Comic MTL: optimized multi-task learning for comic book image analysis

Nhu-Van Nguyen, Christophe Rigaud, Jean-Christophe Burie

Noname manuscript No.
(will be inserted by the editor)

Multi-task model for comic book image analysis

Nhu-Van Nguyen[1] · **Christophe Rigaud**[1] · **Jean-Christophe Burie**[1]

Received: date / Accepted: date

Abstract Comic book image analysis methods often propose multiple algorithms or models for multiple tasks like panel and character (body and face) detection, balloon segmentation, text recognition, etc. In this work, we aim to reduce the processing time for comic book image analysis by proposing one model which can learn multiple tasks called Comic MTL instead of using one model per task In addition to the detection task and segmentation task, we integrate the relation analysis task for balloons and characters into the Comic MTL model. The experiments are carried out on DCM772 and eBDtheque public datasets which contain the annotations for panels, balloons, characters and also the associations between balloon and character. We show that the Comic MTL model can detect the association between balloons and their speakers (comic characters) and handle other tasks like panels, characters detection and balloons segmentation with promising results.

Keywords Comic book image analysis · Association balloon-character · Multi-task learning · CNN · Deep learning

1 Introduction

Digital comic content is mainly produced to facilitate transport, to reduce cost and to allow reading on screens of devices such as computers, tablets, and mobile phones. To access digital comics in an accurate and user-friendly experience on all mediums, it is necessary to extract and identify comic book elements [5]. Accordingly, the relations between these elements could be investigated further to assist the understanding of the digital form of comic books

[1] Laboratoire L3i
Université de La Rochelle
17042 La Rochelle CEDEX 1, France
E-mail: {nhu-van.nguyen, christophe.rigaud, jean-christophe.burie}@univ-lr.fr

by a computer. This strategy will help the user to retrieve information very precisely in the image corpus.

Comic book images are composed of different elements such as panel, speech balloon, text, comic character and their relations (e.g., read before, said by, thought by, addressed to). One research field focuses on analyzing automatically these elements aiming at automatic comics understanding. In early studies, each of them has been first addressed separately and then, in more recent studies, combined all together to get a deeper understanding of the story.

The image analysis community has investigated comics elements extraction for almost ten years, and vary from low-level analysis such as text recognition [3] to high-level analysis such as style recognition [6]. Text detection and recognition in comic book images is one of the most studied elements. The authors from [3,4,32] proposed several methods based on image processing techniques. In [20], the authors introduced techniques based on deep learning models to recognize text without the segmentation step of characters. Comic characters (protagonists) detection is one of the most challenging tasks because of the comic book creators are entirely free in the drawing of their comic characters, hence their appearance can change a lot from one comic book to another and even within individual comic books. Several methods have been proposed for recognizing comic characters based on deep neural network or hand-crafted feature related techniques [7,15,19,21,37]. The speech balloon is a key element in comics, which can have various shapes and contours. Current existing methods for extracting balloons are based on conventional techniques in image processing such as contours, region detection, [3,17,18,23,29]. The method presented in [33] can also associate speech balloons and comic characters. Panel extraction has been studied for a long time [39]. The evolution of screen quality and size of mobile devices such as smartphones and tablets has placed higher demands on accuracy panel extraction recently. Methods in [2, 14,16,34,38] rely on white line cutting, connected component labeling, morphological analysis or region growing. More recently, new methods based on watershed [26], line segmentation using Canny operator and polygon detection [16], region of interest detection [36], and recursive binary splitting [25] have been proposed.

All the existing methods based on deep learning models or conventional techniques treat each comic's element separately. This approach has been used because elements from comic book images are different and hence there is hardly an algorithm that can extract all elements at the same time. For more details about existing methods for comic book analysis, the readers are encouraged to read the survey works in [5,20].

In our work, we investigate a multiple tasks learning based approach which processes multiple elements simultaneously. Our approach can help to reduce the process duration of the comic analysis pipeline. Moreover, we propose a new neural network architecture which can detect the relationship between balloons and comic characters (body or face). In other words, our model can

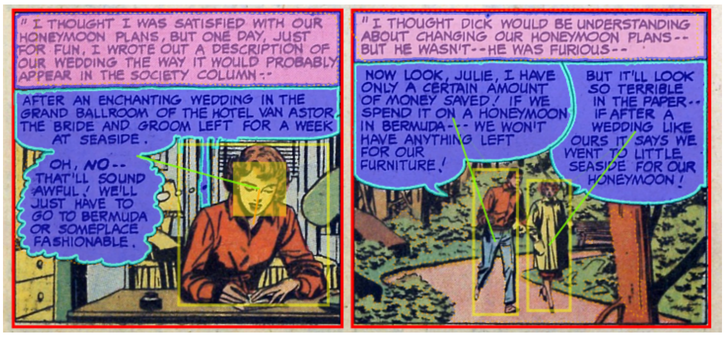
"I THOUGHT I WAS SATISFIED WITH OUR HONEYMOON PLANS, BUT ONE DAY, JUST FOR FUN, I WROTE OUT A DESCRIPTION OF OUR WEDDING THE WAY IT WOULD PROBABLY APPEAR IN THE SOCIETY COLUMN--
AFTER AN ENCHANTING WEDDING IN THE GRAND BALLROOM OF THE HOTEL VAN ASTOR, THE BRIDE AND GROOM LEFT FOR A WEEK AT SEASIDE.
OH, NO-- THAT'LL SOUND AWFUL! WE'LL JUST HAVE TO GO TO BERMUDA OR SOMEPLACE FASHIONABLE.
"I THOUGHT DICK WOULD BE UNDERSTANDING ABOUT CHANGING OUR HONEYMOON PLANS-- BUT HE WASN'T--HE WAS FURIOUS--
NOW LOOK, JULIE, I HAVE ONLY A CERTAIN AMOUNT OF MONEY SAVED! IF WE SPEND IT ON A HONEYMOON IN BERMUDA--- WE WON'T HAVE ANYTHING LEFT FOR OUR FURNITURE!
BUT IT'LL LOOK SO TERRIBLE IN THE PAPER-- IF AFTER A WEDDING LIKE OURS IT SAYS WE WENT TO LITTLE SEASIDE FOR OUR HONEYMOON!

Fig. 1 The elements in comic book images that we are dealing with in our work, best view in color. Panels are surrounded by a red line; speech balloons are highlighted in blue; narrative text boxes in violet, faces and characters is in yellow (highlighted and framed respectively); relations balloon-character are illustrated with a green line.

associate a balloon with its speaker(s). Our main contributions are summarized as follows.

- We propose Comic MTL, an end-to-end CNN model which perform multi-tasks learning to simultaneously detect, segment and classify comic elements;
- We analyze the relation between speech balloons and comic characters (body or face) by developing a binary classifier into the Comic MTL model;
- We extend the DCM772 public dataset [20] by adding more annotations (segmentation masks for balloons, bounding boxes for narrative text boxes, links for balloon-character relations);
- We experiment with our model on the two public datasets DCM772 & eBDtheque [11] and demonstrate the Comic MTL can achieve stat-of-the-art or better performance for the detection and segmentation task of comic elements plus promising results for balloon-character relation analysis;
- We show that our multi-task model is clearly saving a lot of time for comic book image analysis process compared to the analysis process which relies one model for each task.

In the next section, we present the works related to MTL models and balloon-character relation analysis. In Section 3, we describe our approach, the architecture and the implementation details of the proposed model (Comic MTL). Section 4 details how we extend the DCM772 dataset. In Section 5 we present and discuss the results of our experiments and conclude this work in Section 6.

2 Related works

With the success of deep learning approaches, recent works, mostly based on neural network models, have been proposed to extract comic elements [7, 10, 21, 28, 35]. These works include tasks to detect each element separately such as balloon segmentation, panel detection, text recognition or character detection. Hence, the processing pipeline from global images analysis to precise content extraction takes much time. In order to reduce the overall processing time for comic book images analysis, we investigate an approach which can handle multiple elements in one deep learning model. Besides extracting comic elements, our model can also identify the association between speech balloons and comic characters. In this section we present the existing MTL models and existing works on the association balloon-character.

2.1 Multi-Task Learning

MTL models aim at learning multiple related tasks jointly to improve the generalization performance of all the tasks [41]. The work in [41] gives a detailed survey for MTL models. We summarize below some popular MTL works based on Convolutional Neural Network (CNN) for computer vision domain. Zhang *et al.* [42] proposed a deep CNN model which learns jointly different tasks like facial landmark detection, head pose estimation, gender classification, age estimation, facial expression recognition and facial attribute inference using shared CNN layers. Abdulnabi *et al.* [1] proposed a CNN model to predict attributes in images using individual CNNs for each task and fuses different CNNs in a common layer via a sparse transformation. In [40], the authors proposed a multi-task model to learn the rotate facial task with an auxiliary task as the reconstruction of original images based on generated images. Another popular model is the Mask R-CNN [12] which jointly learns the object detection task and object segmentation task by using a shared CNN layers and shared object proposal network. This model achieves state-of-the-art performances for both segmentation and detection tasks. However, the Mask R-CNN model requires the segmentation masks of the objects.

In order to accomplish the training in our work, we have some elements (classes) that do not have object masks (for example annotations for panels, comic characters and text corresponds to bounding boxes) and elements (classes) that have segmentation masks (balloons). Hence, we extend the Mask R-CNN model to learn both detection and segmentation tasks for panels, comic characters (detection), and balloons (segmentation) using both segmentation masks and bounding boxes.

2.2 Relationship balloon-character

The association between balloons and comic characters can create annotations corresponding to story understanding (dialog analysis, situation retrieval). However, whether scanned or digital-born, these relations are not directly encoded in the image but the reader understands them according to other information present in the image. There are few papers in the literature on the topic of relation analysis among comic elements. From our knowledge, only [33] has proposed a method to associate a balloon with its speaker (comic characters). The authors firstly detected panels, balloons and tails of balloons (arrow pointing toward the speaking character), and comic characters; then they used a geometric graph for each panel where vertices are spatial positions of tail and comic character centroids. Edges are straight-line segments (associations). They formulated an optimization problem by searching for the best pairs (2-tuples) of tail and character corresponding to associations.

In our work, we integrate the association balloon-character into the multi-task Comic MTL model. While the method in [33] requires prior knowledge about the positions of panels, balloons, characters, and especially the balloon tails, our method does not require any prior information about characters or balloons. Finally, our new model can learn from different kind of annotations (balloon masks, panels and characters bounding boxes, and balloon-character associations) to detect panels and characters, segment balloons, and detect the associations between detected characters and segmented balloons.

3 Proposed model: Comic-MTL

In this section, we present our proposed Comic-MTL model for comic image analysis which aims at extracting characters, panels, speech balloons, narrative text boxes, and the associations between characters and balloons. Our model is based on the state-of-the-art Mask R-CNN model in instance segmentation [12]. Firstly, in order to learn simultaneously the detection and the segmentation tasks, we modify the loss function of the mask branch in Mask R-CNN to take into account the origin of annotations (masks or bounding boxes). Secondly, we add an additional branch which contains a PairPool layer and a binary classifier to detect associations from all possible pairs (a pair contains a balloon and a comic character). The classifier outputs the probabilities of a pair to be ‘‘has-a-link” and ‘‘has-not-a-link”. The additional branch requires an extraction step of relevant features for the pairs of balloon-character. We describe our two modifications in the next sub-sections and illustrate them in Figure 4.

3.1 Multi-task learning from bounding boxes and object masks

We consider a detection/segmentation problem where there are classes with bounding box and segmentation mask annotations. In our work, we have to deal with comic datasets where panels and characters are annotated with bounding boxes and balloons are annotated with masks. While the state-of-the-art model Mask R-CNN [12] can predict both bounding boxes and masks of the objects, it requires the mask annotations for all classes to train the model. In order to learn jointly the detection task and segmentation task from bounding boxes and object masks, we enhance the model Mask R-CNN where

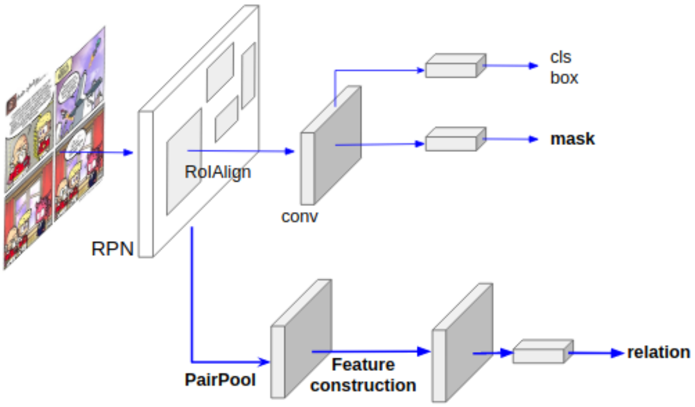

Fig. 2 The Comic MTL framework for comic book image analysis.

only the object from classes with mask annotations can contribute to the loss in the mask branch.

In the Mask R-CNN model, the multi-task loss on each sampled Region of Interest (RoI) in the total N sampled RoIs is defined as

$$L = L_{cls} + L_{box} + L_{mask} \tag{1}$$

where L_{cls} and L_{box} are the loss for the detection branch as in Faster R-CNN [27]. L_{mask} is the binary cross-entropy loss in the mask branch of Mask R-CNN [12]. In our model, we simply apply L_{mask} only for RoIs associated with ground-truth classes M which have the mask annotations. For RoIs associated with ground-truth classes $K = N - M$ which have only bounding box annotations, we optimize only the detection branch. Note that for classes that have segmentation masks, we can always extract the bounding box from a segmentation mask.

The binary cross-entropy loss L_{mask} in Comic MTL is defined in the Equation 2, only including k^{th} mask if the RoI is associated with the ground truth class k which has ground truth segmentation masks.

$$L_{mask} = -\frac{I^k}{m^2} \sum_{1 \leq i,j \leq m} \left[y_{ij} \log \hat{y}^k_{ij} + (1 - y_{ij}) \log(1 - \hat{y}^k_{ij}) \right] \tag{2}$$

where y_{ij} is the label of a pixel (i, j) in the ground-truth mask for the RoI of size mxm; y^k_{ij} is the predicted value of the same pixel in the mask output for the ground-truth class k. $I^k = 1$ if class k has ground truth segmentation masks, and $I^k = 0$ if class k has only ground truth bounding boxes.

3.2 Relation analysis for balloon-character pairs

We address the relationship analysis between balloons and comic characters by considering it as a binary classification problem. Each pair of balloon-character needs to be classified as ‘‘has a link” or ‘‘has not a link”. A pair of balloon-character which has a link means that the character says the text from the corresponding balloon. We suppose that a character may link to multiple balloons but a balloon can only link to one character (even if we know that this is not always the case). We do not pay attention to balloon and a character being in the same panel or not. We assume that the model will learn these features automatically from training examples. In order to make a binary classifier for the relation analysis, we add a new branch to the model Mask R-CNN.

Mask R-CNN summary : Similar to Faster R-CNN, in the first stage, Mask R-CNN use the region proposal network (RPN) [27] to get candidate object bounding boxes. Then in the next stage, it uses RoIAlign [12] to extract the embedded features of these candidate boxes to feed in parallel into a categorical classifier, a bounding-box regressor, and a mask predictor.

Comic MTL : We add an additional branch to the model Mask R-CNN that takes N pair combinations of top anchor candidates from the RPN network as inputs, then output the relation class (“has a link” or “has not a link”) of each pair. We optimize the branch with the binary cross-entropy loss, in parallel to other branches. The additional relation classifier output is distinct from the class, box and mask outputs, requiring a pairs combination step and a feature construction step for all pairs. We present these two important steps in the following.

Balloon-character pairs combination - PairPool : Instead of taking into account all combinations of candidate bounding boxes from the RPN stage, we sample all combination between the candidate boxes which have the best overlap (by a threshold α) with the ground-truth balloons and the candidate boxes which have the best overlap (by the same threshold α) with the ground-truth characters. This step can reduce a lot of possible pairs by removing the bad candidate boxes (boxes with overlap smaller than α). A pair is considered positive if its corresponding balloon and character in the ground truth has a link. We add all positive pairs to the pool and randomly add up-to the same number of positive pairs for negative pairs to the pool, the rest of negative pairs are ignored in the optimization process. Note that in most cases, the total number of negative pairs is bigger than the total number of positive pairs. Next, we pad the set of pairs with zero or trim the set of pairs to N pairs, where N is configurable. We need to fix the number of pairs because it is the input size for the additional branch.

In our proposed model, we define a multi-task loss on each sampled RoI as

$$L = L_{cls} + L_{box} + L_{mask} + L_{rel} \tag{3}$$

The first three components were presented in the previous section. The binary classifier in the new branch (L_{rel}) is optimized with the binary cross-entropy loss.

$$L_{rel} = -\frac{1}{N}\sum_{i}^{N}\left[y_i' \log(y_i) + (1 - y_i')\log(1 - y_i)\right] \tag{4}$$

where N is the number of input pairs; y_i' is the relation class of the pair i (positive or negative) in ground truth; y_i is the relation class predicted of the pair i.

Features construction : To feed the classifier, a simple reuse of the shared features as other branches to feed into this new relation branch is not enough. Indeed, the relation of a balloon and a character does not depend on the individual features of each bounding box but rather depend on multiple features such as individual visual features, visual features of the union of the two boxes, features related to the positions of the two boxes.

We need to encode the visual layout of the pair balloon-character and also the spatial layout of these two elements. Thus, unlike other branches which use a shared feature, we propose to use a combined feature of 1) the visual features of the two bounding boxes and their union, 2) the spatial features.

For the visual features of the two bounding boxes and their union, we reuse the shared features as in Mask R-CNN; in addition to each individual bounding box, we take into account the box equal to the union of the two bounding boxes in the pair balloon-character. This feature allows preserving the global visual information of a balloon and its speaker.

For the spatial features, let $b = [x_b, y_b, w_b, h_b]$ and $c = [x_c, y_c, w_c, h_c]$ denote two bounding boxes of a pair, where (x, y) are the coordinates of the center of the box, and (w, h) are the width and height of the box, respectively. We encode the spatial features with 4-dimensional vectors which are invariant to translation and scale transformations:

$$\left[\frac{x_b - x_c}{w_b}, \frac{y_b - y_c}{y_b}, \frac{x_b - x_c}{w_c}, \frac{y_b - y_c}{y_c}, \frac{b \cap c}{b \cup c}\right] \tag{5}$$

The first four features represent the normalized translation between the two boxes, the fifth feature is the overlap between boxes. In this paper, these features are used directly and are concatenated with the visual features at the last layer to form the final features.

Network architecture : To experiment with the Comic MTL model, we instantiate it with the default architecture used in Mask R-CNN. There are four parts of the architecture: (1) the convolutional backbone architecture used for feature extraction over an entire image, (2) the network head for bounding-box recognition (classification and regression), (3) the network head for mask prediction that is applied separately to each RoI and (4) the network head for the relation classifier. We experiment with the backbone architecture using ResNet-50 together with the Feature Pyramid Network (FPN). This backbone with ResNet-50 and FPN is a common choice used in many works [12, 27]. For the network heads, we closely follow the architecture of Mask R-CNN and we add the additional relation branch as in Figure 3.

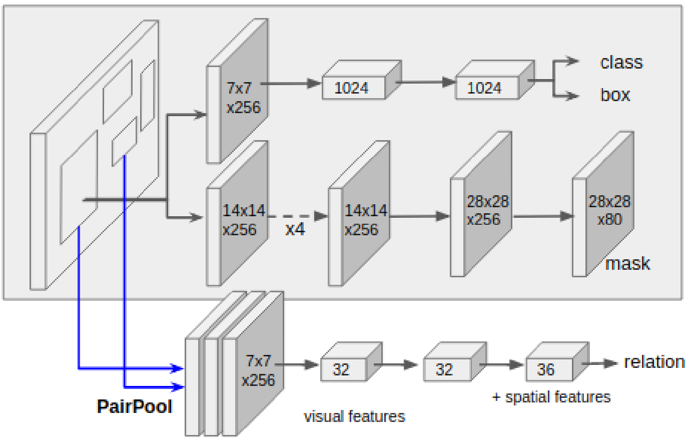

Fig. 3 The Comic MTL head architecture. We extend the existing Mask R-CNN heads [12] to which a relation branch is added. After the PairPool step, a stack of 3 features from the first box, the second box and the union of 2 boxes, is constructed for each pair. In the figure, numbers denote spatial resolution and channels. Arrows denote either convolutional or fully connected layers. The last 3 layers in the relation branch are fully connected layers.

3.3 Implementation details

Training: As in Fast R-CNN, a RoI is considered positive if it has an Intersection over Union (IoU) with a ground-truth box of at least 0.5 and negative otherwise. It is hard to train from scratch the model using small datasets, we leverage transfer learning to train our model from a pre-trained network on the ImageNet dataset [8] for the backbone part of the architecture. We firstly train the head layers for 10 epochs then fine-tune all layers for 40 epochs with a learning rate of 0.001 which is decreased by 10 at the 30^{th} epoch. We use a weight decay of 0.0001 and a momentum of 0.9. The parameter α is set to 0.6, obtained from experiments. In a comic page, the number of balloons and characters are about 15 and 10 respectively (see Section 4.2) so we set the parameter N to 150 which can cover almost all possible combinations and keeps the computation cost low for the relation branch. The training is done with one Nvidia TitanX GPU.

Inference : During the tests, instead of considering the outputs of the RPN network as in training, we take the outputs of the detection branch to construct the pairs of detected balloons and characters to feed the relation classifier. We run the relation prediction branch on these combinations (pairs of balloon-character). This is similar to the approach for the mask branch in Mask R-CNN; although this differs from the parallel computation used in training, it speeds up inference and improves accuracy (due to the use of fewer, more accurate RoIs). Because we only classify on the pairs from detected balloons and characters (we set maximum $N = 150$ pairs for each comic book image),

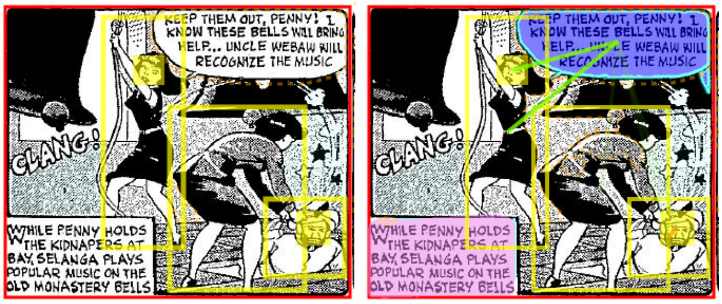
KEEP THEM OUT, PENNY! I KNOW THESE BELLS WILL BRING HELP.. UNCLE WEBAW WILL RECOGNIZE THE MUSIC
CLANG!
WHILE PENNY HOLDS THE KIDNAPERS AT BAY, SELANGA PLAYS POPULAR MUSIC ON THE OLD MONASTERY BELLS
KEEP THEM OUT, PENNY! I KNOW THESE BELLS WILL BRING HELP.. UNCLE WEBAW WILL RECOGNIZE THE MUSIC
CLANG!
WHILE PENNY HOLDS THE KIDNAPERS AT BAY, SELANGA PLAYS POPULAR MUSIC ON THE OLD MONASTERY BELLS

Fig. 4 The DCM772 dataset, before and after the extension. The additional elements are balloons (highlight in blue), narrative text boxes (highlight in violet), associations balloon-character (green line), and associations balloon-faces (green line).

Comic MTL adds marginal run-time to its Mask R-CNN counterpart **(see Todo)**.

4 Proposed dataset: extended DCM772

From our knowledge, eBDtheque is the only public dataset to date which satisfies our needs: has bounding box annotations for comic characters, segmentation masks for panels and balloons, and the association between balloons and its speakers (characters). However, because of the limit size of the eBDtheque dataset, we would like to experiment our approach on a bigger public dataset. We have extended the DCM772 dataset [22] for this purpose.

The dataset DCM772 is composed of 772 images from 27 golden age comic books collected the free public domain collection of comic books: Digital Comic Museum[1]. Images, annotations, and ground-truth tool are freely available here[2] to allow interested people to reproduce results or extend the dataset.

Because the DCM772 dataset contains only ground-truth bounding boxes for panels and characters (body + faces), we extended it by adding segmentation masks for speech balloons, bounding boxes for narrative text and also the association between balloons and its speakers (both face and characters). In order to realize this extension, we have applied a semi-automatic annotation process to add balloon masks and the associations.

4.1 Semi-automatic annotation process

We have used the balloon segmentation algorithm presented in [22] which combines a segmentation model [12] trained on the eBDtheque dataset and a traditional balloon segmentation method [31] to segment the balloon of the

[1] http://digitalcomicmuseum.com

[2] https://git.univ-lr.fr/crigau02/dcm_dataset

Table 1 Statistic of the DCM772 dataset

	Panels	Faces	Characters	Narrative boxes	Balloons	Links BC
772 pages	4500	5469	10902	1557	6415	5585
one page	5.83	7.08	14.12	2.02	8.31	7.23

DCM772 dataset. The generated balloons are stored in the CBML (Comic Book Markup Language) format and are then validated by human using an interactive tool which allows an expert clicking on a generated balloon to validate or invalidate it. After the balloon masks validation, all links between validated balloons and faces/characters are automatically generated, presented as a line connecting the balloon and the character or face, and stored in CBML files. The links (or associations) are then validated by the interactive tool. Note that due to the semi-automatic process, some segmentation masks has small quantity of error pixels, There are also some missing links because of the missing balloons or missing characters in the existing ground-truth of DCM772.

4.2 Statistic of the DCM772 dataset

Table 1 shows the statistic of the DCM772 dataset. There are 4500 panels in the dataset, which correspond to 5.83 panels per page. There are 10902 characters in the dataset, corresponding to 14.12 characters per page. There are 5469 face (7.08 per page), the number of face is smaller than the number of characters because there are many small characters with unclear faces and characters from behind. The number of associations in each page is about 7.23 while on the dataset it is 5585.

4.3 Train, valid, and test sets

In order to experiment with our Comic-MTL model, the DCM772 dataset is divided into 3 sets: train set contains 650 images, valid set contains 50 images and test set contain 72 images. In the DCM772 there are some ground truth problems such as missing bounding boxes, masks or associations. We consider it is a part of the challenge in real-case scenarios. However, to correctly evaluate the model we manually fix all annotation problems in the test set.

5 Experiments

In additional to the DCM772 dataset, we experiment with the Comic MTL model on the the public eBDtheque dataset [11]. This is the only public dataset to date which has bounding box annotations for comic characters, segmentation masks for panels and balloons, and the association between balloons and its speakers (characters). The eBDtheque dataset is composed of one hundred comic book images containing 850 panels, 1550 comics characters, 1092 balloons and 4691 text lines in total. More description can be found in the original paper [11]. Because of the limit size of the eBDtheque dataset, we run the cross-validation tests on five different training and testing sets. Each training set contains 90 images and each testing test contains 10 remaining images. The reported results are the average of these five validations.

In our experiment, we compare the results of the Comic MTL model with the results of the Mask R-CNN for balloons segmentation and Faster R-CNN for panel detection, narrative text detection, face, and character detections. We also compare with existing works which reported the results on DCM772 and eBDtheque datasets. Next, we report the result of relation analysis for balloons and characters using Comic MTL model. For the segmentation task, we compute the precision and recall on pixel-level. For the detection task, we follow the evaluation method used in the Pascal VOC challenge [9]. To a fair comparison between the Comic MTL model and Mask R-CNN model or Faster R-CNN model. We used the same configuration, and the same train/test sets without any augmentation to train and test three models: Faster R-CNN, Mask R-CNN and Comic MTL. All models use ImageNet pre-trained weights.

In the next three sub-sections, we discuss the comparison between the Comic MTL model with other methods on the DCM772 and eBDtheque datasets. Firstly we present the results for segmentation task (balloon); then we compare the results for detection task (panels, characters, faces, narrative boxes); next, we show the results for relation analysis; and finally we demonstrate the generalizability of the Comic MTL by testing the trained model DCM772 on the eBDtheque dataset and compare with other works.

5.1 Segmentation task

To evaluate the Comic MTL model for balloon segmentation, we train the state-of-the-art Mask R-CNN model for balloon segmentation and compare it to the Comic MTL model. We used the same configuration for both models Mask R-CNN and Comic MTL except that we use ground truth of balloons to train the Mask R-CNN model, while we use all ground truth (panels, balloons, narrative boxes, faces, characters, relations) to train the Comic MTL model. In order to compare with existing methods in [20], we use the recall (R), the precision (P), and the F-measure (F1) at pixel-level as metrics, therefore we chose the threshold that maximized the F-measure for Mask R-CNN and Comic MTL models. The result details are given in Table 9.

The neural network models Mask R-CNN and Comic MTL outperform all existing methods with a large margin on the eBDtheque, 19.32% in the F-measure. Compare to the Mask R-CNN, we can see a slightly lower value for Comic MTL of about 0.08% in the F-measure on eBDtheque and better value of 0.16 on DCM772 dataset. Overall, the Mask R-CNN model and the Comic MTL model have a similar performance. However note that the Mask

Table 2 Speech balloon segmentation performance (in percent).

Method	**DCM772**			**eBDtheque**		
	Recall	Precision	F1	Recall	Precision	F1
Arai [3]	-	-	-	18.70	23.14	20.69
Ho [13]	-	-	-	14.78	32.37	20.30
Rigaud [31]	-	-	-	69.81	32.83	44.66
Rigaud [29]	-	-	-	62.92	62.27	63.59
Mask R-CNN [12]	**90.07**	93.91	91.95	**75.31**	92.42	**82.99**
Comic MTL	89.50	**94.87**	**92.11**	74.94	**92.77**	82.91

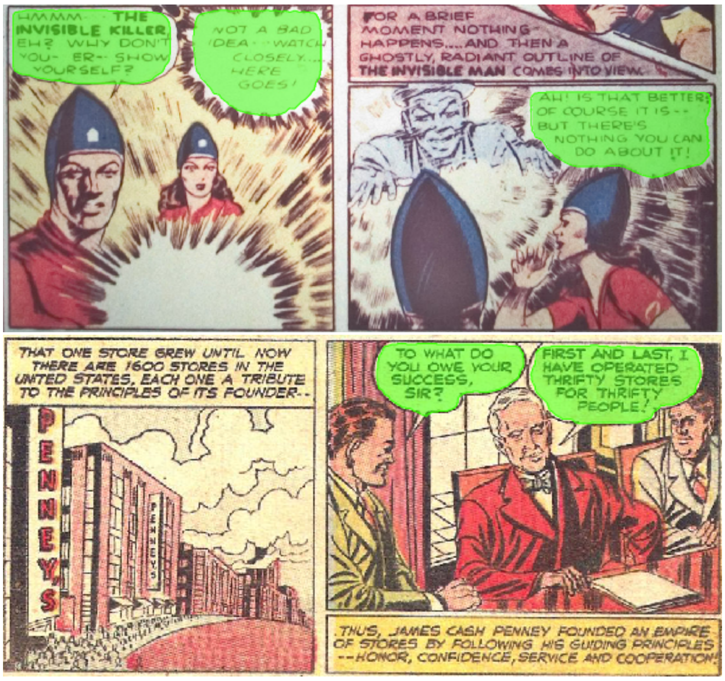
HMMM... THE INVISIBLE KILLER, EH? WHY DON'T YOU- ER-- SHOW YOURSELF?
NOT A BAD IDEA---WATCH CLOSELY.... HERE GOES!
FOR A BRIEF MOMENT NOTHING HAPPENS....AND THEN A GHOSTLY, RADIANT OUTLINE OF THE INVISIBLE MAN COMES INTO VIEW.
AH! IS THAT BETTER? OF COURSE IT IS-- BUT THERE'S NOTHING YOU CAN DO ABOUT IT!
THAT ONE STORE GREW UNTIL NOW THERE ARE 1600 STORES IN THE UNITED STATES, EACH ONE A TRIBUTE TO THE PRINCIPLES OF ITS FOUNDER--
PENNEYS
TO WHAT DO YOU OWE YOUR SUCCESS, SIR?
FIRST AND LAST, I HAVE OPERATED THRIFTY STORES FOR THRIFTY PEOPLE!
THUS, JAMES CASH PENNEY FOUNDED AN EMPIRE OF STORES BY FOLLOWING HIS GUIDING PRINCIPLES --HONOR, CONFIDENCE, SERVICE AND COOPERATION!

Fig. 5 Balloon segmentation by Comic MTL, the goods (two balloon in the right) and the bads (two segmented balloons on the left which have big gap compare to the boundaries of the ground truth

R-CNN can do the balloon segmentation only (because we do not have mask annotations for panels and characters), while balloons segmentation is one of the four tasks that one Comic MTL model can do. In Fig. 5, we illustrate an example of balloon segmentation by the Comic MTL model.

5.1.1 Detection task

We compare our Comic MTL to the baseline model Faster R-CNN (a start-of-the-art object detection model) for all four elements. To train the Faster R-CNN model, we use the same configuration as the training of the Comic MTL model, but the Faster R-CNN is trained for only the detection task

Table 3 Panels detection performance in percent. Note that the detection model of Ogawa [24]* is trained on Manga109 dataset, and the first three methods does not require training

Method	**DCM772**			**eBDtheque**		
	Recall	Precision	F1	Recall	Precision	F1
Arai [3]	-	-	-	58.0	75.3	65.6
Rigaud [34]	-	-	-	78.0	73.2	75.5
Rigaud [30]	-	-	-	81.2	86.6	83.8
Ogawa [24]*	-	-	-	73.3	76.4	74.8
Nguyen [20]	86.62	84.75	85.65	-	-	-
Faster R-CNN [27]	93.21	92.10	92.65	90.77	91.52	91.14
Comic MTL	**98.71**	**96.84**	**97.76**	**90.91**	**92.11**	**91.50**

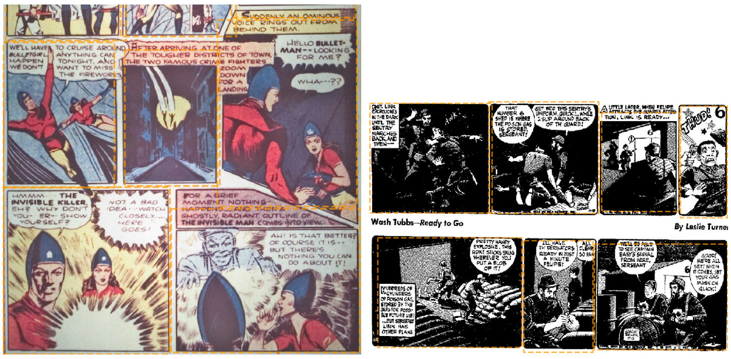
SUDDENLY AN OMINOUS VOICE RINGS OUT FROM BEHIND THEM.
WE'LL HAVE BULLETGIRL HAPPEN WE DON'T
TO CRUISE AROUND ANYTHING CAN TONIGHT, AND WANT TO MISS THE FIREWORKS
AFTER ARRIVING AT ONE OF THE TOUGHER DISTRICTS OF TOWN, THE TWO FAMOUS CRIME FIGHTERS ZOOM DOWN FOR A LANDING
HELLO BULLET-MAN-- LOOKING FOR ME?
WHA--??
HMMM-- THE INVISIBLE KILLER, EH? WHY DON'T YOU-- ER-- SHOW YOURSELF?
NOT A BAD IDEA-- WATCH CLOSELY..... HERE GOES!
FOR A BRIEF MOMENT NOTHING HAPPENS.... AND THEN A GHOSTLY, RADIANT OUTLINE OF THE INVISIBLE MAN COMES INTO VIEW.
AH! IS THAT BETTER? OF COURSE IT IS-- BUT THERE'S NOTHING YOU CAN DO ABOUT IT!
SGT. LINK CROUCHES IN THE DARK UNTIL THE SENTRY MARCHES BACK, AND THEN—
THAT NUMBER 6 SHED IS WHERE THE POISON GAS IS STORED, SERGEANT!
GET INTO THIS SENTRY'S UNIFORM, QUICK!... WHILE I SLIP AROUND BACK OF TH' GUARD!
A LITTLE LATER, WHEN FELIPE ATTRACTS THE GUARD'S ATTENTION, LINK IS READY...
THUD!
6
Wash Tubbs—Ready to Go
By Leslie Turner
PRETTY HANDY EXPLOSIVE, THIS RDX! STICKS SNUG WHEREVER YOU PUT A BLOB OF IT!
HUNDREDS OF CYLINDERS OF POISON GAS, STORED BY THE JAPS FOR POSSIBLE FUTURE USE! ...BUT SERGEANT LINK HAS OTHER PLANS.
I'LL HAVE TH' DETONATORS READY IN JUST A MINUTE FELIPE!
ALL CLEAR SO FAR
WE'LL BE ABLE TO SEE CAPTAIN EASY'S SIGNAL FROM HERE, SERGEANT
GOOD! WE'RE ALL SET! WHEN IT COMES, GET YOUR GAS MASK ON QUICK!

Fig. 6 Panel detection by Comic MTL, the goods and the bads. The bad detected panels are mostly the panels which are not in the form of a rectangle (while we use the bounding box to annotate a panel, we supposed a panel is in form of a rectangle)

of the four elements (without balloons). We also compare with other existing methods for panels and characters detection. In order to compare with existing methods [20,3,34,30,24], we use the recall (R), the precision (P), and the F-measure (F1) and chose the threshold that maximized the F-measure for the Comic MTL model. We followed PASCAL VOC criteria and used IoU >= 0.5 as threshold for good detections [9].

5.1.2 Panels and characters detection

Both datasets eBDtheque and DCM772 provide the ground truth for panels and characters detection. Table 3 shows the scores of existing methods (copied from [20]), the Faster-RCNN model and the Comic MTL model for panels and characters detection. On the DCM772, Comic MTL model comes at first place with 5.11% better compared to the second place of Faster-RCNN model. On the eBDtheque dataset, with small number of training examples and more complex panels, the two models give almost the same performance.

Table 4 shows the scores of existing methods and the model Comic MTL for characters detection. For this task, Comic MTL model gives similar performance as Faster R-CNN model. Both models outperform traditional methods, and 17.8% better than the CNN model in [24]. However, in [24], the authors

Table 4 Characters detection performance in percent. Note that the detection model of Ogawa [24]* is trained on Manga109 dataset, and the first three methods does not require training

Method	DCM772			eBDtheque		
	Recall	Precision	F1	Recall	Precision	F1
Arai [3]	-	-	-	-	-	-
Rigaud [34]	-	-	-	-	-	-
Rigaud [30]	-	-	-	21.6	40.5	28.2
Ogawa [24]*	-	-	-	42.2	58.0	48.8
Faster R-CNN [27]	65.25	78.93	71.44	61.56	71.23	66.04
Comic MTL	67.56	76.21	71.62	62.17	71.79	66.63

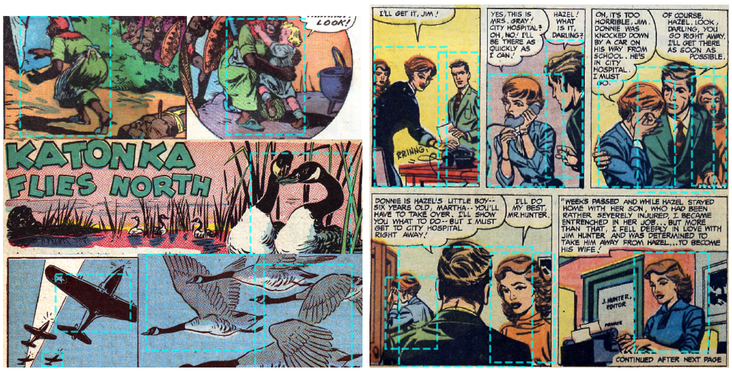
LOOK!
KATONKA FLIES NORTH
I'LL GET IT, JIM!
RRINNG!
YES, THIS IS MRS. GRAY! CITY HOSPITAL? OH, NO! I'LL BE THERE AS QUICKLY AS I CAN!
HAZEL! WHAT IS IT, DARLING?
OH, IT'S TOO HORRIBLE, JIM. DONNIE WAS KNOCKED DOWN BY A CAR ON HIS WAY FROM SCHOOL. HE'S IN CITY HOSPITAL. I MUST GO.
OF COURSE, HAZEL. LOOK, DARLING, YOU GO RIGHT AWAY. I'LL GET THERE AS SOON AS POSSIBLE.
DONNIE IS HAZEL'S LITTLE BOY-- SIX YEARS OLD, MARTHA-- YOU'LL HAVE TO TAKE OVER. I'LL SHOW YOU WHAT TO DO-- BUT I MUST GET TO CITY HOSPITAL RIGHT AWAY!
I'LL DO MY BEST, MR. HUNTER.
"WEEKS PASSED AND WHILE HAZEL STAYED HOME WITH HER SON, WHO HAD BEEN RATHER SEVERELY INJURED, I BECAME ENTRENCHED IN HER JOB... BUT MORE THAN THAT, I FELL DEEPLY IN LOVE WITH JIM HUNTER AND WAS DETERMINED TO TAKE HIM AWAY FROM HAZEL...TO BECOME HIS WIFE!
J. HUNTER, EDITOR
PRIVATE
CONTINUED AFTER NEXT PAGE

Fig. 7 Character detection by Comic MTL, the goods and the bads. The bad detected characters are often come from the very small characters (at the middle in the right part) or attached characters (at the top of the right part) or due to the small number of similar training examples (at the bottom of the right part)

Table 5 Faces and narrative boxes detection performance on DCM772

Method	Faces			Narrative boxes		
	Recall	Precision	F1	Recall	Precision	F1
Faster R-CNN [27]	70.99	73.12	72.04	81.85	80.72	81.28
Comic MTL	72.39	82.12	76.95	84.38	91.22	87.66

test a model trained on another dataset than the eBDtheque dataset (we also compare our Comic MTL model trained on DCM772 and the model in [24] on the eBDtheque dataset in Section 5.3). In Fig. 7, we show an example of character detection, the bad cases and the good cases.

5.1.3 Faces and narrative text boxes detection

The eBDtheque dataset does not provide the ground truth for character's faces and narrative text boxes. Hence, we compare the Comic MTL model to the Faster R-CNN model on the DCM772 dataset.

Table 5 compares the performance of the Faster R-CNN model and the Comic MTL model for faces and narrative text boxes detection. We can see that Comic MTL has better performance for both categories. If we examine

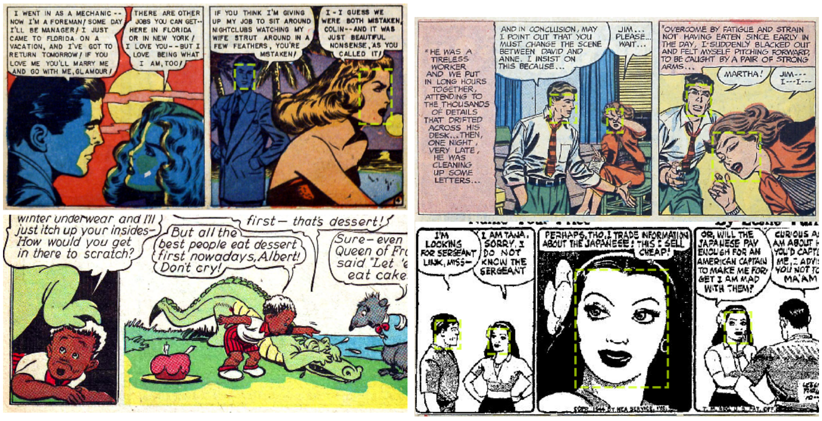
I WENT IN AS A MECHANIC -- NOW I'M A FOREMAN! SOME DAY I'LL BE MANAGER! I JUST CAME TO FLORIDA ON A VACATION, AND I'VE GOT TO RETURN TOMORROW! IF YOU LOVE ME YOU'LL MARRY ME AND GO WITH ME, GLAMOUR!
THERE ARE OTHER JOBS YOU CAN GET-- HERE IN FLORIDA OR IN NEW YORK! I LOVE YOU--BUT I LOVE BEING WHAT I AM, TOO!
IF YOU THINK I'M GIVING UP MY JOB TO SIT AROUND NIGHTCLUBS WATCHING MY WIFE STRUT AROUND IN A FEW FEATHERS, YOU'RE MISTAKEN!
I-I GUESS WE WERE BOTH MISTAKEN, COLIN--AND IT WAS JUST BEAUTIFUL NONSENSE, AS YOU CALLED IT!
"HE WAS A TIRELESS WORKER AND WE PUT IN LONG HOURS TOGETHER, ATTENDING TO THE THOUSANDS OF DETAILS THAT DRIFTED ACROSS HIS DESK...THEN, ONE NIGHT, VERY LATE, HE WAS CLEANING UP SOME LETTERS...
AND IN CONCLUSION, MAY I POINT OUT THAT YOU MUST CHANGE THE SCENE BETWEEN DAVID AND ANNE. I INSIST ON THIS BECAUSE...
JIM... PLEASE... WAIT...
"OVERCOME BY FATIGUE AND STRAIN NOT HAVING EATEN SINCE EARLY IN THE DAY, I SUDDENLY BLACKED OUT AND FELT MYSELF PITCHING FORWARD, TO BE CAUGHT BY A PAIR OF STRONG ARMS...
MARTHA!
JIM--- I---I---
winter underwear and I'll just itch up your insides– How would you get in there to scratch?
first– that's dessert!
But all the best people eat dessert first nowadays, Albert! Don't cry!
Sure– even Queen of Fr said "Let 'e eat cake
I'M LOOKING FOR SERGEANT LINK, MISS–
I AM TANA. SORRY, I DO NOT KNOW THE SERGEANT
PERHAPS, THO, I TRADE INFORMATION ABOUT THE JAPANESE! THIS I SELL CHEAP!
OR, WILL THE JAPANESE PAY ENOUGH FOR AN AMERICAN CAPTAIN TO MAKE ME FORGET I AM MAD WITH THEM?
CURIOUS AS AM ABOUT YOU'D CAPT ME, I ADVI YOU NOT TO MA'AM
COPR. 1944 BY NEA SERVICE, INC.
T. M. REG. U. S. PAT. OFF.

Fig. 8 Faces detection by Comic MTL, the goods and the bads.

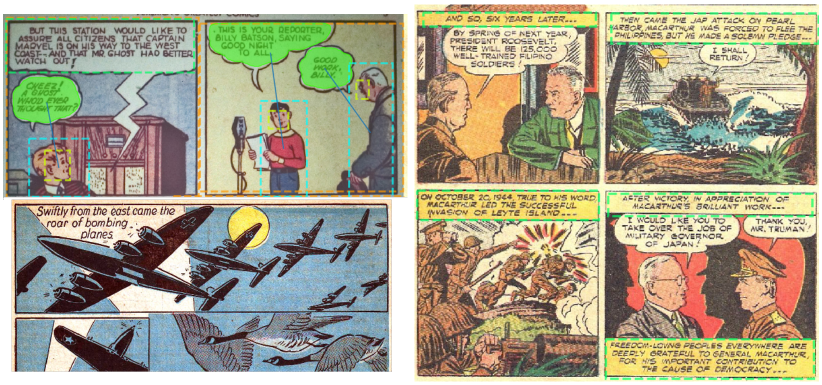
BUT THIS STATION WOULD LIKE TO ASSURE ALL CITIZENS THAT CAPTAIN MARVEL IS ON HIS WAY TO THE WEST COAST--AND THAT MR. GHOST HAD BETTER WATCH OUT!
CHEEZ! A GHOST. WHO'D EVER THOUGHT THAT?
THIS IS YOUR REPORTER, BILLY BATSON, SAYING GOOD NIGHT TO ALL
GOOD WORK, BILLY.
Swiftly from the east came the roar of bombing planes.
AND SO, SIX YEARS LATER...
BY SPRING OF NEXT YEAR, PRESIDENT ROOSEVELT, THERE WILL BE 125,000 WELL-TRAINED FILIPINO SOLDIERS!
THEN CAME THE JAP ATTACK ON PEARL HARBOR, MACARTHUR WAS FORCED TO FLEE THE PHILIPPINES, BUT HE MADE A SOLEMN PLEDGE...
I SHALL RETURN!
ON OCTOBER 20, 1944, TRUE TO HIS WORD, MACARTHUR LED THE SUCCESSFUL INVASION OF LEYTE ISLAND...
AFTER VICTORY, IN APPRECIATION OF MACARTHUR'S BRILLIANT WORK---
I WOULD LIKE YOU TO TAKE OVER THE JOB OF MILITARY GOVERNOR OF JAPAN!
THANK YOU, MR. TRUMAN!
FREEDOM-LOVING PEOPLES EVERYWHERE ARE DEEPLY GRATEFUL TO GENERAL MACARTHUR, FOR HIS IMPORTANT CONTRIBUTION TO THE CAUSE OF DEMOCRACY...

Fig. 9 Narrative text boxes detection by Comic MTL, the goods and the bads.

together all four elements (faces, characters, panels, narrative boxes) in the detection task, we can see that the Comic MTL has the same performance with the Fast R-CNN model for characters but it has better performance for other three classes. This evaluation shows that the Comic MTL can achieve the state of art or better performance for detection task. While both models have lower performance for complex object such as characters, the Comic MTL can give better performance for easier objects such as narrative boxes or panels.

5.2 Relation analysis

In this section, we evaluate the relation analysis between balloons and characters (body or face). The Comic MTL proposes a number of pairs balloon-character and classifies these pairs into two classes: has-link or not-has-link. A pair is in class "has-link" if the association between the corresponding balloon

Table 6 Balloon-character association performance in percent.

Method	**DCM772**			**eBDtheque**		
	Recall	Precision	F1	Recall	Precision	F1
Setting 1 [33]	-	-	-	-	18.01	-
Setting 2 [33]	-	-	-	-	93.32	-
Comic MTL	45.64	71.01	55.57	28.40	71.93	40.35

and character exists in the ground truth. We compare the Comic MTL model to the work in [33].

5.2.1 Association between balloons and characters body

In the work of [33], panels, balloons, balloon tails, and characters are necessary. The authors use two different settings: 1) panels, balloons, balloon tails, and characters are extracted automatically by some conventional extraction methods; 2) panels, balloons, and characters are available prior.

Table 6 shows the performance of Comic MTL model compared to [33]. We can see that with more training examples, the model trained on DCM772 gives better results than the model trained on eBDtheque. Compare to the work in [33] on the eBDtheque dataset, with the same settings where panels, balloons, and characters are extracted automatically (Setting 1 of [33]), the model Comic MTL gives better performance. When panels, balloons, and characters are available for the work of [33] (Setting 2), the Comic MTL is behind. One of the reasons is that the measured error in the Comic MTL model and the Setting 1 of [33] will be a combination of errors from the proposed method and other element extractions (e.g. missed speech balloons, missed comic characters or over-segmentation of panels for [33]). However, we believe that we can investigate further on the features extraction step of the proposed model Comic MTL to improve the results of relation analysis. There is useful information that has not been integrated into the model such as a balloon and a character should be in the same panel to have a link between them, or learning the direction of the balloon tail may help to improve the learning of its association with characters.

In Fig. 10, we show two bad associations (in the right) and some good associations (in the left). The reason for the first bad association is that the model can not detect and take into account the tail of the balloon. For the second wrong association, can be filtered in the post-processing if we take into account the good detected panels, with the apsumption that balloon and its speaker should be on the same panel.

5.2.2 Association between balloons and characters face

The eBDtheque does not provide the links for face-balloon. We report the results of a Comic MTL model which is trained to detect associations between faces (instead of characters body like the model in precedent sections) and balloons on the DCM772 dataset.

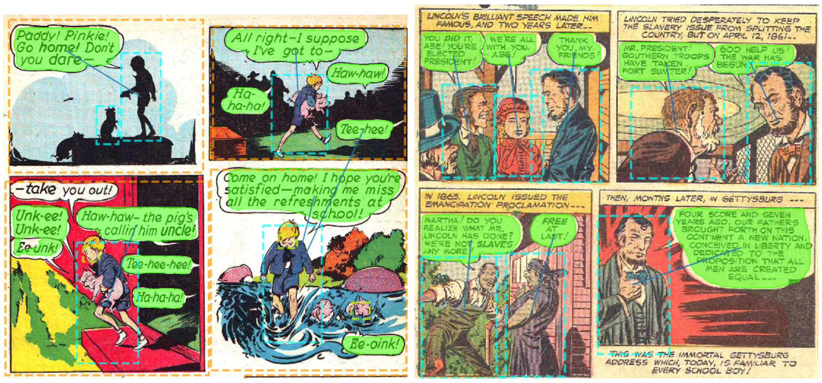
Paddy! Pinkie! Go home! Don't you dare—
All right—I suppose I've got to—
Haw-haw!
Ha-ha-ha!
Tee-hee!
—take you out!
Unk-ee! Unk-ee!
Ee-unk!
Haw-haw- the pig's callin' him uncle!
Tee-hee-hee!
Ha-ha-ha!
Come on home! I hope you're satisfied—making me miss all the refreshments at school!
Ee-oink!
LINCOLN'S BRILLIANT SPEECH MADE HIM FAMOUS, AND TWO YEARS LATER...
YOU DID IT, ABE! YOU'RE ELECTED PRESIDENT!
WE'RE ALL WITH YOU, ABE!
THANK YOU, MY FRIENDS!
LINCOLN TRIED DESPERATELY TO KEEP THE SLAVERY ISSUE FROM SPLITTING THE COUNTRY, BUT ON APRIL 12, 1861...
MR. PRESIDENT! SOUTHERN TROOPS HAVE TAKEN FORT SUMTER!
GOD HELP US! THE WAR HAS BEGUN!
IN 1863, LINCOLN ISSUED THE EMANCIPATION PROCLAMATION ---
MARTHA! DO YOU REALIZE WHAT MR. LINCOLN HAS DONE? WE'RE NOT SLAVES ANY MORE!
FREE AT LAST!
THEN, MONTHS LATER, IN GETTYSBURG ---
FOUR SCORE AND SEVEN YEARS AGO, OUR FATHERS BROUGHT FORTH ON THIS CONTINENT A NEW NATION, CONCEIVED IN LIBERTY AND DEDICATED TO THE PROPOSITION THAT ALL MEN ARE CREATED EQUAL ---
THIS WAS THE IMMORTAL GETTYSBURG ADDRESS WHICH, TODAY, IS FAMILIAR TO EVERY SCHOOL BOY!

Fig. 10 Associations detection by Comic MTL, the goods and the bads.

Table 7 Balloon-face association performance on DCM772 (in percent).

Method	Recall	Precision	F1
Comic MTL	44.83	69.04	54.36

Table 8 Detection task performance on eBDtheque (F1 scores). The Comic MTL gives better performance compare to traditional method. The CNN model of Ogawa [24] is trained on another dataset (the Manga109 dataset) as the Comic MTL model. Even if the Manga109 has more data than the DCM772, Comic MTL trained on DCM772 is still perform better.

Method	Panels	Faces	Characters	Narrative boxes
Arai [3]	65.6	-	-	-
Rigaud [34]	75.5	-	-	-
Rigaud [30]	83.8	-	28.2	-
Ogawa [24]	74.8	-	48.8	-
Comic MTL	84.23	-	66.06	-

The performance of balloon-face association detection is shown in the Table 7. Overall, the model can give better results in balloon-face association detection than balloon-character association detection. One of the reason that the results of faces detection is better than characters detection (see Tables 4 and 5).

5.3 Generalizability of the Comic MTL model

In the comic image analysis domain, we can have different datasets with very different drawing styles, color schemes and structures. We can use the model trained from one dataset to detect objects in a different dataset. That is the reason why we would like to know if the Comic MTL model can learn robust representations of comic elements and the relation balloon-character. To evaluate the generalization capacity of the Comic MTL model, we test the Comic MTL model trained with DCM772 on the unseen dataset eBDtheque and compare with other works in [3,13,24,29,31].

Table 9 Speech balloon segmentation performance (in percent). Comic MTL model is trained on the DCM772 train set, other methods does not require training. The Comic MTL model has the best performance in both recall and precision.

Method	Recall	Precision	F1
Arai [3]	18.70	23.14	20.69
Ho [13]	14.78	32.37	20.30
Rigaud [31]	69.81	32.83	44.66
Rigaud [29]	62.92	62.27	63.59
Comic MTL	63.11	**92.36**	74.98

Table 10 Balloon-character association performance of the Comic MTL model (in percent). In setting 1, the model is trained on DCM772 train set and tested on DCM772 test set; the models in setting 2 are trained on eBDtheque and tested on eBDtheque using 5 cross-validations; setting 3 has the model trained on DCM772 train set and tested on 100 images of the eBDtheque.

Method	Characters recall	Characters precision	F1
Comic MTL setting 1	45.64	71.01	55.57
Comic MTL setting 2	28.40	71.93	40.35
Comic MTL setting 3	25.22	66.88	36.62

Table 11 Inference and training time evaluation

Method	Comic MTL	Faster R-CNN [27]	Mask R-CNN [12]
Inference time on 1 image	0.273s	0.177	0.210
Training time on DCM772 train set	7h	5h25	4h30
Training time on eBDtheque train set	2h	1h35	1h20

5.4 Process time evaluation

We report the inference time for comic book image analysis using the Comic MTL model compare to the analysis using multiple models, one model for one task (so we need 3 models for detection, segmentation and relation analysis). The results are measured on one Nvidia TitanX GPU.

5.5 Visual results

We visualize an example of the model Comic MTL which includes all elements: panels, faces, characters detection, balloons segmentation and relation analysis. Fig. 11 show the results of a model trained on DCM772 train set on an image of DCM772 test set.

6 Conclusion

In this paper, we proposed the Comic MTL model which can handle multiple tasks on one CNN model: characters detection, panels and balloons segmentation, and balloon-character association analysis for comic book images. Another advantage of this model is that we can reduce the processing time

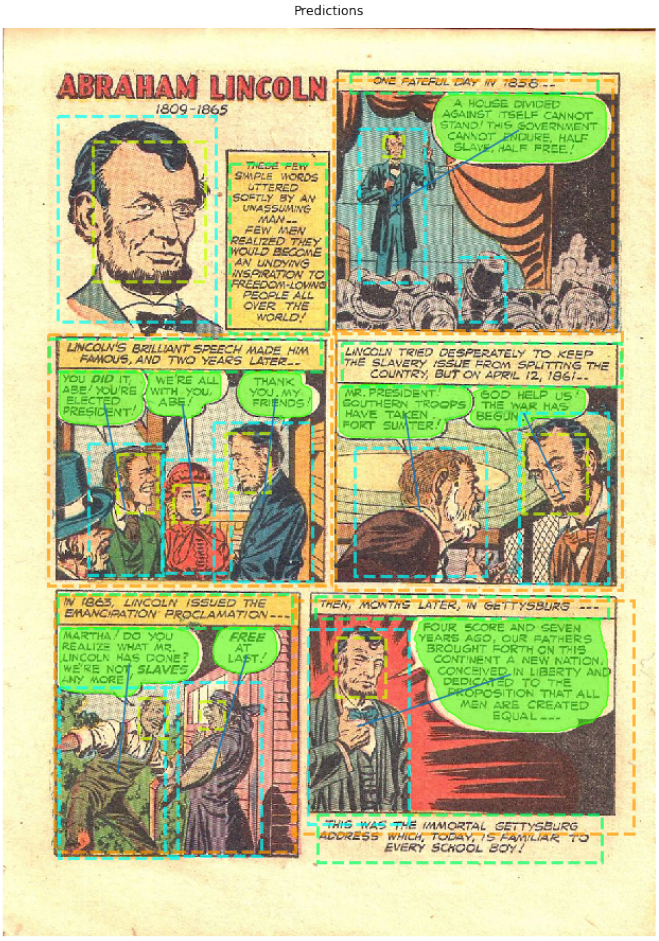
Predictions
ABRAHAM LINCOLN
1809-1865
THESE FEW SIMPLE WORDS UTTERED SOFTLY BY AN UNASSUMING MAN-- FEW MEN REALIZED THEY WOULD BECOME AN UNDYING INSPIRATION TO FREEDOM-LOVING PEOPLE ALL OVER THE WORLD!
ONE FATEFUL DAY IN 1858--
A HOUSE DIVIDED AGAINST ITSELF CANNOT STAND! THIS GOVERNMENT CANNOT ENDURE, HALF SLAVE, HALF FREE!
LINCOLN'S BRILLIANT SPEECH MADE HIM FAMOUS, AND TWO YEARS LATER--
YOU DID IT, ABE! YOU'RE ELECTED PRESIDENT!
WE'RE ALL WITH YOU, ABE!
THANK YOU, MY FRIENDS!
LINCOLN TRIED DESPERATELY TO KEEP THE SLAVERY ISSUE FROM SPLITTING THE COUNTRY, BUT ON APRIL 12, 1861--
MR. PRESIDENT! SOUTHERN TROOPS HAVE TAKEN FORT SUMTER!
GOD HELP US! THE WAR HAS BEGUN!
IN 1863, LINCOLN ISSUED THE EMANCIPATION PROCLAMATION---
MARTHA! DO YOU REALIZE WHAT MR. LINCOLN HAS DONE? WE'RE NOT SLAVES ANY MORE!
FREE AT LAST!
THEN, MONTHS LATER, IN GETTYSBURG ---
FOUR SCORE AND SEVEN YEARS AGO, OUR FATHERS BROUGHT FORTH ON THIS CONTINENT A NEW NATION, CONCEIVED IN LIBERTY AND DEDICATED TO THE PROPOSITION THAT ALL MEN ARE CREATED EQUAL---
THIS WAS THE IMMORTAL GETTYSBURG ADDRESS WHICH, TODAY, IS FAMILIAR TO EVERY SCHOOL BOY!

Fig. 11 An eBDtheque image analyzed by the Comic MTL model trained on DCM772. The figure shows the balloon masks (filled green mask), the bounding boxes of narrative text boxes (dashed green boxes), panels (dashed brown boxes) , faces (dashed yellow boxes), characters (dashed cyan boxes) and the lines that connect a balloon and a character (blue). There are 6 correct relations detected over a total of 9 relations. There is one wrong character detected at top right panel (stair railing) over a total of 9 characters. There is one missing narrative text box at the top over total of 4 narrative boxes. All balloons are detected.

to analyze comic book images. We compared the Comic MTL model with the model Mask R-CNN and other existing methods on the public eBDtheque dataset. Experiments confirm that the Comic MTL can handle multiple tasks in comic book images analysis (3 compared to 1 of existing models) with better results compared to the state-of-the-art performance.

7 Acknowledgement

This work is supported by the CPER NUMERIC programme funded by the Region Nouvelle Aquitaine, CDA, Charente Maritime French Department, La Rochelle conurbation authority (CDA) and the European Union through the FEDER funding".

References

1. Abdulnabi, A.H., Wang, G., Lu, J., Jia, K.: Multi-task CNN model for attribute prediction. IEEE Transactions on Multimedia **17**(11), 1949–1959 (2015)
2. Arai, K., Tolle, H.: Method for automatic e-comic scene frame extraction for reading comic on mobile devices. In: 7th Int. Conf. on Information Technology: New Generations, ITNG, pp. 370–375. IEEE Computer Society, Washington DC, USA (2010)
3. Arai, K., Tolle, H.: Method for real time text extraction of digital manga comic. International Journal of Image Processing (IJIP) **4**(6), 669–676 (2011)
4. Aramaki, Y., Matsui, Y., Yamasaki, T., Aizawa, K.: Text detection in manga by combining connected-component-based and region-based classifications. In: 2016 IEEE International Conference on Image Processing (ICIP), pp. 2901–2905 (2016)
5. Augereau, O., Iwata, M., Kise, K.: A survey of comics research in computer science. Journal of Imaging **4** (2018)
6. Chu, W.T., Cheng, W.C.: Manga-specific features and latent style model for manga style analysis. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1332–1336 (2016)
7. Chu, W.T., Li, W.W.: Manga facenet: Face detection in manga based on deep neural network. In: Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, pp. 412–415. ACM (2017)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR09 (2009)
9. Everingham, M., Eslami, S.M., Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. Int. J. Comput. Vision **111**(1), 98–136 (2015)
10. Fujino, S., Mori, N., Matsumoto, K.: Recognizing the order of four-scene comics by evolutionary deep learning. In: Distributed Computing and Artificial Intelligence, pp. 136–144 (2015)
11. Guérin, C., Rigaud, C., Mercier, A., Ammar-Boudjelal, F., Bertet, K., Bouju, A., Burie, J.C., Louis, G., Ogier, J.M., Revel, A.: eBDtheque: A representative database of comics. In: 2013 12th International Conference on Document Analysis and Recognition, pp. 1145–1149 (2013)
12. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. CoRR **abs/1703.06870** (2017)
13. Ho, A.K.N., Burie, J.C., Ogier, J.M.: Panel and Speech Balloon Extraction from Comic Books. 2012 10th IAPR International Workshop on Document Analysis Systems pp. 424–428 (2012)
14. In, Y., Oie, T., Higuchi, M., Kawasaki, S., Koike, A., Murakami, H.: Fast frame decomposition and sorting by contour tracing for mobile phone comic images. International journal of systems applications, engineering and development **5**(2), 216–223 (2011)
15. Khan, F.S., Anwer, R.M., van de Weijer, J., Bagdanov, A.D., Vanrell, M., Lopez, A.M.: Color attributes for object detection. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3306–3313 (2012)
16. Li, L., Wang, Y., Tang, Z., Gao, L.: Automatic comic page segmentation based on polygon detection. Multimedia Tools Appl. **69**(1), 171–197 (2014)
17. Liu, X., Li, C., Zhu, H., Wong, T.T., Xu, X.: Text-aware balloon extraction from manga. The Visual Computer **32**(4), 501–511 (2016)

18. Liu, X., Wang, Y., Tang, Z.: A clump splitting based method to localize speech balloons in comics. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 901–905 (2015)
19. Matsui, Y., Ito, K., Aramaki, Y., Yamasaki, T., Aizawa, K.: Sketch-based manga retrieval using Manga109 dataset. CoRR **abs/1510.04389** (2015)
20. Nguyen, N., Rigaud, C., Burie, J.: Digital comics image indexing based on deep learning. J. Imaging **4**(7), 89 (2018)
21. Nguyen, N.V., Rigaud, C., Burie, J.: Comic characters detection using deep learning. In: 2nd International Workshop on coMics Analysis, Processing, and Understanding, MANPU 2017, Kyoto, Japan, November 9-15, 2017, pp. 41–46 (2017)
22. Nguyen, N.V., Rigaud, C., Burie, J.C.: Digital comics image indexing based on deep learning. Journal of Imaging **4**(7), 89 (2018)
23. Obispo, S.L., Kuboi, T.: Element detection in Japanese comic book panels (2014)
24. Ogawa, T., Otsubo, A., Narita, R., Matsui, Y., Yamasaki, T., Aizawa, K.: Object detection for comics using manga109 annotations. CoRR **abs/1803.08670** (2018)
25. Pang, X., Cao, Y., Lau, R.W., Chan, A.B.: A robust panel extraction method for manga. In: Proceedings of the 22nd ACM International Conference on Multimedia, MM '14, pp. 1125–1128. ACM, New York, NY, USA (2014)
26. Ponsard, C., Ramdoyal, R., Dziamski, D.: An OCR-enabled digital comic books viewer. In: Computers Helping People with Special Needs, pp. 471–478. Springer (2012)
27. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In: C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, R. Garnett (eds.) Advances in Neural Information Processing Systems 28, pp. 91–99. Curran Associates, Inc. (2015)
28. Rigaud, C., Burie, J., Ogier, J.: Segmentation-free speech text recognition for comic books. In: 2nd International Workshop on coMics Analysis, Processing, and Understanding, 2017, Kyoto, Japan, November 9-15, pp. 29–34 (2017)
29. Rigaud, C., Burie, J.C., Ogier, J.M.: Text-independent speech balloon segmentation for comics and manga. In: Graphic Recognition. Current Trends and Challenges: 11th International Workshop, GREC 2015, Nancy, France, pp. 133–147. Cham (2017)
30. Rigaud, C., Guérin, C., Karatzas, D., Burie, J.C., Ogier, J.M.: Knowledge-driven understanding of images in comic books. International Journal on Document Analysis and Recognition (IJDAR) **18**(3), 199–221 (2015)
31. Rigaud, C., Karatzas, D., Van de Weijer, J., Burie, J.C., Ogier, J.M.: An active contour model for speech balloon detection in comics. In: Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR), pp. 1240–1244 (2013)
32. Rigaud, C., Karatzas, D., Van de Weijer, J., Burie, J.C., Ogier, J.M.: Automatic text localisation in scanned comic books. In: Proceedings of the 8th International Conference on Computer Vision Theory and Applications (VISAPP) (2013)
33. Rigaud, C., Thanh, N.L., Burie, J.., Ogier, J.., Iwata, M., Imazu, E., Kise, K.: Speech balloon and speaker association for comics and manga understanding. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 351–355 (2015)
34. Rigaud, C., Tsopze, N., Burie, J.C., Ogier, J.M.: Robust frame and text extraction from comic books. In: Graphics Recognition. New Trends and Challenges, vol. 7423, pp. 129–138. Springer Berlin Heidelberg (2013)
35. Singh, S.P., Markovitch, S. (eds.): Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA (2017)
36. Stommel, M., Merhej, L.I., Müller, M.G.: Segmentation-free detection of comic panels. In: Computer Vision and Graphics, pp. 633–640. Springer (2012)
37. Sun, W., Burie, J.C., Ogier, J.M., Kise, K.: Specific comic character detection using local feature matching. In: 12th Int. Conf. on Document Analysis and Recognition, pp. 275–279. Washington, DC, USA (2013)
38. Tanaka, T., Shoji, K., Toyama, F., Miyamichi, J.: Layout analysis of tree-structured scene frames in comic images. In: IJCAI'07, pp. 2885–2890 (2007)
39. Yamada, M., Budiarto, R., Endo, M., Miyazaki, S.: Comic image decomposition for reading comics on cellular phones. IEICE Transactions **87-D**(6), 1370–1376 (2004)

40. Yim, J., Jung, H., Yoo, B., Choi, C., Park, D., Kim, J.: Rotating your face using multi-task deep neural network. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 676–684 (2015)
41. Zhang, Y., Yang, Q.: A survey on multi-task learning. CoRR **abs/1707.08114** (2017). URL `http://arxiv.org/abs/1707.08114`
42. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Facial landmark detection by deep multi-task learning. In: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (eds.) Computer Vision – ECCV 2014, pp. 94–108. Cham (2014)