Text recognition on images from social media

Presenter: Belyaeva Oksana Vladimirovna
Authors: Akopyan M. S., Belyaeva O. V., Plechov T. P., Turdakov D. Y.

September, 2019

Introduction

Optical character recognition (OCR) — translation of handwritten, printed text from images into an editable representation of characters on a computer (for example, in a text editor).
OCR on social networks is used for:

- analysis of emotional color;
- analysis of user requirements and wishes, etc.

Problem

Input data - images from social networks, in many cases characterized by:

- the presence of a complex background;
- low quality;
- distortion as a result of poor shooting conditions (taken from a smartphone);

OCR systems **do not work well** on such data.

Tasks

- Development and implementation of a text information extraction tool;
  - Analysis of existing OCR systems;
  - Analysis and selection of image pre-processing methods to increase the accuracy of OCR systems;

Existing OCR Systems

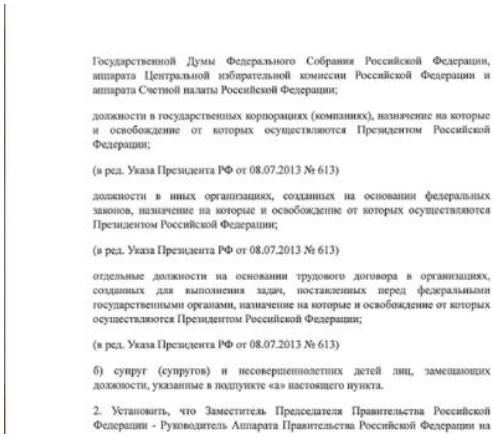
OCR systems	Availability	Support	Advanced processing features
Finreader OCR	-	ABBYY	Camera OCR
OCRopus	+	-	-
Tesseract OCR	+	Google	-

Social Media Images

Demotivators

Certificates

Scanned



Государственной Думы Федерального Собрания Российской Федерации, аппарата Центральной избирательной комиссии Российской Федерации и аппарата Счетной палаты Российской Федерации;

должности в государственных корпорациях (компаниях), назначение на которые и освобождение от которых осуществляются Президентом Российской Федерации;

(в ред. Указа Президента РФ от 08.07.2013 № 613)

должности в иных организациях, созданных на основании федеральных законов, назначение на которые и освобождение от которых осуществляются Президентом Российской Федерации;

(в ред. Указа Президента РФ от 08.07.2013 № 613)

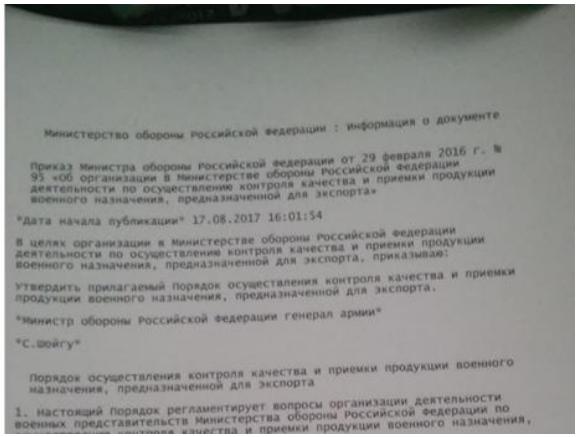
отдельные должности на основании трудового договора в организациях, созданных для выполнения задач, поставленных перед федеральными государственными органами, назначение на которые и освобождение от которых осуществляются Президентом Российской Федерации;

(в ред. Указа Президента РФ от 08.07.2013 № 613)

б) супруг (супругов) и несовершеннолетних детей лиц, замещающих должности, указанные в подпункте «а» настоящего пункта.

2. Установить, что Заместитель Председателя Правительства Российской Федерации - Руководитель Аппарата Правительства Российской Федерации на

Smartphone



Министерство обороны Российской Федерации : Информация о документе

Приказ Министра обороны Российской Федерации от 29 февраля 2016 г. № 95 «Об организации в Министерстве обороны Российской Федерации деятельности по осуществлению контроля качества и приемки продукции военного назначения, предназначенной для экспорта»

Дата начала публикации 17.08.2017 16:01:54

В целях организации в Министерстве обороны Российской Федерации деятельности по осуществлению контроля качества и приемки продукции военного назначения, предназначенной для экспорта, приказываю:

Утвердить прилагаемый Порядок осуществления контроля качества и приемки продукции военного назначения, предназначенной для экспорта.

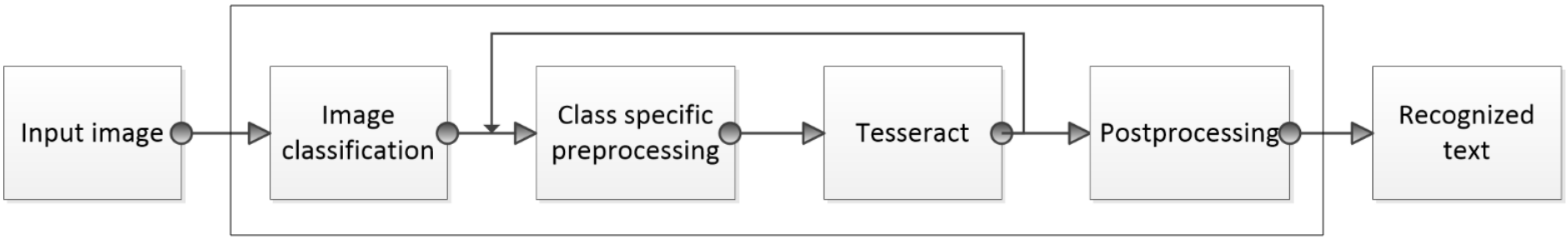
Министр обороны Российской Федерации генерал армии

С.Шойгу

Порядок осуществления контроля качества и приемки продукции военного назначения, предназначенной для экспорта

1. Настоящий Порядок регламентирует вопросы организации деятельности военных представительств Министерства обороны Российской Федерации по

Text Extraction Pipeline



- 1 Step: Image classification
- 2 Step: Preprocessing
- 3 Step: Tesseract OCR
- 4 Step: Delete characters with low confidence

1 Step - Image Classification

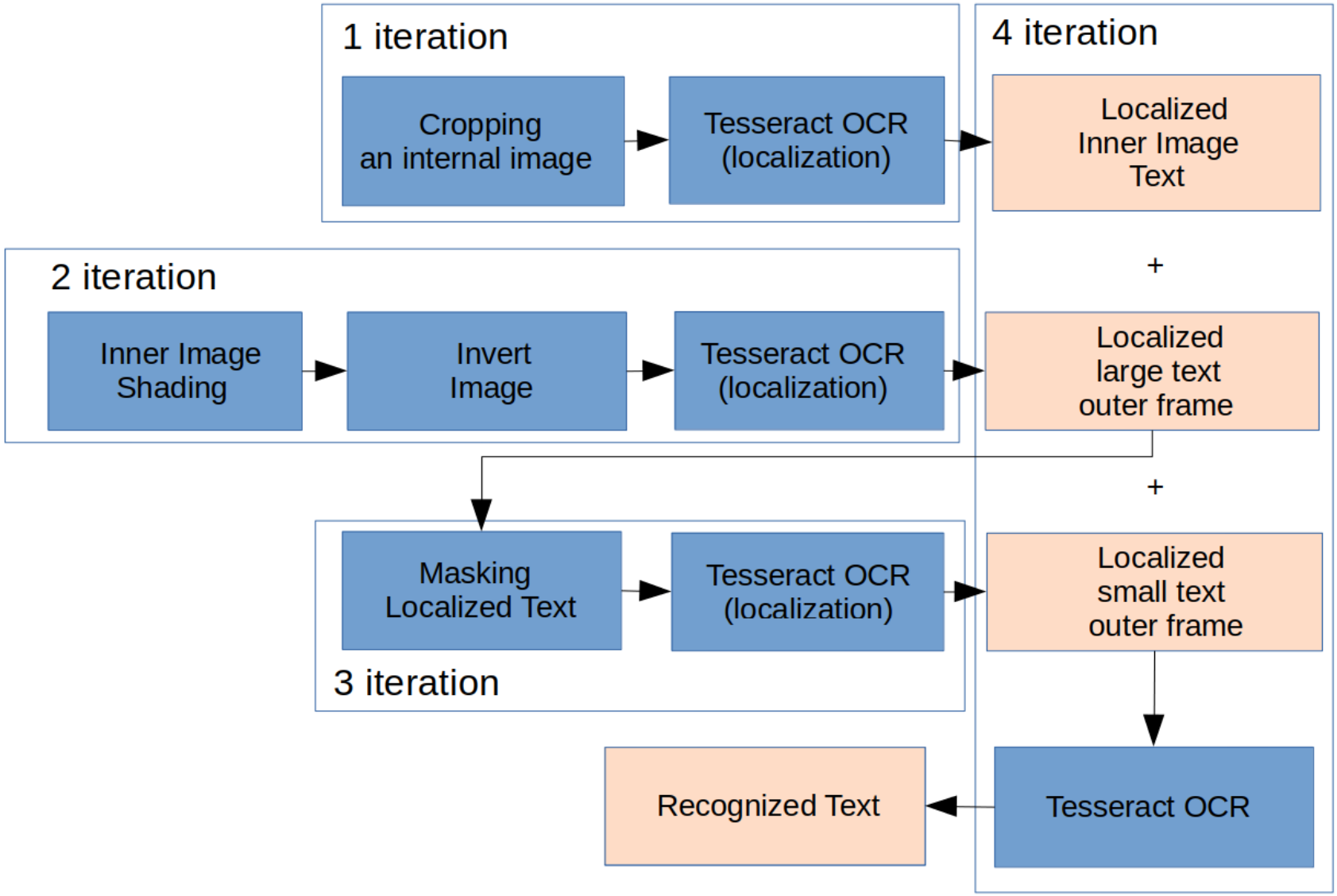
Power training and testing dataset: 11600 images.

Classificator	Accuracy
ResNet50	70
MobileNet	75
Gradient Boosting	95

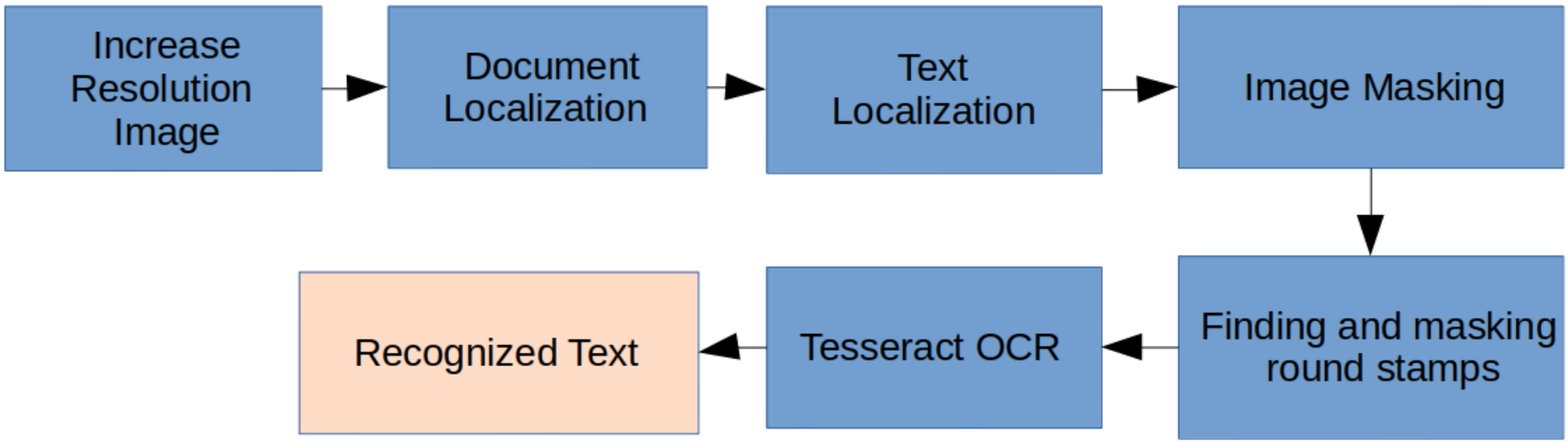
2 Step. Preprocessing methods

- Image Resolution Enhancement (IRE) - restoration of small-sized details (characters).
- Text Localization - the calculation of text blocks.
- Document Localization - calculating the position of a document in an image.

Demotivators class preprocessing



Certificates class preprocessing



Preprocessing classes Scanned, Smartphone

- Scanned images are of high quality and do not require pre-processing;
- The method of resolution enhancement is applied to images of the Smartphone class, and perspective distortion is searched and corrected.

Accuracy metric

$$\mathbf{Character\_accuracy} = \frac{\mathit{Number\_of\_characters} - \mathit{errors}}{\mathit{Number\_of\_characters}}$$

- Number of charachters - number of characters in the document;
- Errors - minimum number of character edits.

Results

#	Classification	Document localization	IRE	Text localization	Accuracy	GPU(CPU) seconds
1	-	-	-	-	*29.5%*	*--(2)*
2	manual	manual	NN2	NN3	87%	3.5(54)
3	manual	NN1	NN2	NN3	86%	5(58)
4	manual	NN1	-	NN3	78%	2.9(5.3)
5	manual	NN1	-	-	72%	2.7(4.6)
6	**GradBoost**	**NN1**	**NN2**	**NN3**	**84.5%**	**5(58)**

Thanks for attention