Segmentação de Captchas

Primeiro Seminário

BCC448

Reconhecimento de Padrões

Alunos:

Filipe Eduardo Mata dos Santos
Pedro Henrique Lopes Silva

Paper

- "Text-based CAPTCHA strengths and weaknesses.“
- Elie Bursztein, Matthieu Martin e John Mitchell
- *Proceedings of the 18th ACM conference on Computer and communications security*.
- ACM, 2011

Captcha

- "Completely Atomated Public Turing tests to tell Computers and Humans Apart"
- "Reverse Turing tests"
- Uses

Real-world Captcha Security Features

- Anti-recognition
- Anti-segmentation

Real-world Captcha Security Features

- Anti-recognition
  1. **Multi-fonts**

Real-world Captcha Security Features

- Anti-recognition
  1. Multi-fonts
  2. **Charset**

Real-world Captcha Security Features

- Anti-recognition
  1. Multi-fonts
  2. Charset
  3. **Font Size**

Real-world Captcha Security Features

- Anti-recognition
  1. Multi-fonts
  2. Charset
  3. Font Size
  4. **Distortion**

Real-world Captcha Security Features

- Anti-recognition
  1. Multi-fonts
  2. Charset
  3. Font Size
  4. Distortion
  5. **Blurring**

Real-world Captcha Security Features

- Anti-recognition
  1. Multi-fonts
  2. Charset
  3. Font Size
  4. Distortion
  5. Blurring
  6. **Tilting**

Real-world Captcha Security Features

- Anti-recognition
  1. Multi-fonts
  2. Charset
  3. Font Size
  4. Distortion
  5. Blurring
  6. Tilting
  7. **Waving**

Real-world Captcha Security Features

- Anti-segmentation
  8. **Complex Background**

Real-world Captcha Security Features

- Anti-segmentation
  8. Complex Background
  9. **Lines**

Real-world Captcha Security Features

- Anti-segmentation
  8. Complex Background
  9. Lines
  10. **Collapsing**

Segmentation

- Background Confusion
  - Complex Background

Segmentation

- Background Confusion
  - Complex Background
  - Color Similarity

Segmentation

- Background Confusion
  - Complex Background
  - Color Similarity
  - Noise*

Segmentation

- Using Lines
  - Small Lines

Segmentation

- Using Lines
  - Small Lines
  - Big Lines

Segmentation

- Using Lines
  - Small Lines
  - Big Lines
- Collapsing
  - Predictable Collapsing

Segmentation

- Using Lines
  - Small Lines
  - Big Lines
- Collapsing
  - Predictable Collapsing
  - Unpredictable Collapsing

Data Set

- **Authorize(50%)**

Data Set

- Authorize(50%)
- **Baidu(1-10%)**

Baidu

[Baidu]

Data Set

- Authorize(50%)
- Baidu(1-10%)
- **Blizzard(50%)**

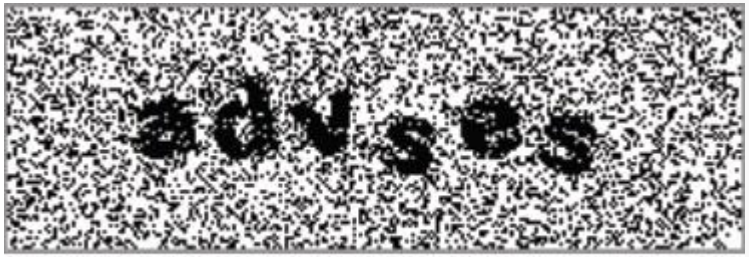
Blizzard

[Blizzard]

Data Set

- Authorize(50%)
- Baidu(1-10%)
- Blizzard(50%)
- **Captcha.net(50%)**

Captcha.net



[Captcha.net]

Data Set

- Authorize(50%)
- Baidu(1-10%)
- Blizzard(50%)
- Captcha.net(50%)
- **CNN(10-24%)**

CNN

[CNN]

Data Set

- Authorize(50%)
- Baidu(1-10%)
- Blizzard(50%)
- Captcha.net(50%)
- CNN(10-24%)
- **Digg(10-24%)**

Digg

[Digg]

Data Set

- Authorize(50%)
- Baidu(1-10%)
- Blizzard(50%)
- Captcha.net(50%)
- CNN(10-24%)
- Digg(10-24%)
- **eBay(25-49%)**

eBay

[eBay]

Data Set

- Authorize(50%)
- Baidu(1-10%)
- Blizzard(50%)
- Captcha.net(50%)
- CNN(10-24%)
- Digg(10-24%)
- eBay(25-49%)
- **Google(0%)**

Google

Data Set

- Authorize(50%)
- Baidu(1-10%)
- Blizzard(50%)
- Captcha.net(50%)
- CNN(10-24%)
- Digg(10-24%)
- eBay(25-49%)
- Google(0%)
- **Megaupload (50%)**

Megaupload

[Megaupload]

Data Set

- Authorize(50%)
- Baidu(1-10%)
- Blizzard(50%)
- Captcha.net(50%)
- CNN(10-24%)
- Digg(10-24%)
- eBay(25-49%)
- Google(0%)
- Megaupload(50%)
- **NIH(50%)**

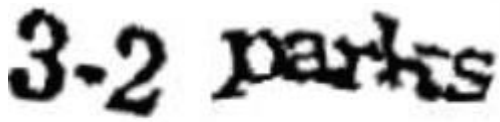
NIH

[NIH]

Data Set

- Authorize(50%)
- Baidu(1-10%)
- Blizzard(50%)
- Captcha.net(50%)
- CNN(10-24%)
- Digg(10-24%)
- eBay(25-49%)
- Google(0%)
- Megaupload(50%)
- NIH(50%)
- **Recaptcha(0%)**

Recaptcha



[Recaptcha]

Data Set

- Authorize(50%)
- Baidu(1-10%)
- Blizzard(50%)
- Captcha.net(50%)
- CNN(10-24%)
- Digg(10-24%)
- eBay(25-49%)
- Google(0%)
- Megaupload(50%)
- NIH(50%)
- Recaptcha(0%)
- **Reddit(25-49%)**

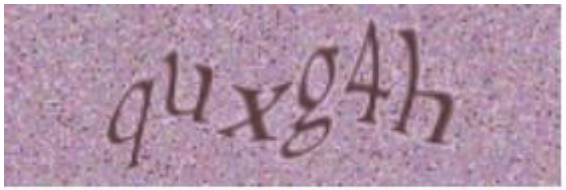
Reddit

[Reddit]

Data Set

- Authorize(50%)
- Baidu(1-10%)
- Blizzard(50%)
- Captcha.net(50%)
- CNN(10-24%)
- Digg(10-24%)
- eBay(25-49%)
- Google(0%)
- Megaupload(50%)
- NIH(50%)
- Recaptcha(0%)
- Reddit(25-49%)
- **Skyrock(1-10%)**

Skyrock



[Skyrock]

Data Set

- Authorize(50%)
- Baidu(1-10%)
- Blizzard(50%)
- Captcha.net(50%)
- CNN(10-24%)
- Digg(10-24%)
- eBay(25-49%)
- Google(0%)
- Megaupload(50%)
- NIH(50%)
- Recaptcha(0%)
- Reddit(25-49%)
- Skyrock(1-10%)
- **Slashdot(25-49%)**

Slashdot

[Slashdot]

Data Set

- Authorize(50%)
- Baidu(1-10%)
- Blizzard(50%)
- Captcha.net(50%)
- CNN(10-24%)
- Digg(10-24%)
- eBay(25-49%)
- Google(0%)
- Megaupload(50%)
- NIH(50%)
- Recaptcha(0%)
- Reddit(25-49%)
- Skyrock(1-10%)
- Slashdot(25-49%)
- **Wikipedia(25-49%)**

Wikipedia

Data Set

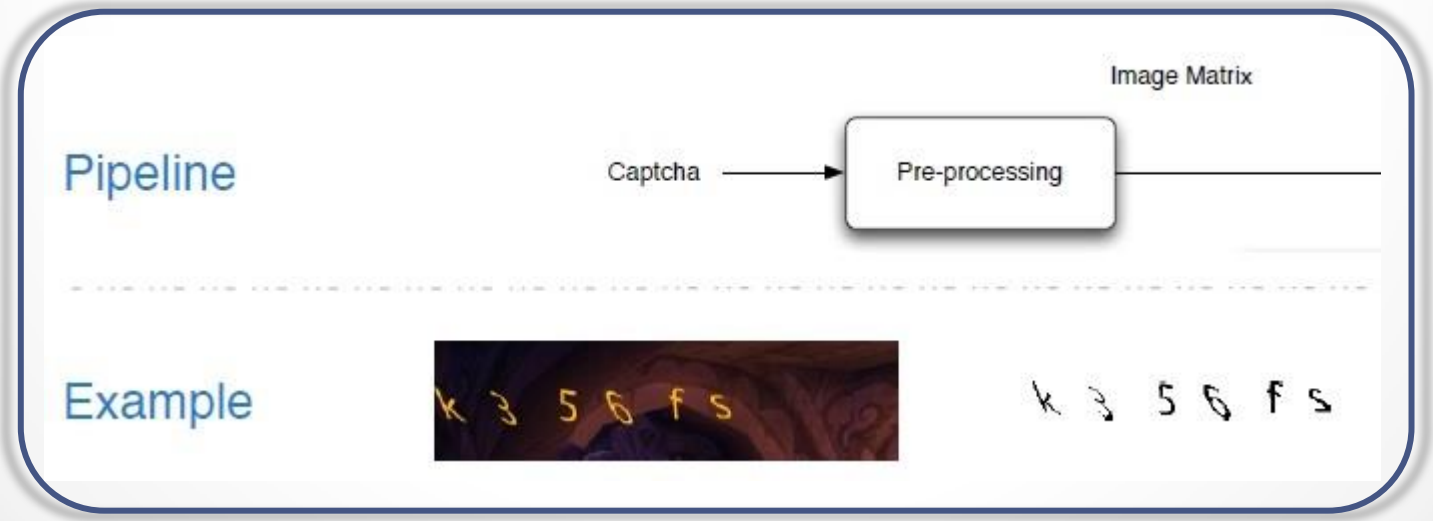
- Authorize(50%)
- Baidu(1-10%)
- Blizzard(50%)
- Captcha.net(50%)
- CNN(10-24%)
- Digg(10-24%)
- eBay(25-49%)
- Google(0%)
- Megaupload(50%)
- NIH(50%)
- Recaptcha(0%)
- Reddit(25-49%)
- Skyrock(1-10%)
- Slashdot(25-49%)
- Wikipedia(25-49%)

Decaptcha

- Code in C#
  - Speed
  - Robustness
  - Availability of AI/Vision Libraries
- Visual Studio
- Framework
  - aForge
  - Accord
- Machine Learning
  - SVM

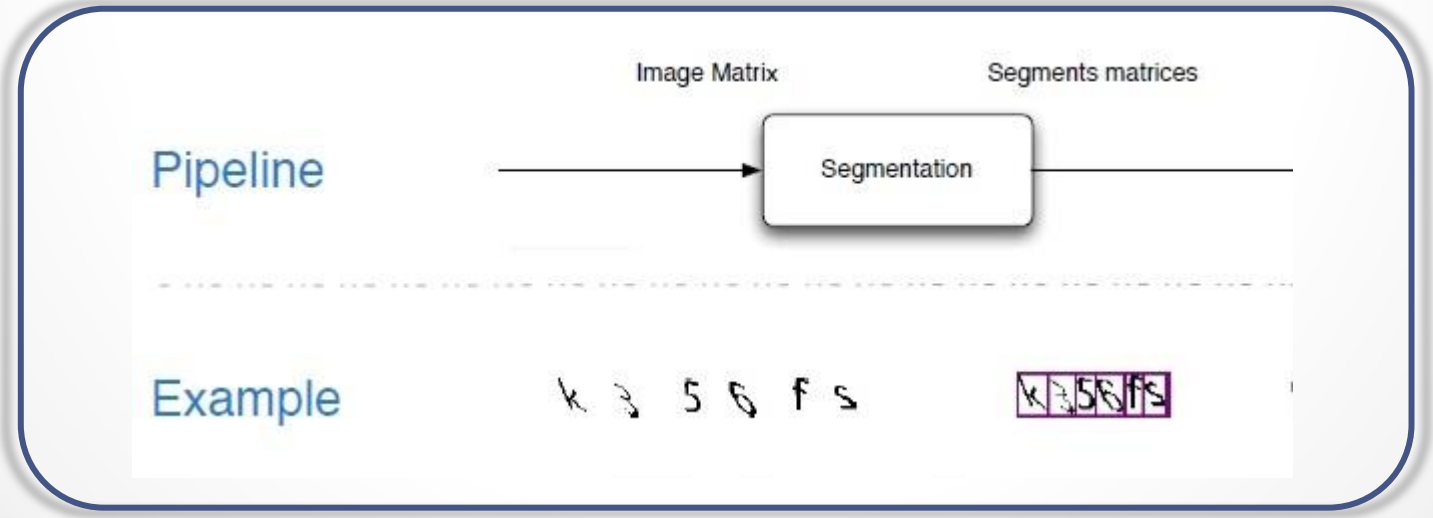
Decaptcha Pipeline

- Method
  1. Preprocessing



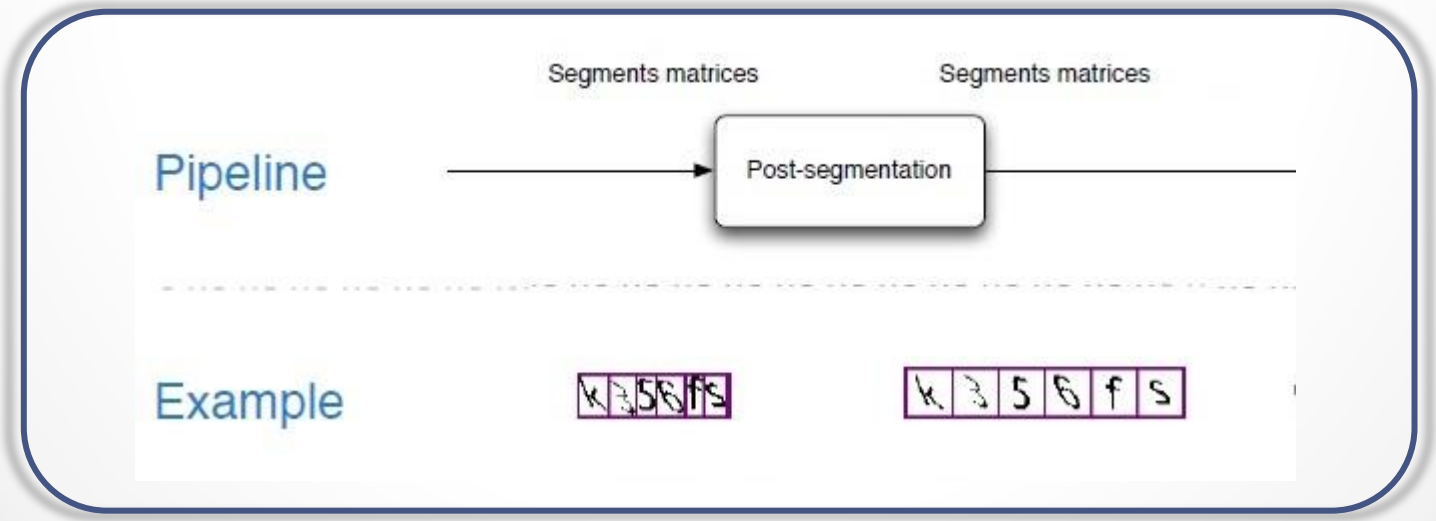
Decaptcha Pipeline

- Method
  1. Preprocessing
  2. Segmentation



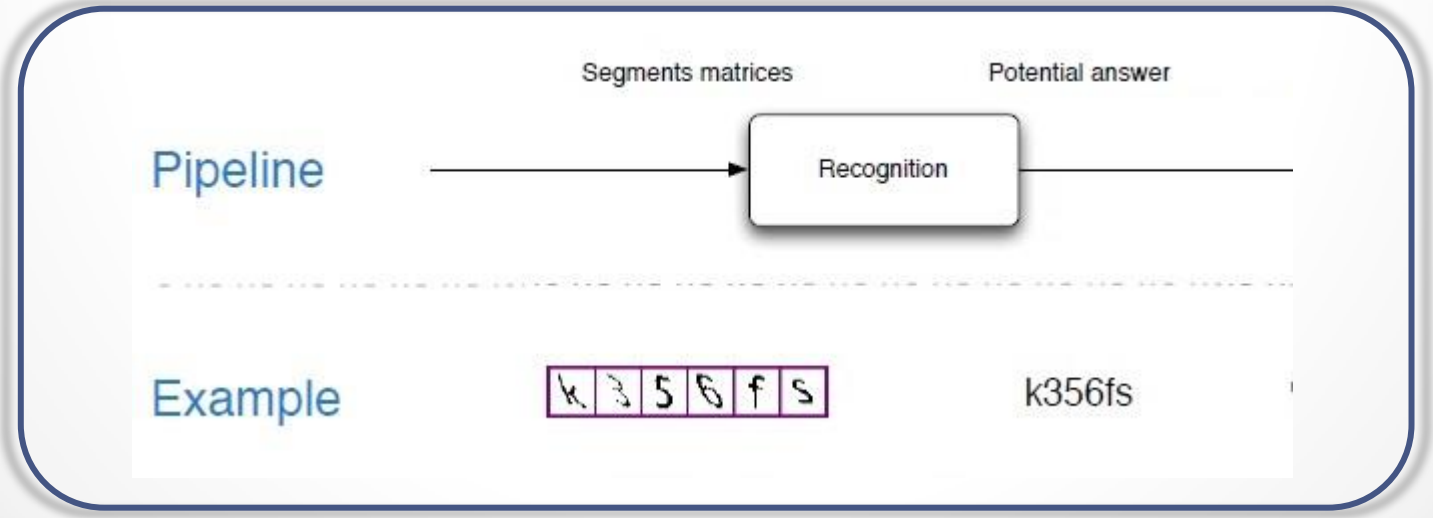
Decaptcha Pipeline

- Method
  1. Preprocessing
  2. Segmentation
  3. Post-segmentation



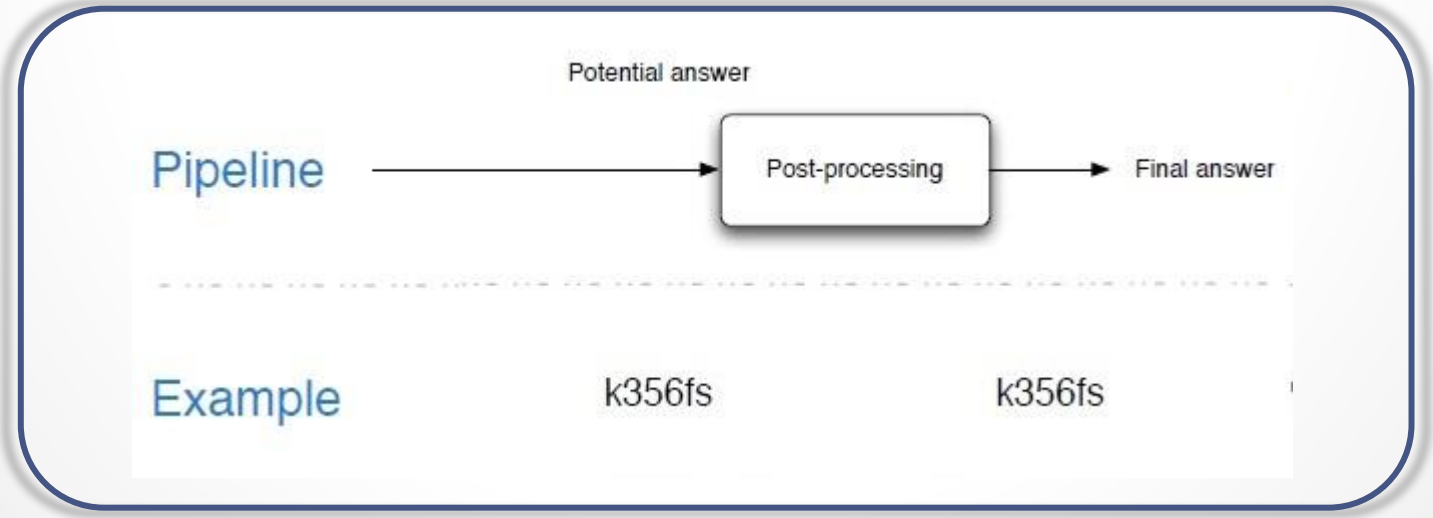
Decaptcha Pipeline

- Method
  1. Preprocessing
  2. Segmentation
  3. Post-segmentation
  4. Recognition



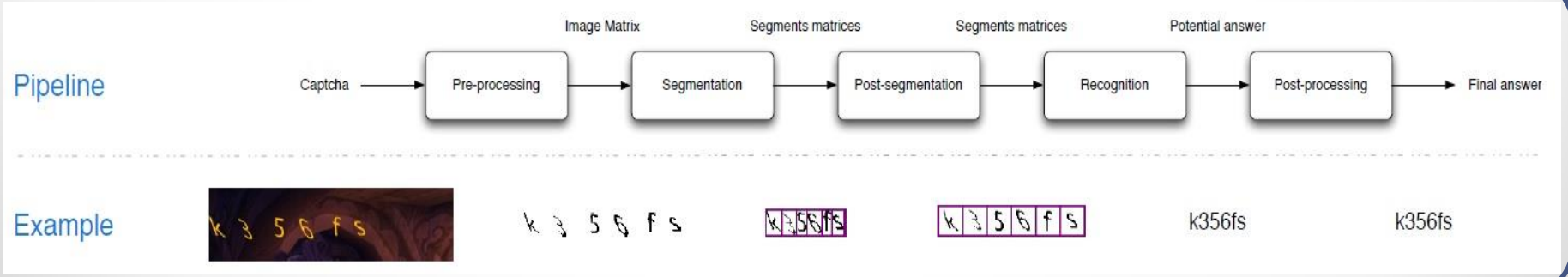
Decaptcha Pipeline

- Method
  1. Preprocessing
  2. Segmentation
  3. Post-segmentation
  4. Recognition
  5. Post-processing



Decaptcha Pipeline

- Method
  1. Preprocessing
  2. Segmentation
  3. Post-segmentation
  4. Recognition
  5. Post-processing



References

- Bursztein, Elie, Matthieu Martin, and John Mitchell. "Text-based CAPTCHA strengths and weaknesses." *Proceedings of the 18th ACM conference on Computer and communications security*. ACM, 2011