Extraction of Semantic Content and Styles in Comic Books

Damitha Lenadora#1, Rakhitha Ranathunge#2, Chamath Samarawickrama#3, Yumantha De Silva#4, Indika Perera#5, Anuradha Welivita#6

***Abstract*— Digitisation of comic books would play a crucial role in identifying new areas in which digital comics can be used. Currently, existing systems in this domain lack the capacity to achieve complete digitisation. Digitisation requires a thorough analysis of the semantic content within comic books. This can be further sub-categorised as detection and identification of comic book characters, extraction and analysis of panels as well as texts, derivation of associations between characters and speech balloons, and analysis of different styles of reading. This paper provides an overview of using several object-detection models to detect semantic content in comics. This analysis showed that, under the constraint of limited computational capacity, YOLOv3 was the best-suited model out of the models evaluated. A study of text extraction and recognition using Optical Character Recognition, a method for determining associable speech balloons, as well as a distance-based approach for associations between characters and speech balloons are also presented here. This association method provides an increased accuracy compared to the Euclidean distance-based approach. Finally, a study on comic style is provided along with a learning model with an accuracy of 0.89 to analyse the reading order of comics.**

***Keywords* — comics, digitisation, content detection, text recognition, speech balloon to character association, comic styles**

I. Introduction

A barrier that comic books face to expand their horizons regarding usages as well as to acquire more readers is that they are traditionally a paper-based medium of entertainment. A remedy for this would be the digitisation of comic books. Although this is possible by simply taking pictures of the comic book in question, one can argue that this form of digitisation is not complete as it would not contain any information of the content within the comic book. It would also retain a significant amount of noise. Digitising to such a complete extent would open the doors for a wide variety of usages for comic books. To perform indexing and plot analysis with ease as well as to view the content within comic books in an immersive environment are examples of such potential use-cases.

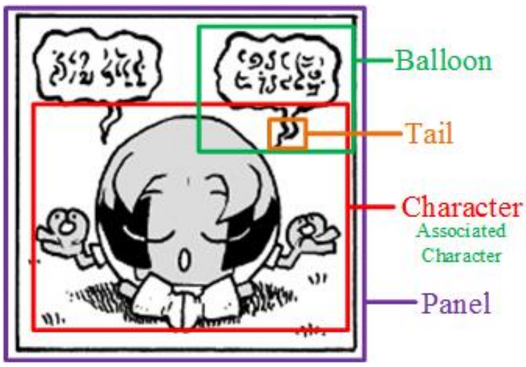


Fig. 1 Semantic content in comics.

However, let alone to completely digitise comic books, systems which are capable of extracting the semantic content within comic books without any extensive human intervention are limited. A reason for this may be due to the high diversity of the origins of comic books, authors as well as artists, which leads to an extensive number of styles of artwork as well as methods of storytelling via images being created. Nonetheless, there are certain semantic elements within comic books that are considered to be common across most comic books. These are panels, characters and balloons, as shown by Fig. 1. Panels are sections of a comic book page that depict a certain scene in the story that is being told, characters in comic books are those who play a role in the story, while balloons are used to display various texts in the comic book. In addition, if a balloon has an associated character that utters or thinks the content within the balloon, it can be referred to as a speech balloon. In such instances, it may also contain a segment called a tail that point towards the said character. This can also be seen in Fig. 1. All of these combine to form the narrative intended by the authors of the comic and present it to the readers. Thus, with the final goal of achieving the monstrous task of digitising comic books, the main focus of the research mentioned in this article is to analyse, locate and extract the aforementioned semantic content.

Though there exist many conventional semantic content extraction methods applied in the domain of comic books, the usage of learning models, especially Deep Learning which requires a lot of data, is only starting to raise its head within this domain. However, with multiple comic book datasets now available, the time is nigh to apply such techniques to aid in the task of extracting semantic content from comic books. Thus, the prime focus of this research is the application of learning techniques such as object detection to extract the semantic content in comic books and the analysis regarding the need for its usage. In addition, the extension of existing conventional techniques is also looked into as there can be certain instances where such techniques give acceptable results.

The remaining sections of the paper are as follows: Section II contains the review regarding the existing literature relevant

Manuscript received on 06 Nov. 2019. Recommended by Dr. D.D. Karunaratna on 07 March 2020.

This paper is an extended version of the paper “Comic Digitization through the Extraction of Semantic Content and Style Analysis” presented at the ICTer 2019.

Damitha Lenadora, Rakhitha Ranathunge, Chamath Samarawickrama, Yumantha De Silva and Indika Perera are from the Department of Computer Science & Engineering, University of Moratuwa Sri Lanka. (damitha.15@cse.mrt.ac.lk, rakhitha.15@cse.mrt.ac.lk, chamath.15@cse.mrt.ac.lk, yumantha.15@cse.mrt.ac.lk, indika@cse.mrt.ac.lk).

Anuradha Welivita is from the Computer & Communication Sciences, École polytechnique fédérale de Lausanne, Lausanne, Switzerland (kalpani.welivita@epfl.ch).

to this research. The datasets that have been used for this work are presented in Section III. Section IV presents the discussion regarding the panel and character extraction as well as character recognition. The content of Section V is that of the analysis of the text extraction processes used as well as text recognition using Optical Character Recognition (OCR). Section VI is the analysis of the character and speech balloon association. Afterwards, in Section VII is the analysis of the different styles found in comic books. Section VIII, which is located at the end, gives the conclusion to this research mentioned in this article.

II. Literature Review

The accurate identification of semantic content in comic books is a must for a reader to properly understand the story shown through comic book images, let alone even to consider their digitisation. Numerous techniques have been developed to extract such semantic content from comics. However, these methods, which often have vast differences among one another, are generally only able to extract a single type of semantic content.

When traditional panel extraction techniques are considered, the usage of density gradient [1], as well as line segment combination [2] can be observed. Although the results of the tests conducted for these methods prove that they work well, for comics with border-free panels and those with no distinct separation between the background and the foreground, the performance of these methods are not very good since they have not been taken into consideration when devising these methods. In addition, as a solution for comic books with panels which are not separated by white backgrounds, techniques that utilize region of interest detection [3] as well as recursive binary splitting [4] have been developed.

On the other hand, techniques such as binarization [5] and morphological operations [6] are prime examples for the extraction of speech balloons. The use of the two approaches stated above can be seen in the researches [7] and [8]. Another common technique that can be seen in many of the related works is Connected Component Labelling [9]. It is an algorithm which uses graph theory and is capable of detecting regions within binarized images. This technique can be seen being used in both [10] and [11]. The active contour model [12] is utilized successfully for the extraction of speech balloons in [13].

For face detection in Manga, Viola-Jones detection framework [14] is used in the research done in [15]. However, [16] explains that the traditional face detection and face recognition techniques do not perform well when applied to comic book characters due to the fact that they possess considerable differences when compared to real humans. As a solution to this, a method which utilizes skin colour and regions has been suggested by the authors. The work in [17] has extended the aforementioned method for the detection and identification of comic book characters. Furthermore, [18] suggests the usage of a graph-theory based method for identifying main characters in comics. This is done by representing each panel as an attributed adjacency graph and locating the most common colour patterns. The authors in [19] have carried out similar research by using a SIFT descriptor, which shows good potential for the task of character identification.

Content extraction of comic books through the utilisation of deep learning has also seen a sharp rise throughout the past few years. The ever-popular Convolutional Neural Networks (CNN) and its derivations are often used in these said usages as they show great potential in dealing with images. The researches carried out in [20], [21] and [22] through the usage of the object detection models of YOLOv2 [23], a customised Faster R-CNN [24] model and Mask R-CNN [25] respectively, stand as a testimony for this fact.

Text identification in comics varies depending on whether the text is typewritten or handwritten. Handwritten text is more difficult to identify due to the different styles of writing used. Tesseract [26] OCR is one of the prominent tools used in identifying texts, which is capable of recognising the text of a similar style once trained. [27] uses this with a Long Short-Term Memory (LSTM) [28] algorithm called OCRopus to identify text automatically in a segmentation-free manner. A different method is proposed in [29] using LSTM. Rather than using pre-trained OCR, this approach focuses on developing an OCR system with token dictionaries using k-means clustering. The advantage of this is a low error rate in text recognition and the potential to outperform existing OCR tools.

The final step is the character to speech balloon association in comic books. The method of association proposed in [30] via the usage of geometric graph analysis is the earliest research found to be done on this particular topic. This method requires the identification of panels, balloons as well as characters in comic books, and predicts a relation per speech balloon solely by relying on the minimum Euclidian distance to a character. Considering association recognition as a binary classification problem, the proposed model in [22] also trains a portion of the model to predict these associations in comics. A notable observation in both methods mentioned above is that they do not utilise the tail of a balloon which comic book authors use specifically to associate a speech balloon with a character. However, the authors intend to utilise balloon tails as a plausible future improvement for their proposed methods.

III. Datasets

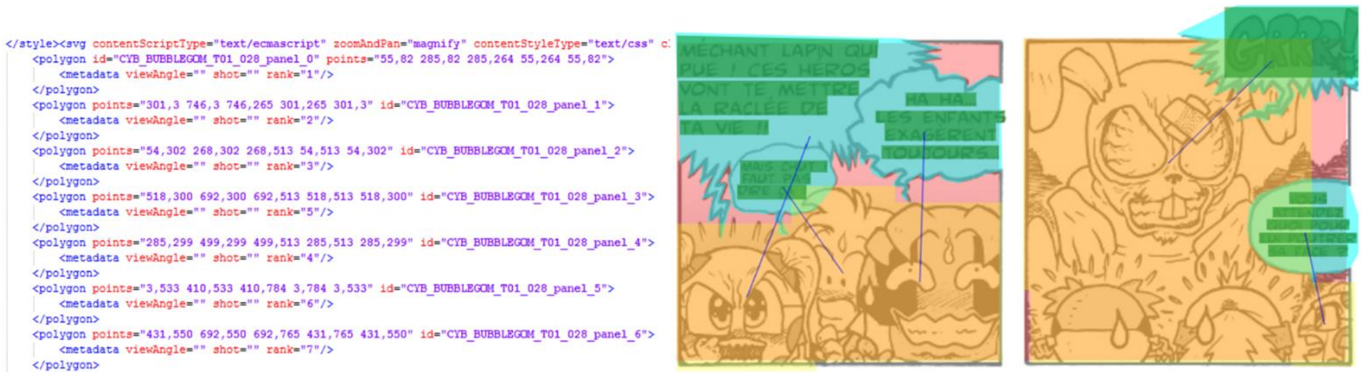
A. *eBDtheque*

The training of object detection models, analysis of text recognition and speech balloon-character association was done using the eBDtheque [31] dataset. The eBDtheque dataset is a selection of comic pages from America, Japan (manga) and Europe. Each comic page consists of three categories of visible objects; characters, balloons and panels. Apart from that, the dataset also comprises of speech balloon to character associations and general metadata related to the comic pages such as the ISBN, language, author and editor names, which are included in the ground truth files. To elaborate further on the statistics of the annotated elements;

- Pages – 100
- Frames (panels) – 850
- Balloons – 1,081
- Characters – 1,620
- Text lines – 4,693

The ground truth files are in Scalable Vector Graphics (SVG) format and follow the underlying XML and encoding information, as shown in Fig. 2. Semantic and visual annotations on a given page are gathered in such files making the files in the database simple and easily shareable.

The original SVG format of the ground truths is unique to the eBDtheque dataset, and none of the object detection models is made to support this format as an input. We had two



```
</style><svg contentScriptType="text/ecmascript" zoomAndPan="magnify" contentStyleType="text/css" c:
    <polygon id="CYB_BUBBLEGOM_T01_028_panel_0" points="55,82 285,82 285,264 55,264 55,82">
        <metadata viewAngle="" shot="" rank="1"/>
    </polygon>
    <polygon points="301,3 746,3 746,265 301,265 301,3" id="CYB_BUBBLEGOM_T01_028_panel_1">
        <metadata viewAngle="" shot="" rank="2"/>
    </polygon>
    <polygon points="54,302 268,302 268,513 54,513 54,302" id="CYB_BUBBLEGOM_T01_028_panel_2">
        <metadata viewAngle="" shot="" rank="3"/>
    </polygon>
    <polygon points="518,300 692,300 692,513 518,513 518,300" id="CYB_BUBBLEGOM_T01_028_panel_3">
        <metadata viewAngle="" shot="" rank="5"/>
    </polygon>
    <polygon points="285,299 499,299 499,513 285,513 285,299" id="CYB_BUBBLEGOM_T01_028_panel_4">
        <metadata viewAngle="" shot="" rank="4"/>
    </polygon>
    <polygon points="3,533 410,533 410,784 3,784 3,533" id="CYB_BUBBLEGOM_T01_028_panel_5">
        <metadata viewAngle="" shot="" rank="6"/>
    </polygon>
    <polygon points="431,550 692,550 692,765 431,765 431,550" id="CYB_BUBBLEGOM_T01_028_panel_6">
        <metadata viewAngle="" shot="" rank="7"/>
    </polygon>
```

MECHANT LAPIN QUI PUE ! CES HÉROS VONT TE METTRE LA RACLÉE DE TA VIE !!
MAIS CHUT FAUT PAS DIRE ÇA
HA HA.. LES ENFANTS EXAGÈRENT TOUJOURS.
GRRR!

Fig. 2 SVG files content and its representation.

alternatives to overcome this problem; change the model's input-output pipeline to support these formats or convert SVG files to a supported format like PascalVOC, COCO or TFRecord. We chose the latter alternative and were able to convert the SVG files to the COCO format as well as TFRecord format using several Python scripts. Minor changes were done later to support specific models. Thereby, we managed to train and validate different models using the data in the eBDtheque dataset.

The dataset was divided into two parts: i.e., training dataset and validation dataset. We decided that the most suitable way to divide the dataset is by allocating 80% of the data to the training set and the rest to the validation set, i.e., 80 images to the training set and 20 images to the validation set. Due to the presence of images of different comic types, we decided to put at least one of each kind to the validation set to assert that mean Average Precisions (mAPs) obtained covered all types of comic pages in the dataset.

B. *Manga109*

The Manga109 [32] is a dataset that consists of 109 manga titles with a total of 10,130 pages. It is made publicly available for academic research purposes with proper copyright notation. This dataset comprises of 109 manga volume drawn by professional manga artists in Japan.

The annotations of this dataset are given in XML format with bounding boxes over character faces, character bodies, text parts and frames (panels). One of the significant differences between this dataset and the eBDtheque dataset is that in Manga109, the bounding boxes are drawn over the specific text portion rather than the balloons, which consist the text parts. We can still use the conversion methods stated above, with slight modifications, to convert these annotations to a standard format which can be used to train and test an existing model.

One of the special features in this dataset is that the characters are annotated with a character ID, which is essential for the task of character identification. The statistics regarding the numbers of each element of the dataset are as follows;

- Manga volumes – 109
- Pages – 10,130
- Character faces – 118,715
- Character bodies – 157,152
- Text portions – 147,918
- Frames (panels) – 103,900

We attempted to do character identification from this dataset using FaceNet [33], but due to the usage of Triplet Loss by FaceNet which calculates a distance between faces, this attempt did not yield the results as expected. The reason for this was determined to be the similarities between faces especially due to the black and white nature of the images in the dataset as well as the vast differences in the faces of the same character which was a result of the nature of manga drawings. This resulted in faces of different characters being similar to each other than those of the same character. It was verified through observation that these were not mistakes of the neural network or the code since many characters could not be identified even with the naked eye.

C. *COMICS Dataset*

COMICS dataset [34] is a massive dataset with a size of over 120GB comprised of over 1.2 million panels paired with automatic textbox transcriptions. The main objective of this dataset is to demonstrate that neither text nor image alone can tell a comic book story. Therefore, a computer must understand both modalities to keep up with the plot.

This dataset consists of comic book panels and text boxes. Panels are marked with rectangular bounding boxes in 500 randomly selected pages. Each bounding box encloses both the panel artwork and the textboxes within the panel. 1500 selected panels are annotated for textboxes. To get the text in textboxes, they have applied OCR to the selected text boxes.

The statistics of the dataset are as follows;

- Books – 3,948
- Pages – 198,657
- Panels – 1,229,664
- Textboxes – 2,498,657

Other than these, this dataset has elements called text cloze, visual cloze and character coherence, which are aimed towards reaching the main objective of creating the dataset.

IV. Content Extraction via Object Detection

A. *Faster R-CNN*

The R-CNN family is a set of object detection frameworks that performs selection as well as the classification of objects in images through proposing candidate object regions. Predicting bounding boxes for the objects detected, the evolved Faster R-CNN framework performs significantly better than its predecessors.

A Faster R-CNN object detection network that uses Inception v2 as its backbone was trained on the eBDtheque dataset. For the implementation, the TensorFlow Object Detection API [35] was used with the configurations of the model being, a batch size of 1, an initial learning rate of 0.002, and momentum optimiser value of 0.9. After approximately 60,000 steps, signs of overfitting began to emerge, and further training was halted. The best results of the model were an mAP of 62% with a 50% intersection and a mean Average Recall (mAR) of 41% per 100 detections. The model's reliance on whitespace to detect the boundaries of a panel could be observed due to the incorrect detection of additional panels in black and white images.

B. Single Shot Multibox Detector

Single Shot Multibox Detector (SSD) [36], is a method for encapsulating object detection processes and computation into a single network. In the SSD architecture, there are two types of deep neural networks. First, a base network for high-quality image classification tasks and then, a Multibox detection network which detects objects through bounding boxes.

With the TensorFlow Object Detection API [35], SSD Mobilenet model was implemented. The label map and the trainer were configured to have three classes to accommodate the eBDtheque dataset. The input images were resized to 300x300 with the use of a fixed shape resizer since the SSD network ends with fully connected layers.

The trainer was set to have a batch size of 1, an initial learning rate of 0.004 and a momentum optimiser value of 0.9. After training the network for 200,000 steps, the obtained mAP values were shown to be extremely poor, standing at less than 0.4. The mAR values stood at around 0.3, which was still, far below acceptable. The main reasons for this underwhelming performance could be the lack of training data and the initial resizing of the input that happens in the model.

C. YOLO

YOLO stands for You Only Look Once [37]. It is an object detection framework, which spatially separates bounding boxes and associates with class probabilities. Through this, it is able to predict both the bounding boxes as well as the class probabilities from full images in one evaluation while using a single neural network. YOLOv3 [38], which is the 3rd version of the YOLO network, is an incremental improvement of its predecessor, YOLOv2(also known as YOLO9000). The neural network used for this version is a network with 53 convolutional layers, named Darknet-53. This model predicts an objectness score for each bounding box using logistic regression. Class prediction is made using multilabel classification, and independent logistic classifiers are used instead of softmax. Binary cross-entropy loss for class predictions is used during training. Bounding boxes are predicted at three different scales. The final convolutional layer predicts a 3D tensor which encodes bounding box, objectness and class prediction. Bounding box priors are determined using k-means clustering, which is similar to the previous version of YOLO. YOLOv3, the latest model, which is explained above, was used in our experiments.

First, we adjusted the Darknet-53 configuration files to be in line with the eBDtheque database. There are three main classes in eBDtheque dataset; character, panel and balloon. We also changed the number of filters in various convolutional layers to suit the dataset. We used the aforementioned conversion methods to obtain the labels for the training set and the validation set. The network, Darknet-53 was trained for 250 epochs under a batch size of 1, and the following results were obtained. The confidence threshold was set to 0.3, NMS threshold was set to 0.45 and IOU threshold was set to 0.5. The two graphs which are shown in Fig. 3 and Fig. 4 depict how the total mAP and the mAPs of each class change with respect to the number of epochs, respectively.

It can be clearly observed that there is a shortage of training data from the results shown in Fig. 3 and Fig. 4. The maximum total mAP that could be obtained was slightly below 0.6, even though the network was trained for a considerable number of epochs. When looking at the graph as a whole, the total mAP can be observed to arrive at a slightly stable state after about 80 epochs. Thereby, we can conclude this to be a rough estimation for the optimum number of epochs the model should be trained for.

The mAP per class graph, which is shown by Fig. 4 clearly indicates the low mAP for the "character" class, which maxes out at around 0.35. The main reason for this is the large variations among different characters in different comics, as well as between the same characters when drawn by different artists, who have their own drawing styles. Due to this and the low amount of data, the model is currently unable to generalize well for detecting different comic characters.

On the other hand, the speech balloons and panels give considerably good mAPs, which are above 0.7. The reason for this is the consistent styles of these which do not differ within the same comic and even among different comics. Therefore, the model is able to generalize well for these classes, even with a limited amount of data.

We used YOLO to test the performance in the Manga109 dataset as well. The training and testing processes were carried out in a similar fashion to those done using eBDtheque dataset. The model was able to reach a total mAP value of 0.682, a precision value of 0.755 and a recall value of 0.738. However, Manga109 dataset purely consists of black and white pictures. Due to this, if the model is trained solely using this dataset, it would fail to generalize to other types of comics, of which a majority are coloured.

The first aspect we considered when tuning the hyperparameters was finding out which optimizer to use to train the model. The tests were carried out using the eBDtheque dataset. The dataset was trained for 100 epochs, with a learning rate of 0.001, an NMS threshold of 0.45, an IOU threshold of 0.5 and a confidence threshold of 0.3. The batch size used for the purpose was 16. The tests were carried out with the hyperparameters stated above while varying the optimizers. The following tables show the total results obtained and the class-wise results obtained for each of the tested optimizers, respectively.

TABLE I

RESULTS FROM THE TOTAL DATASET FOR DIFFERENT OPTIMIZERS

	Total mAP	**Mean Precision**	**Mean Recall**
Adadelta	0.0554	0.0148	0.4024
Adagrad	0.2636	0.3694	0.4179
Adam	**0.5587**	**0.5900**	**0.6542**
RMSprop	0.2942	0.4187	0.3887
SGD	0.5072	0.5253	0.6254

TABLE II

RESULTS PER EACH CLASS IN THE DATASET FOR DIFFERENT OPTIMIZERS

	Character mAP	**Panel mAP**	**Balloon mAP**
Adadelta	0.0275	0.0709	0.0821
Adagrad	0.1797	0.2584	0.3743
Adam	0.4106	**0.6458**	**0.6862**
RMSprop	0.1948	0.3734	0.3331
SGD	**0.4230**	0.5178	0.6166

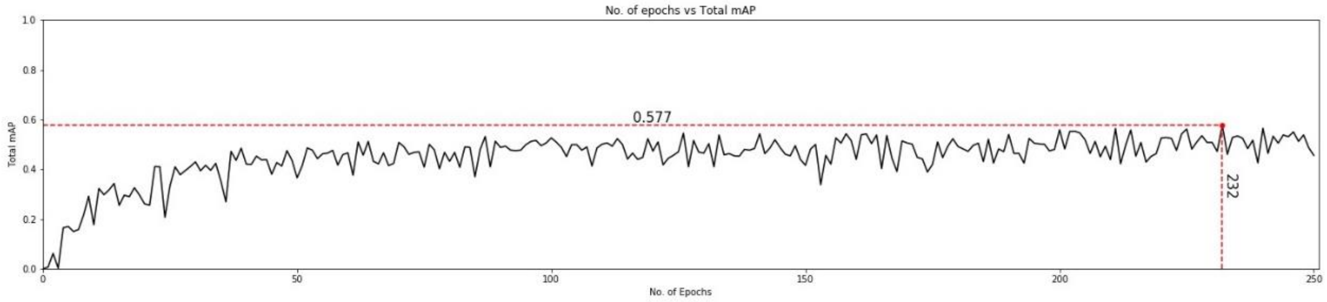


Fig. 3 Total mAP of YOLO.

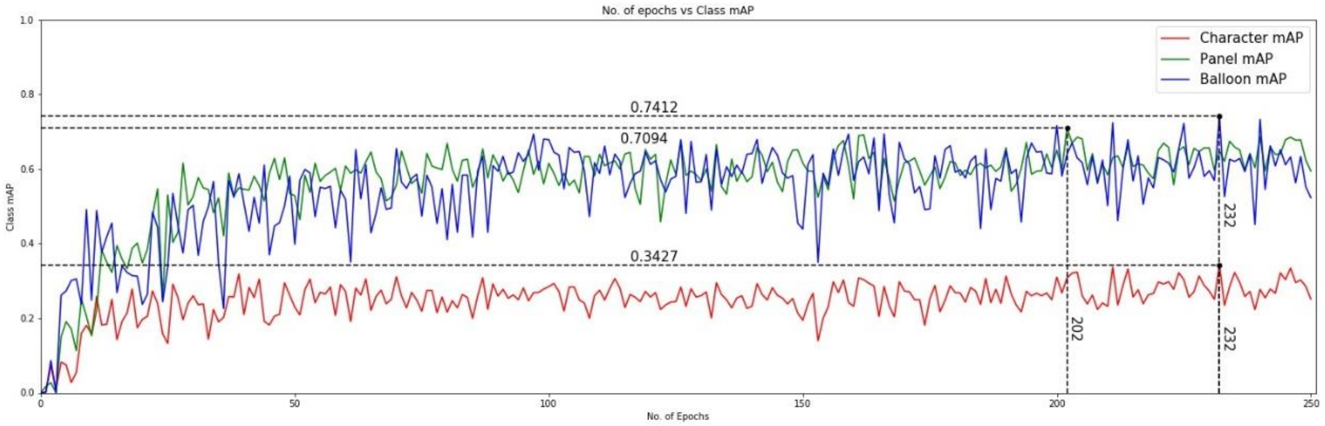


Fig. 4 mAP per class; mAP for the Character class – Red, mAP for the Panel class – Green, mAP for the Balloon class – Blue.

By observing the above results, Adam was concluded to be the best optimizer fit for our purpose. SGD was a close second and also managed to outperform Adam for the "character" class. However, through further testing, we discovered that Adam was able to give better results than SGD, even for the "character" class when trained further (the test was done for 250 epochs, using the same hyperparameters used for the previous test).

Another test was carried out to determine the optimum learning rate for the model. This was also done using the eBDtheque dataset. The dataset was trained for 100 epochs, with Adam set as the optimizer, an NMS threshold of 0.45, an IOU threshold of 0.5 and a confidence threshold of 0.3. A batch size of 16 was used for this task. The tests were carried out with the above-stated hyperparameters while varying the learning rates. The learning rates were increased by roughly threefold from the previous test case. The tested leaning rates started at 0.001 and ended at 0.1, which was deemed to be a worthy range for carrying out the tests. Table III shows the accuracies corresponding to the tested learning rates.

TABLE III

ACCURACIES CORRESPONDING TO DIFFERENT LEARNING RATES

Learning Rate	**Accuracy**
0.001	**0.4309**
0.003	0.3907
0.01	0.3454
0.03	0.2959
0.1	0.2000

It was concluded that 0.001 is the best learning rate to train the model in order to get the most accurate results, considering the above.

Out of the tested models which we have stated above, YOLO was able to get considerably better results compared to the SSD model and was able to do so whilst using much better memory usage compared to the R-CNN family, in the task of object detection. Due to the better results as well as the limited computation capabilities available, YOLO was decided to be the best model for the task of object detection. With more training data and slight hyperparameter adjustments, this model has the best potential to generalize well for the task at hand.

V. TEXT EXTRACTION AND ANALYSIS

Text extraction and analysis from panels involve two main steps: (1) extracting the speech balloons from the panel, and (2) processing the text within the speech balloons using OCR software. Here, we will discuss the methodology we followed in order to extract the text from the comic book panels and process them for further analysis.

Even though we use models such as YOLO, Faster R-CNN, and SSD for content detection, we decided to take a different approach to speech balloon detection. The model we have used here is Mask-RCNN, which is an improved version of its predecessor; Faster-RCNN. We used a TensorFlow and Keras implementation [39] of Mask RCNN instead of using the original version to reduce the usage of resources.

The main reason for using Mask-RCNN is the generation of masks that mark the object boundary as opposed to just

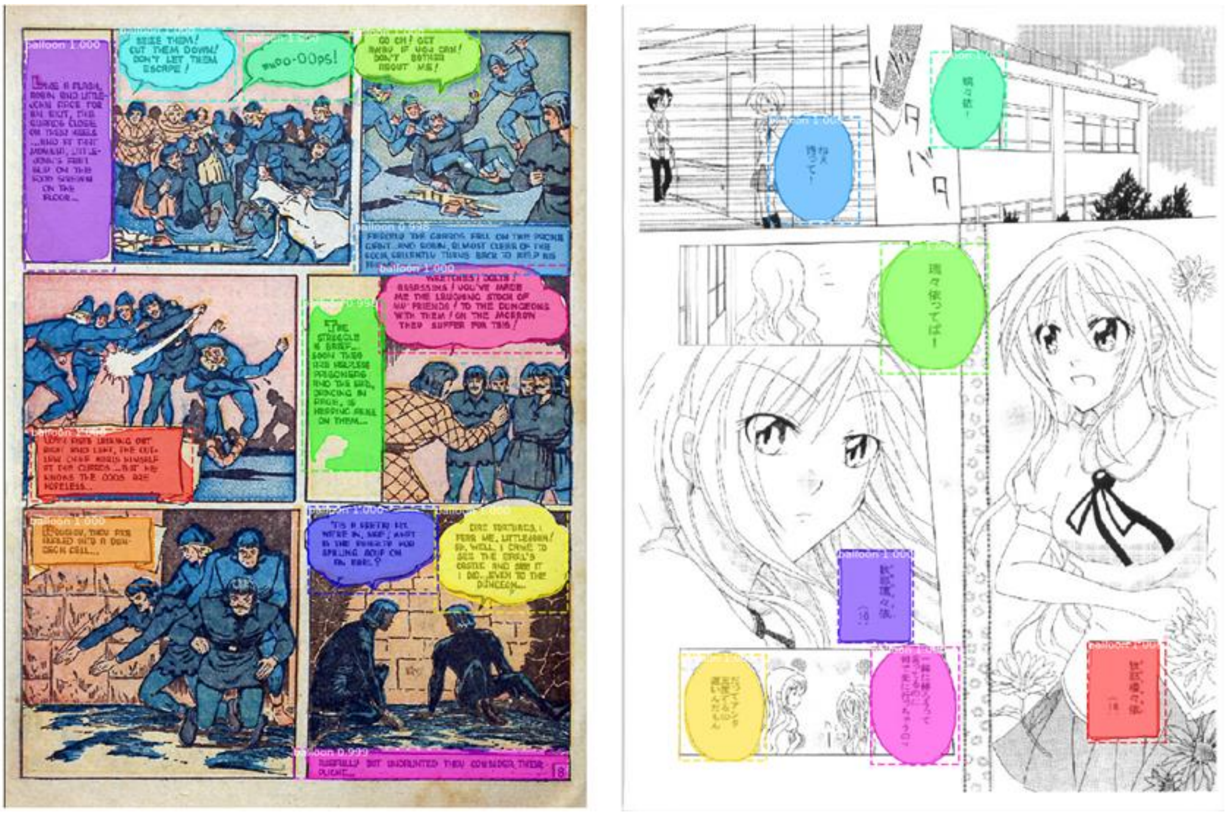

balloon 1.000
SEIZE THEM! CUT THEM DOWN! DON'T LET THEM ESCAPE!
Whoo-OOps!
GO ON! GET AWAY IF YOU CAN! DON'T BOTHER ABOUT ME!
balloon 0.998
balloon 1.000
WRETCHES! DOLTS! ASSASSINS! YOU'VE MADE ME THE LAUGHING STOCK OF MY FRIENDS! TO THE DUNGEONS WITH THEM! ON THE MORROW THEY SUFFER FOR THIS!
balloon 0.999
8

Fig. 5 Results of speech balloon detection of different comics by Mask-RCNN.

locating the object. Mask-RCNN facilitates instance segmentation by providing the exact coordinates of the objects, unlike other object detection models. i.e. the exact pixels of the detected speech balloon in the comic book page will be returned by the model. The generated mask will be vital in detecting the tail of the speech balloon. This will be used in making associations between the speech balloons and characters, as mentioned in Section VI.

The eBDtheque dataset was used to train the model in order to detect speech balloons. Since we were trying to distinguish only the speech balloons, adjustments were made to accommodate the model to a single class. The model we chose uses a resnet50 [40] backbone architecture. The default configurations of the model were for the COCO dataset, which has about 80 different classes. The model was trained for 100 epochs with 80 steps per epoch while having a batch size of 1. We decided to use a transfer learning approach to train the model by initializing the model with COCO pre-trained weights and fine-tuning the layer heads with eBDtheque data. The heads include the Region Proposal Network, classifier, and mask heads of the network. The model was trained with a learning rate of 0.001, a momentum of 0.9, and with Stochastic Gradient Descent as the optimiser.

However, if we increase the number of epochs, the validation loss would increase, causing the model to overfit. So, the model obtained in this scenario is more generic to all data. The mAP of the model over the validation set is around 0.65. The main reason for this would be the lack of diverse training data.

Fig. 5 shows some of the detections done using the model we trained. The coloured speech balloons show the masks generated when a detection is made by the model. The model returned Average Precision (AP) of 0.758 and 0.905, respectively for the above two figures for the IOU threshold range 0.5 - 0.95.

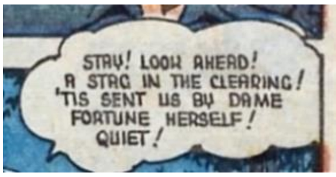
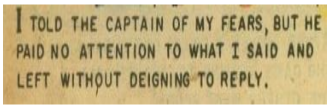
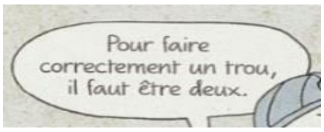
We decided to use the Tesseract OCR for the analysis of text found in extracted speech balloons. The latest version of Tesseract 4 options in the OCR engine. These are Legacy engine, LSTM engine, Legacy + LSTM engine, and default engine. We decided to use the neural network-based LSTM engine for our purpose. This also provides different page segmentation options which can be crucial under different conditions you come across. The results we obtained with both Tesseract Command-Line and pytesseract, which is a Tesseract package in Python, are similar.

Table IV contains results obtained through Tesseract for three different styles of text found in the comic books. The first two instances are in English, but one is Old English while the other is Modern English. The third instance is in French. It is clear from the results that the text recognition in the 1st instance is somewhat weak compared to the second instance. The style of the letters and characters is a huge factor here. Some characters are difficult to be recognized even though the naked eye. The Character Error Rate is quite high in this case due to that. In both cases, the Word Error Rate is lower than the Character Error Rate. In the third instance; which is in French, the recognised text is identical to what's in the extracted speech balloon. The error rate will increase if the style of the text varies a lot from the standard style. This is due to the fact that Tesseract is trained using characters which belong to a standard format.

One reason for the errors in recognition would be the noise in the extracted parts. There is a possibility to improve the results by pre-processing the extraction before processing them using Tesseract. Examples for such pre-processing techniques are noise removal, blur, and binary thresholding.

TABLE IV

TESSERACT OCR RESULTS

 STAY! LOOK AHEAD! A STAG IN THE CLEARING! 'TIS SENT US BY DAME FORTUNE HERSELF! QUIET!	STAY! LOON AHEAD! A STAG IN THE CLEABING / 'TIS SENT US BY DA FORTUNE JNERSELE | QUIET /
 I TOLD THE CAPTAIN OF MY FEARS, BUT HE PAID NO ATTENTION TO WHAT I SAID AND LEFT WITHOUT DEIGNING TO REPLY.	T ToLD THE CAPTAIN OF MY FEARS, BUT HE PAID NO ATTENTION TO WHAT T SAID AND_ LEFT WITHOUT DEIGNING-TO REPLY.
	Pour faire correctement un frou, il faut être deux.

VI. CHARACTER AND SPEECH BALLOON ASSOCIATION

The association between a speech balloon in a comic and the relevant characters that uttered the content in the said balloon play a significant role in presenting the story intended by the author to the readers of the comic book. For a human reader, such associations are easily understood by the hints given through the artwork within comic books. One such clue which is primarily used is the tail that can be seen in speech balloons. These tails are usually represented by an elongated section of a speech balloon that ends in a point facing towards the associated comic book character who is within the same panel as the balloon. However, when performing this association via a program, it is not as simple as for humans due to intuition also playing a pivotal role in accurately identifying a correct association.

A. Association Methods

Building upon the works of Rigaud et al. [30], a method of character to speech balloon association given the location of panels, characters, speech balloons and the tail of the speech balloons was looked into. Thus, for obtaining an association, each speech balloon was checked to locate its associated character rather than checking each character. Furthermore, each speech balloon and character were assigned to a panel within the comic. This association was done by selecting the panel containing the maximum area of the associating speech balloon or character. If either a balloon or character does not have any area in any of the panels in the page, it is then assigned to a separate outlier panel created to hold such objects.

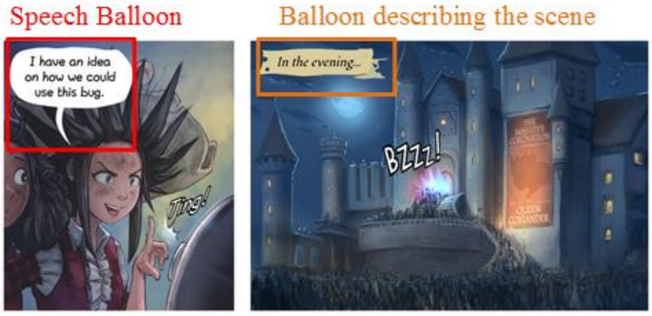

Speech Balloon
Balloon describing the scene
I have an idea on how we could use this bug.
Tjing!
In the evening...
BZZZ!

Fig. 6 Types of comic balloons

As shown in Fig. 6, not all balloons have an associated character as some balloons are used to describe scenes. Thus, the selection of balloons that are assumed to have an association with a character is necessary. For this, three different methods were evaluated. These methods are selecting all balloons, selecting balloons that possess tails and selecting balloons that possess tails or are dissimilar to a square shape. The intentions behind the final two selection methods are respectively, due to speech balloons with associations usually possessing tails and due to most balloons that describe scenes are of a shape similar to a square. The dissimilarity from a square shape is assumed if the area of the balloon itself is 90% or less of the bounding box that surrounds the balloon.

Afterwards, the following methods were tested to identify a way of associating speech balloons to characters with considerably high accuracy.

1) *Method A:* The character with the minimum Euclidean distance from the tail of the speech balloon to the centre of the character. This is a method that has shown the highest results in [30].

2) Method B: The character with the minimum perpendicular distance from its centre to the line drawn in the direction pointed by the tail of a balloon, which is also intersected by the line extended from the tail in the direction the said tail points towards.

3) Method C: The character with the minimum Euclidean distance from the tail of the speech balloon to the centre of the character, which is also intersected by the line extended from the tail in the direction the said tail points towards.

4) *Method D:* The combination of *Method B* and *Method C*. This is done by associating the character with the minimum sum of squares of the results of the aforementioned two methods.

Out of these four methods, *Method A* was chosen as a baseline to evaluate the performance of the other methods. The line drawn in the direction of the tail is obtained by utilizing the point where the tail is located in a balloon and the mean point of the two adjacent points to the tail. In the latter three methods, checking that if the line extended from the tail intersects any part of the bounding box of the associated character is done to avoid associating a speech balloon to a character in other directions than the one the tail is pointing towards. In the situation that a character is not intersected by the line extended from the tail, infinity would be returned as a result.

B. Experiments and analysis of results

For experimenting the speech balloon to character association methods, the eBDtheque v3 dataset along with its annotations was used as a test set. Though the acquisition of the previously mentioned prerequisites would be possible to a certain degree utilizing object detection, for the sake of experimentation, the annotated data of the panels, characters, and balloons of the eBDtheque dataset were used.

TABLE V

PERFORMANCE OF THE SELECTION OF A BALLOON HAVING A SPEAKER ASSOCIATION

Method of balloon selection with associations	Performance	
	Precision	Recall
All balloons	81.96%	100%
Balloons with tails	96.09%	97.07%
Balloons with tails or dissimilar to a square shape	93.64%	99.67%

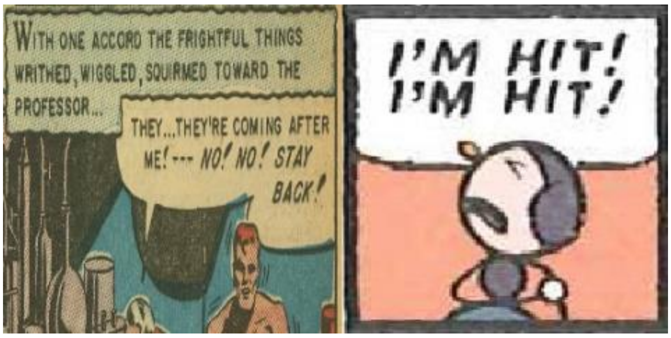

WITH ONE ACCORD THE FRIGHTFUL THINGS WRITHED, WIGGLED, SQUIRMED TOWARD THE PROFESSOR...
THEY...THEY'RE COMING AFTER ME!--- NO! NO! STAY BACK!
I'M HIT! I'M HIT!

Fig. 7 Balloons with differing styles to the norm.

When considering the performance of the selection of speech balloons having an association with a character as shown by Table V, if selecting all balloons, the 195 balloons out of 1081 in the dataset that have no association to a character were also selected. Though it is possible to recall all the balloons with associations through this method, the precision drops considerably due to the balloons with no association. With the method of selecting only the balloons with tails, the precision rises significantly yet is not 100% due to the presence of speech balloons where the characters who make the utterance in the said balloons are not depicted in the same panel. The recall drops to 97.07% as speech balloons with no tails are ignored in this method. The final method of selection, which is to select balloons that either has a tail or are dissimilar to a square shape, gives a more balanced result compared to the prior two methods. However, an issue in this method is that it would still take into account non-speech balloons that are not square, as well as ignore speech balloons with no tails and shaped similar to squares as shown by Fig. 7.

The initial evaluation of the association methods alone is done by taking the percentage of correct associations out of all the speech balloons with tails as well as associations to characters that are mentioned in the annotations of the dataset. This is done since all the association methods experimented upon requiring the identification of the tail of the speech balloon. The content of Table VI shows the results of the evaluation.

TABLE VI

PERFORMANCE OF ASSOCIATION METHODS

Character to balloon association	Performance
Method A	93.95%
Method B	86.62%
Method C	94.53%
Method D	95%

Using *Method A* as a baseline, it can be seen that *Method B*'s performance is significantly poorer. The reason for this drop was because this method tends to ignore characters who are closer to the balloon and associated characters who are more matching in regard to the direction the tail of the balloon points to. This is especially true for panels that are crowded with characters. Thus, through the evaluation of this method, it is possible to speculate that the distance from a speech balloon also plays great importance, sometimes even more so than the balloon tail for associating speech balloons and characters.

When evaluating *Method C*, a considerable increase in accuracy can be observed. This is due to this method rectifying the weaknesses of both *Method A* and *Method B*. The intuition behind this method is to select the closest character who is in the direction the balloon points to. A weakness that could be observed in this method was when the tail of a speech balloon is located within the bounding box of a character not associated with the balloon, as shown in Fig. 8. The balloon, which is highlighted in blue, is truly associated with the character on the left although its tail is located within the bounding box of the character that is highlighted in yellow.

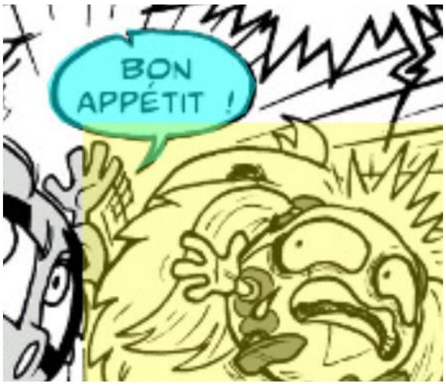


Fig. 8 Speech balloon located in the bounding box of another character.

It is possible to circumvent this weakness by making the conditions to check if a given character is in the direction that the tail of a speech balloon points to stricter via using the centroid of the character rather than simply intersecting the bounding box. However, this results in a lower overall accuracy since the margin to determine if a character is in the direction that the tail points to reduces. This is especially true for balloons where the associated character is directly below or above it with the tails being represented vertically.

The intuition behind *Method D* is to combine both the results of *Method B* and *Method C,* as a single distance measure. As shown in Table VI, this gives the highest accuracy out of all the methods considered. Upon analysing the results of this method, the weakness of *Method C* is also negated due to considering the distance from the line extended in the direction the tail of the balloon points to. This again brings up the weakness of *Method B*, wherein certain situations, the direction the tail point towards is considered at a higher degree and incorrectly predicts the associations that *Method C* predicts correctly.

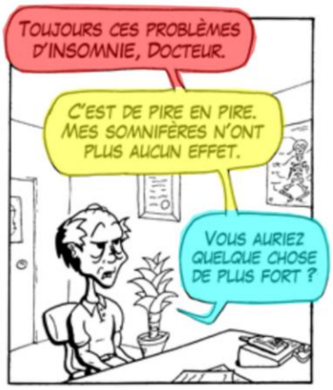

TOUJOURS CES PROBLÈMES D'INSOMNIE, DOCTEUR.
C'EST DE PIRE EN PIRE. MES SOMNIFÈRES N'ONT PLUS AUCUN EFFET.
VOUS AURIEZ QUELQUE CHOSE DE PLUS FORT ?

Fig. 9 Balloons, which are pointing to other balloons.

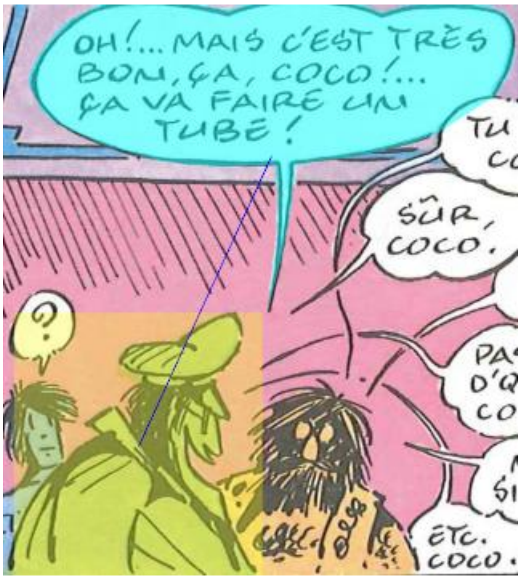

OH!... MAIS C'EST TRÈS BON, ÇA, COCO!... ÇA VA FAIRE UN TUBE!
TU CO
SÛR, COCO.
?
PA O'Q CO
SI
ETC. COCO.

Fig. 10 Highlighted balloon pointing to its respective character vaguely.

Analysis of the results of the four associations reveals the presence of balloons that point toward other balloons in the eBDtheque v3 dataset, as shown in Fig. 9. These are situations where the utterances a character makes are spread across several speech balloons. When combined as a single balloon, it possible to observe an increase in accuracy in the *Methods A*, *B*, *C*, and *D* respectively to 94.26%, 87.8%, 95.41% and 95.7%. Another situation where the latter three methods gave incorrect results were when the tails of the balloons did not point towards characters directly, but rather vaguely, as shown in Fig. 10. Here, the highlighted balloon points towards the edge of its associated character and not directly.

When analysing the results of the observations of the association presence selector via balloon tail or dissimilarity to a square shape, as well as the character association methods as a whole, a way of measuring the performance would be the collective correct decisions, which are the correct character associations and the correctly detected non-associated balloons, divided by the total number of balloons. A point to note would be that for the balloons that did not have a tail but are dissimilar to a square shape was associated to characters utilizing the minimum Euclidean distance from the centroid of the balloon to the centroid of the associable characters. Furthermore, in addition to the balloons which were deemed not to possess an association by the tail selection, certain balloons were deemed not to have an association by the association methods themselves in situations where an identifiable associated character could not be located. The collective performances using the unmodified eBDtheque v3 dataset, of the selection by tail or dissimilarity to a square shape, and *Method A*, *Method B*, *Method C* and *Method D*, were 93.52%, 89.18%, 95.47% and 95.84% respectively. These results, which are not at the level of perfection, are acceptable to a certain degree with the method based on *Method D* being the best.

However, a critical point of concern would be for speech balloons that possess multiple associations and other different styles to that of the norm. The methods contained in the evaluation done prior does not consider such situations. One cannot ignore these situations as comics are comprised of the artwork of the authors. This art can be abstract and can be used to give different meanings from author to author as well as require intuition rather than following a simple set of rules or patterns to grasp its true meaning. Thus, the need for the incorporation of learning methods, as well as different learning models per art style seem to be hinted through these observations. In addition, the requirement of the detection of panels, characters, speech balloons and the tails of the speech balloons would also cause a considerable limitation if extracted from a computer-based system instead of directly from human annotations. This is due to each said detection practically possessing a considerable chance of being incorrect as well as imprecise. This would collect more and more and leave an overall low accuracy at the end of the association process. Thus, a high accuracy must be achieved in this detection or circumventing the requirement would be a necessity.

VII. Style Analysis

In the process of analysing semantic content of comics, determining the reading style of panels is a crucial step. Comics from different origins have shown to have differences in the art style, formatting, panelling as well as the reading style. To determine the reading style and the order, a human reader should take cues from the art style, formatting, panelling, and the origin of a comic. Furthermore, comics which follow a reading style can have diverse art styles depending on multiple attributes like the region, artist, published year, etc. Hence, a significant level of expertise is required to determine the reading order of a comic.

Reading style of a comic can mainly take one of the following three flavours. Left-to-right, right-to-left or top-to-bottom. Generally, the left-to-right reading style is practised in western comics and in Chinese Manhua. Right-to-left reading style is heavily popularized by Manga, which is the Japanese version of comics. The top-to-bottom reading style is adopted by Korean Manhwa and webtoons [41]. Fig. 11 shows the aforementioned reading styles in comics. Here, the numbers in the panels indicate the order in which they should be read.

Distinguishing between the different art styles, formatting and panelling in comics that belong to Manga, Manhwa, Manhua and Western origins help to determine the reading style of the comics, whether it'd be left-to-right, right-to-left or top-to-bottom.

A. The difference in Art Style

The art style in a comic refers to the specific ways the colouring, the character drawing and the background is done.

In terms of colour, Manga is generally black and white although the cover pages and the first pages of chapters are usually coloured. Manhwa, Manhua and Western comics are generally coloured since they are digitally drawn and are available in digital formats.

Style of character drawing contributes immensely to distinguish the art styles between Manga, Manhwa, Manhua and Western comics. Here, Western comics differ from the rest considerably with its wide variety where differences between Manga, Manhua and Manhwa are rather minute. Chinese Manhua mostly uses slender characters with largely drawn muscles and narrow waists, where Japanese Manga uses more slender characters with no overgrown muscle types [42]. Korean Manhwa is a combination of Manhua and Manga with the focus lying on drawing the characters realistically [41].

Manga, Manhwa and Manhua also use more of a cinematic style than Western comics, portraying characters in dramatic angles more in sync with a film than a comic book [43].

The background style is another factor which aids to distinguish between different comic styles. Manhua and Western comics differentiate from the rest with their more detailed backgrounds. Due to its limitation of not using colour, Manga usually has much simpler backgrounds compared to others.

B. The difference in Formatting and Panelling

In Western comics, the establishing shot is generally centred on occupying the first scene in the comic. Manga, however, places its establishing shot at the bottom of the page. Manga, Manhua and Manhwa structures their scenes frame-by-frame, representing a snapshot of the action and in sync with the dialogue. Western comics are graphic novels, and as such, the stories and visuals don't necessarily sync with the dialogue and visual action [43].

C. Experiments and analysis of results

The defining characteristics of Manga, Manhwa, Manhua and Western comic genres were used to identify to which genre the comic belongs to. The reading styles were then inferred as left-to-right for Western comics and Manhwa, right-to-left for Manga and top-to-bottom for Manhua.

A custom dataset with comic images belonging to Manga, Manhwa, Manhua and Western genres was used to experiment with the above approach. Three approaches that try to determine the reading style of the comics were tested on this dataset.

The first approach was to determine the reading style of comic images based on its colour. With the presumption being most of the black and white comics belong to the genre Manga, this approach attributed a right-to-left reading style, which is the general reading style for Manga comics, to every black and white image. This approach was tested on the dataset, and it yielded an accuracy of 0.89. However, the result can be biased due to the limited dataset and the imbalanced nature of it.

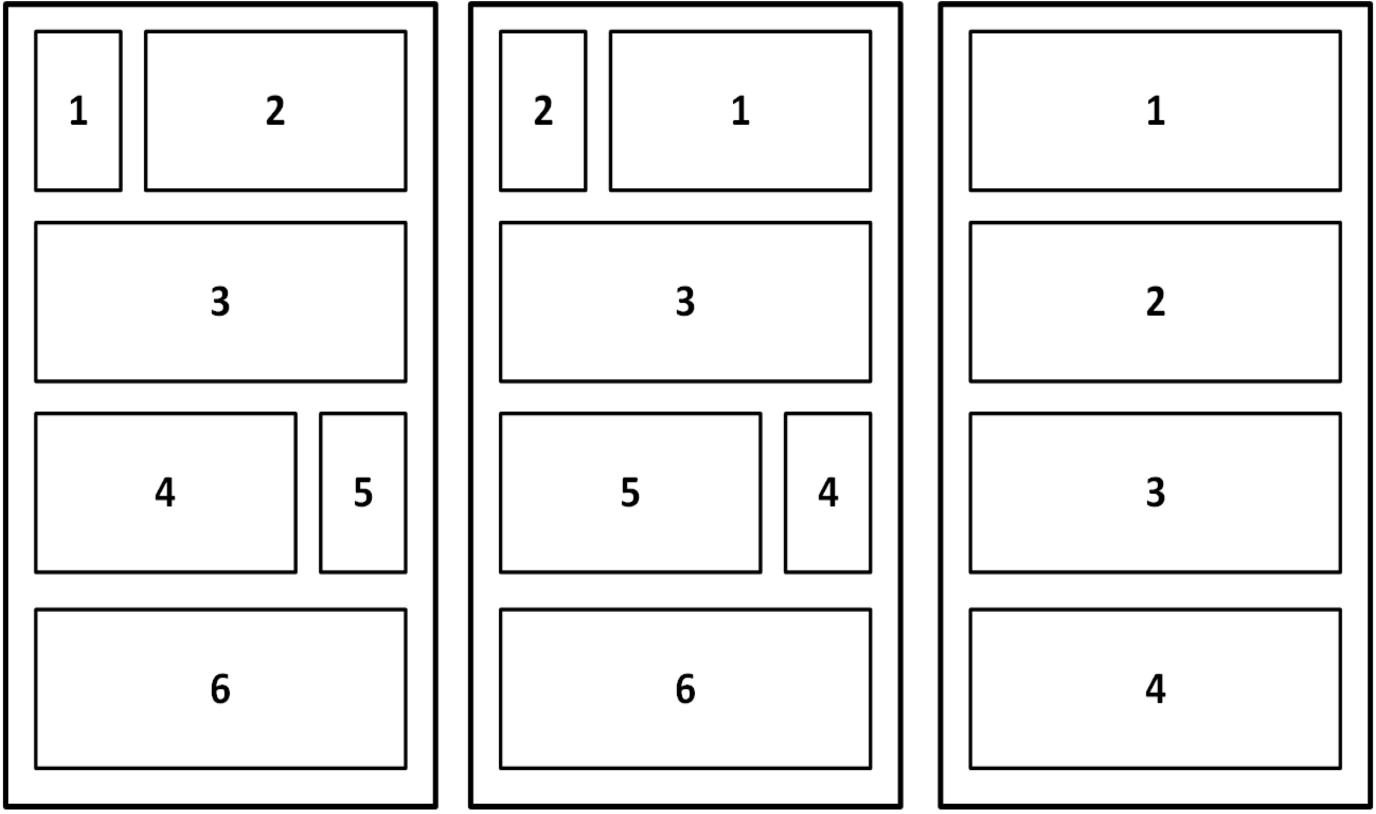


Fig. 11 Left-to-right, right-to-left and top-to-bottom reading styles.

The second approach was to classify the comic images into the main genre classes; Western, Manga, Manhwa and Manhua, using a neural network model based on the mobilenet v2 [44].

In this approach, the model, trained on the custom dataset, would classify the image to one of the genre classes. Afterwards, the reading style is attributed to the image, following the reading style of that genre. With the first step being classifying comics to their genres, this can later be extended to develop genre-specific extraction models. This approach yielded an accuracy of 0.80 on the tested dataset.

The third approach was an improvement from the second approach, where the model used for the classification changed whilst the methodology remained unchanged. There, a Convolutional Neural Network model was trained using the custom dataset to classify comic images into their genre classes. This approach yielded an accuracy of 0.89 superseding the performance of the mobilenet v2 based model.

A summary of the results for the approaches that were tested is presented in Table VII.

TABLE VII

SUMMARY OF THE RESULTS

Approach	Accuracy
Colour-based classification	0.89
mobilenetv2_1.00_224	0.80
CNN conv2D	0.89

Out of the three approaches, the third approach with the CNN model was selected because of its high accuracy and the ability to classify the genres of comics which can be extended to develop genre-specific extraction models in future developments.

VIII. Conclusion

The initial component of this research consisted of an analysis regarding the semantic content extraction of comic books with the utilization of object detection techniques. For this, the object detection models; Faster R-CNN, SSD Mobilenet as well as YOLOv3 were used. From these models, it was deemed that YOLOv3 performed the best under the constraint of limited computational power, due to its low memory usage as well as high accuracy in making detections in the test set.

Afterwards, focusing on text extraction, test results on speech balloon detection with a Mask R-CNN model were presented. Furthermore, test results on analysing the extracted speech balloons with the use of the Tesseract OCR engine were also shown.

Subsequent to extracting the semantic content in comics, this research presents a method for character-speech balloon association. An evaluation of three methods capable of selecting balloons for an association as well as four methods for predicting associations with characters and speech balloons was performed in this section. With test results for each method, it was concluded that for the selection for an association, selecting balloons that possess tails or are dissimilar to a square shape performed the best. In addition, the best results for character association was given when both the distance from the balloon as well as the distance from the direction the tail of the balloon points towards was taken as a distance measure.

Finally, after analysing the variations of reading styles in comics, a method is proposed by this research for determining the reading style for different genres of comics. A model-based approach was presented, with a data set that is currently being developed. Thus, by combining the various components of this research, it would be possible to yield a promising result for digitising comics.

Acknowledgement

We appreciate the support of the Department of Computer Science and Engineering at the University of Moratuwa, Sri Lanka. We are also grateful for the publishers of the public domain datasets and comics used for this research.

References

[1] T. Tanaka, K. Shoji, F. Toyama and J. Miyamichi, "Layout Analysis of Tree-Structured Scene Frames in Comic Images," in *20th International Joint Conference on Artificial Intelligence*, Hyderabad, India, 2007.

[2] Y. Wang, Y. Zhou and Z. Tang, "Comic Frame Extraction via Line Segments Combination," in *International Conference on Document Analysis and Recognition*, Nancy, France, 2015.

[3] M. Stommel, L. I. Merhej and M. Müller, "Segmentation-Free Detection of Comic Panels," in *Computer Vision and Graphics: Proc. Int. Conf. ICCVG, 2012*, Poland, 2012.

[4] X. Pang, Y. Cao, R. W. Lau and A. B. Chan, "A Robust Panel Extraction Method for Manga," in *Proceedings of the 22nd ACM international conference on Multimedia*, Orlando, Florida, USA, 2014.

[5] F. Niklas, "Image Processing - Binarization," 2019. [Online]. Available: http://felixniklas.com/imageprocessing/binarization.

[6] "Morphological Dilation and Erosion," The Mathworks Inc., 2019. [Online]. Available: https://www.mathworks.com/help/images/morphological-dilation-and-erosion.html.

[7] C. Rigaud, J.-C. Burie and J.-M. Ogier, *Text-independent speech balloon segmentation for comics and manga,* 2017.

[8] H. Jomaa, M. Awad and L. Ghaibeh, "Panel Tracking for the Extraction and the Classification of Speech Balloons," in *International Conference on Image Analysis and Processing*, Genoa, 2015.

[9] R. Fisher, S. Perkins, A. Walker and E. Wolfart, "Image Analysis - Connected Component Labeling," 2003. [Online]. Available: https://homepages.inf.ed.ac.uk/rbf/HIPR2/label.htm.

[10] C. Rigaud, N. Tsopze, J.-C. Burie and J.-M. Ogier, "Robust frame and text extraction from comic books," *HAL,* no. hal-00841493, 2013.

[11] A. Ghorbel, J.-M. Ogier and N. Vincent, "Text extraction from comic books," in *GREC 2015 Eleventh IAPR International Workshop on Graphics Recognition*, Nancy, 2015.

[12] M. Kass, A. Witkin and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision,* vol. 1, no. 4, p. 321–331, 1988.

[13] C. Rigaud, J. Burie, J. Ogier, D. Karatzas and J. V. D. Weijer, "An Active Contour Model for Speech Balloon Detection in Comics," in *2013 12th International Conference on Document Analysis and Recognition*, Washington, DC, 2013.

[14] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," *International Journal of Computer Vision,* vol. 57, no. 2, pp. 137-154, May 2004.

[15] W. Sun and K. Kise, "Detection of exact and similar partial copies for copyright protection of manga," *International Journal on Document Analysis and Recognition (IJDAR),* vol. 16, no. 4, pp. 331-349, December 2013.

[16] K. Takayama, H. Johan and T. Nishita, "Face Detection and Face Recognition of Cartoon Characters Using Feature Extraction," in *IEEJ Image Electronics and Visual Computing Workshop*, Kuching, Malaysia, 2012.

[17] F. S. Khan, R. M. Anwer, J. v. d. Weijer, A. D. Bagdanov, M. Vanrell and A. M. Lopez, "Color attributes for object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, 2012.

[18] H. N. Ho, C. Rigaud, J.-C. Burie and J.-M. Ogier, "Redundant structure detection in attributed adjacency graphs for character detection in comics books," in *10th IAPR International Workshop on Graphics Recognition*, Washington, D.C., USA, 2013.

[19] W. Sun, J.-C. Burie, J.-M. Ogier and K. Kise, "Specific Comic Character Detection Using Local Feature Matching," in *12th International Conference on Document Analysis and Recognition*, Washington, DC, USA, 2013.

[20] N.-V. Nguyen, C. Rigaud and J.-C. Burie, "Digital Comics Image Indexing Based on Deep Learning," *Journal of Imaging,* 2017.

[21] X. Qin, Y. Zhou, Z. He, Y. Wang and Z. Tang, "A Faster R-CNN based Method for Comic Characters Face Detection," in *14th IAPR International Conference on Document Analysis and Recognition*, 2017.

[22] N.-V. Nguyen, C. Rigaud and J.-C. Burie, "Multi-task Model for Comic Book Image Analysis," in *MMM 2019*, Thessaloniki, 2019.

[23] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017.

[24] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *TPAMI*, 2017.

[25] K. He, G. Gkioxari, P. Doll´ar and R. Girshick, "Mask R-CNN," *CoRR,* vol. abs/1703.06870, 2017.

[26] Google, "Tesseract Open Source OCR Engine," *Github Repository,* 2018.

[27] C. Rigaud, J.-C. Burie and J.-M. Ogier, "Segmentation-free speech text recognition for comic books," *HAL,* Vols. hal-01719619, 2018.

[28] C. Olah, "Understanding LSTM Networks," August 2015. [Online]. Available: http://colah.github.io/posts/2015-08-Understanding-LSTMs/.

[29] T. Breuel, A. Ul-Hasan, M. A. Al-Azawi and F. Shafait, "High-Performance OCR for Printed English and Fraktur Using LSTM

Networks," in *International Conference on Document Analysis and Recognition*, Washington, DC, 2013.

[30] C. Rigaud, N. L. Thanh, J.-C. Burie, J.-M. Ogier, M. Iwata, E. Imazu and K. Kise, "Speech balloon and speaker association for comics and manga understanding," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, Tunis, Tunisia, 2015.

[31] C. Guérin, C. Rigaud, A. Mercier, F. Ammar-Boudjelal, K. Bertet, A. Bouju, J.-C. Burie, G. Louis, J.-M. Ogier and A. Revel, "eBDtheque: a representative database of comics," in *12th International Conference on Document Analysis, Aug 2013*, Washington D.C., 2013.

[32] Aizawa-Yamasaki Lab, "Manga109," [Online]. Available: http://www.manga109.org. [Accessed 29 04 2019].

[33] F. Schroff, D. Kalenichenko and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015.

[34] M. Iyyer, V. Manjunatha, A. Guha, Y. Vyas, J. Boyd-Graber, H. D. III and L. Davis, "The Amazing Mysteries of the Gutter: Drawing Inferences Between Panels in Comic Book Narratives," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[35] "tensorflow/models/research/object_detection/," Github, [Online]. Available: https://github.com/tensorflow/models/tree/master/research/object_detection. [Accessed 25 February 2019].

[36] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu and A. C. Berg, "SSD: Single Shot MultiBox Detector," in *European Conference on Computer Vision*, Amsterdam, The Netherlands, 2016.

[37] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016.

[38] J. Redmon and A. Farhadi, *YOLOv3: An Incremental Improvement*, 2018.

[39] W. Abdulla, "Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow," *Github Repository*, 2017.

[40] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 2016.

[41] GodAnimeReviews, "Difference and origin of Manga, Manhua and Manhwa," 19 August 2018. [Online]. Available: https://godanimereviews.com/difference-origin-manga-manhua-manhwa/. [Accessed 28 April, 2019].

[42] L. Graillat, "America vs Japan: the influence of American Comics on Manga," *Refractory: a Journal of Entertainment Media*, vol. 10, 2006.

[43] C. Sager, "What's up with manga? A comics fan's deep dive," 19 March 2012. [Online]. Available: http://geekout.blogs.cnn.com/2012/03/29/whats-up-with-manga-a-comics-fans-deep-dive/. [Accessed 28 April 2019].

[44] M. Sandler, A. Howard, M. Zhu, M. Zhu and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA, 2018.