Large-scale Analysis and Comparison of Web Page Segmentation Approaches

Defense of Master's Thesis

Lars Meyer

January 28th, 2020

Watertown Daily Times
Serving the communities of Jefferson, St. Lawrence and Lewis counties, New York
In print daily. Online always.

"Bitten by Magic" by Kelly Ashley published today

By SEAN EWART
PUBLISHED: SATURDAY, SEPTEMBER 7, 2013 AT 4:30 AM

"Bitten by Magic," the second book in Kelly M. Ashley's Sisters of Fate series, is out today.

A follow up to her fantasy romance novel, "The Fur Flies," which was released in May, Mrs. Ashley's new book centers on the passion of Alexa, a witch, and Wilhelm, a vampire.

In "Bitten by Magic" Alexa is required to marry Wilhelm in order to settle a dispute.

"It's a paranormal romance," Mrs. Ashley, who writes under the pen name Kelliea Ashley, said. "If anybody likes the witches and the vampires and the werewolves that's what it is."

But Mrs. Ashley cautioned, "It's not 'Twilight,'" referring to the popular book and movie series. "My vampires don't sparkle."

Born in Gouverneur, Mrs. Ashley currently lives in Ogdensburg and said she draws inspiration from 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."

Mrs. 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.

"It's open ended," she said, so readers can expect more tales of forbidden love between witches, vampires and werewolves.

Mrs. Ashley said she has been writing since she was 13 but it was her husband, Joshua, who finally convinced her to find a publisher for her novels in 2012.

Her latest book is available as an e-book at breathlesspress.com, amazon.com, allromanceebooks.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," Mrs. Ashley said. "I'm not looking to get rich; I'm just looking to put my work out there."

Related Stories
OGD NEWS
- North Country Library System launches mobile app
- Buffalo Soldiers to be Subject of Remington Lecture
- St. Lawrence County has second-highest August unemployment rate in the state
- Ogdensburg man pleads guilty to meth possession
- Community Gallery Project on display at Ogdensburg International Airport

Watertown Daily Times
Serving the communities of Jefferson, St. Lawrence and Lewis counties, New York
In print daily. Online always.

"Bitten by Magic" by Kelly Ashley published today

By SEAN EWART
PUBLISHED: SATURDAY, SEPTEMBER 7, 2013 AT 4:30 AM

"Bitten by Magic," the second book in Kelly M. Ashley's Sisters of Fate series, is out today.

A follow up to her fantasy romance novel, "The Fur Flies," which was released in May, Mrs. Ashley's new book centers on the passion of Alexa, a witch, and Wilhelm, a vampire.

In "Bitten by Magic" Alexa is required to marry Wilhelm in order to settle a dispute.

"It's a paranormal romance," Mrs. Ashley, who writes under the pen name Kelliea Ashley, said. "If anybody likes the witches and the vampires and the werewolves that's what it is."

But Mrs. Ashley cautioned, "It's not 'Twilight,'" referring to the popular book and movie series. "My vampires don't sparkle."

Born in Gouverneur, Mrs. Ashley currently lives in Ogdensburg and said she draws inspiration from 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."

Mrs. 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.

"It's open ended," she said, so readers can expect more tales of forbidden love between witches, vampires and werewolves.

Mrs. Ashley said she has been writing since she was 13 but it was her husband, Joshua, who finally convinced her to find a publisher for her novels in 2012.

Her latest book is available as an e-book at breathlesspress.com, amazon.com, allromanceebooks.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," Mrs. Ashley said. "I'm not looking to get rich; I'm just looking to put my work out there."

Related Stories
OGD NEWS
- North Country Library System launches mobile app
- Buffalo Soldiers to be Subject of Remington Lecture
- St. Lawrence County has second-highest August unemployment rate in the state
- Ogdensburg man pleads guilty to meth possession
- Community Gallery Project on display at Ogdensburg International Airport

What is a segment?

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

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

Use cases

- 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
  - …

Use cases

- 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
  - …

Use cases

- 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
  - ...

Use cases

- 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
  - …

Use cases

- 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
  - …

Use cases

- 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*
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

Evaluation setup

Webis Web Segments 2020

- first crowd-sourced dataset for Web Page Segmentation

Kiesel et al., "Web Page Segmentation from First Principles", 2020

Webis Web Segments 2020

- first crowd-sourced dataset for Web Page Segmentation

Kiesel et al., "Web Page Segmentation from First Principles", 2020

Webis Web Segments 2020

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

Kiesel et al., "Web Page Segmentation from First Principles", 2020

Webis Web Segments 2020

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

- 8490 pages, 5 annotators per page
  $\longrightarrow$ 42450 human segmentations

Kiesel et al., "Web Page Segmentation from First Principles", 2020

Webis Web Segments 2020

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

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

Kiesel et al., "Web Page Segmentation from First Principles", 2020

Evaluation metrics

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

Kiesel et al., "Web Page Segmentation from First Principles", 2020

Evaluation metrics

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

- *pixels*

Kiesel et al., "Web Page Segmentation from First Principles", 2020

Evaluation metrics

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

- *pixels*
- $edges_{fine}$ and $edges_{coarse}$ (Canny edge detection)

Kiesel et al., "Web Page Segmentation from First Principles", 2020

Evaluation metrics

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

- *pixels*
- $edges_{fine}$ and $edges_{coarse}$ (Canny edge detection)
- *nodes* (in the DOM tree)

Kiesel et al., "Web Page Segmentation from First Principles", 2020

Evaluation metrics

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

- *pixels*
- $edges_{fine}$ and $edges_{coarse}$ (Canny edge detection)
- *nodes* (in the DOM tree)
- *characters*

Kiesel et al., "Web Page Segmentation from First Principles", 2020

Evaluation metrics

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

- *pixels*
- $edges_{fine}$ and $edges_{coarse}$ (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

Kiesel et al., "Web Page Segmentation from First Principles", 2020

Evaluation metrics

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

- *pixels*
- $edges_{fine}$ and $edges_{coarse}$ (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

$\longrightarrow$ 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

Consistency and reproducibility

Consistency and reproducibility

- *Webis Web Segments 2020* based on *Webis Web Archive 17*

Consistency and reproducibility

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

Consistency and reproducibility

- *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

Consistency and reproducibility

- *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 → highly consistent input

Consistency and reproducibility

- *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 → highly consistent input

⟶ Contribution: TypeScript/JavaScript port of VIPS

CSSBox (VIPS-Java)

CSSBox (VIPS-Java)

Safari

Summary: Methodology

Summary: Methodology

Webis Web Archive 17

Summary: Methodology

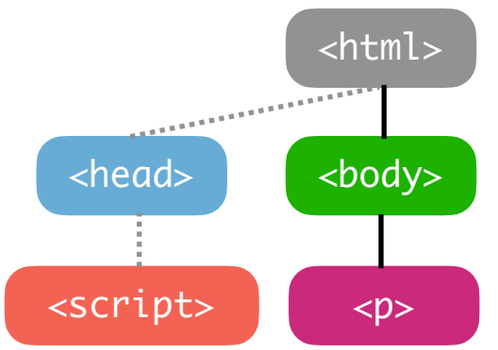
Webis Web Archive 17

algorithm →

```
{
"id":000000,
"width":1366,
"height":750,
"segmentations":
{
"single":[[[[0,0],
[0,750],[1366,750],
[1366,0],[0,0]]]]
}
}
```

JSON

Summary: Methodology



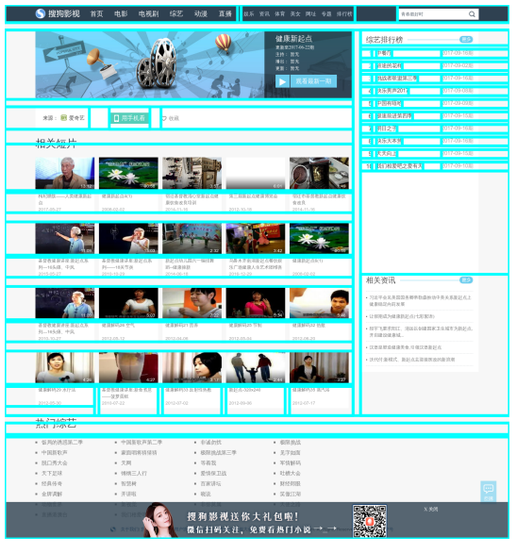
Webis Web Archive 17

algorithm


```
{
"id":000000,
"width":1366,
"height":750,
"segmentations":
{
"single":[[[[0,0],
[0,750],[1366,750],
[1366,0],[0,0]]]]
}
}
```


JSON

algorithm
segmentation



ground truth

Summary: Methodology

Webis Web Archive 17

algorithm

```
{
"id":000000,
"width":1366,
"height":750,
"segmentations":
{
"single":[[[[0,0],
[0,750],[1366,750],
[1366,0],[0,0]]]]
}
}
```

JSON

algorithm
segmentation

ground truth

P_{B^3}

R_{B^3}

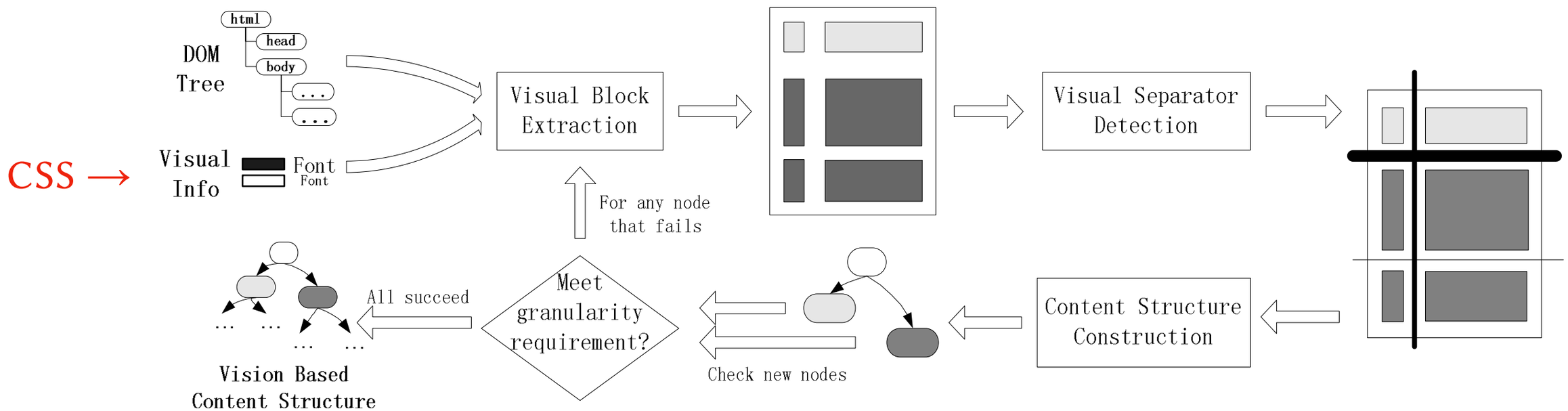
F_{B^3}

Approaches, Evaluations and Results

Overview

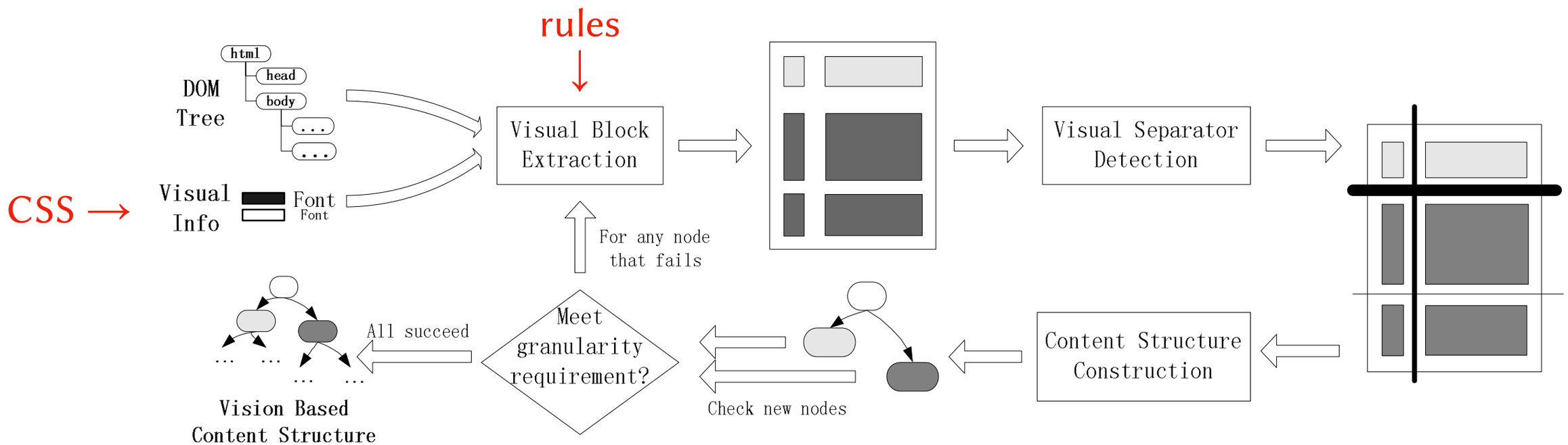
1. Evaluation of all algorithms + single-segment baseline against the ground truth
2. Parameter analyses: VIPS and Cormier et al.
3. Visual/hybrid segmentations fit to DOM nodes
4. Cross-evaluation (algorithm similarity)
5. *Min-vote* ensemble (combining algorithm segmentations)

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

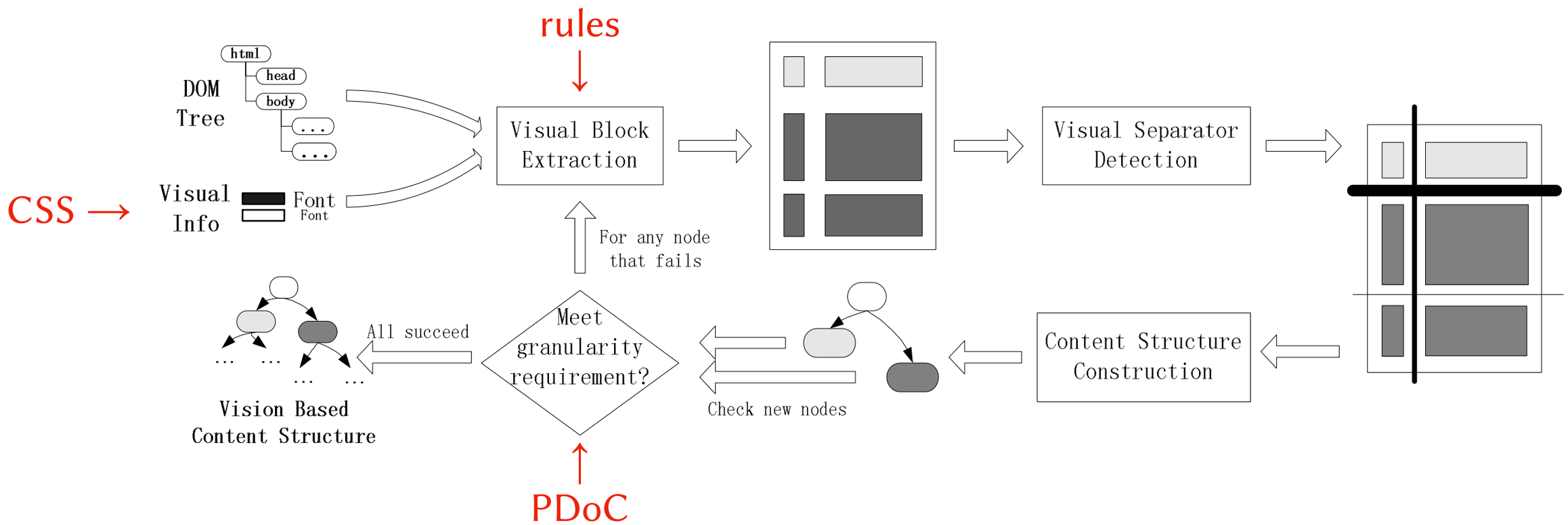
DOM-only approach: VIPS



- 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

Results: VIPS

		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)

Results: VIPS

		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)

$\longrightarrow$ high precision, low recall, low F-score

Results: VIPS

		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

Results: VIPS

		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:

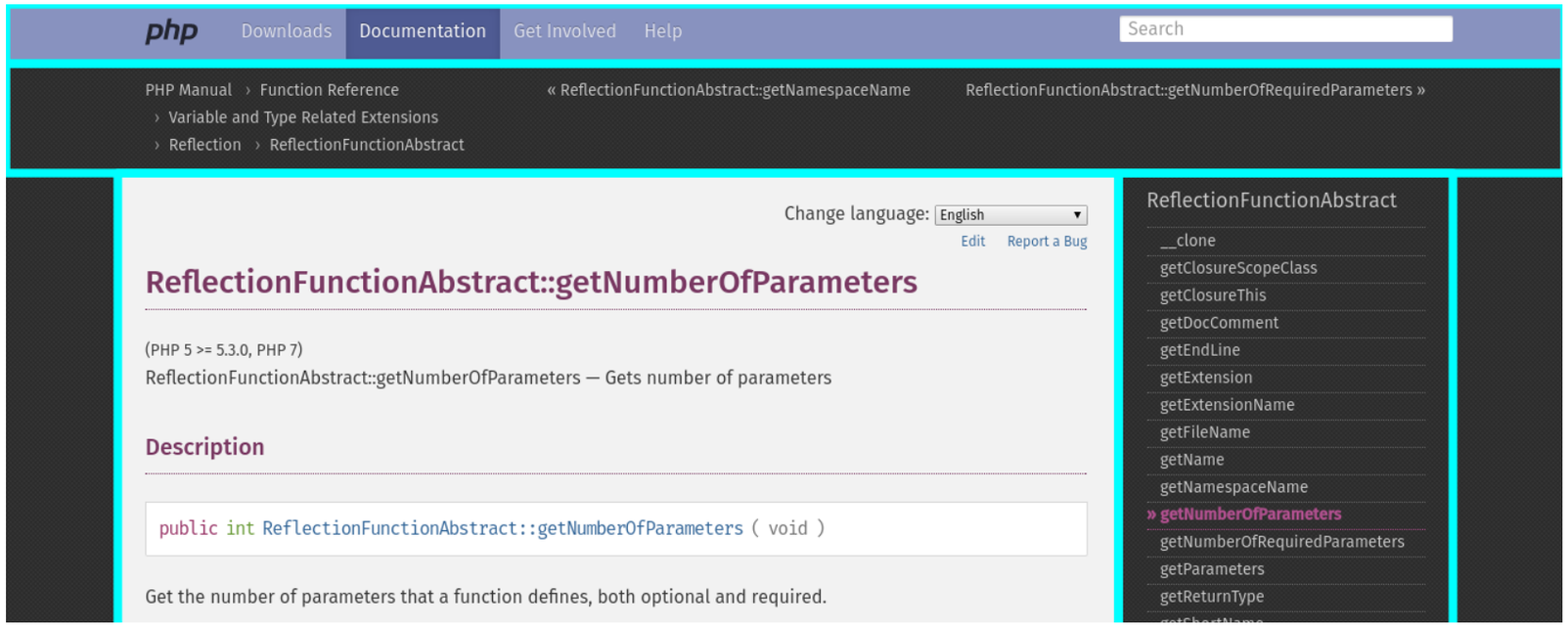
- PDoC > 6 applies rules targeting specific element types

$\longrightarrow$ outdated, detrimental to segmentation quality

Results: VIPS

		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

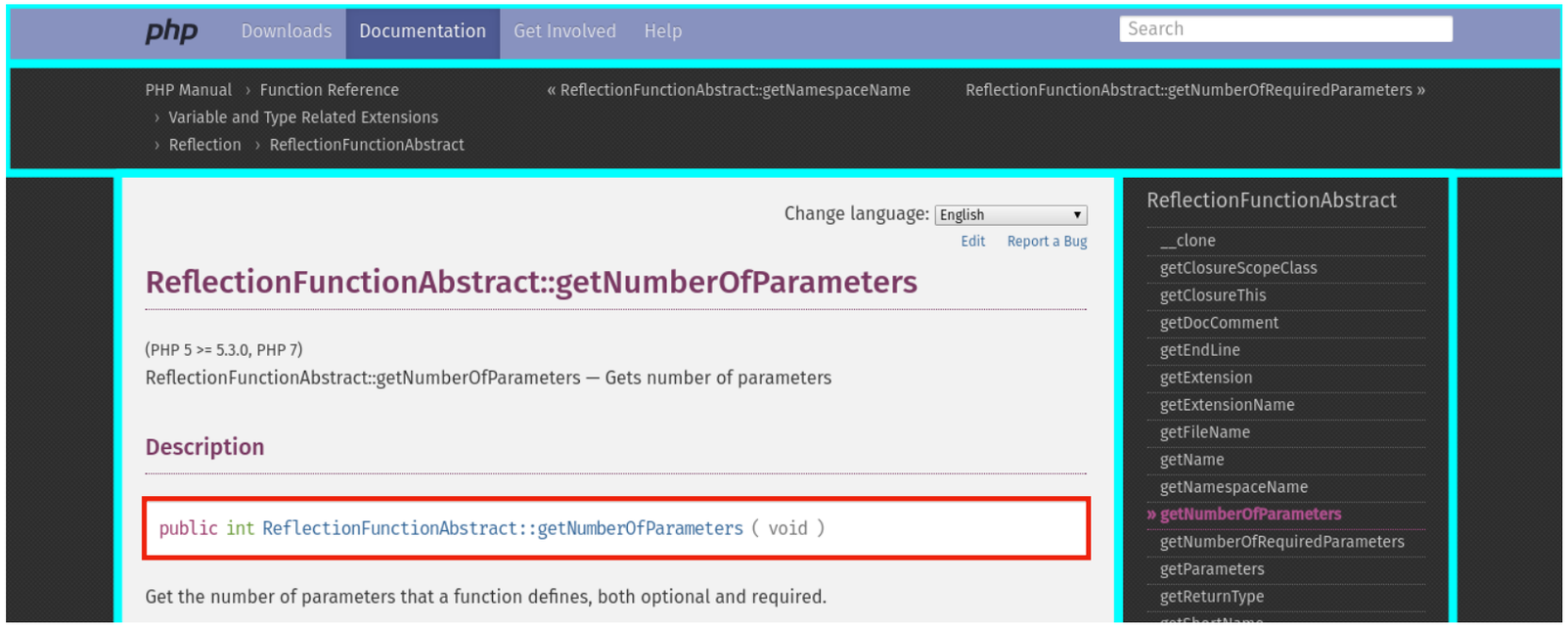
Example:
HTML <code> element



Results: VIPS

		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

Example:
HTML <code> element



Results: VIPS

		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

Results: VIPS

		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

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

Results: VIPS

		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

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

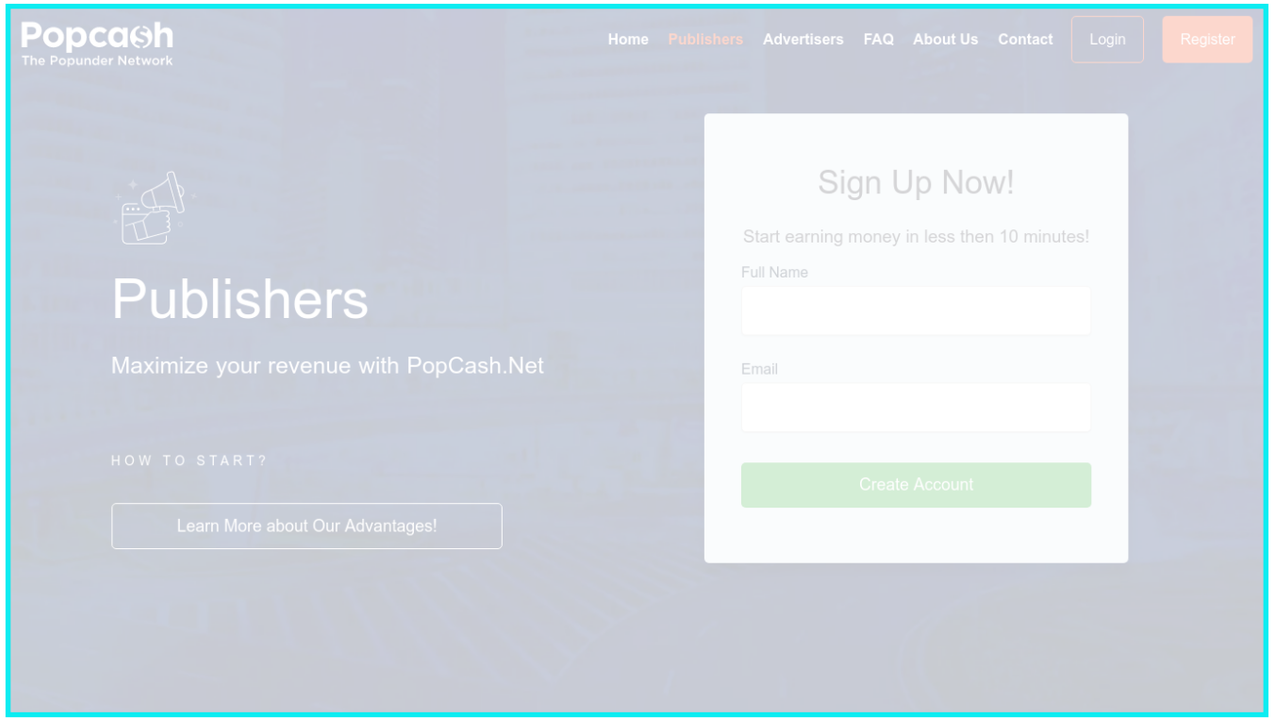
$\longrightarrow$ **VIPS (PDoC 5) is best single algorithm**

PDoC 8

PDoC 5

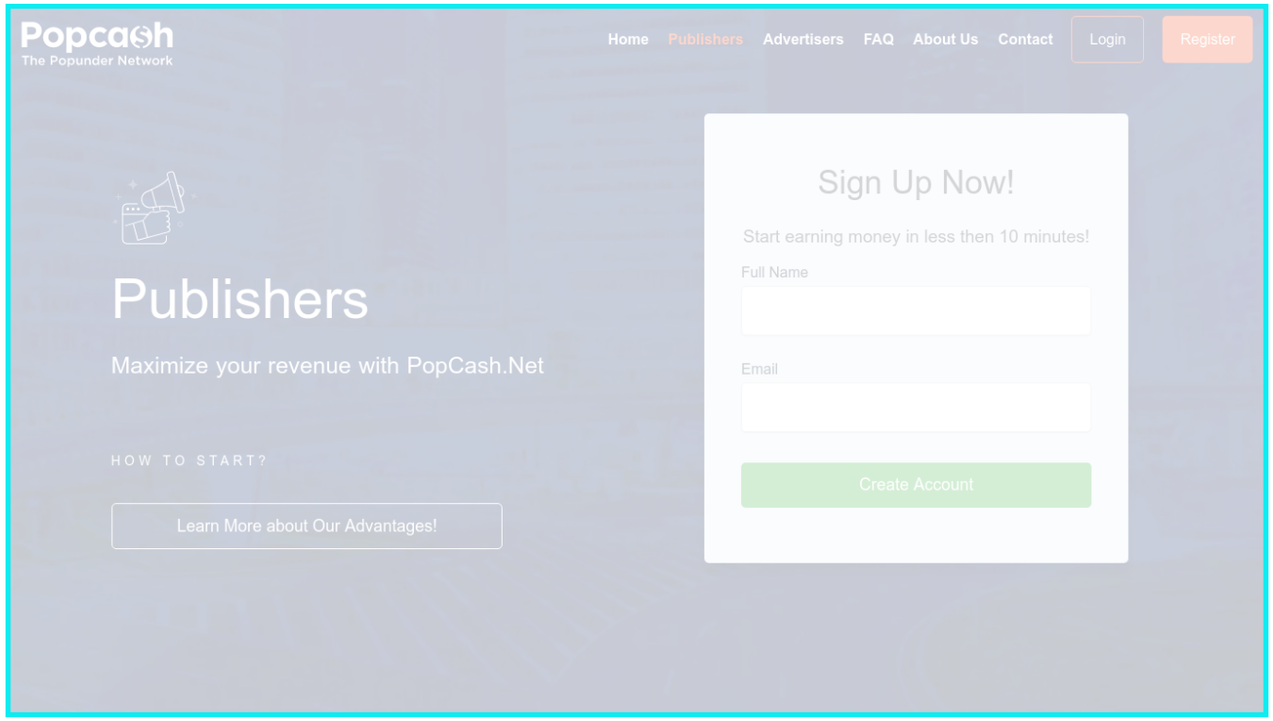
ground truth

Visual approach: Cormier et al.



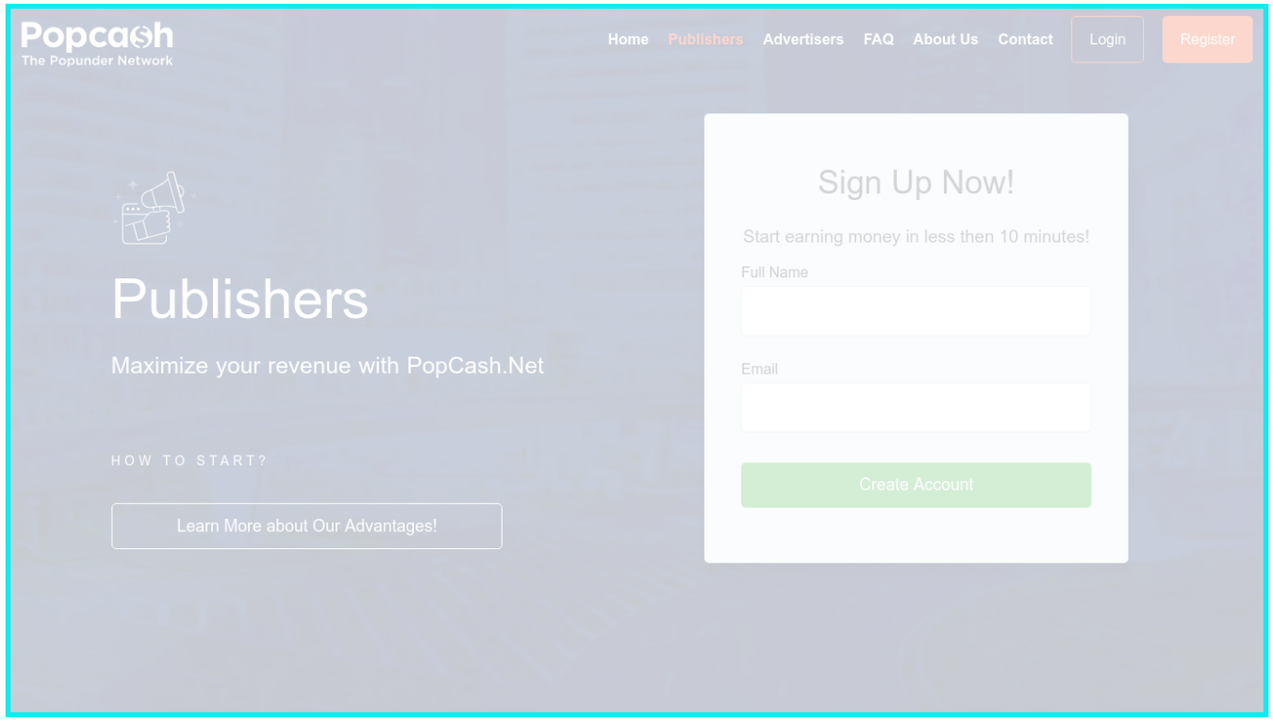
Visual approach: Cormier et al.

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



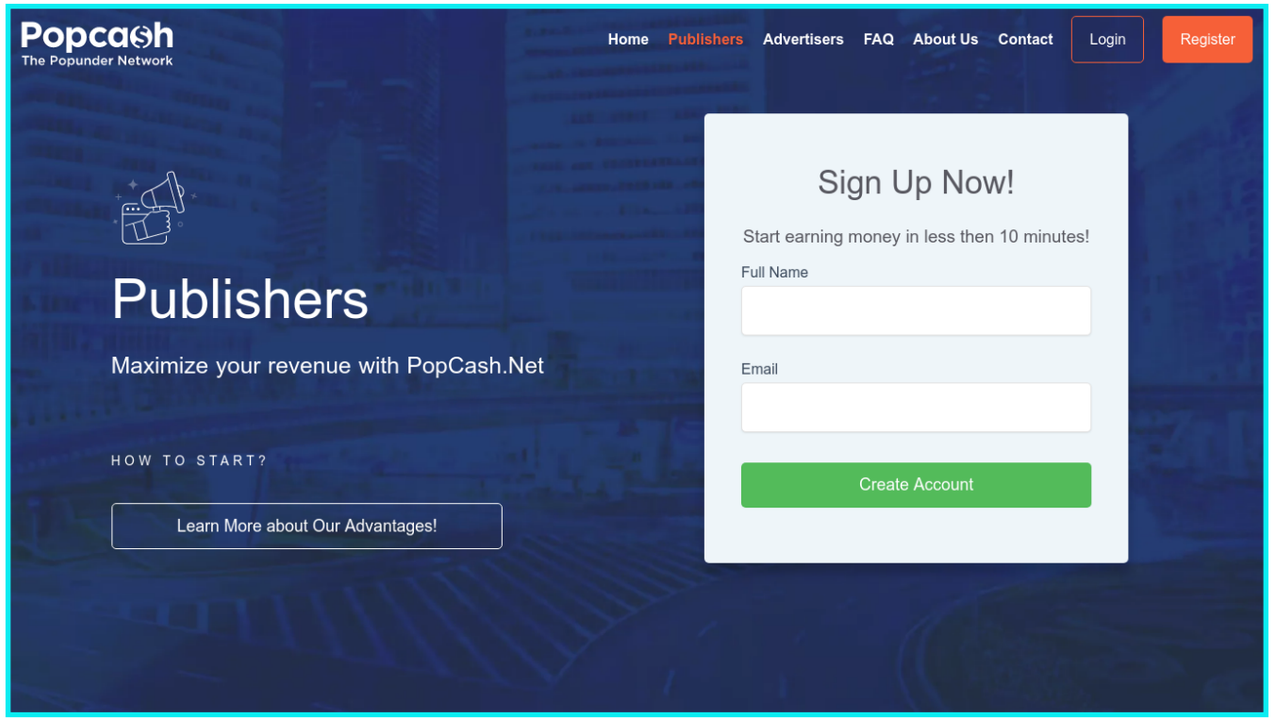
Visual approach: Cormier et al.

- *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)



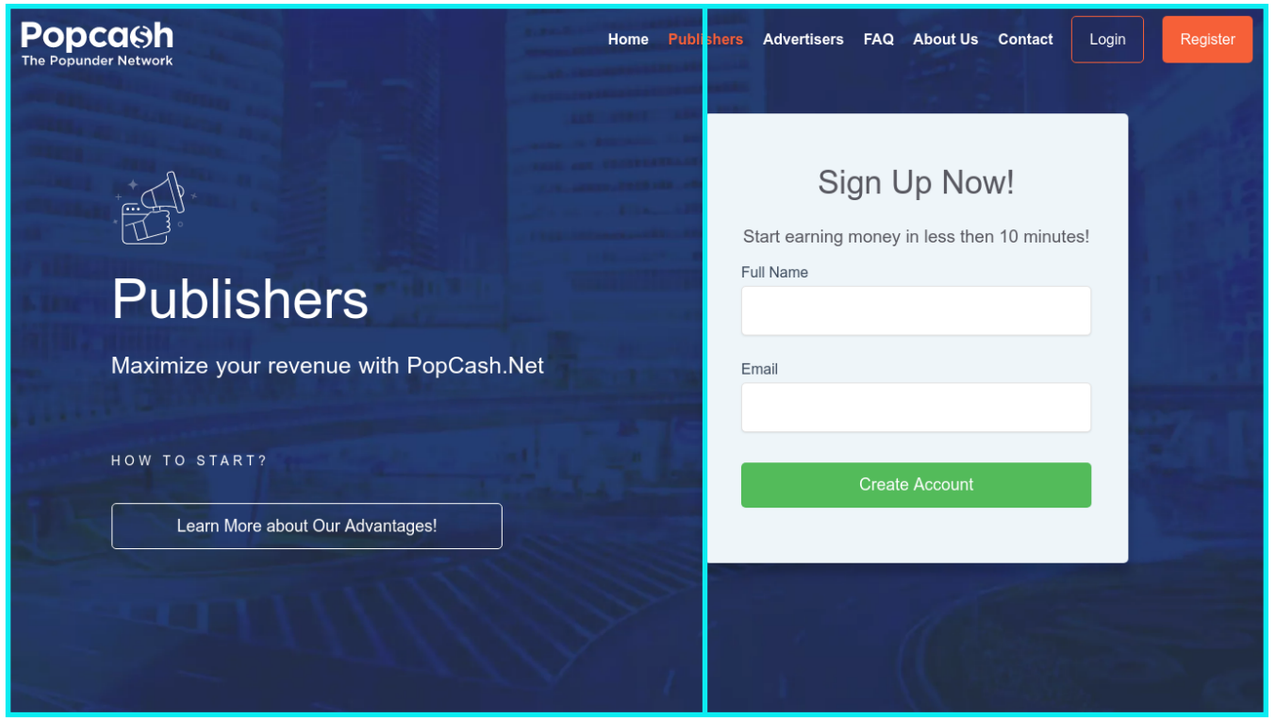
Visual approach: Cormier et al.

- *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)



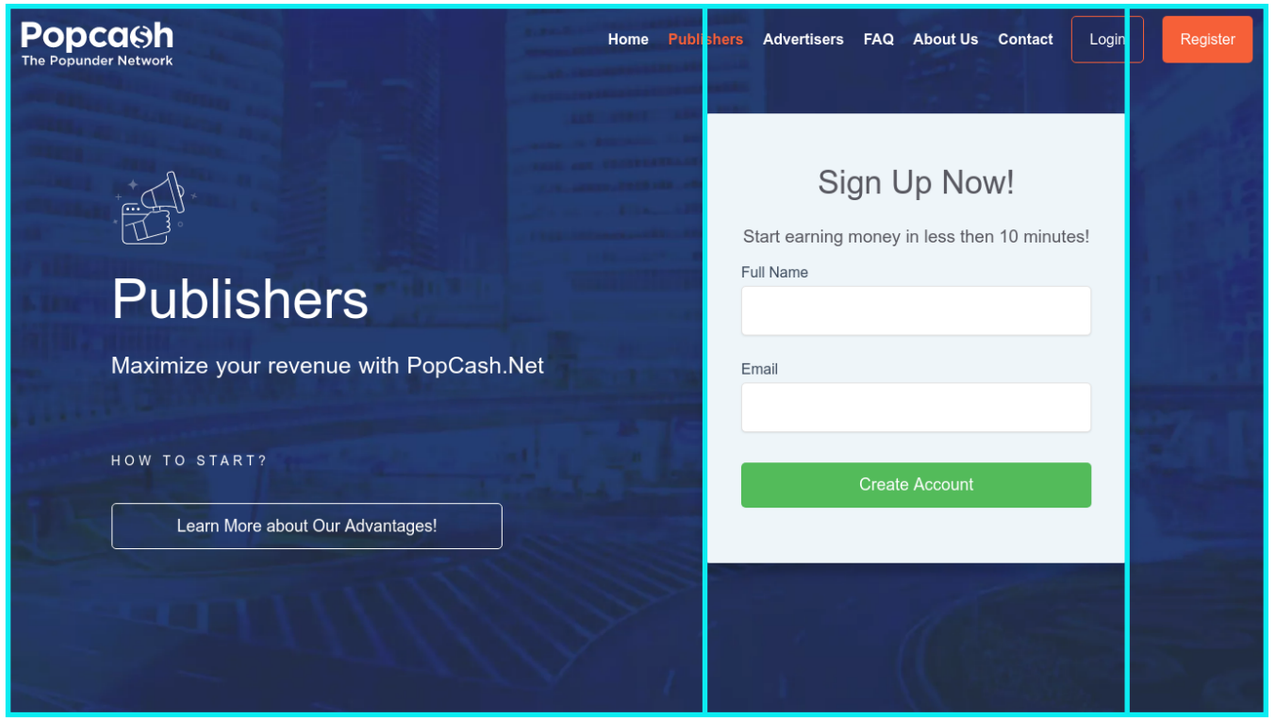
Visual approach: Cormier et al.

- *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)



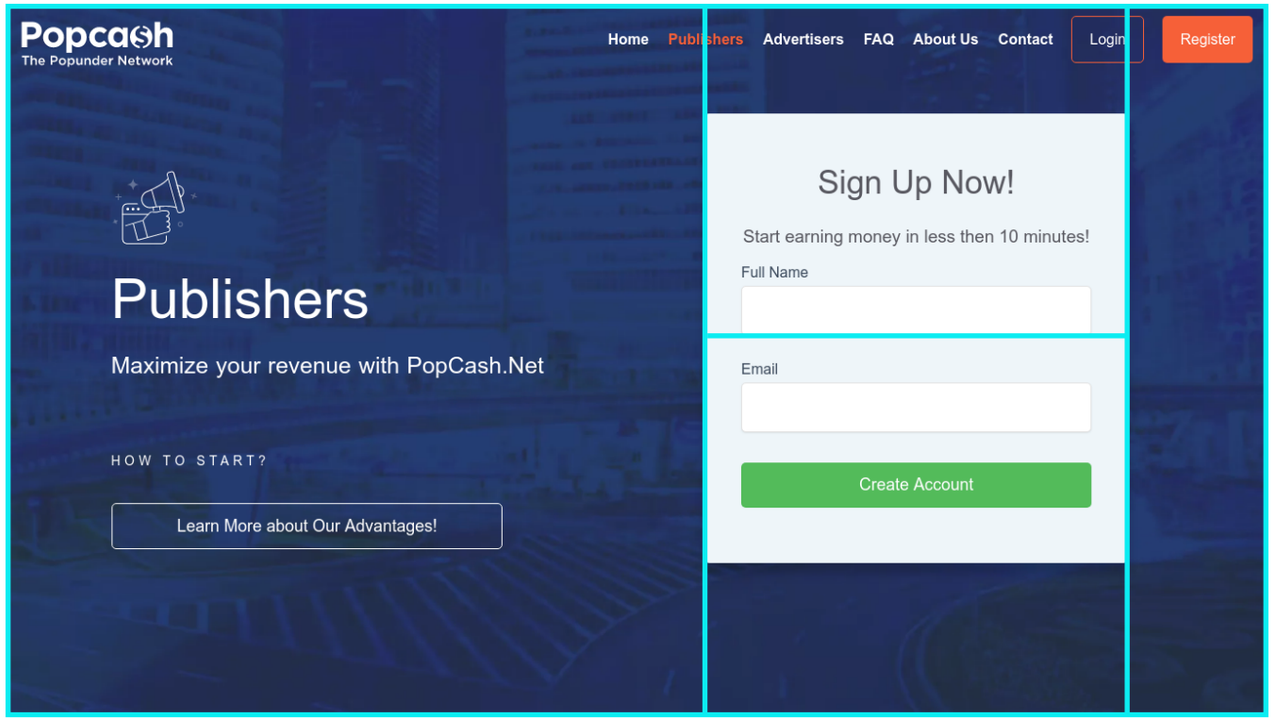
Visual approach: Cormier et al.

- *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)



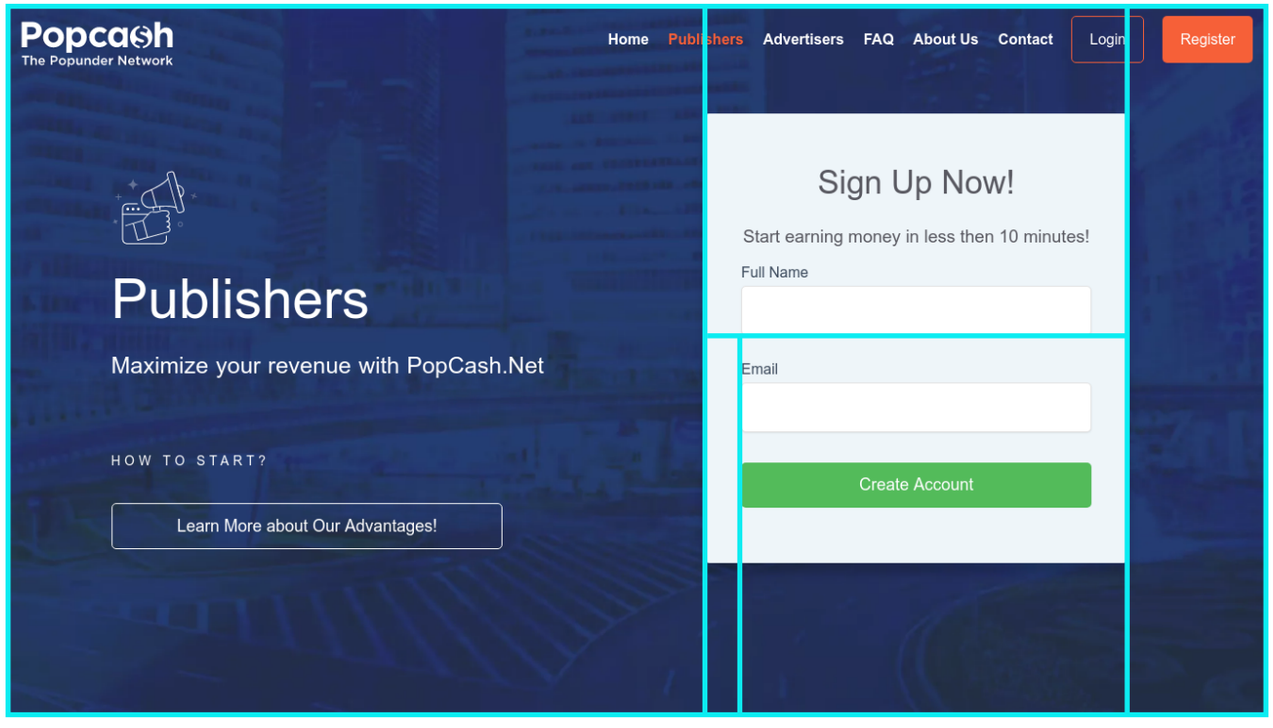
Visual approach: Cormier et al.

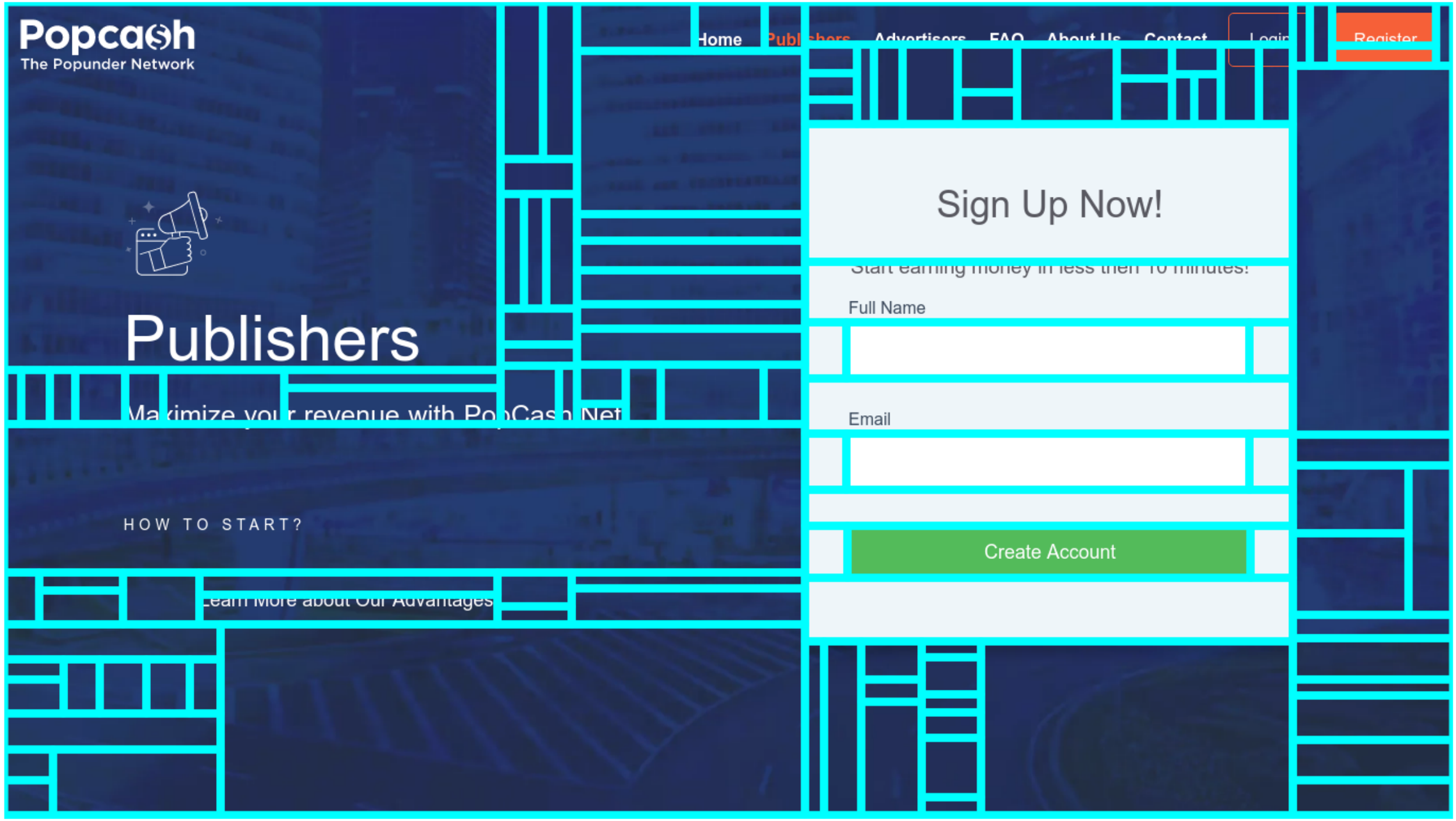
- *probabilistic* algorithm based on edge detection, optimized for *locally significant* edges
- designed to detect *extended lines* (visually non-continuous lines that may form segment borders)



Visual approach: Cormier et al.

- *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)





Cormier et al.

$(s_{min} = 45,\ t_l = 512)$

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_{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

Primary observations:

- Purely visual approach comes close to VIPS' performance

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_{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

Primary observations:

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

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_{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

Primary observations:

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

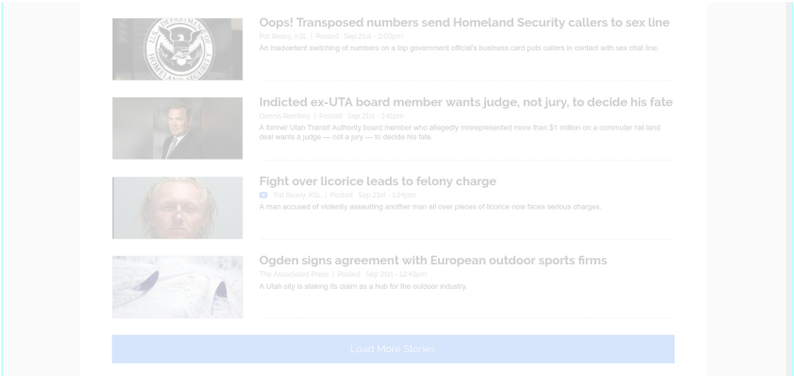
$\longrightarrow$ expresses fundamentally different operation (visual vs. DOM-based)

Results: Cormier et al.

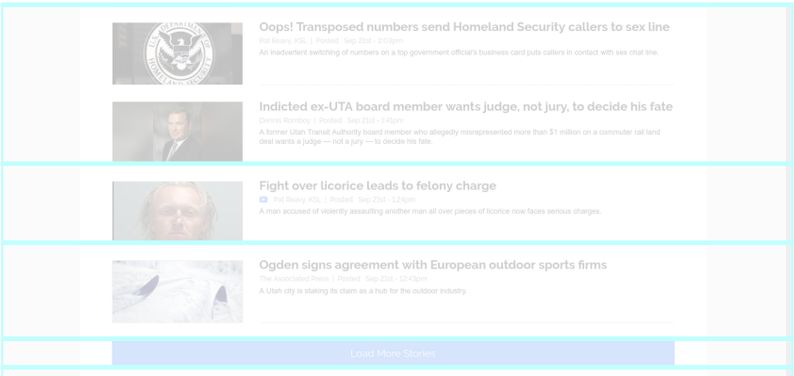
			pixels			*characters*		
	Parameters	*# segments*	P_{B^3}	R_{B^3}	F_{B^3}	P_{B^3}	R_{B^3}	F_{B^3}
(worst)	s_{min} = 90 px t_l = 256 px	18.4	0.29	0.86	0.35	0.60	0.87	0.63
(best)	s_{min} = 45 px t_l = 512 px	38.0	0.34	0.77	0.36	0.67	0.78	0.62
	Δ	+ 19.6	+ 0.05	- 0.09	+ 0.01	+ 0.07	- 0.09	- 0.01
	VIPS (PDoC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

Parameters:

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



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



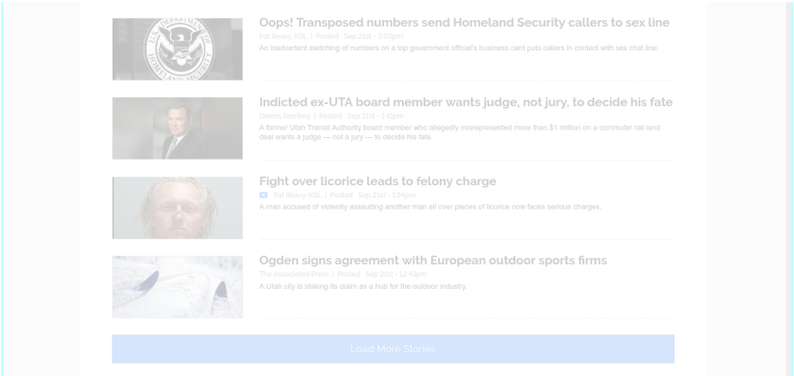
(b) $s_{min} = 45, t_l = 512$ px

Results: Cormier et al.

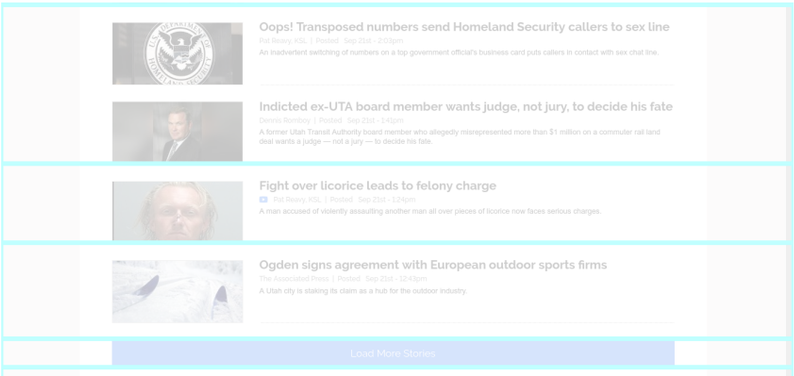
			pixels			*characters*		
	Parameters	*# segments*	P_{B^3}	R_{B^3}	F_{B^3}	P_{B^3}	R_{B^3}	F_{B^3}
(worst)	s_{min} = 90 px t_l = 256 px	18.4	0.29	0.86	0.35	0.60	0.87	0.63
(best)	s_{min} = 45 px t_l = 512 px	38.0	0.34	0.77	0.36	0.67	0.78	0.62
	Δ	+ 19.6	+ 0.05	- 0.09	+ 0.01	+ 0.07	- 0.09	- 0.01
	VIPS (PDoC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

Parameters:

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



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



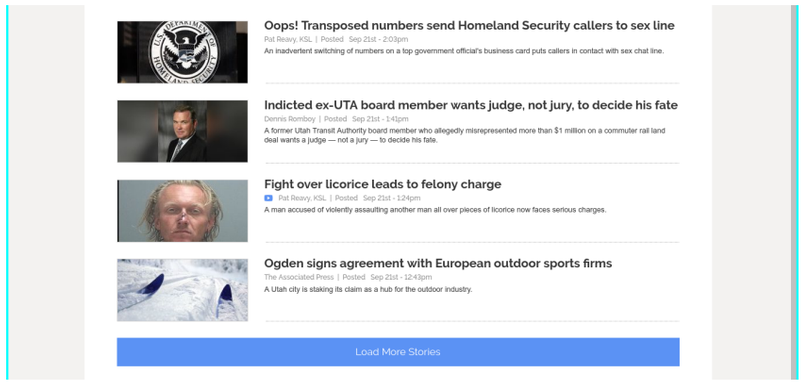
(b) $s_{min} = 45, t_l = 512$ px

Results: Cormier et al.

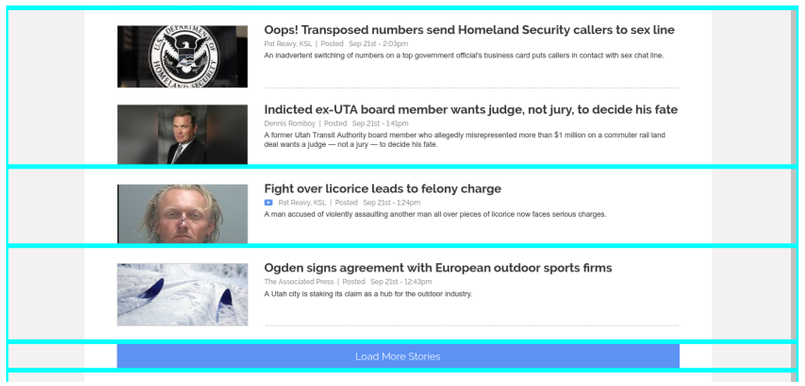
			pixels			*characters*		
	Parameters	*# segments*	P_{B^3}	R_{B^3}	F_{B^3}	P_{B^3}	R_{B^3}	F_{B^3}
(worst)	s_{min} = 90 px t_l = 256 px	18.4	0.29	0.86	0.35	0.60	0.87	0.63
(best)	s_{min} = 45 px t_l = 512 px	38.0	0.34	0.77	0.36	0.67	0.78	0.62
	Δ	+ 19.6	+ 0.05	- 0.09	+ 0.01	+ 0.07	- 0.09	- 0.01
	VIPS (PDoC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

Parameters:

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



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



(b) $s_{min} = 45, t_l = 512$ px

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_{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

(a) $s_{min} = 90, t_l = 256$ px

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

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_{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
- influences segmentation granularity directly

(a) $s_{min} = 90, t_l = 256$ px

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

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_{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
- influences segmentation granularity directly
- *interaction* with t_l

(a) $s_{min} = 90, t_l = 256$ px

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

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_{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
- influences segmentation granularity directly
- *interaction* with t_l

(a) $s_{min} = 90, t_l = 256$ px

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

Results: Cormier et al. - DOM fitting

		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

Results: Cormier et al. - DOM fitting

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

Results: Cormier et al. - DOM fitting

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

Results: Cormier et al. - DOM fitting

		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*: 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)

Results: Cormier et al. - DOM fitting

		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*: 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*

Results: Cormier et al. - DOM fitting

		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

- human segmentations are fit to DOM nodes (containment threshold θ_c = 0.75)

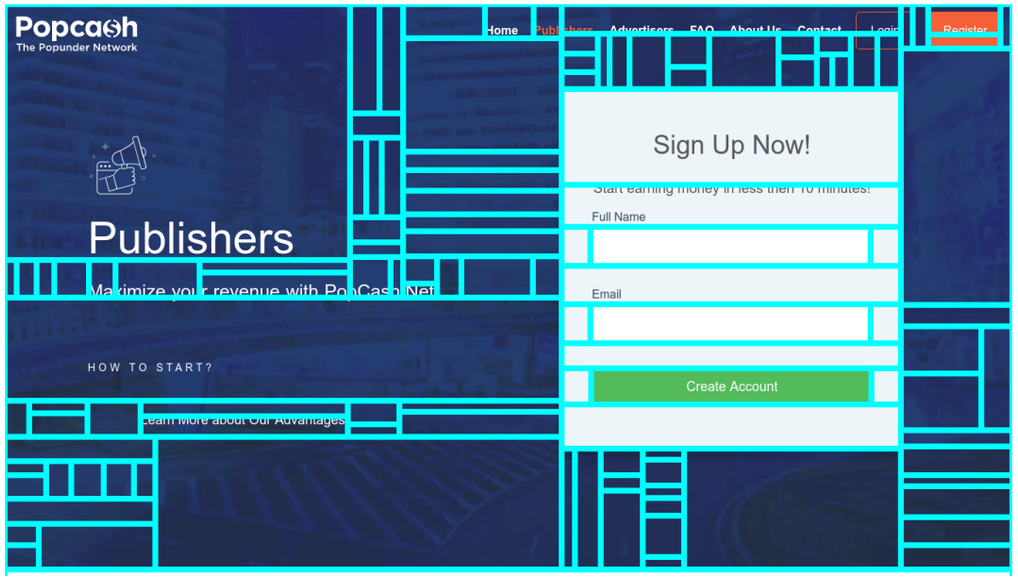
Results: Cormier et al. - DOM fitting

		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

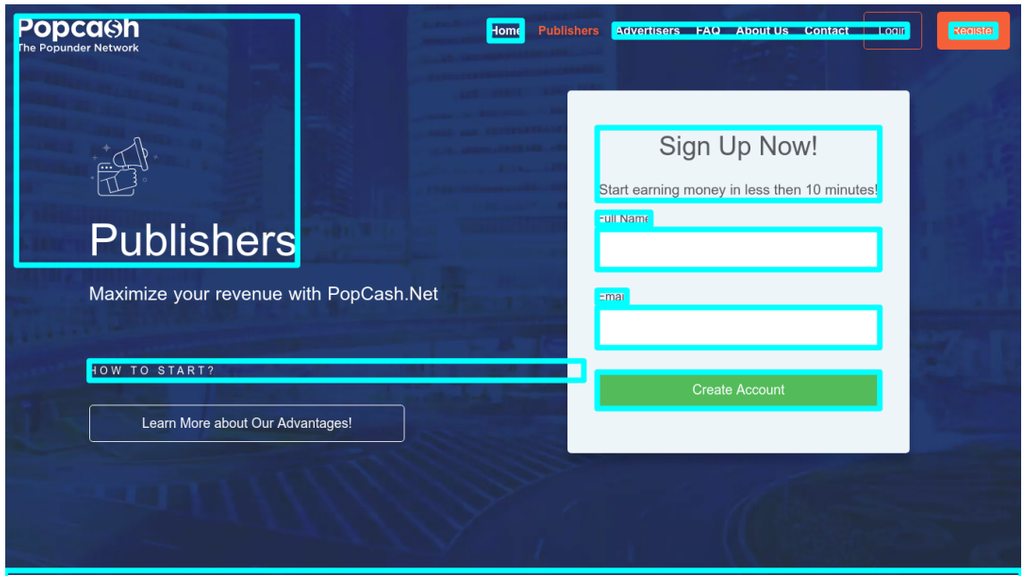
- human segmentations are fit to DOM nodes (containment threshold $\theta_c = 0.75$)

$\longrightarrow$ fair treatment: fit visual/hybrid algorithms to DOM nodes, too

DOM fitting: example



original



fitted

Results: Cormier et al. - DOM fitting

		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*)

Results: Cormier et al. - DOM fitting

		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*

Visual approach: MMDetection

Visual approach: MMDetection

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

Visual approach: MMDetection

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

- ... in fact, not an algorithm by itself
  $\longrightarrow$ machine learning toolkit

Visual approach: MMDetection

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

- ... in fact, not an algorithm by itself
  $\longrightarrow$ machine learning toolkit

- designed for real-world object detection
  and *instance segmentation* (i.e. segmenting real-world images)

Visual approach: MMDetection

- **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)

Visual approach: MMDetection

- **NOT** a Web Page Segmentation algorithm!
- ... in fact, not an algorithm by itself
  $\longrightarrow$ 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

Visual approach: MMDetection

- **NOT** a Web Page Segmentation algorithm!
- ... in fact, not an algorithm by itself
  $\longrightarrow$ 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

Visual approach: MMDetection

- **NOT** a Web Page Segmentation algorithm!
- ... in fact, not an algorithm by itself
  $\longrightarrow$ 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

$\longrightarrow$ transfer to Web Page Segmentation possible?

Results: MMDetection

		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

Results: MMDetection

		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

Results: MMDetection

		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:

Results: MMDetection

		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

Results: MMDetection

		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

Results: MMDetection - DOM fitting

		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

Results: MMDetection - DOM fitting

		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

Results: MMDetection - DOM fitting

		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

Results: MMDetection - DOM fitting

		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

Algorithm cross-evaluation

Algorithm cross-evaluation

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

Algorithm cross-evaluation

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

$\longrightarrow$ possibility of expressing similarity between algorithms

Algorithm cross-evaluation

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

$\longrightarrow$ possibility of expressing similarity between algorithms

F_{B^3}	S			
S^*	**VIPS**	**HEPS**	**Cormier**	**MMDet.**
VIPS	1.00	0.41	**0.51**	0.31
HEPS	0.41	1.00	0.50	0.31
Cormier	**0.51**	0.50	1.00	0.37
MMDet.	0.31	0.31	0.37	1.00

pixels

F_{B^3}	S			
S^*	**VIPS**	**HEPS**	**Cormier**	**MMDet.**
VIPS	1.00	0.48	**0.60**	0.41
HEPS	0.48	1.00	0.43	0.36
Cormier	**0.60**	0.43	1.00	0.40
MMDet.	0.41	0.36	0.40	1.00

characters

Min-vote ensemble

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

Min-vote ensemble

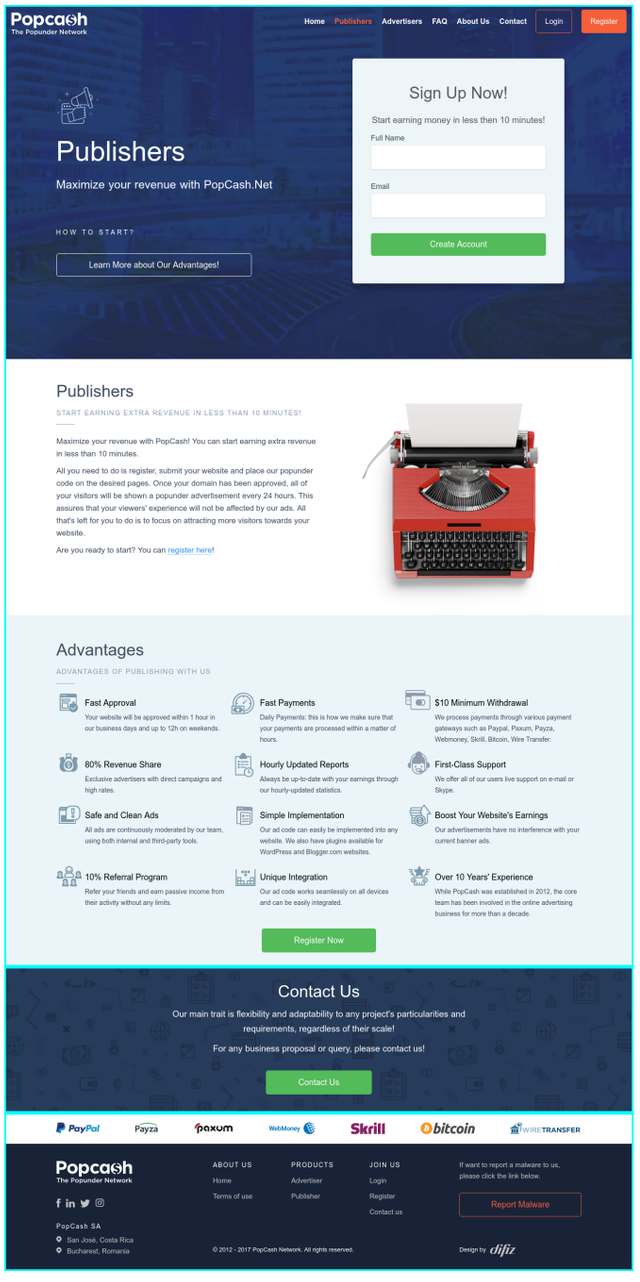
- Applying segmentation fusion (used in ground truth creation) to algorithm segmentations
- *Min-vote@n*, $n \in [1, 2, 3, 4]$ = number of algorithm segmentations that put a given element in a segment

Min-vote ensemble

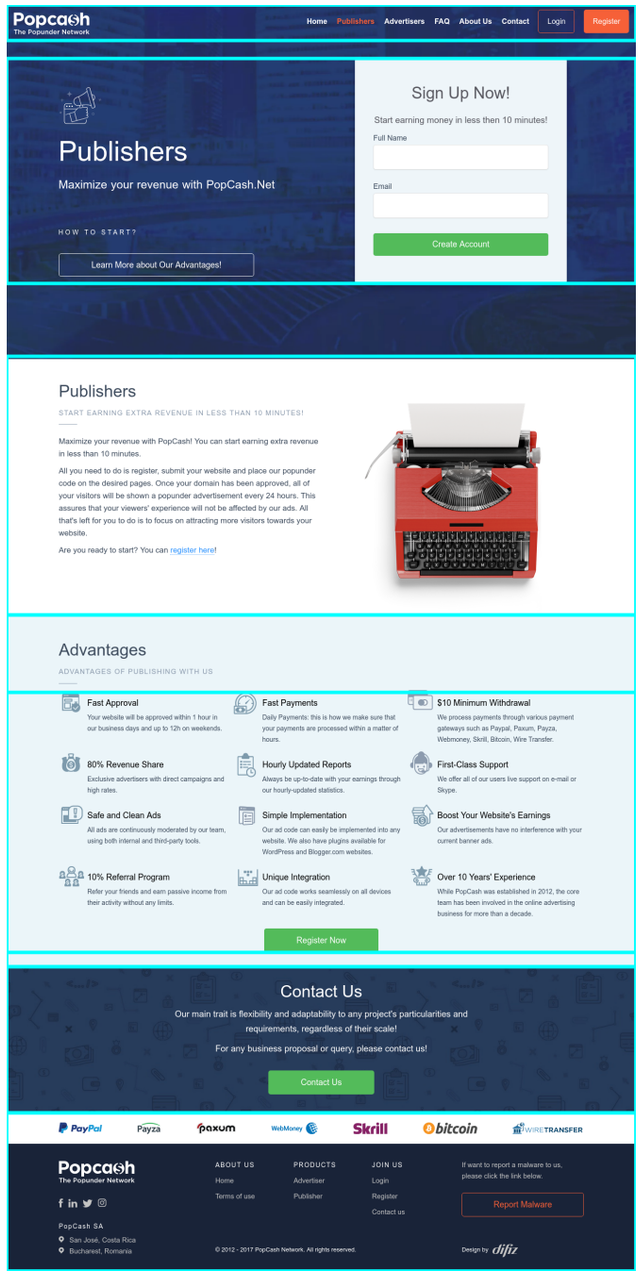
- Applying segmentation fusion (used in ground truth creation) to algorithm segmentations
- *Min-vote@n*, $n \in [1, 2, 3, 4]$ = number of algorithm segmentations that put a given element in a segment
- Initially evaluated for paper with unoptimized parameters

Min-vote ensemble

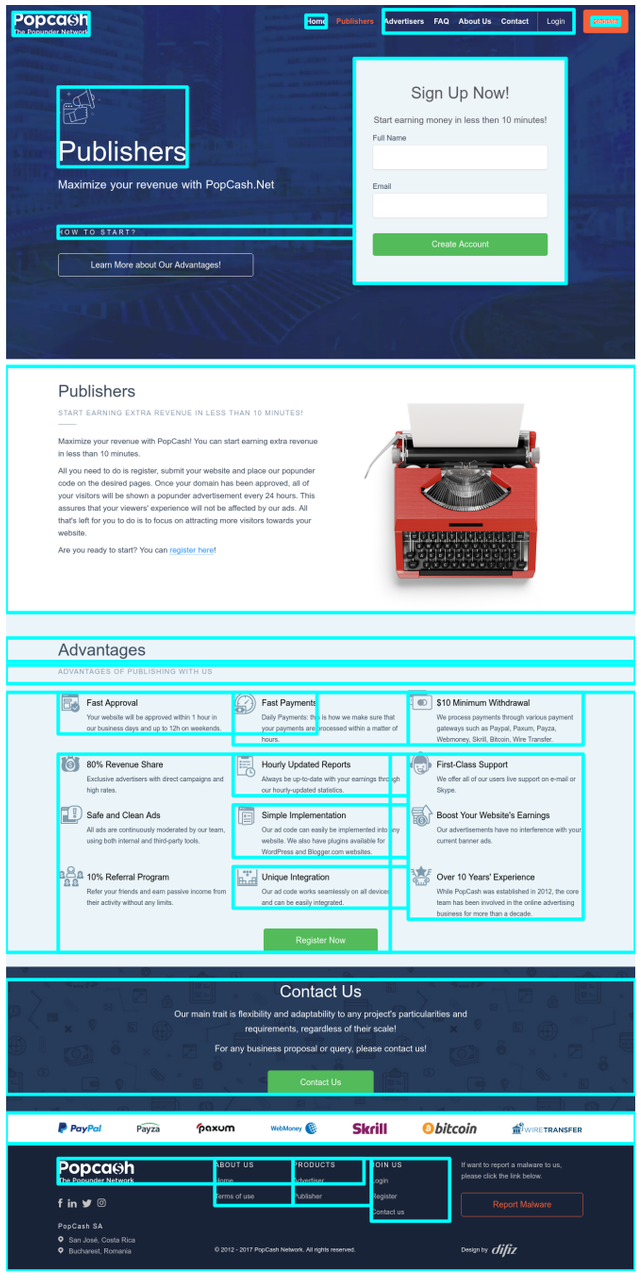
- Applying segmentation fusion (used in ground truth creation) to algorithm segmentations
- *Min-vote@n*, $n \in [1, 2, 3, 4]$ = number of algorithm segmentations that put a given element in a segment
- Initially evaluated for paper with unoptimized parameters
- now: optimized parameters, fitted segmentations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


Popcash
The Popunder Network
Home Publishers Advertisers FAQ About Us Contact Login Register
Publishers
Maximize your revenue with PopCash.Net
HOW TO START?
Learn More about Our Advantages!
Sign Up Now!
Start earning money in less then 10 minutes!
Full Name
Email
Create Account
Publishers
START EARNING EXTRA REVENUE IN LESS THAN 10 MINUTES!
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 desired pages. Once your domain has been approved, all of your visitors will be shown a popunder advertisement every 24 hours. This assures that your viewers' experience 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!
Advantages
ADVANTAGES OF PUBLISHING WITH US
Fast Approval
Your website will be approved within 1 hour in our business days and up to 12h on weekends.
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 Paypal, Paxum, Payza, Webmoney, Skrill, Bitcoin, Wire Transfer.
80% Revenue Share
Exclusive advertisers with direct campaigns and high rates.
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.
Boost Your Website's Earnings
Our advertisements have no interference with your current banner ads.
10% Referral Program
Refer your friends and earn passive income from their activity without any limits.
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
Contact Us
Our main trait is flexibility and adaptability to any project's particularities and requirements, regardless of their scale!
For any business proposal or query, please contact us!
Contact Us
PayPal Payza paxum WebMoney Skrill bitcoin WIRETRANSFER
Popcash
The Popunder Network
ABOUT US
Home
Terms of use
PRODUCTS
Advertiser
Publisher
JOIN US
Login
Register
Contact us
If want to report a malware to us, please click the link below.
Report Malware
PopCash SA
San José, Costa Rica
Bucharest, Romania
© 2012 - 2017 PopCash Network. All rights reserved.
Design by difiz

$n = 1$


Popcash
The Popunder Network
Home Publishers Advertisers FAQ About Us Contact Login Register
Publishers
Maximize your revenue with PopCash.Net
HOW TO START?
Learn More about Our Advantages!
Sign Up Now!
Start earning money in less then 10 minutes!
Full Name
Email
Create Account
Publishers
START EARNING EXTRA REVENUE IN LESS THAN 10 MINUTES!
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 desired pages. Once your domain has been approved, all of your visitors will be shown a popunder advertisement every 24 hours. This assures that your viewers' experience 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!
Advantages
ADVANTAGES OF PUBLISHING WITH US
Fast Approval
Your website will be approved within 1 hour in our business days and up to 12h on weekends.
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 Paypal, Paxum, Payza, Webmoney, Skrill, Bitcoin, Wire Transfer.
80% Revenue Share
Exclusive advertisers with direct campaigns and high rates.
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.
Boost Your Website's Earnings
Our advertisements have no interference with your current banner ads.
10% Referral Program
Refer your friends and earn passive income from their activity without any limits.
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
Contact Us
Our main trait is flexibility and adaptability to any project's particularities and requirements, regardless of their scale!
For any business proposal or query, please contact us!
Contact Us
PayPal Payza paxum WebMoney Skrill bitcoin WIRETRANSFER
Popcash
The Popunder Network
ABOUT US
Home
Terms of use
PRODUCTS
Advertiser
Publisher
JOIN US
Login
Register
Contact us
If want to report a malware to us, please click the link below.
Report Malware
PopCash SA
San José, Costa Rica
Bucharest, Romania
© 2012 - 2017 PopCash Network. All rights reserved.
Design by difiz

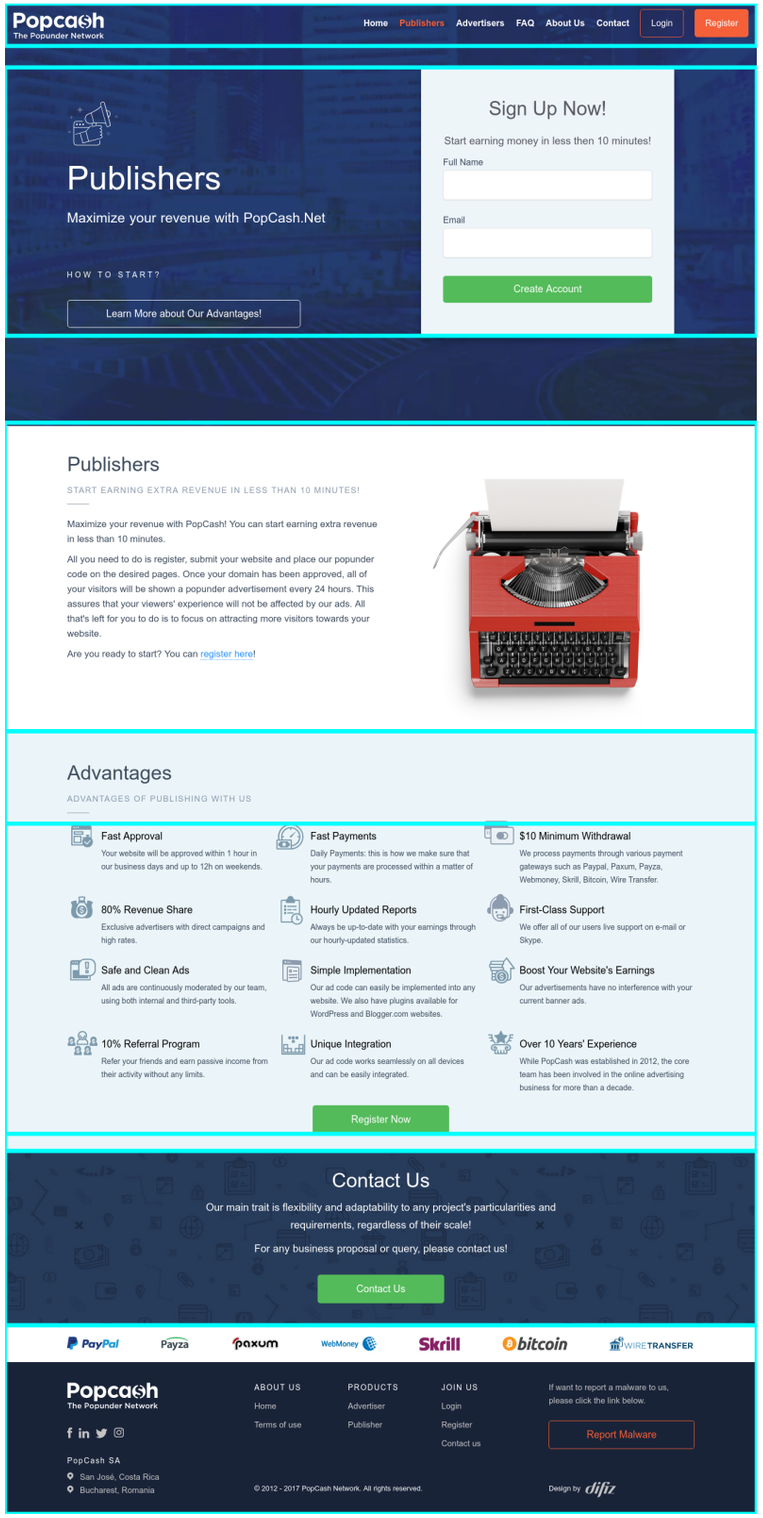
$n = 2$


Popcash
The Popunder Network
Home Publishers Advertisers FAQ About Us Contact Login Register
Publishers
Maximize your revenue with PopCash.Net
HOW TO START?
Learn More about Our Advantages!
Sign Up Now!
Start earning money in less then 10 minutes!
Full Name
Email
Create Account
Publishers
START EARNING EXTRA REVENUE IN LESS THAN 10 MINUTES!
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 desired pages. Once your domain has been approved, all of your visitors will be shown a popunder advertisement every 24 hours. This assures that your viewers' experience 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!
Advantages
ADVANTAGES OF PUBLISHING WITH US
Fast Approval
Your website will be approved within 1 hour in our business days and up to 12h on weekends.
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 Paypal, Paxum, Payza, Webmoney, Skrill, Bitcoin, Wire Transfer.
80% Revenue Share
Exclusive advertisers with direct campaigns and high rates.
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.
Boost Your Website's Earnings
Our advertisements have no interference with your current banner ads.
10% Referral Program
Refer your friends and earn passive income from their activity without any limits.
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
Contact Us
Our main trait is flexibility and adaptability to any project's particularities and requirements, regardless of their scale!
For any business proposal or query, please contact us!
Contact Us
PayPal Payza paxum WebMoney Skrill bitcoin WIRETRANSFER
Popcash
The Popunder Network
ABOUT US
Home
Terms of use
PRODUCTS
Advertiser
Publisher
JOIN US
Login
Register
Contact us
If want to report a malware to us, please click the link below.
Report Malware
PopCash SA
San José, Costa Rica
Bucharest, Romania
© 2012 - 2017 PopCash Network. All rights reserved.
Design by difiz

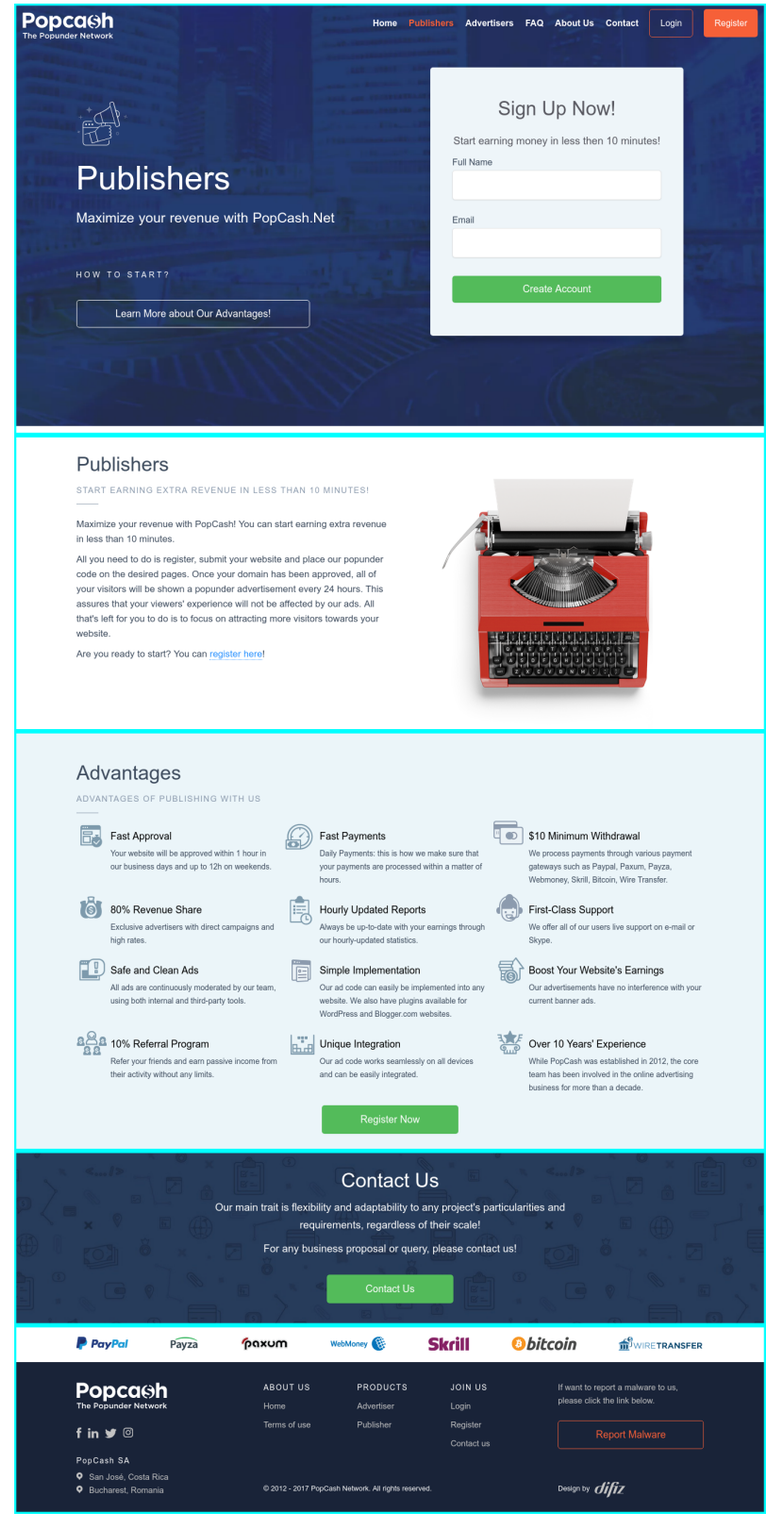
$n = 3$


Popcash
The Popunder Network
Home Publishers Advertisers FAQ About Us Contact Login Register
Publishers
Maximize your revenue with PopCash.Net
Learn More about Our Advantages!
Sign Up Now!
Start earning money in less then 10 minutes!
Full Name
Email
Create Account
Publishers
START EARNING EXTRA REVENUE IN LESS THAN 10 MINUTES!
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 desired pages. Once your domain has been approved, all of your visitors will be shown a popunder advertisement every 24 hours. This assures that your viewers' experience 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!
Advantages
Fast Approval
Your website will be approved within 1 hour in our business days and up to 12h on weekends.
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 Paypal, Paxum, Payza, Webmoney, Skrill, Bitcoin, Wire Transfer.
80% Revenue Share
Exclusive advertisers with direct campaigns and high rates.
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.
Boost Your Website's Earnings
Our advertisements have no interference with your current banner ads.
10% Referral Program
Refer your friends and earn passive income from their activity without any limits.
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
Contact Us
Our main trait is flexibility and adaptability to any project's particularities and requirements, regardless of their scale!
For any business proposal or query, please contact us!
Contact Us
PayPal Payza paxum WebMoney Skrill bitcoin WIRETRANSFER
Popcash
The Popunder Network
ABOUT US
Home
Terms of use
PRODUCTS
Advertiser
Publisher
JOIN US
Login
Register
Contact us
Report Malware
PopCash SA
San José, Costa Rica
Bucharest, Romania
© 2012 - 2017 PopCash Network. All rights reserved.
Design by difiz

$n = 4$



$n = 2$



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**

Summary - Ranking

		pixels			*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. (**fitted**)	7.0	0.56	0.39	0.26	0.66	0.48	0.42

Summary - Ranking

		pixels			*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. (**fitted**)	7.0	0.56	0.39	0.26	0.66	0.48	0.42

Summary - Ranking

		pixels			*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. (**fitted**)	7.0	0.56	0.39	0.26	0.66	0.48	0.42

Summary - Ranking

		pixels			*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. (**fitted**)	7.0	0.56	0.39	0.26	0.66	0.48	0.42

Summary - Ranking

		pixels			*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. (**fitted**)	7.0	0.56	0.39	0.26	0.66	0.48	0.42

Summary - Ranking

		pixels			*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. (**fitted**)	7.0	0.56	0.39	0.26	0.66	0.48	0.42

Summary: Contributions

Summary: Contributions

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

Summary: Contributions

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

Summary: Contributions

- 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:

Summary: Contributions

- 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

Summary: Contributions

- 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

Summary: Contributions

- 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

Outlook

Outlook

- There are more approaches to be evaluated!

Outlook

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

Outlook

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

Outlook

- 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!