Using Multi-level Segmentation Features for Document Image Classification

15TH IAPR INTERNATIONAL WORKSHOP ON
DOCUMENT ANALYSIS SYSTEMS
May 22-25, La Rochelle, France

Panagiotis Kaddas[1,2] and Basilis Gatos[1],

[1]Computational Intelligence Laboratory,
Institute of Informatics and Telecommunications,
National Research Center "Demokritos",
153 10 Athens, Greece
{pkaddas, bgat}@iit.demokritos.gr

[2]Department of Informatics and Telecommunications
University of Athens, GR 157 84, Athens, Greece

Abstract

Document Image classification is a crucial step in the processing pipeline for many purposes (eg. indexing, OCR, keyword spotting) and is being applied at early stages. At this point, textual information about the document (OCR) is usually not available and additional features are required in order to achieve higher recognition accuracy. On the other hand, one may have reliable segmentation information (eg. Text block, paragraph, line, word, symbol segmentation results), extracted also at pre-processing stages. In this paper, visual features are fused with segmentation analysis results in a novel integrated workflow and end-toend training can be easily applied. Significant improvements on popular datasets (Tobacco-3482 and RVL-CDIP) are presented, when compared to state-of-the-art methodologies which consider visual features.

Introduction

The proposed work focuses on the Document Image Classification problem by combining a Convolutional Neural Network (CNN) architecture with multi-level information provided by image segmentation techniques (text block, paragraph, line, word, symbol segmentation results). Textual information is not considered in the proposed work, under the assumption that document image classification usually takes place in pre-processing stages where textual information (OCR) is not available.

The contributions of this paper are as follows:

a) A novel integrated architecture is described and end-to-end training can be applied by only using a document image and one or more image masks that correspond to the segmentation levels that are mentioned above.

b) An experimental study is being conducted in order to determine which segmentation levels should be used and decide whether multi-level segmentation features contribute to the task at hand.

c) We present competitive results when compared to the state-of-the-art techniques, evaluated on commonly-used datasets (Tobacco-3482, RVL-CDIP).

d) An additional proof-of-concept is presented for a new private dataset from The Library of the Piraeus Bank Group Cultural Foundation (PIOP) used in the CULDILE project.

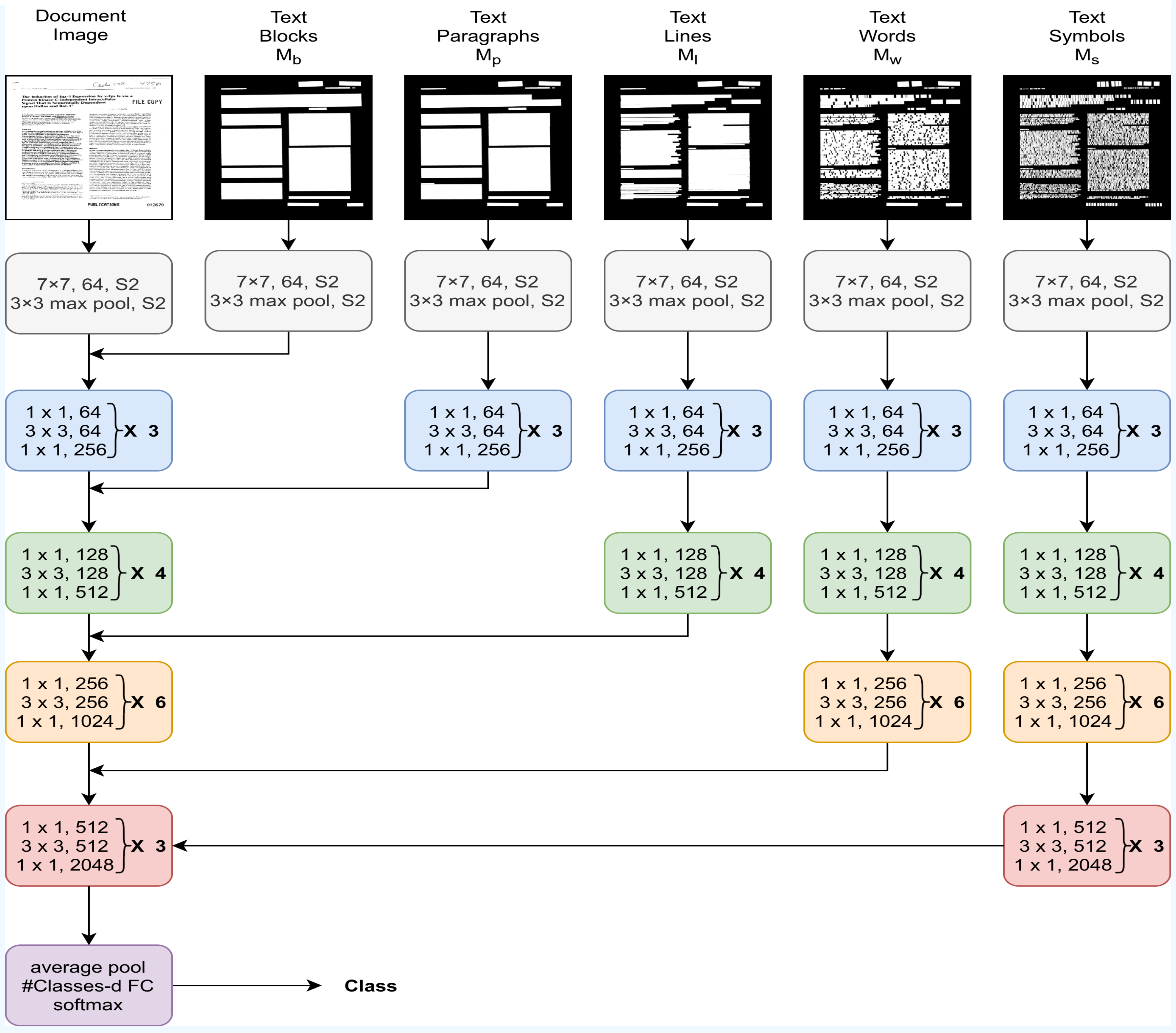
Overall Architecture

- Document image and segmentation masks are forwarded in parallel network streams using ResNet50 as backbone network.
- Each segmentation stream is "deeper" than the previous one in order to be able to learn higher level of information.
- Each output stream is added to the corresponding layer of the backbone (left branch) and finally
- A Fully Connected layer yields class probabilities.
- Input images of size 256×256
- We use "bottleneck" residual blocks for convolutions, followed by Batch Normalization (BN) and ReLU activation layers.
- Cyclic learning rate with Stochastic Gradient Descent (SGD), with values ranging in [0.0001, 0.1].
- Before the final Fully Connected (FC) layer, we apply Dropout
- Augmentation using random cropping (80% of the original size at most) and mirroring over $y-$ axis.
- ImageNet initialization

Baseline Experiments

- A: we train a single ResNet50.
- B: we train several classic ResNet50 models using only segmentation information (and not the original document image), by stacking masks in a single image of depth n, where n is the number of the stacked masks. We do this for every possible combination over the segmentation masks.
- C: we just concatenate the output probabilities of already trained models of A and B for every combination of the latter.
- D: we train A and B models (all combinations again) in a parallel scheme, where we concatenate the convolution outputs for both models and we use an FC layer of 1024 neurons before the output FC layer. This can be considered as a simple ensemble scheme.

Proposed Architecture


Document Image
Text Blocks M_b
Text Paragraphs M_p
Text Lines M_l
Text Words M_w
Text Symbols M_s
7×7, 64, S2 3×3 max pool, S2
1 x 1, 64 3 x 3, 64 1 x 1, 256 X 3
1 x 1, 128 3 x 3, 128 1 x 1, 512 X 4
1 x 1, 256 3 x 3, 256 1 x 1, 1024 X 6
1 x 1, 512 3 x 3, 512 1 x 1, 2048 X 3
average pool #Classes-d FC softmax
Class

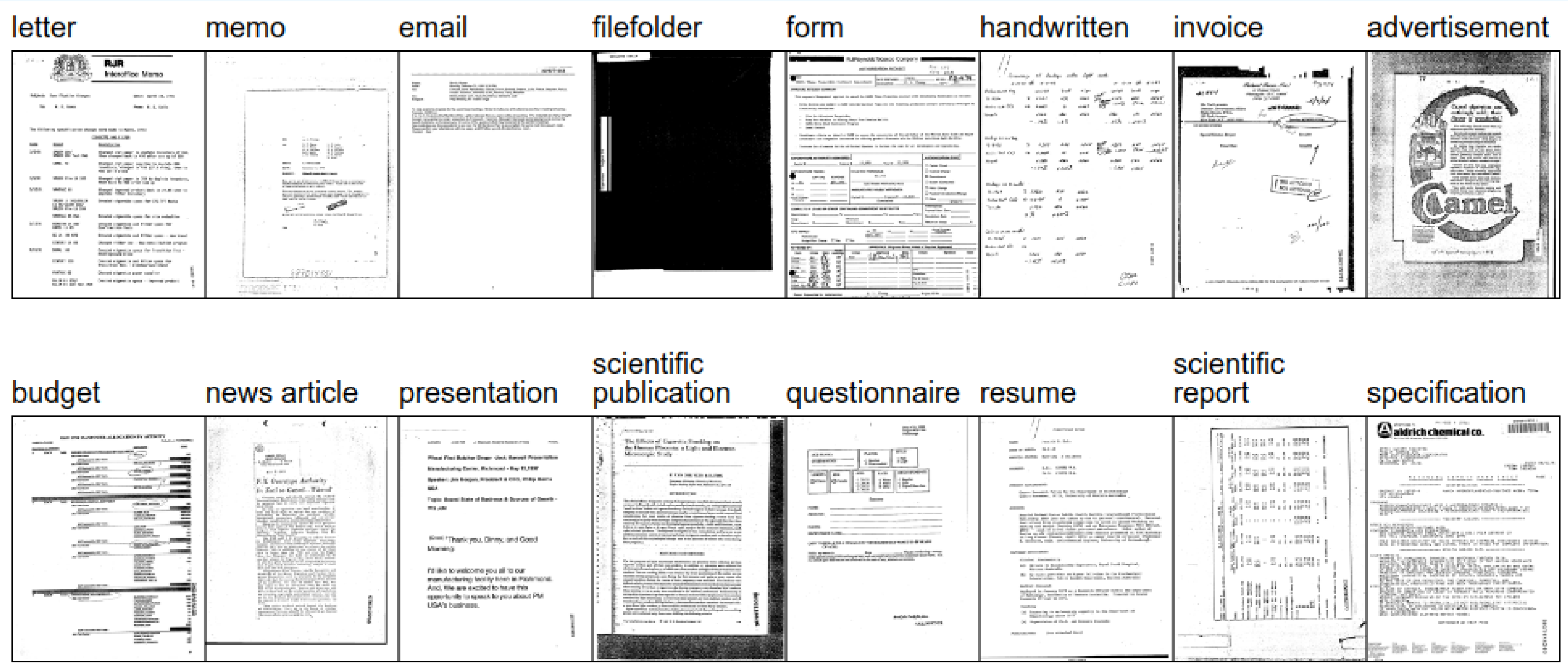


Table 1. Accuracy of combinations over multi-level segmentation masks for document image classification using F-Measure for Tobacco-3482 (%).

Method	Image	Block	Line	Word	Symbol	Accuracy (%)
Baseline$_A$	✓					68.78
Baseline$_B$			✓	✓	✓	75.40
Baseline$_C$	✓		✓	✓	✓	79.86
Baseline$_D$	✓	✓	✓	✓		78.63
Proposed Method	✓		✓	✓	✓	**80.64**
Harley et eal. - Ensemble of regions[4]	✓					79.90
Afzal et al. - VGG-16 [11]	✓					77.60
Afzal et al. - ResNet50 [5]	✓					67.93
Audebert et al. - MobileNetV2 [7]	✓					**84.50**
Fernando et al. - EfficientNet [9]	✓					**85.99**

Table 2. Accuracy of combinations over multi-level segmentation masks for document image classification using F-Measure for RVL-CDIP and PIOP datasets. (%).

Method	Accuracy on RVL-CDIP(%)	Accuracy on PIOP(%)
Harley et el. - Ensemble of regions[4]	89.80	-
Csurka et al. - GoogleNet [12]	90.70	-
Afzal et al. - ResNet50 [5]	90.40	-
Afzal et al. - VGG-16 [5]	90.97	-
Das et al. - Ensemble of VGG-16 models [6]	92.21	-
Fernando et al. - EfficientNet [9]	92.31	-
Baseline$_A$	90.55	84.28
Proposed Method	**92.95**	**86.31**

This research has been co-financed by the program of Industrial Scholarships of Stavros Niarchos Foundation as well as the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH-CREATE-INNOVATE (project code: T1EΔK-03785) http://culdile.bookscanner.gr