Document Understanding: Problems and Technological Solutions

Kristina Arnaoudova[1] and Maria Nisheva[2, 3]

[1] AI Lab at IBS Bulgaria, 4 Pimen Zografski Str., Sofia, Bulgaria
kristina.arnaoudova@icloud.com

[2] Faculty of Mathematics and Informatics, Sofia University St. Kliment Ohridski, 5 James Bourchier Blvd., 1164 Sofia, Bulgaria

[3] Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, acad. G. Bonchev Str., Block 8, 1113 Sofia, Bulgaria
marian@fmi.uni-sofia.bg

Abstract. The paper analyzes the most significant issues related to the information extraction from documents and discusses some theoretical models suitable for recognition. It presents, in brief, the main components of a hybrid approach, which combines an application of a domain ontology and deep learning techniques to recognize and classify the document structure efficiently. The approach is based on the use of a symbolic logic inference model to utilize the knowledge about the document semantics. The document's segments are referring to the conceptual purpose of the document, which is recognized by applying modern architecture for object detection. An appropriate domain ontology solves the issues related to the semantic completeness of the document's data. Several experiments have been carried out and the results obtained have been analyzed in terms of the applicability of various modern technologies for the implementation of document understanding system.

Keywords: document processing, document layout understanding, ontology, deep learning.

1 Introduction

Automatic information extraction from documents is not a new issue. Manual document processing is a major cost driver in organizations. Meaningful results are achieved in text recognition, using a variety of Artificial Intelligence methods. The traditional way to digitize documents is optical character recognition (OCR), which recognizes text and is a proven and workable approach. The OCR, though, does not retain the formatting elements of the template and fails at recognizing symbols with enough precision due to the presence of tables and other items. In the modern versions of optical recognition, neural networks are also used

with radical improvements in performance. At the same time, however, the problem with text documents formatted in a specific way is substantial and not entirely resolved. When tables or other graphical elements are used, recognition libraries not only cannot retain formatting, but the presence of such elements can dramatically worsen the results. Even when the format may be given, just small deviations could be challenging.

Moreover, a hierarchical structure needs additional knowledge about the relationships among the image components. The processing of formatting elements may need a detailed definition of the used template in terms of relations between particular regions on the documents. The convolutional neural networks are revolutionary in recognition but still not design for interpreting the spatial relations.

Therefore, we consider a hybrid approach using a domain ontology for the description of the document sections and their relations as very suitable to that gap. The particular approach we propose is a top-down one and envisages three fundamental phases: *concept recognition*, *template recognition*, and *semantic understanding*. It enables inference complementary to annotated image segments and formatting elements, considering that the recognition of the formatting elements of the document is a key point to a successful text extraction using OCR.

2 Main components and phases

2.1 Conceptual knowledge – domain ontology

An ontology [1] is a formalization of a conceptual representation of a domain. Human recognition capabilities can be seen as a combination of visual perception and inference based on complex sampled data. Computer vision is inspired by human perception and performs the classification task with basic recognition by repeated examples. The idea to use ontologies for image interpretation is not new, involving classification following the concept's definitions, which should be provided as done by humans [2]. Some of the first efforts in this direction are discussed in [3] and [4].

Ontologies enable the application of symbolic logic inference mechanisms as classification. The ontological approach is managing and integrating different sources and generally is a widely adopted approach to structuring unstructured data. The document content as an unstructured amount of information could be formalized using appropriate ontologies. On the other hand, the document is an image, and the data may be presented in a hybridlike manner using the best of both approaches – typical computer vision and knowledge-based inference. The significance of this hybrid approach is explained by the potential reduction of the cost of generating recognition models by reducing the training data annotation.

The ontology's global definition can be reused without significant changes. Therefore, our attempt to fully understanding the document is following the idea of a hybrid approach, using the ontology as a primary domain knowledge source and deep learning that allows to:

- map the formatting elements,
- perform template classification,
- semantically validate the document content.

Using an ontology with mapped concepts could significantly reduce the needed resources for document type classification. Following a hybrid approach, the defined classes are mapped to the semantic elements of the documents. So, enable the performing of additional inference steps for image template recognition. The ontology defines the concepts and formatting elements. For example, an ontology of documents could describe concepts like person, organization, document, issuer, entitled, etc. and formatting elements as title, table, paragraph, box, n-dimensional grid, point, coordinates, etc.

We have validated the concepts with a sick leave documents dataset. Examples of semantic concepts as elements of the experimental domain ontology being under development for this research are shown in Table 1 and Fig. 1.

Table 1. Examples from the domain ontology concept's map.

Concepts	Specific Domain	Format	Measures
Document	Sick Leave		Ordered coordinates, length, height, aspect ratio
Issuer	Medical organization, Doctor	table, girds	
Person	Person	box, label, field	
Entitled	Person	grid	
…	Regime	…	

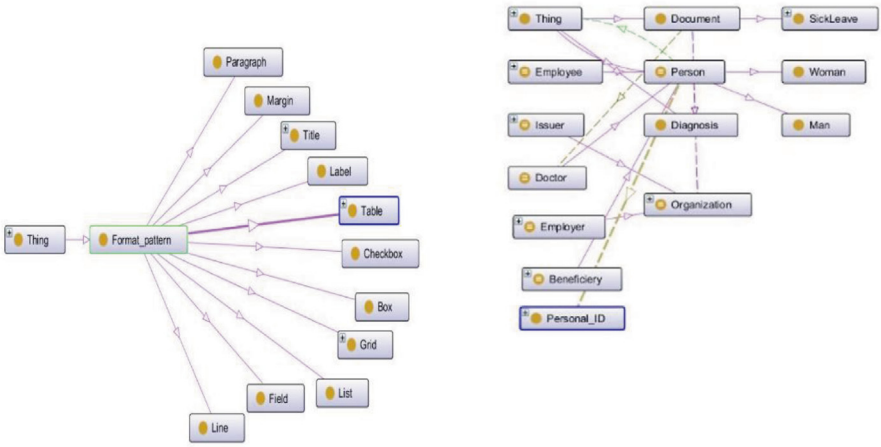

Fig. 1. Example ontology concepts.

2.2 Template classification

An essential step in the successful implementation of information extraction is the definition and recognition of the template. The template is understood as a model, grouping similar primitive patterns of formatting. It is an essential step for processing the image realized with image processing techniques. According to the identified formatting elements and as a preprocessing step in the training and inference stage, various image-processing techniques are applied, relative to the type of formatting elements. The formatting pattern can be classified using the spatial relation described in the ontology as metadata. The template identification is represented as spatial relationships between concepts. The spatial relation within the documents refers to their relative position on the document grid. The recognition of the used template is related to the common document format pattern classification described in the ontology. Fig. 2 shows the main steps of the process.

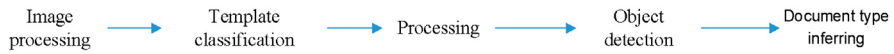

Fig. 2. Template processing.

The final document template is the combination of formatting elements and detected objects of the document type. The document type may improve significantly the object's recognition confidence score by applying different processing. The classification is based on the metadata including concepts and spatial relation by their grid relative position on the document image.

2.3 Semantic understanding

The convolutional neural network object detectors identify the regions of interest, and the OCR extracts the text successfully. The template mapping uses the recognized formatting elements embedded in the regions. The understanding, however, comes with the inferred knowledge about the concepts of the documents. Our hybrid approach suggests additional processing of semantic truthfulness and explanation of the extracted data using a symbolic logic model, and inference algorithms, using diferent types of information, in particular data from the detector and conceptual knowledge. The metadata is used as semantical validation of the recognized objects using the ontology axioms and rules. The expected effect is leveraging the recognition by the deep learning algorithm and generating an explanation of the results in terms of the used template. The concepts recognized by the object detector will be processed by linking to existing axioms of such basic concepts to form other, more complicated, defined concepts. An example in the context of the sick leave document could be the validation if the recommended regime is a logically valid result for the given diagnosis.

2.4 Convolutional neural networks

The origin of convolutional neural networks (CNN) was introduced by Kunihiko Fukushima (Neocognitron) in 1980 [5] and Yann LeCun at al. (LeNet-5) in 1998 [6]. The name CNN comes from one of the most important operations in the network, which is the convolution. The convolution is performed on the input data with the use of a kernel to produce a feature map. It is executed by sliding the filter over the input. At every location, matrix multiplication is performed and summation of the result onto the feature map. Convolutional neural nets have revolutionized speech and object detection. CNN leverages three important fundaments as sparse connectivity, shared parameters, and invariance to translation. Sparse weights refer to the reduced number of parameters, and parameter sharing is the factor used for more than one function, making the parameters significantly computationally efficient. The invariance to translation means that if the input translates to some extent, the output changes in the same way. Among the very successful methods for object detection being of the focus in AI, one can mention Faster R-CNN, RPN, Mask-RCNN, and FCN.

Our experiments successfully apply a convolutional neural network, implementing automatic document processing for detecting the main concepts and the embedded elements. Based on the concepts, identifying the main classes, which represent the basic concepts of the document's domain, the segments of the document image are annotated. We have experimented with several convolutional models; among them is Mask R-CNN.

3 Object detection

An object recognition algorithm identifies which objects are present in an image. It takes the entire image as an input and outputs class labels and class probabilities of objects present in that image. For example, a class label could be "cat," and the associated class probability could be 97%. The object detection methods add the location with coordinates of outputs bounding boxes; predict where on the image is the object.

A convolutional neural network for object detection classifies the objects and finds their location within the document in terms of detected regions. Annotations may be different according to the required semantics. We have conducted various approaches to conceptualize the image area segmentation used for recognition through annotation. We have chosen to label regions, following the predefined concepts, such as a person, experts, purpose, subject, etc. Also, the formatting elements, such as a box, table, grid, etc. The recognized regions represent complex concepts within the document, a data structure comprising different attributes. Along with our experiments, we encounter quite a few challenges, some of which are fonts, scale, quality, small objects, and the number of recognized objects. To overcome the various limitation, we have applied different image processing

techniques. One of the main challenges encountered are the recognition of the small symbols and overlapping elements. Another obstacle is the learning algorithm limitation in the number of recognized objects. The results are substantial, though the task for document understanding needs objects and relations among them. It is not one of the designed purposes of CNN, here we singled out the usage of ontology engineered concepts and reasoning as an integrated step document understanding solution.

In the experiments, we have used IBM Power AI Vision, a software that implements the most modern computer vision convolutional neural network architectures. We have conducted experiments with different architectures; among them, Mask R-CNN implemented by Detectron (FAIR).

3.1 Detectron FAIR

The software system Facebook AI Research implements advanced algorithms for detecting objects, including Mask R-CNN architecture [7]. Some results from Mask R-CNN are shown in Fig. 3.

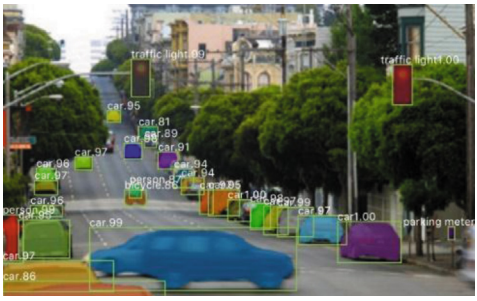

Fig. 3. Mask R-CNN.

The purpose of Detectron is to provide a high quality, high-performance base for researching an object. It is designed to be flexible and to support rapid implementation and evaluation of new research. Detectron includes implementations of a variety of algorithms for detecting objects like Mask R-CNN:
Mask R-CNN,
RetinaNet,
Faster R-CNN,
RPN,
Fast R-CNN,
R-FCN.

Mask R-CNN

Mask R-CNN efficiently detects objects in an image while simultaneously generating a high-quality segmentation mask for each instance. The architecture of Mask R-CNN extends Faster R-CNN by adding a branch for predicting an object mask in parallel with the existing branch for bounding box recognition. Mask R-CNN is a deep neural network targeted to solve instance segmentation problems in computer vision. In other words, it can separate different objects and returns the object bounding boxes, classes, and masks. There are two stages of Mask R-CNN. Firstly, it generates proposals about the regions where there might be an object based on the input image. Secondly, it predicts the class of the object, refines the bounding box, and generates a mask at the pixel level of the object based on the already existing proposals.

Our experience shows that Mask R-CNN is the most efficient document-processing model. The achieved result in segmentation performance is 96% accuracy.

We have significantly reduced the implementation time with the IBM Visual Insights (prev. IBM PowerAI Vision), a software system for computer vision and deep learning implementing the most modern CNN architectures and among them Detectron.

FCN – Fully convolutional network

Mask R-CNN is a Region convolutional neural network which has been improved by the Fully convolution networks (FCN) branch [8]. FCN is built only from locally connected layers, such as convolution, pooling, and upsampling. No dense layer is used in this kind of architecture. This reduces the number of parameters and computation time. In addition, the network can work regardless of the original image size, without requiring any fixed number of units at any stage, given that all connections are local. Fig. 4 presents the Mask R-CNN architecture.

R-CNN: Regions with CNN features

A group of architectures based on several different components, each of which is a network. The main components are the proposed region classification of classes. R-CNN [9] and is one of the first architectures used for object detection. R-CNN first try to construct different proposed regions. The proposed region is the area on the image, which has a high probability of containing the object. To construct proposed regions, external region proposal methods like Selective Search are used.

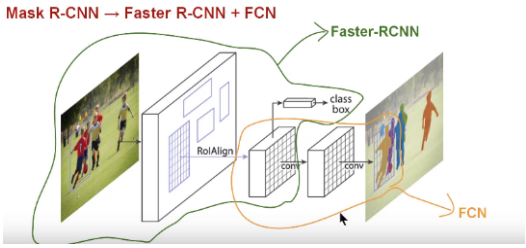

Fig. 4. Mask R-CNN.

Selective Search Algorithm:
1. Generate initial sub-segmentation, generating many candidate regions
2. Use a greedy algorithm to combine similar regions into larger ones recursively
3. Use the generated regions to produce the final candidate region proposals

In R-CNN, the input image is fed to a region proposal algorithm like selective search, and CNN is performed on each proposed region. The output of CNN is given to SVMs for the classification of objects detected. Therefore, if there are 2000 proposed regions, the needed runs are 2000 CNN networks.

Fast R-CNN [10]

This model is an improvement of R-CNN. It is 25x more effective than R-CNN. To reduce the overhead of multiple CNN networks in R-CNN, first, the input image is fed to CNN, which gives an insight on the features of the image, and then the selective search is performed to get proposed regions.

Faster R-CNN [11]

Faster R-CNN is an improvement of Fast R-CNN where Region Proposal Network is used as a proposed region generator instead of selective search; Fig. 4 shows R- CNN, Fast R-CNN, and Faster R-CNN. Faster R-CNN is a combination of Fast R-CNN and RPN. The first stage here is to prepare the features in the respective map and then suggest coordinates of the presumed location of the site. The proposed regions are given to a classifier for object classification. Faster R-CNN is a highly effective approach for object detection. It is 250x more effective than R-CNN.

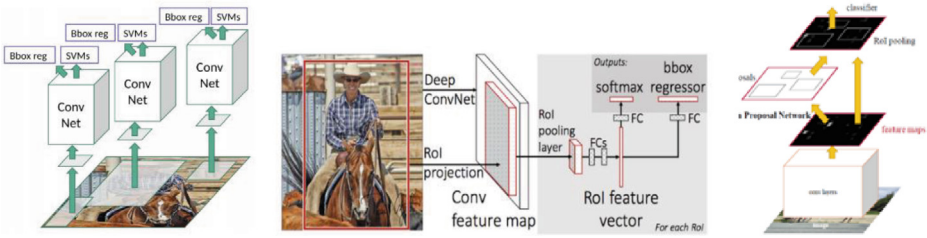

Fig. 5. R-CNN**,** Fast R-CNN, Faster R-CNN.

RPN (Region Proposal Network)

RPN provides a time-effective way of generating region proposals/regions of interest. It is more effective than the selection search used in R-CNN/Fast R-CNN. RPN ranks region boxes, called anchors, and proposes the ones most likely containing objects. Each RPN has a classifier and a regressor. To generate proposals for the region where the object is, a small network is a slide over a convolutional feature map that is the output by the last convolutional layer. The anchor is the central point of the sliding window. The anchors are of different aspect ratios, and the proposals are based on significant Intersection-over-union overlap with a ground truth box. RPN is an algorithm that needs to be trained and has defined the loss function.

From the group of architectures discussed in this paper, we have experimented with Faster R-CNN in the context of the sick leave document type. It is not so accurate in determining the coordinates of the object and cannot recognize small objects with enough accuracy.

YOLO – You only look once [12]

One of the fastest CNN, working in real-time**.** It is highly efficient because of not repeating a segment of the picture looking for different objects (see Fig. 6). The algorithm is not optimized for accuracy, but for time. It is not sufficient to be able to recognize many small objects and full overlapping elements. Some elements in the document sample for sick leave are low for this type of network.

4 Conclusion

An end-to-end implementation of an effective, comprehensive document information extraction could be a combination of different inference mechanisms, integrating several approaches. The fundamental recognition and localization of different concepts are based on the modern object detectors, and further understanding is grounded on logically inferred relationships using ontological engineering. We have experimented, organizing the recognition process in

semantic blocks. Significant stages in the architecture are the ontology layer, object detection, and post-processing layer, based on the template classification. We use inferred knowledge from the ontology axioms and rules, and accordingly, the system undertakes different actions. The result of the object detection is reliable with sufficient accuracy. Still, the information extracted should be additionally processed or confirmed semantically. Therefore, the proposed hybrid approach could be effectively applied.

Our experiments were conducted with a limited set of data; the small amount of data assumes no balance of classes. Upon accumulating data and continuous training, the object detection accuracy and the variety of the concepts are expected to improve. Document's element recognition is a widely used task that meets difficulties in formatting recognition. The experiments show that the document can be successfully read automatically using a hybrid approach based on the utilization of suitable domain ontology and deep learning object detection neural networks. The next steps are to develop further the ontology semantics and to improve the extraction of features with maximum preservation of spatial information so that we can recognize small objects.

Acknowledgments

The AI laboratory at IBS Bulgaria has supported the presented research.

References

1. Gruber, T.: Toward Principles for the Design of Ontologies Used for Knowledge Sharing. International Journal of Human-Computer Studies, Vol. 43 (1995), pp. 907-928.
2. Porello, D., Cristani, M., Ferrario, R.: Integrating Ontologies and Computer Vision for Classification of Objects in Images. In Proceedings of the Workshop on Neural - Cognitive Integration (NCI @ KI 2015), PICS Publications of the Institute of Cognitive Science Vol. 3 (2015).
3. Straccia, U., Visco, G.: Dlmedia: an ontology mediated multimedia information retrieval system. In Proceedings of the 2007 International Workshop on Description Logics DL2007 (Brixen-Bressanone, Italy, 8-10 June 2007).
4. Town, C.: Ontological inference for image and video analysis. Mach. Vis. Appl., 17(2):94–115 (2006).
5. Fukushima, K.: Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics, 36(4): 93-202 (1980).
6. LeCun, Y., Hafnner, P., Bottou, L., Bengio, Y.: Object Recognition with Gradient-based Learning. LNCS, Vol. 1681, pp 319-345 (1999).
7. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. arXiv:1703.06870 (2017).
8. Long, J., Shelhamer, E., Darrell, T.: Fully Convolutional Networks for Semantic Segmentation. arXiv:1411.4038v2 [cs.CV] 8 Mar 2015
9. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. *2014 IEEE Conference on Computer Vision and Pattern Recognition (*Columbus, OH, 2014), pp. 580-558.

10. Girshick, R.: Fast R-CNN. *2015 IEEE International Conference on Computer Vision (*Santiago, 2015), pp. 1440-1448.
11. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with 10. Region Proposal Networks. arXiv:1506.01497 (2015).
12. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV, 2016), pp. 779-788.