.50	0.50
Meier et al. ( <b>fitted</b> )	7.0	0.56	0.39	0.26	0.66	0.48	0.42

• First implementation of VIPS in TypeScript/JavaScript (ported from Java)

- First implementation of VIPS in TypeScript/JavaScript (ported from Java)
- First comparison of segmentation performance on large dataset

- First implementation of VIPS in TypeScript/JavaScript (ported from Java)
- First comparison of segmentation performance on large dataset
- Analyses of all algorithms, with in-depth parameter discussions:

- First implementation of VIPS in TypeScript/JavaScript (ported from Java)
- First comparison of segmentation performance on large dataset
- Analyses of all algorithms, with in-depth parameter discussions:
  - VIPS: Highlighting relationship between PDoC and rules

- First implementation of VIPS in TypeScript/JavaScript (ported from Java)
- First comparison of segmentation performance on large dataset
- Analyses of all algorithms, with in-depth parameter discussions:
  - VIPS: Highlighting relationship between PDoC and rules
  - Cormier et al.: Revealing parameter interactions

- First implementation of VIPS in TypeScript/JavaScript (ported from Java)
- First comparison of segmentation performance on large dataset
- Analyses of all algorithms, with in-depth parameter discussions:
  - VIPS: Highlighting relationship between PDoC and rules
  - Cormier et al.: Revealing parameter interactions
- Promising combination of DOM information with visual segmentation, has benefits beyond fair evaluation treatment

• There are more approaches to be evaluated!

- There are more approaches to be evaluated!
- Analysis of performance by web page genre

- There are more approaches to be evaluated!
- Analysis of performance by web page genre
- DOM-based segmentations: incorporate HTML5/ARIA tags

- There are more approaches to be evaluated!
- Analysis of performance by web page genre
- DOM-based segmentations: incorporate HTML5/ARIA tags
- New hybrid approaches combining visual segmentation strategies with DOM information

# Thank you!