IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

# Detecting Components in Manga and Comics: A Survey

[1]Saikumar S, [2]Raghavendra R

[1]PG Student, [2]Assistant Professor

[1]NDepartment of MCA, School of CS and IT

[1]JAIN(Deemed-to-be-University), Bangalore, India

***Abstract:*** Comics and manga is becoming popular over the years, thanks to a slew of popular series luring newcomers to the medium. This has increased the necessity of digitizing these mediums to deliver to the consumers. Digitization which can preserve these pieces of art, can also help in extracting useful information from them. This can be addressed by using machine learning algorithms to extract components from the mediums. The methodologies explored by several scholars to achieve - text detection and voice bubble detection – are discussed in this work. We examine these tactics to determine their merits and downsides in order to comprehend existing procedures and forecast future advancements in this field. This paper intends to provide researchers interested in contributing to this subject with an overview of the various techniques in the aforementioned domains.

***Index Terms*** **- Manga, Comic, Text Detection, Speech Balloon Detection, Panel Extraction.**

## I. INTRODUCTIONS

The popularity of comic books and manga has grown throughout recent years. Despite the fact that traditional hard copies remain the primary source of sales in these mediums, they are no longer the primary source of revenue. This is due to the growing global market for these mediums, as well as how popular their works have become. As a result, numerous works have been digitized to make them more accessible.

While the notion of digitizing media is not new, it is novel for these mediums. We can maintain these materials in good shape by digitizing them, which is more difficult with a physical copy. We haven't found that digitizing them is really extremely valuable. Scholars have utilized novels, movies, and even audio samples published to the internet to discover further capabilities for technologies like machine learning. Scholars have been able to teach machine learning models to deep fake another person's face on a target, compose entirely new scripts, and synthesis someone's voice.

The challenge of obtaining usable information from images is an impediment to such data utilization in comics and manga. While data models in other medium employ the material itself, any text, speech bubble, or object in comics and manga must first be recognized and retrieved before any form of model can be trained on the resulting data.

Researchers have discovered that they can extract a few components from these media. While text detection as an OCR is pretty widespread, it is influenced by picture noise, which is quite a bit due to the nature of these mediums, hence specific approaches have been developed. Many studies have been conducted in the domain of speech balloon extraction, with diverse methodologies providing variable but great results. While some employ theoretical approaches such as an active contour, others employ more sophisticated techniques like as CNN (Convoluted Neural Networks).

## II. RELATED WORKS

During our survey, we found many studies that focus on extracting various components from the images. For each component there were various methods that were being employed to extract them, since all of these had similar results, we decided to survey different methods rather than stick to one. To the best of our knowledge there is only one dataset for manga – Manga109 [1], but many different ones do exist for comics.

The following sections are organized as follows Sections A through C provide an insight about text detection and extraction in the aforementioned mediums. Sections D and E focus solely on Speech Balloon Detection, while Section F talks about Panel Extraction in addendum to Speech Balloon detection.

### 2.1 Detecting Text in Manga using Stroke Width Transform

During our research, we discovered that using the Stroke Width Transform methodology in conjunction with a Support Vector Machine was incredibly efficient, outperforming some of the more traditional methods.

The SWT technique [2] is employed in this study to identify "stroke" properties of an item that can distinguish between textual and non-textual image sections. It's utilised in conjunction with the Histogram of Oriented Diagrams, which is a computer vision object detection mechanism. Before we go into the mechanics of this proposed solution, let's have a look at how SWT works.

SWT begins by creating an output image and assigning a value of ∞ to each pixel. The algorithm then calculates the width of the stroke element it detects on the original image by measuring the distance between two locations on opposite sides of the stroke.

If the detected width is less than the pixel's value, the pixel is reassigned to the lesser width value. Finally, a matrix with each element's stroke-width ratio is returned.

To find letters, each element with a stroke-width ratio that does not exceed 3.0 is grouped together. However, if the connected elements are too big or small they are eliminated. After a letter is found, chains are formed to merge different letters based on the stroke-width value, distance between letters and their height. After grouping a chain, a word is formed. Chains can be merged if two chains have shared letters and are in the same direction. Finally, the group or chain for a word is obtained from the image. Each element with a stroke-width ratio of less than 3.0 is combined together to find letters. If the connecting pieces are excessively large or small, they are removed. Following the discovery of a letter, chains are built to combine different letters based on the stroke-width value, the distance between letters, and the height of the letters. A word is generated once a chain has been grouped. If two chains share letters and are in the same direction, they can be consolidated. Lastly, the group or chain for words are extracted from the image.

The proposed method of the paper starts off with conducting the SWT technique, with the size of the output image same as the input image. For narrowing down the letters, since manga have a lot of background and natural scenes, the rules for grouping the elements are altered in this method, using the following function (1).

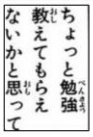
$$f(d,h,w,s) = \begin{cases} 1, if\ 1 << 15\ and\ s \leq 80\ and \\ \qquad 5 < h,w < 50 \\ 0, otherwise \end{cases}, (1)$$

where d is the diameter of the connected component, h is the height in pixel, s is the median stroke width, w is the width of the connected component in pixels.

This method ensures that all the letters are captured that would have not been possible just with the baseline SWT technique. However, there are many non-letter components that are captured too. Through an SVM classifier, image patches and non-image patches are classified. Image patches are then passed through an HOG descriptor, that is able to extract the letters. The distance between letters are used to form groups. A line is then created by including characters that are closer to another than 1.5 times the narrower character's width. This ratio comes from the experiments conducted in the paper. This model has a F-measure score of 0.506 which is significantly better than the baseline SWT method which has a F-measure score of 0.113.

### 2.2 Text detection in Manga by Combining Connected-Component-Based and Region-Based Classifications

Due to its discrimination capacity and high precision metrics, Deep-Learning is employed in this paper to distinguish between the words and the scenery in the image [3]. In contrast to the previous work, the purpose of this one is to detect texts by first identifying text regions. A manga's text can take many forms, as shown in Figure 1.

| | Clean (Text only) | Dirty (With other objects) |
|---|---|---|
| Typical font | TC  | TD  |
| Atypical font | AC  | AD  |

The suggested method consists of two steps: creation of region proposals and classification of regions. The technique of region proposal generation is used to find text regions quickly by classifying related components. Deep learning is used in region categorization, which improves the precision of the procedure. The extraction of features of related components is done first in region proposal generation, and then they are categorised and organised into rectangular areas based on their features. Deep learning is used in Region classification to categorise them based on deep features. The procedure is illustrated in Figure 2.

Region proposal generation — all connected components are retrieved from a page of manga, and each component is categorised as either a character or a non-character component because each component in manga is drawn independently. Geometric aspects of each component, such as area, perimeter, slope of approximation ellipse, and Euler characteristics, are determined for this classification. Positive samples of connected components from the TC and AC categories, as well as negative samples of connected components outside of all regions in either category, are utilised to classify these components using a Random Forest. This minimizes the classification's noise. The areas are then grouped according to their distance from one another, and these grouped regions can be sent to region categorization as region proposals.

Region Classification is done on a region-wide scale rather than on a component-by-component basis in this step. This removes the text sections that were incorrectly recognised. Deep features are utilised as region features to classify the regions, and these are extracted using deep learning. In this stage, the SVM model is utilised to classify the data. Two models are used to extract deep features: the first for ImageNet classification. This model was trained on large-scale natural pictures and comprises five convolutional and three fully connected (FC) layers. The deep features are recovered as a 4096-dimensional vector from the sixth layer.

The second model is made using Il- lustration2vec (I2v). I2v is based on VGG models, except it uses convolutional layers instead of FC layers to adapt to more detailed sections. SVM classifier is given these deep features, with text regions as positive samples and all other regions outside of text regions as negative samples. The final results are the regions labelled as text regions.

The greatest F-measure score was 0.466 when the ImN (ImageNet) model was used to extract deep features in conjunction with the other steps, whereas the best accuracy score was 0.15 when the I2v (Illustration2Vec) model was used. When the regions are exhaustively grouped, the approach achieves a recall of 0.851. The computational expenses of training deep learning models for two different tasks and then using two different classifiers build up, despite the method's excellent results.

### 2.3 Text Extraction from Digital English Comic Image Using Two Blobs Extraction Method

In this study, comic information such as speech balloons and the text contained within those balloons is extracted for data mining and translation applications. Text extraction methods based on region are used to extract the text. The focus of this work is solely on extracting English texts from comic books. The methods are further separated into two categories: connected components and edge-based approaches [4]. In the prior paper, connected components and region-based approaches were also utilised [3].

Connected components is a bottom-up strategy in which tiny components are combined into bigger and larger ones until all of the areas in the image are found. The picture is inputted, ideally in colour, and RGB band values are applied to the image for band selection. The RGB images are then transformed to a binary image, i.e. black and white, using threshold values ranging from 0 to 1. The average pixel value of the comic picture is used to calculate the threshold value T. The cut-off point has been set at 0.9. The RGB images (which contain noise) are pre-processed to reduce that noise, which improves the effectiveness of text extraction. The Median Filter is used to achieve this. The image is smoothed as a consequence of the technique, which averages all surrounding pixels.

Next, the picture is subjected to a CCL (Linked Components Labelling) method, which aids in the detection of connected components. Region borders are then discovered after CCL, and areas that are not separated by a boundary will be removed [5]. Text and non-text blobs are created through balloon detection. The picture now contains three bands: R-band, B-band, and G-band; the technique picks B-band since it has the fewest related components. Figure 3. shows the various bands in CCL.

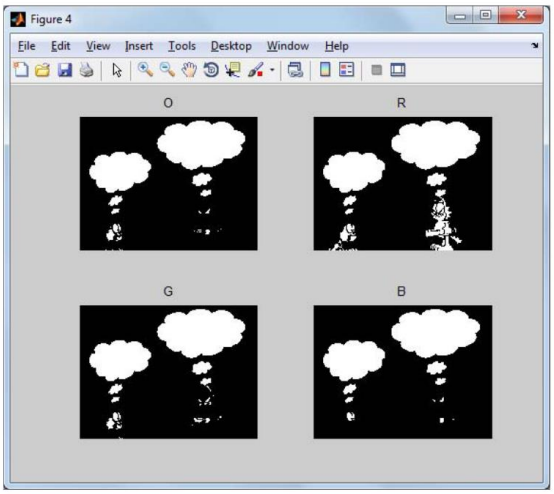



Fig. 3. Connected Component Labelling (CCL)

The picture now contains both textual and non-textual blobs, which are correctly identified using the blob size characteristics. The blob's area is estimated using the equation (2) below.

$$A.TB[i] = TB[i].Width * TB[i].Height \quad ,(2)$$

Where A is Area of Text Blob, TB is an array containing all the text blobs, Width and Height are properties of the Text Blob.

If the blob covers 10% or 8% of the original picture, it is classified as a text blob, whereas all other blobs are classified as non-textual blobs. The text is recovered from an OCR once the blob has been recognized. The goal of optical character recognition (OCR) is to categorize optical patterns that match to alphanumeric or other characters. Segmentation, feature extraction, and classification are all part of the OCR process, and the retrieved text is saved in a text file for later use.

With a Text Extraction Ratio of 94.82 percent, the system can accurately recognize text blobs and extract text from a digital English comic picture using the Median Filter.

### 2.4 An active contour model for speech balloon detection in comics

The notion of active contours is employed to detect speech balloons in this survey article [6]. An active contour is a snake-like deformable model that travels through image space to reduce the energy of a function. The snake is made up of a number of points that cover the image's specified region.

The snake draws the route of the speech balloon using three different energy concepts. The equation for external energy $E_{ext}$ internal energy $E_{int}$, and text energy $E_{text}$ are illustrated in equation (3):

$$E = E_{ext} + E_{int} + E_{text} , \qquad z(3)$$

External Energy $E_{ext}$ is given by equation (4),

$$E_{ext} = \gamma \min A(i,j) = \gamma \min \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}, \qquad (4)$$

where $E_{ext}$ is the minimum Euclidean distance (A) between a point $i$ and the nearest edge point $j$, $\gamma$ is a weighting parameter. This strategy aims to draw the snake from a distance distant from the speech balloon's targeted edges. Because edges matching to the text in the picture are undesirable, they are deleted before the image is subjected to an equivalent distance transform, which prevents the text from contributing to the external energy.

$E_{cont}$ and $E_{curv}$ are two energy concepts that make up Internal Energy $E_{int}$. $E_{cont}$ is the force that keeps all points of the snake at the same distance from one another, forcing the contour/snake to be continuous. As a result, when the distance between the sites is close to the average of all distances, the energy $E_{cont}$ will be the smallest. By punishing high contour curvatures, $E_{curv}$ enforces smoothness and prevents snake oscillations.

Text Energy $E_{text}$, communicates information about the relative placements of text regions and the balloon contours that go with them. Given the text's location, $E_{text}$ seeks to push the snake outward toward the most likely balloon localization. This ensures that the contour/snake follows the route of the speech balloon correctly.

The speech balloon detection technique begins by recognizing text inside the image; this is a critical stage because the system's overall outcomes are dependent on it. After that, the results of the text detection are post-processed to organize text lines into text paragraphs. The snake's movement and the accuracy of the final paragraphs are affected by the initial amount of points. The method swiftly localizes the global form of the text area by spreading out an equal number of points in the first step.

The algorithm then proceeds to smooth out the curve by adding more points to the contour. This causes the contour to move away from the text area and toward the balloon boundary when the weights of the $E_{ext}$ and $E_{text}$terms increase. If there is an explicit border, $E_{text}$ will draw the snake to it; otherwise, the snake will be pushed to the snake's suggested location.

The snake is then fitted to the coarse sections of the balloon, such as the peaks and tails. This is accomplished by adding intermediate points to the snake, adjusting the weighting values of the energy function, and then fitting the snake again. The procedure relaxes $E_{curv}$ term to allow the snake to fit into the coarse regions of the boundary while keeping $E_{cont}$ strong enough to maintain a consistent inter-point distance across the contour.

By just going through the picture through a one-stage contour fitting procedure without intervening spikes, the approach proved to be exceptionally accurate, with up to 93.4 percent recall and 92.8 percent accuracy. The contour has a 97.7% recall value and an 88.7% accuracy rating after a second step of additional fitting.

### 2.5 A clump splitting based method to localize speech balloons in comics

The notion of region extension is utilized in this study. Candidate areas are selected and extended, and depending on whether or not they meet the real speech balloon, they are either approved or divided at one of the area's points to fit into that speech balloon. [7]

This approach works on the assumption that the text sections of the pictures are known and delivered in bounding form before the comic image is processed. The technique for creating the image is shown in Figure 4. below.

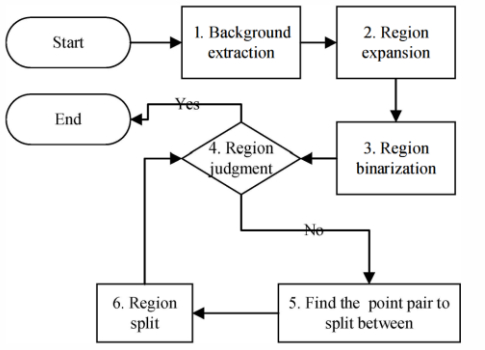



Fig. 4. Procedure to localize speech balloons

The background of the image is extracted using a component connected analysis approach. The background is then expanded into a suitable balloon region using a region growing approach. The region judgement technique then runs each candidate balloon region through a set of heuristic rules to see if it's a viable balloon. If that's the case, the text for that area is extracted. If the candidate is false, the algorithm locates a point pair on the contour of the region that splits it in half. Only one of the two child areas is preserved. This region is subjected to a recursive process of region judgement and region split until a viable candidate balloon region is discovered.

To take a closer look at the procedure. A connected components analysis is used to start the background extraction process. The backdrop is chosen as the component's largest bounding box. Points that are four-connected and have a grey value difference of less than 40 points are regarded to be part of the same component.

Traditionally, region expansion entails using an algorithm to extend a seed point according to specified rules. However, the region-growing mechanism is used in a different way in this study. The area expansion is repeated as many times as the number of background points once all points on the background bounding box are marked as unused. A different unused backdrop point is utilized as the seed point each time, and the region is extended.

To make calculation easier, the area is binarized. The value of all grey points is 255, whereas the value of all other points is 0. We receive the potential balloon areas after binerization.

Heuristic rules are used to find valid candidate balloon areas in the region judgement. The narrowest component of the potential balloon region is chopped off at the end. The bottleneck of the region is the visually narrowest section, which must be eliminated

for efficient speech balloon extraction containing the text. The suggested technique has an F1 score of 88.5 percent, an accuracy of 90.1 percent, and a recall of 86.9%.

### 2.6 A clump splitting based method to localize speech balloons in comics

This approach works on the assumption that the text sections of the pictures are known and delivered in bounding form before the comic image is processed. The technique for creating the image is shown in Fig. 4. below.

For panel extraction and speech balloon extraction, the research presents two distinct models. Similar to the previous study [7], the approach employs region growth for panel extraction and candidate area selection for speech balloon extraction [8]. However, ideas for deep learning models that use CNNs to extract panels from comics have been made [9]. CNN models are also available for recognizing manga objects such as a character's face [10]. However, training them is computationally intensive.

The assumption for panel extraction is that all of the panels have a white background. A region-growing technique is used to extract the background. This entails the algorithm assessing the original seed points' surrounding pixels to see if they should be included to the expanding area. Iteratively, the procedure results in the backdrop being white and the remainder of the page being black. This ensures that all of the panels are black as shown in Figure 5.

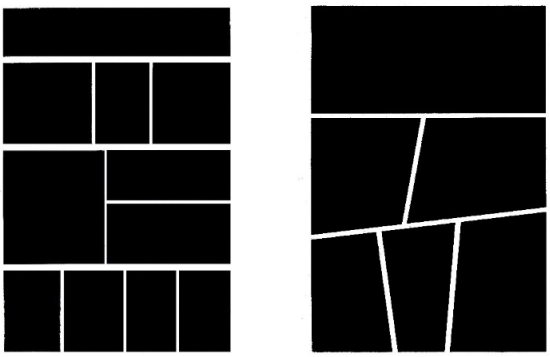
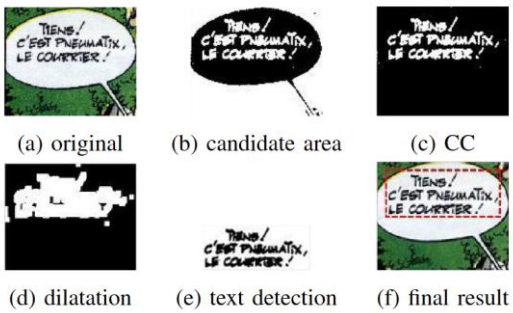


Fig. 5. Panels after detection are colored completely black

Although this isolates panels, it connects them when extended components are present, such as a speech bubble drawn on top of two panels. To avoid this, the image is subjected to mathematical morphology. Mathematical morphology is a method of analysing an image's geometric structure. In mathematical morphology, dilation allows black areas of a binary picture to shrink while white regions expand. Erosion, on the other hand, allows for the expansion of black regions and the contraction of white parts. N dilation operations are used to break the linkages between panels, and once a threshold is achieved, N erosion operations are used to allow the black sections to restore their original size. The image is then filtered for noise and each panel is labelled using CCL [3].

The first step in the Balloon Extraction method is to transform the RGB picture to an HSV (Hue, Saturation, Value) colour space. The Value (V) and Saturation (S) of all bright and white hues will be high (S). The initial pick is made based on the aforementioned parameters. Then, depending on the size and form of the potential regions, a second selection is made. Small areas have been eliminated. The areas where the ratio of number of pixels to the ones of the bounding box is 60% are kept. The next step is to see if any text is present in the potential locations. Text elements are connected using connected components to make a text block.

After then, dilation is used to enlarge white pixels and link together related components, such as text block elements. After that, all of the image's small related components are deleted. The image's speech balloons are the only remaining potential locations. An example of the procedure on a speech balloon is shown in Figure 6.




(a) original (b) candidate area (c) CC

(d) dilatation (e) text detection (f) final result

Fig. 6. Detection of Text in a Speech Balloon

## III. Conclusion

This paper examines several ways for extracting the various elements found in comics and manga. To summarize, this initiative will enable us to digitize comic books and manga, which will aid future technologies such as automatic translation. Computer vision technologies will also be improved as part of the project to aid digitization. It will also aid in the preservation of these works of literature and art, as well as act as a database for text mining, multimedia indexing, and other similar purposes. This project still has a lot of potential for development, such as onomatopoeia extraction, which is now being worked on by a number of researchers.

## REFERENCES

**[1]** Ali, A. 2001.Macroeconomic variables as common pervasive risk factors and the empirical content of the Arbitrage Pricing Theory. Journal of Empirical finance, 5(3): 221–240.

**[2]** Basu, S. 1997. The Investment Performance of Common Stocks in Relation to their Price to Earnings Ratio: A Test of the Efficient Markets Hypothesis. Journal of Finance, 33(3): 663-682.

**[3]** Bhatti, U. and Hanif. M. 2010. Validity of Capital Assets Pricing Model.Evidence from KSE-Pakistan.European Journal of Economics, Finance and Administrative Science, 3 (20).

**[1]** K. Aizawa et al., "Building a Manga Dataset 'Manga109' With Annotations for Multimedia Applications," IEEE multimed., vol. 27, no. 2, pp. 8–18, 2020, doi: 10.1109/mmul.2020.2987895.

**[2]** B. Piriyothinkul, K. Pasupa, and M. Sugimoto, "Detecting text in Manga using stroke width transform," in 2019 11th International Conference on Knowledge and Smart Technology (KST), 2019, pp. 142–147.

**[3]** Y. Aramaki, Y. Matsui, T. Yamasaki, and K. Aizawa, "Text detection in manga by combining connected-component-based and region-based classifications," in 2016 IEEE International Conference on Image Processing (ICIP), 2016, pp. 2901–2905.

**[4]** M. Sundaresan and S. Ranjini, "Text extraction from digital English comic image using two blobs extraction method," in International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME-2012), 2012, pp. 449–452.

**[5]** W.-T. Chu and C.-C. Yu, "Text detection in Manga by deep region proposal, classification, and regression," in 2018 IEEE Visual Communications and Image Processing (VCIP), 2018, pp. 1–4.

**[6]** C. Rigaud, J.-C. Burie, J.-M. Ogier, D. Karatzas, and J. Van De Weijer, "An active contour model for speech balloon detection in comics," in 2013 12th International Conference on Document Analysis and Recognition, 2013, pp. 1240–1244.

**[7]** X. Liu, Y. Wang, and Z. Tang, "A clump splitting based method to localize speech balloons in comics," in 2015 13th International Conference on Document Analysis and Recognition (ICDAR), 2015, pp. 901–905.

**[8]** A. K. N. Ho, J.-C. Burie, and J.-M. Ogier, "Panel and speech balloon extraction from comic books," in 2012 10th IAPR International Workshop on Document Analysis Systems, 2012, pp. 424–428.

**[9]** V. Nguyen Nhu, C. Rigaud, and J.-C. Burie, "What do we expect from comic panel extraction?," in 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), 2019, vol. 1, pp. 44–49.

**[10]** H. Yanagisawa, T. Yamashita, and H. Watanabe, "A study on object detection method from manga images using CNN," in 2018 International Workshop on Advanced Image Technology (IWAIT), 2018, pp. 1–4.