“D3.2”

Set of publicly available algorithms to separate compound images

Version: 1.0

Last Update: 22/6/20

Distribution Level: *PU*

Distribution level

PU = Public,

RE = Restricted to a group of the specified Consortium,

PP = Restricted to other program participants (including Commission Services),

CO= Confidential, only for members of the ExaMode Consortium (including the Commission Services)

Project supported by European Union

Horizon 2020 grant agreement 825292

The ExaMode Project Consortium groups the following Organizations:

Partner Name	**Short name**	**Country**
HAUTE ECOLE SPECIALISEE DE SUISSE OCCIDENTALE	HES-SO	Switzerland
UNIVERSITA DEGLI STUDI DI PADOVA	UNIPD	Italy
SIRMA AI	SIRMA	Bulgaria
STICHTING KATHOLIEKE UNIVERSITEIT	RADBOUDUMC	Netherlands
MICROSCOPEIT SP ZOO	MICROSCOPEIT	Poland
AZIENDA OSPEDALIERA PER L'EMERGENZA CANNIZZARO	AOEC	Italy
SURFSARA BV	SURFSARA BV	Netherlands

Document Identity

Creation Date:	15/5/20
Last Update:	22/6/20

Revision History

Version	**Edition**	**Author(s)**	**Date**
0	1	Manfredo Atzori	15/5/2010
Comments:	Defined deliverable structure		
0	2	Niccolò Marini, Gaétan Lepage, Stefano Marchesin	8/6/2010
Comments:	Wrote the deliverable preliminary version		
0	3	Manfredo Atzori	9/6/2010
Comments:	Proofread the preliminary version, added parts and refined it		
0	4	Gianmaria Silvello	15/6/2010
	Review of the deliverable		
1	0	Manfredo Atzori	22/6/2010
	Answered to the comments of the reviewer and performed final control		

Executive summary

Figures represent a fundamental part of scientific literature. They represent a valuable source of knowledge and they allow humans to understand the content scientific works. Developing methods that can help to extract knowledge from the figures and text included in scientific articles is part of the ExaMode objectives, and particularly of Deliverable 3.2.

Deliverable 3.2 "Set of publicly available algorithms to separate compound images" describes the complete pipeline to handle compound figure separation. The algorithms to separate compound figures and link them to related text are publicly released on Github and a link to them is available on the ExaMode website. The publication describing the algorithms and the results of their evaluation (which corresponds to the text of this deliverable in its preliminary form) is planned to be released by the end of July, since before submitting, it would be beneficial to test few more changes of the pipeline that might allow to improve the results.

Table of Contents

Table of Figures

Index of Tables

List of abbreviations

CNN	Convolutional Neural Network
PMC	PubMed Central
CCA	Connected Components Analysis
HOG	Histograms of Oriented Gradient
SVM	Support Vector Machine
NMS	Non Max Suppression
FPN	Feature Pyramid Networks
POS	Part of Speech
MAP	mean Average Precision

1 Introduction

Figures represent a fundamental part of scientific literature. They allow humans to better understand the content described in the text of articles and books, thus representing a valuable source of knowledge. Developing methods that can help to extract knowledge from the figures and text included in scientific articles and books is a problem that is currently unsolved [1] and that is part of the ExaMode objectives. There are many problems that make it difficult to extract knowledge from scientific literature. Among those, compound figures separation is one of the biggest and still less studied ones. This document clarifies the problem of compound figures separation, it describes what has been done in literature to face it, it proposes a procedure to fully approach it and evaluates it on concrete data. The code for performing compound figure separation was publicly released. Our team is currently working on improving the results before submitting a final publication to describe the results. Submission is foreseen by the end of July.

Compound figures can be defined as images that include several sub-figures (panels), identified by panel labels (that can be letters, arabic or roman numerals, or combinations of them). Compound figures are usually associated with a textual description (the caption), that in most cases refers to each panel via the labels.

The scientific literature stores a very large amount of available medical knowledge in terms of text and figures. This is particularly evident in the biomedical domain. For instance, the PubMed Central (PMC) repository[1] includes more than 2 million articles in total with an average of 3.5 figures per article, including 1.5 compound figures of 4 sub-figures each. This sums up to approximately seven million figures, including 3 million compound figures with 12 million sub-figures for a total of >16 million figures available in 2018 if separated correctly. With an expected increase of nearly 3 million figures in 2019 and more in the following years, it promises in the near future very large amounts of training data in various applications and modalities. Difficulties related to scientific literature data include the heterogeneity of the images, the presence of compound images and the automation of ground truth labels extraction from the text [1].

Automatic methods have been proposed to access to biomedical images and their associated metadata using text-based approaches [2, 3, 4] and content-based methods [5, 6, 7, 8]. Considering the amount of images available in scientific literature, it is important that these methods can access the information in an efficient way. The performance of the methods can be improved with pre-processing operation on the data (e.g. classifying the images according to their modality) or with post-processing operation on the results (e.g. using criteria for filtering the results).

Compound images represent almost 50% of the figures within open access biomedical literature [9, 10, 11]. Their characteristics raise some open challenges, [10, 12]. The sub-figures within a compound image usually represent different concepts, therefore it is necessary to separate them in order to apply content-based image analysis methods [12]. Similarly, the caption also includes a textual description for each of the images. Therefore, also in this case, the content related to its sub-parts needs to be identified and to be associated to the sub-images.

The problem of separating compound figures can be decomposed in two main phases, namely panels segmentation and caption splitting [12]. Panel segmentation consists in separating the compound image into the sub-figures that compose it and associating the right label to each sub-figure. Thus, panel segmentation can be itself divided into two sub-tasks: panel splitting (targeting the division of the figure into the sub-figures) and label recognition (targeting the recognition of the labels within each sub-figure and their association to the sub-figure). Caption splitting consists in identifying within the caption the textual information related to each sub-figure.

In literature, panel splitting is faced with approaches based on traditional computer vision algorithms and approaches based on machine learning. Traditional computer vision approaches usually exploit the gap between the different panels, using techniques adopted in detection tasks [10, 13, 10].

1 http://www.ncbi.nlm.nih.gov/pmc/

Müller et al. [10] proposed an approach for panel splitting based on two phases: detection and analysis. In the detection phase, vertical and horizontal lines are identified, applying recursive operations to separate the panels. In the analysis phase, the lines classified as false positive are removed. Li et al. [å13] developed a method based on Connected Components Analysis (CCA) for removing small objects within an image. The algorithm maintains only the main components (the panels) inside the image. However, they noticed that the process is not working well with images that are not well-connected or blurry. Therefore, the author used an edge detector to improve blurry components detection and applied dilation on the edge image, in order to increase the connectivity between panels. Also in Cheng et al. [14], the panels are split using techniques for detecting vertical and horizontal lines. In this work, a pipeline of operations is applied for detecting the lines: a Sobel filter is applied to the image and then candidate bounding boxes are generated. A filter is then applied to remove false positive bounding boxes. The approaches based on machine learning algorithms mainly exploit the results reached by Convolutional Neural Networks (CNN) in computer vision tasks [15, 12, 16]. In Tsutsui et al. [16], a CNN is trained for separating the panels. The problem is organized as an object detection problem. The "You Only Look Once version 2 (YOLOv2)" system [17] is applied in order to detect the sub-figures and to define their bounding boxes. In Zou et al. [15] both the sub-tasks of panel segmentation (panel splitting and label recognition) are explored. The panels are detected using the features extracted from the figures with Histograms of Oriented Gradient (HOG) in order to train an Support Vector Machine (SVM) classifier. The label recognition is proposed as a classification problem: a 50 classes (alphanumerical characters and digits) SVM classifier estimates the posterior probabilities for each one of the candidate labels. Finally, the panels are linked with the corresponding labels using a beam-search algorithm, discarding false-positive elements. The same authors improved the task in [12], performing both the sub-tasks using deep convolutional neural networks. The neural network works in this case in a single step, separating the figure in panels and detecting their labels simultaneously. The backbone neural network is organized in various layers, so that the features are organized in a pyramid. The features feed a classifier for generating candidates, for both panels and labels. Also in this case, the candidates for panels and labels are fused using a beam search algorithm.

Caption splitting can be considered as a sub-task of the information extraction domain. Two main kinds of approaches are described in literature to perform caption splitting: approaches based on hand-coded rules and approaches based on machine learning methods In the hand-coded rules approaches, a set of rules is applied to the caption. Cohen et al. [18], apply a set of hand-coded rules in order to extract and classify image labels and design two methods for the evaluation. Each of the image labels is classified into three classes: bulleted list indicators, proper noun indicators and reference indicators. Two hand-coded approaches are tested. The first method achieves high precision (98.5%) but very low recall (45.6%), while the second achieves lower precision (74.5%) but higher recall (98.0%) and a higher F-score. Apostolova et al. [19] adopted similar rules for their work, focusing it on the extraction of bulleted list indicators. In both the previous works, false-positive labels are eliminated, using filter rules. In [20], a semi-automatic method, based on different hand-coded rules is proposed. In this work, the algorithm needs the image label type (e.g. (A), A), or A:) as input. The information is used in order to extract the image labels and the sub-figure captions from the figure caption. Freitag and Kushmerick [21] introduce a machine learning method to learn string patterns, in order to split captions. However, the algorithm needs several input parameters: the starting index, the ending index and the length of each text label within the caption. The approach takes inspiration from [18], but it differs in the approach.

As described in the previous paragraphs, several works in literature describe separately panel segmentation, panel splitting and caption splitting. Instead, works describing the combination of the different phases together were not described in literature until now to the best of our knowledge. The following sections describe in detail the comprehensive procedure developed for this aim (including panel segmentation and caption splitting) and the evaluation of each phase. Currently, our team is working on improving the results before submitting a final publication to describe the results, which is planned to be submitted by the end of July.

2 Methods

Compound figures can be defined as figures that are composed of several panels. Usually, labels (such as letters, arabic or roman numerals, or combinations of them) identify each panel. Compound images are usually associated with a textual description (the caption), that in most cases refers to each panel via the labels. Compound figure separation can be described as composed of several sub-tasks: panel segmentation (including panel splitting and label recognition) and caption splitting. Those different steps are summarized in Figure 1. This work presents a complete pipeline to deal with all the mentioned tasks, required to perform compound figure separation. Each step is detailed in the following subsections.

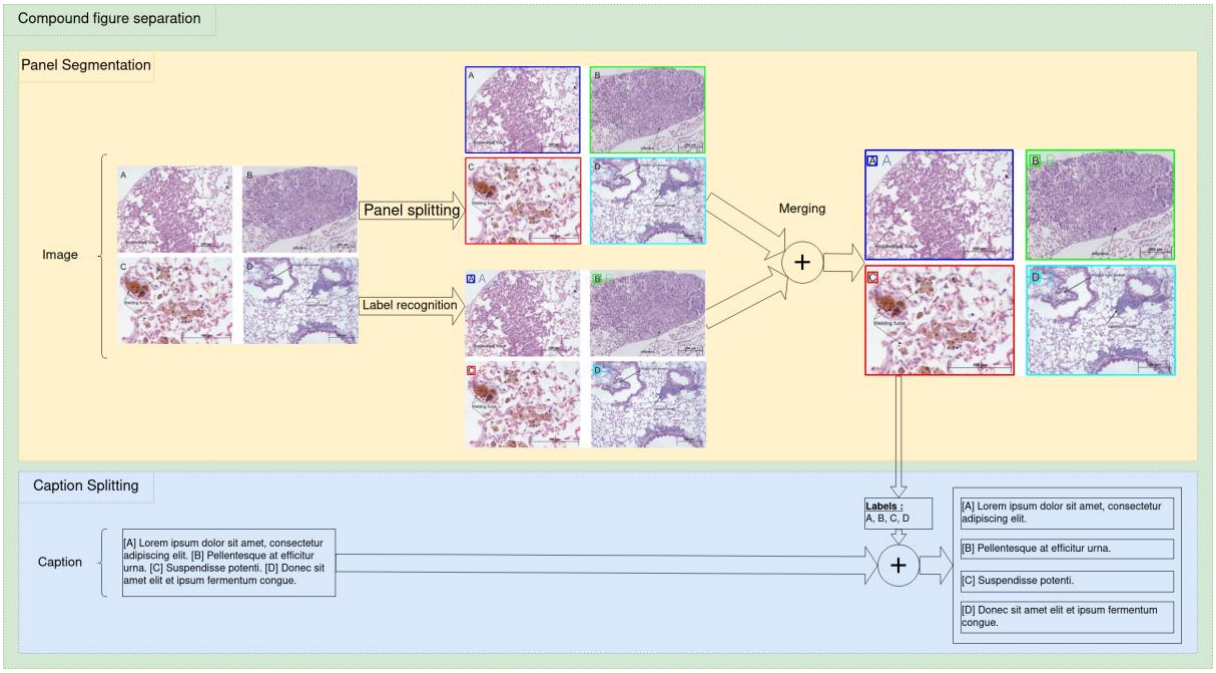

Figure 1 : Overview of the procedure to separate compound images.

2.1 Panel segmentation

Panel Segmentation targets both panel splitting and label recognition. The results from those two sub-tasks are then merged using the beam search algorithm described in [12]. Even though two independent networks were trained to achieve panel splitting and Label Recognition, a single unified network handles both tasks. While the work in this section is inspired to the recent work by Zou et al. [15], that proposed a single unified network handling both tasks, the code was consistently re-written.

2.1.1 Panel splitting

Panel splitting consists in dividing the compound figure into the sub-figures that compose it. Panel splitting is achieved using RetinaNet [22] for object detection. This model is based on the *Focal Loss* for dealing with extreme foreground- background class imbalance. Labels are completely ignored when dealing with panel splitting. We implemented a neural network aiming at achieving this task specifically. The raw output of the neural network goes through a Non Max Suppression (NMS) procedure. This approach is fairly common in modern object detection pipelines.

It is used to filter out the overlapping predictions made by the model. It is important to note, though, that NMS was configured specifically for our task. Indeed, contrary to conventional object detection in photos, panels will almost never overlap with each other. By choosing a significantly low IoU (Intersection over Union) threshold, predictions involving overlapping proposals are rejected. This solution was chosen after experimental tests, since it lead to a noticeable improvement of performance.

2.1.2 Label recognition

The task identified as label recognition is independent from panel splitting. The goal is to localize and identify the labels possibly present in a compound figure. Currently, this phase targets single character labels. Zou et al. [12] proposes 50 different classes representing single alphanumerical character labels. For sake of convenience, we use the same network architecture as for panel splitting.

2.1.3 Fusion of panel splitting and label recognition

To simultaneously detect panels and labels from a single compound image, a single CNN feature extractor is shared by the two sub-networks that specialize into either panel detection or label detection. The architecture was previously described in [12] and it is illustrated in Figure 2. More precisely, a single Resnet50 [23] feature extractor is shared as a common backbone. Then, two different subsets of the backbone output are feeding two different Feature Pyramid Networks (FPN). The convolutional features used are C3, C4 and C5 for panel detection and C2, C3 and C4 for label recognition. The ResNet50 architectures offers several features reflecting different receptive fields. C2, C3, C4 and C5 denote the residual blocks constituent the ResNet 50 architecture. The higher the number, the deeper the layer and so, the larger the receptive field. Hence, to detect the labels, that are smaller, we use features from lower layers. A classification head and a regression head follow each FPN. The panel classification head is trained to binary classify the presence of a panel in each proposed region. The regression head is trained to discriminate over the 50 classes presented by [12] (which correspond to single alphanumerical characters). At this point, the presented model outputs, for a single image, a set of panel boxes and a set of labeled label boxes. However, a key step needed to achieve panel segmentation is to match the splitted panels to the recognised labels. In order to do this, we applied the beam search algorithm proposed by Zou et al. [12], which is a greedily approach aiming at matching the detected panels and labels. This algorithm also helps eliminating panel or label false positives as it outputs panel-label pairs.

2.2 Caption splitting

Captions convey important information that help to understanding each sub-figure. The caption splitting component separates the caption, linked to the compound image, into sub-captions, each one linked to

a different panel (sub-figure) of the image. This caption splitting component has two input parameters, it is composed by a pipeline of operations and it was tested on a manually annotated ground truth. The two input parameters of the algorithm are the caption (associated to the compound image) and the labels detected by the label recognition step of the pipeline (described in section 2.2. The "caption" input parameter is represented by a string. The "labels" input parameter is stored within a list. In this version of the algorithm, the labels are alphanumeric characters, both lower-case (e.g. a, b, c, d) or upper-case (e.g. A, B, C, D). The caption splitting algorithm is composed of a pipeline of operations applied to the "caption" input. The operations are applied in order to identify and filter the labels, to identify the relative text snippets and to separate them into sub-captions. The first step of the procedure detects the positions of the candidate labels. The operation identifies the labels within the input captions and the relative label positions. The output of the operation is a tuple including the label and its position. The second step of the procedure filters the detected labels. The filtering is made considering the position of the label within the sentence. It is important in order to determinate if it is a real label or if it is an element of the sentence. Three-position classes are detected: 1) the labels that precede the panel description (e.g. a) ....); 2) labels that follow the panel description (e.g. ... (a)); and 3) labels that are contained in a Part of Speech (POS) description (e.g. labels preceded by words like in, from, and panel). The labels classified as Part of Speech are not considered as actual labels since they are used for reference or as proper names within the sentences. The third step of the pipeline detects the snippets of the caption associated with the labels. The caption is fragmented in snippets using the filtered tuple (label, position). A set of hand-coded rules is applied to the labels (classified as pre or post description) in order to fragment it in snippets. The rules consider the position within the tuple and the class of the corresponding label. Depending on the class, the text snippet, contained between two labels, is assigned to the preceding label (pre-description class) or to the following label (post-description class). If a text snippet is associated to label ranges or sequences, it is duplicated as many times as the number of labels in the range/sequence (e.g. the range A-D duplicates the text snippet for A, B, C, and D). The example reported below shows the input and output of the algorithm applied to a caption. In this case, the labels belong to the post-description class. INPUT: Immunohistochemical expression of c-MET in human prostate cancer.c-MET is highly expressed in scattered prostate cancer cells (A), and particularly at invasive fronts within peri-prostatic fat tissue (B); arrowheads indicate positive cells. Original magnification 100x.

OUTPUT: A: Immunohistochemical expression of c-MET in human prostate cancer.c-MET is highly expressed in scattered prostate cancer cells. arrowheads indicate positive cells. Original magnification 100x. B: , and particularly at invasive fronts within peri-prostatic fat tissue. arrowheads indicate positive cells. Original magnification 100x.

The caption splitting component is evaluated on a partition of PudMed dataset. The partition includes 193 (originally 250) samples and it comes without ground truth annotations, thus required to manually create it. Each sample within the partition includes a caption and the corresponding compound image. In order to evaluate the performance of the algorithm, the captions were manually annotated by a human operator. For each of the captions, a list of labels and the corresponding text snippets was identified. After the process of manual annotation of the text snippets, only 193 samples were selected to compose the partition. The partition originally included 250 samples, but 57 of the 250 captions were discarded because within it was not possible to identify the labels or the corresponding text snippets.

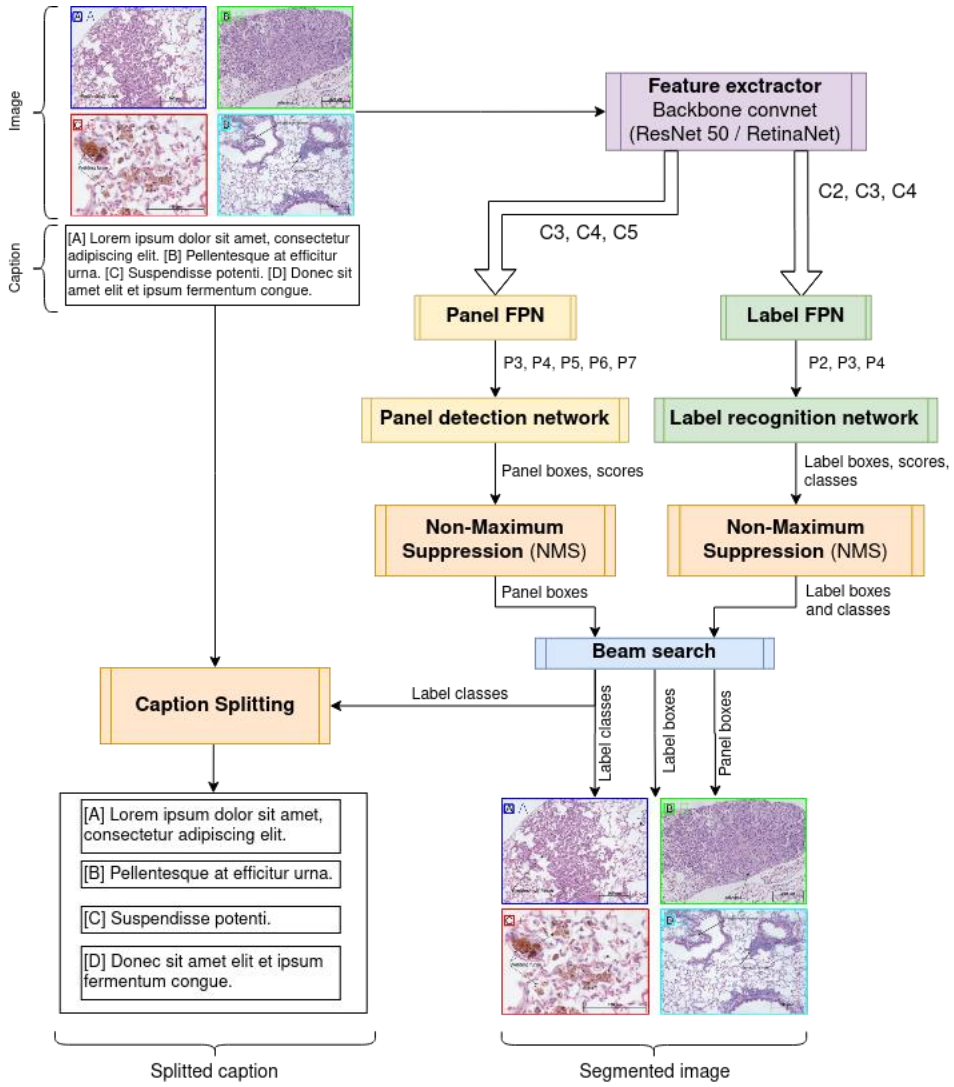

Figure 2 : Detailed view of the full procedure to separate compound images

3 Results

The pipeline to perform all the phases required to separate compound images was completed, tested and made publicly available on GitHub[2] and linked from the ExaMode web page[3]. The implementation code relies on the Pytorch library [24] through Facebook Detectron 2 API [25]. The code for each sub-task was tested independently, providing the results summarized below, in order to allow the comparison of performance with previous works.

The goal of panel splitting is to localize the panels within a compound figure. This task can be described as single class object detection problem. This task was first proposed within the ImageCLEF 2013 challenge [11] and re-proposed within the ImageCLEF 2016 edition. We chose to evaluate our model for panel splitting on the 2016 data set to be able to compare our results to several recent works.

The ImageCLEF 2016 data set is limited by including annotations for panels only. Hence, it could not be used to evaluate performance of the system for the other tasks (caption splitting, label recognition and thus panel segmentation). The ImageCLEF data were divided in the following way to train the neural network: 6783 samples (81% of the data set) were used for training and 1,614 for testing.

To overcome this limitation, Zou et al. [12] gathered a new data set that will now be referenced as the *Panel Seg data set*. The authors have extracted 10,642 figures from the original PubMed Central data set 4 and annotated both the panels and the labels. Evaluation of label recognition and panel segmentation tasks have then been conducted on this data set. The Panel Seg data were divided in the following way to train the neural network: 9,642 samples were used for training 90%) and the remaining 1000 images were used for testing.

The results for the panel splitting task (on the ImageCLEF 2016 data set and on the Panel Seg data set) are reported respectively in Table 1 and in Table 2. The used evaluation metrics are precision, recall, and Mean Average Precision (MAP). Since in the panel splitting task we only detect a single class (panels), the MAP metric corresponds to the average precision. The performance of the algorithm on the ImageCLEF data set was also evaluated with the "ImageCLEF accuracy", a metric defined specifically for the panel splitting task and described in detail [11]. The performance of the model proposed by Zou et al. are also provided.

Model	ImageCLEF accuracy	Precision	Recall	mAP
Tsutsui et al.	84.6	87.5	75.1	77.3
Zou et al. (ResNet 152)	85.1	89.8	**78.9**	**78.4**
Zou et al. (ResNet 50)	83.8	**90.0**	77.7	78.6
Ours (ResNet 50)	**85.2**	88.2	77.4	75.8

Table 1 : Panel splitting results (ImageCLEF 2016 data set)

Model	Precision	Recall	mAP
Zou et al.	82.9	91.1	**88.4**
Ours	68.8	**92.0**	89.3

Table 2 : Panel splitting results (Panel Seg data set)

2 https://github.com/GaetanLepage/panel-seg
3 https://www.examode.eu/software/

Label recognition was evaluated on the *Panel Seg* data set. The results for the label recognition task are reported in Table 3. Also in this case, the used evaluation metrics are precision, recall, and mean Average Precision (MAP). The performance of the model proposed by Zou et al. are also provided. Even though our implementation mainly follows a very similar architecture, the results of the model developed in the context of this work are significantly better in terms of precision, recall and MAP.

Model	Precision	Recall	mAP
Zou et al.	12.4	85.9	55.0
Ours	**52.97**	**88.26**	**53.3**

Table 3 : Label recognition results

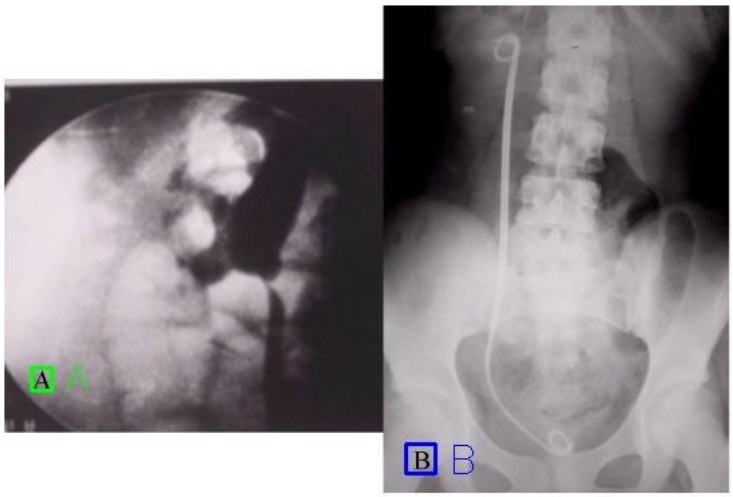

Figure 3 : Example of label recognition output

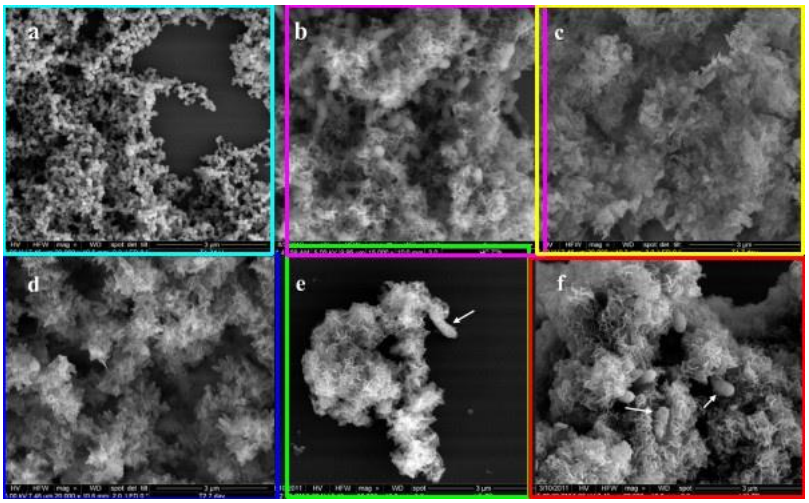

Figure 4 : Example of panel splitting output

A

```
caption:
(A) Transrectal biopsy produced a diagnosis of poorly differentiated adenocarcinoma with small ce
ll NE carcinoma. HE staining produced an initial diagnosis of Gleason pattern 5b poorly different
iated adenocarcinoma (magnification, *100). (B) PSA staining revealed that PSA-positive and -nega
tive cells were intermixed in the biopsy sample (magnification, *100). HE, hematoxylin and eosin;
 NE, neuroendocrine; PSA, prostate-specific antigen.
['A', 'B']
label A
 Transrectal biopsy produced a diagnosis of poorly differentiated adenocarcinoma with small cell
NE carcinoma.. HE staining produced an initial diagnosis of Gleason pattern 5b poorly differentia
ted adenocarcinoma (magnification, *100).
ground_truth A
A: Transrectal biopsy produced a diagnosis of poorly differentiated adenocarcinoma with small cel
l NE carcinoma. HE staining produced an initial diagnosis of Gleason pattern 5b poorly differenti
ated adenocarcinoma (magnification, ×100).
label B
 PSA staining revealed that PSA-positive and -negative cells were intermixed in the biopsy sample
 (magnification, *100).. HE, hematoxylin and eosin; NE, neuroendocrine; PSA, prostate-specific an
tigen.
ground_truth B
B: PSA staining revealed that PSA-positive and -negative cells were intermixed in the biopsy samp
le (magnification, ×100). HE, hematoxylin and eosin; NE, neuroendocrine; PSA, prostate-specific a
ntigen.
```

B

```
caption:
Immunohistochemical expression of c-MET in human prostate cancer.c-MET is highly expressed in sca
ttered prostate cancer cells (A), and particularly at invasive fronts within peri-prostatic fat t
issue (B); arrowheads indicate positive cells. Original magnification 100*.
['A', 'B']
label A
ET is highly expressed in scattered prostate cancer cells .  arrowheads indicate positive cells.O
riginal magnification 100*.
ground_truth A
A: Immunohistochemical expression of c-MET in human prostate cancer. c-MET is highly expressed in
 scattered prostate cancer cells ; arrowheads indicate positive cells. Original magnification 100
×.
label B
, and particularly at invasive fronts within peri-prostatic fat tissue .  arrowheads indicate pos
itive cells.Original magnification 100*.
ground_truth B
B: Immunohistochemical expression of c-MET in human prostate cancer. , and particularly at invasi
ve fronts within peri-prostatic fat tissue ; arrowheads indicate positive cells. Original magnifi
cation 100×.
```

C

```
caption:
Immunostain of left external iliac lymph nodes: (A) Positive prostate-specific antigen stain; (B)
 Positive CD-10 stain.
['A', 'B']
label A

ground_truth A
A: Immunostain of left external iliac lymph nodes: Positive prostate-specific antigen stain;
label B

ground_truth B
B: Immunostain of left external iliac lymph nodes: Positive CD-10 stain.
```

Figure 5 : Example of caption splitting output. Panel A) includes includes an example of caption well split (labels equal to their corresponding ground truth). Panel B) includes an example of caption well split (labels similar to their corresponding ground truth). Panel C) includes an example of caption considered as not well split.

Caption splitting is evaluated on a partition of the PubMed dataset, described in 2.2. The dataset includes captions and the corresponding ground truth (a set of sub-captions that were manually annotated). A metric to evaluate this specific problem was not presented before in literature. Therefore, we decided to assess the performance by measuring the percentage of captions that were correctly split by the algorithm. A sample is correctly split if all the predicted sub-captions are equal (or sufficiently similar) to the ground truth sub-captions (Fig. 5, panel A and panel B). Otherwise, a sample is not correctly split (Fig. 5, panel C). The evaluation is made considering the sentences (components) within a sub-caption according to the following rules. Two sub-captions are equal if both of them include the same components, in the same order (Fig. 5, panel A). Two sub-captions are similar if both of them has a common set of components (not necessarily in the same order) and the non comment components are exclusive for the corresponding label (e.g. the sub-caption of label b does not include text snippets linked to the label a) (Fig. 5, panel B). For each sample, the human operator manually assessed if the caption was well split in the text snippets. According to the mentioned rules, the algorithm was capable to divide the caption into subcaptions in 86% of the considered cases.

4 Conclusion

This work describes the complete pipeline to handle compound figure separation. The algorithms to separate compound figures and link them to related text are publicly released on Github and a link to them was also placed on the ExaMode website. The publication describing the algorithms and the results of their evaluation (which corresponds to the text of this deliverable in its preliminary form) is planned to be released by the end of July, since before submitting, it would be beneficial to test few more changes of the pipeline that might allow to improve the results.

As described in the results section, the fundamentals of the algorithm are promising. Nevertheless, some aspects are currently at an early development stage and will be refined. This is the case for instance of the unified neural network. Concerning label recognition, we plan to leverage the single character hypothesis for the labels. To achieve this, different solutions can be considered. For instance, the unified neural network could be used to infer the location of the labels and not their classes. A separate OCR algorithm might then be used on the extracted label patches to identify the label text. Otherwise, the*structure* of the labels within an image could be exploited. In fact, the labels usually respect some rules, as they are a contiguous sequence of indexes. An image might be classified into several*pre-defined* categories such as ’A, B, C’, ’1a, 1b, 2a, 2b’, ’1, 2, 3’ etc. Identifying the label structure may bring additional reliability to the label detection, thus allowing to improve the results. Panel segmentation might benefit some improvement of the beam search algorithm. Finally, caption splitting might also be improved with more hand-coded rules or with more innovative approaches based on deep learning.

5 References

[1] Henning Müller, Vincent Andrearczyk, Oscar Jimenez del Toro, Anjani Dhrangadhariya, Roger Schaer, and Manfredo Atzori. Studying public medical images from the open access literature and social networks for model training and knowledge extraction. In *International Conference on Multimedia Modeling*, pages 553–564. Springer, 2020.

[2] Kevin W Boyack, David Newman, Russell J Duhon, Richard Klavans, Michael Patek, Joseph R Biberstine, Bob Schijvenaars, André Skupin, Nianli Ma, and Katy Börner. Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PloS one*, 6(3), 2011.

[3] Hagit Shatkay, Nawei Chen, and Dorothea Blostein. Integrating image data into biomedical text categorization. *Bioinformatics*, 22(14):e446–e453, 2006.

[4] Aaron M Cohen and William R Hersh. A survey of current work in biomedical text mining. *Briefings in bioinformatics*, 6(1):57–71, 2005.

[5] Henning Müller, Nicolas Michoux, David Bandon, and Antoine Geissbuhler. A review of content-based image retrieval systems in medical applications—clinical benefits and future directions. *International journal of medical informatics*, 73(1):1–23, 2004.

[6] Sameer K Antani, L Rodney Long, and George R Thoma. Content-based image retrieval for large biomedical image archives. In *Medinfo*, pages 829–833, 2004.

[7] William Hsu, L Rodney Long, Sameer Antani, et al. Spirs: A framework for content-based image retrieval from large biomedical databases. *MedInfo*, 12:188–192, 2007.

[8] Henning Müller, Jayashree Kalpathy-Cramer, Charles E Kahn Jr, and William Hersh. Comparing the quality of accessing medical literature using content-based visual and textual information retrieval. In *Medical Imaging 2009: Advanced PACS-based Imaging Informatics and Therapeutic Applications*, volume 7264, page 726405. International Society for Optics and Photonics, 2009.

[9] Alba Garcia Seco de Herrera, Henning Müller, and Stefano Bromuri. Overview of the imageclef 2015 medical classification task. In *CLEF (Working Notes)*, 2015.

[10] Henning Müller, Fabrice Meriaudeau, Antonio Foncubierta-Rodríguez, Dimitrios Markonis, and Ajad Chhatkuli. Separating compound figures in journal articles to allow for subfigure classification. SPIE, Medical Imaging, 2013.

[11] Alba Garcia Seco De Herrera, Jayashree Kalpathy-Cramer, Dina Demner-Fushman, Sameer Antani, and Henning Müller. Overview of the ImageCLEF 2013 medical tasks. Valencia, Spain, September 2014. CEUR Workshop Proceedings. Meeting Name: CLEF 2013 Conference.

[12] Jie Zou, George Thoma, and Sameer Antani. Unified Deep Neural Network for Segmentation and Labeling of Multipanel Biomedical Figures. *Journal of the Association for Information Science and Technology*, 2019. _eprint: https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.24334.

[13] Pengyuan Li, Xiangying Jiang, Chandra Kambhamettu, and Hagit Shatkay. Compound image segmentation of published biomedical figures. *Bioinformatics*, 34(7):1192–1199, April 2018.

[14] Beibei Cheng, Sameer Antani, R Joe Stanley, and George R Thoma. Automatic segmentation of subfigure image panels for multimodal biomedical document retrieval. In *Document Recognition and Retrieval XVIII*, volume 7874, page 78740Z. International Society for Optics and Photonics, 2011.

[15] Jie Zou, Sameer Antani, and George Thoma. Localizing and Recognizing Labels for Multi-Panel Figures in Biomedical Journals. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 753–758, November 2017. ISSN: 2379-2140.

[16] Satoshi Tsutsui and David Crandall. A Data Driven Approach for Compound Figure Separation Using Convolu- tional Neural Networks. *arXiv:1703.05105 [cs]*, August 2017. arXiv: 1703.05105.

[17] Joseph Redmon and Ali Farhadi. YOLO9000: better, faster, stronger. *CoRR*, abs/1612.08242, 2016.

[18] William W Cohen, Richard Wang, and Robert F Murphy. Understanding captions in biomedical publications. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 499–504, 2003.

[19] Emilia Apostolova, Daekeun You, Zhiyun Xue, Sameer Antani, Dina Demner-Fushman, and George R. Thoma. Image retrieval from scientific publications: text and image content processing to separate multi–panel figures. *Journal of the American Society for Information Science*, 64:893–908, 2013.

[20] Mushtaq Ali, Le Dong, Yan Liang, Zongyi Xu, Ling He, and Ning Feng. A novel algorithm for extracting text labels and subfigure captions from multi-panel figure caption. In *2014 11th International Computer Conference on Wavelet Actiev Media Technology and Information Processing (ICCWAMTIP)*, pages 226–229. IEEE, 2014.

[21] Dayne Freitag and Nicholas Kushmerick. Boosted wrapper induction. In *AAAI/IAAI*, pages 577–583, 2000.