Snap & Hear: Comic Book Analyst for Children Having Literacy and Visual Barriers

R. B. Dias Yapa[1], T. L. Kahaduwa Arachchi[2], V. S. Suriyarachchi[1], U. D. Abegunasekara[1] and S. Thelijjagoda[3]

[1]*Department of Software Engineering, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka*
[2] *Department of Information Technology, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka*
[3]*Department of Information Systems Engineering, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka*

Keywords: Comics, Visual and Literacy Barriers, Recognition, Association, Image Processing, Machine Learning, Audio Story.

Abstract: Comic books are very popular across the world due to the unique experience they provide for all of us in the society without any age limitation. Because of this attraction, which comic books have received, it has proved that comic literature will be able to survive in the twenty first century, even with the existence of multi-dimensional movie theatres as its competitors. While the biggest global filmmakers are busy with making movies from comic books, many researchers have been investigating their time on digitizing the comic stories as it is, expecting to create a new era in the comic world. But most of them have focused only on one or few components of the story. This paper is based on a research which aims to give the full experience of enjoying the comic books for everyone in the world despite of visual and literacy barriers people are having. Proposed solution comes as a web application that translates input image of a comic story into a text format and delivers it as an audio story to the user. The story will be created using extracted components such as characters, objects, speech text and balloons and considering the association among them with the use of image processing and deep learning technologies.

1 INTRODUCTION

Comic is a unique category of entertainment which was initially printed in papers. Later, with the development of the technology combined with immense trend towards reading comic books have made it to the era of digital comic books where people have started reading comics through digital devices, instead of reading traditional ones.

Even today, storytelling is used as one of the best methods on guiding and improving the minds of kids. Out of all the story books, comic books have been playing a contrasting role in this field. But, one needs to have enough literacy to read and understand the association of dialogues and the images contained in these books in order to have the full experience of enjoying comics. Therefore, some people, especially the children who do not have enough literacy level and those who are visually impaired, have always been challenged and kept far away from obtaining this wonderful experience that they deserve. They had to be guided by an external party like parents or a caretaker to make them understand these stories. But unfortunately today everyone has become busy with their own work. Even caretakers might reject on storytelling repeatedly.

This paper is on the research that has been conducted on implementing a prototype by aiming to initiate with the process of giving the opportunity to have this comic book experience by targeting an extended audience by providing a user friendly web application to generate audio story and make it listen to the children through a guardian.

Going further, this research paper will provide information on related work which explains the current status of the related field, procedure and the methodology that has been carried out in order to make this into existence, results and discussion on the work done and the references which guided throughout this research process.

Yapa, R., Arachchi, T., Suriyarachchi, V., Abegunasekara, U. and Thelijjagoda, S.
Snap Hear: Comic Book Analyst for Children Having Literacy and Visual Barriers.
DOI: 10.5220/0008936402090215
In *Proceedings of the 12th International Conference on Computer Supported Education (CSEDU 2020) - Volume 2*, pages 209-215
ISBN: 978-989-758-417-6

1.1 Proposed Solution

Throughout the literature survey conducted, it was discovered that many researches have been conducted related to this field of digitizing the comic books, but almost all of them were focused only on one or few components. Therefore as a solution for this problem, this research bears the aim of implementing a prototype of a web application which could generate the audio story corresponding to an image which will be uploading to this web site, which will make the delivery of storytelling task easier an automated by just one click.

It took a lot of image understanding and processing works for the development and improvement of digitizing and automating the manual experience of comic books for us.

Following are the supported key areas in order to come up with the final prototype solution:

- Detection and identification of the comic characters and objects that exist within the scope.
- Speech text recognition and extraction from the segmented speech balloons.
- Speech balloon extraction and Story building through relationship analysis.
- Panel extraction and Text-to-speech conversion.

Following figure depicts the high level architecture of the proposed solution:

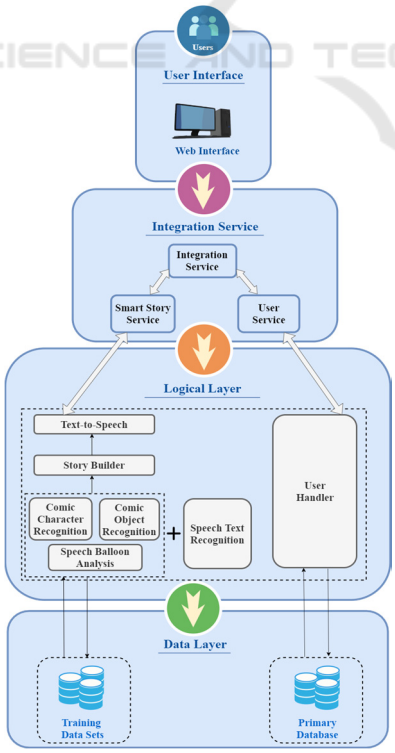

Figure 1: High level diagram.

2 RELATED WORK

A literature survey was conducted for this project to identify the existing work and solutions that have been used in the process of creating an audio story out of the comic images. There the unavailability of a proper commercial product as a complete solution for the proposed idea is observed. But most of the research has been done as separate concerns of extracting content of a comic book. None of the existing applications have built to generate the story by combining these extracted information from panels in a given comic book. Some of the researches which have conducted related to the comic story field are briefly explained below.

2.1 Related Work in Product Level

2.1.1 Smart Teddy Bear *A Vision-based Story Teller*

Smart Teddy Bear is a research conducted as an audio book player concealed as a teddy bear equipped with visual recognition capability Pham et al., 2013 . When a page of a book of interest is opened before the smart teddy bear, the system automatically analyze the images captured by the teddy bear's camera and recognizes which book is interested and plays the corresponding audio story. The proposed system is based on the Bag-Of-Words model in learning and recognizing visual object categories to identify the book corresponding a selected page. Then the corresponding audio story have been retrieved from a database.

2.1.2 Segmentation and Indexation of Complex Objects in Comic Book Images

This research was conducted on different approaches on comic books analysis. They had focused on three such approaches which describes the image content. This research was only conducted on segmenting and indexing the comic images, but not on examining the relationships between them Rigaud, 2016.

2.1.3 Enhancing the Accessibility for All of Digital Comic Books

Main aim of this research was to implement a system to make it possible for mobile users, motor-impaired people, and low-sighted people to access comic books (Ponsard and Fries, 2009). So, a system has been improved in the aspect of providing more improved

images in a way that targeted category can see the contents in comic books clearly.

2.2 Related Work in Component Level

2.2.1 Panel Extraction

Panel contains a key moment of the comic story which is depicted through speech balloons, speech text, characters and other story specific objects.

Panel extraction and ordering processes have been concerned in few researches based on X-Y recursive cut algorithm (Sutheebanjard and Premchaiswadi, 2010). A robust panel extraction method has been proposed to extract irregular panels but fails to separate panels joined by some elements (Pang et al., 2014).

2.2.2 Comic Object and Character Recognitions

Comic characters and objects are important components that are needed to focus on when performing content enhancement of comic books. Identification of comic characters is a challenging task because of the unlimited freedom of comic creators when drawing the comic characters, and also they are not rich with a vast number of features as in real human and other characters (Nguyen et al., 2018a).

Use of deep learning approaches along with machine learning techniques has become popular with the advancement of the technologies (Nguyen et al., 2018b). Analyzing the bounding boxes of the detected comic characters had been used in the extraction of those elements (Nguyen et al., 2018b). Convolutional Neural Networks is the latest technology which has been found as the most successful technique in character and object recognition fields (Ogawa et al., 2018). Although CNN performs good results with neuralistic images it does not provide the same performance with the comic book images (Ogawa et al., 2018).

Another approach had been used to identify the comic characters by using predefined colour thresholds for each character (Siddiqui, 2019). This approach can be useful in identification of characters with a specific costume where those dress codes possessed with a particular colour combination.

2.2.3 Text Recognition

In the process of story building, dialogues of the characters have been identified through speech text recognition.

Several researches have been conducted to recognize and extract speech text using image processing algorithms but only few researches have been done on identifying text fonts. In those researches adaptive thresholding, edge detection operators and mathematical morphology technologies (Shejwal Bharkad, 2017) had been used for text detection and K-nearest classifier (Shejwal Bharkad, 2017) had been used for text classification and fonts have been identified using a simple correlation- based method (Bashir, 2013).

2.2.4 Balloon Segmentation and Speaker Association

Balloons contain most of the text and go pairwise with comic characters. Association between speech balloons and comic characters involved in understanding the comic story. Literature survey was carried out to identify the existing solutions and technologies in balloon segmentation and association determination.

Few work concerns balloon extraction and mainly closed speech balloons have been studied. Only few has been done on open balloon extraction because of its partial outlines. These researches are based on Connected component detection approaches (Tolle and Arai, 2011,Ho et al., 2012,Rigaud et al., 2017) and contour-based classification methods (Rigaud et al., 2013). Anchor point selection approach has been proposed to determine association (Rigaud et al., 2015).

2.2.5 Story Building

Story building has been done through Visual Captioning and Visual Storytelling. In past research concrete content of image which is depicted through sentences are retrieved from visual captioning and summarize the idea and tell a story via visual storytelling.

Novel Adversarial Reward Learning (AREL) algorithm is proposed to generate more human-like stories for a given image sequences (Wang et al., 2018). Reinforcement learning, and adversarial training approach is proposed in a sequence to sequence modelling (Jing et al., 2018). Novel Bidirectional Attention Recurrent Neural Networks framework is proposed for visual storytelling (Liu et al., 2017).

2.2.6 Text-to-Speech

Text to speech conversion can be challenging in determining the real voice of comic characters.

Several researches have been done to convert text to speech and few are based on Microsoft Win32 Speech API (SAPI) (Domale et al., 2012). Few researches have been done to identify the real voice of comic characters and those are based on Amazon Mechanical Turk (Mishra et al., 2012).

3 METHODOLOGY

There are three application components implemented in our proposed solution.

1. End User Application - Angular
2. Integration Service - Node.js
3. Image Processing Service - Python

As the name implies, the Angular application acts as the frontend component for application which is visible for the end user to interact with. The Node.js service is the middle component which is listening to the requests made by the frontend and executes the story building process using the Application Programming Interfaces(APIs) exposed by the Python service

As per the high level diagram, the logical layer consists of several sub-components which have been implemented and powered by a Python flask service. Here are the main functionalities provided by this service.

3.1 Panel Extraction

When an image of a comic book, containing a set of story panels, is provided to this function, it will analyse the source and output the panel images that the source image is created with. In this panel extraction process, component labelling (CCL) algorithm and N erosions followed by N dilations performed on CCL, has been used to identify the panel shapes and cut the connecting elements between the panels.

Then, X-Y recursive cut algorithm has been used to extract the panels and get the extracted panels as separate panels and order the panels by top-down and left-to-right approach.

3.2 Comic Characters and Objects Recognition

This is for Identifying comic objects and characters along with their coordinates once a panel is provided to the component.

In implementation point of view, initially data of both comic characters and objects were gathered for the preparation of data sets. Two separate classifiers were implemented in python language to run sequentially and to work on recognising multiple classes of comic characters and objects from the panel images. Once the targeted components are recognised, those components were marked with the bounding boxes.

As the next step, locations of the recognised components have been extracted and passed along with recognised classes as the input for the association analysis and story building sections.

3.3 Speech Text Recognition

Images of the extracted speech balloons were taken as the input for the text recognition process and subjected to go through the Tesseract OCR model, which was identified as the best text recognition solution for this prototype. The recognized text will be saved as a text file on the deployed environment for further processing of the application.

3.4 Speech Balloon Segmentation and Component Association

This will create the association of the characters, objects and text contents recognized in the previous steps. In order to achieve that, it is necessary to get to know the association between the text and the characters who speak that particular dialogue. That is where the balloon segmentation comes to the play.

A model based on a deep convolutional neural network has been used for balloon segmentation. In this algorithm speech balloon detection is considered as a pixel-wise classification task and fully convolutional network approach has been used. Since this approach lack in precision for a single panel in preparation of the dataset, images were preprocessed to fit into the required image ratio of the model.

An algorithm has been proposed to retrieve the association link between speech balloon and speaker using anchor point selection and euclidean distance techniques with the assumption of character positions have been extracted and comic characters and speech balloons are relatively close. These associations are used in the process of building the story.

3.5 Story Building

In Story building section story will be built through connecting relationships between contents within the panel. String Manipulation has been used to combine the relationships of extracted character, objects and the speaker association and generate sentences. Then

the Sentences are combined to build the story. Built story will be delivered in text format as the output of this section.

3.6 Text-to-Speech

In this section Google Speech API has been used to convert text into audio. Amazon Mechanical Turk (MTurk) has been used to get the real voice of comic characters by data driven approach and using Naive Bayes classifier, the relationship between the character attributes and the voice has been modeled.

4 RESULTS

Results gained throughout the flow of the implemented prototype have been briefed below. The flow initiates with the upload of comic image. Once the image is uploaded to the web application it goes through the panel segmentation process and generates the segmented panels as separated images. Then from each of those images, desired comic characters, objects and speech balloons will be recognized as the results of the subjected character recognition, object recognition and balloon recognition deep learning models.

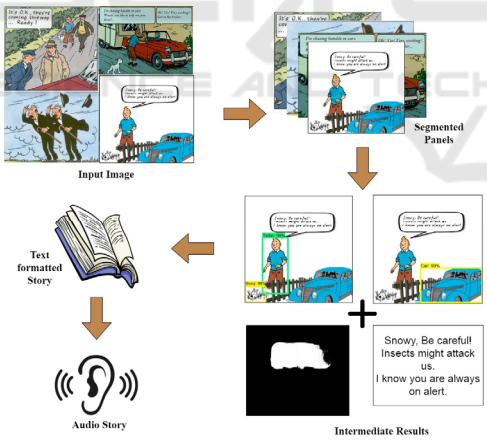

Figure 2: Resulted Product Flow.

Identification of comic characters and objects has been carried out successfully and the accuracy up to 99% can be reached by comparing with the generated inference graph when implementing the classifiers by training data. It is possible to happen misrecognitions due to the difficulties faced when gathering datasets for both characters and objects (specially for the objects). But in most cases, it is able to identify characters and objects accurately even with the challenges that were facing.

Output of the speech balloon extraction has also been another critical element of research which leads to the association of characters and the dialogues. Isolation of speech balloons from the initial image is merged with the original image again and provide the resulted image to the OCR component in order to extract the text of speech balloons from the given image.

As mentioned earlier, the text will be extracted from the output of balloon segmentation component, but not from the original image. Therefore the accuracy of text recognition will depend on the quality of the output of balloon segmentation component.

The extracted text will be saved in a text file in order to provide for the next steps.

For a given input image (such as Fig 3), a story will be generated using the results of intermediate steps which are explained above and will be displayed on the GUI as a plain text as shown in Fig. 4. Then the final result of this prototype will be delivered as an audio story to the user by generating an audio file for the text output which was created at the end of story building phase.

Figure 3: Input Image - in Panel Form.

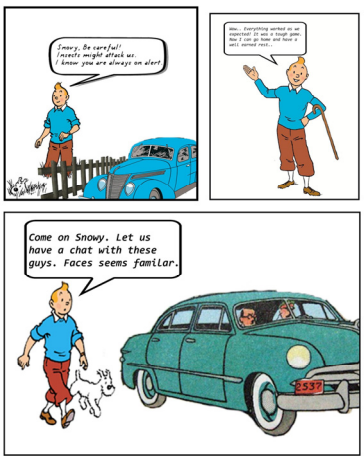

Figure 4: Input Image - in Page Form.

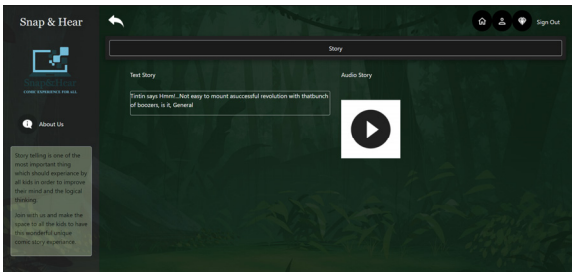

Figure 5: User Interface with Final Result Generated to the Input Image in panel form.

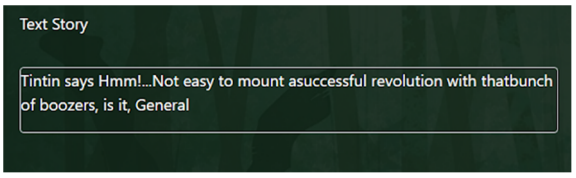

Figure 6: Close up Image of Generated Text Story.

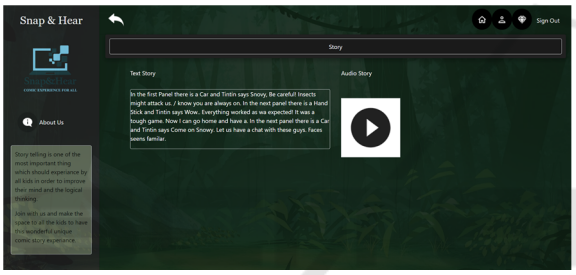

Figure 7: User Interface with Final Result Generated to the Input Image in page form.

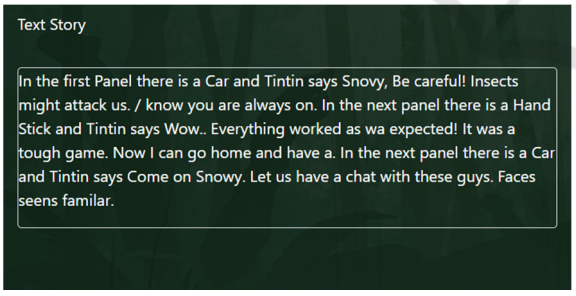

Figure 8: Close up Image of Generated Text Story.

5 DISCUSSION

Proposed approach can recognise specific classes of comic characters and objects as planned in the scope. Actions, speech tones and emotions of the comic characters are not taken into account.

While the dialogues which are boarded by the speech balloons are considered for the story, the graphical text which often used to convey emotions are not considered within our scope. The speech text recognition is only guaranteed for the font styles which has a fair equality to a clear font styles such as Arial, Times New Roman and Calibri.

Out of the different scenarios with the placement of speech balloons, tail and speaker within a panel, unique scenario of all the content included in a single panel, speech text is inside speech balloons, speech balloons are nearby their speakers and related speakers and speech balloons are within the same place is chosen for this work.

Although speaker voice feeding has been taken into consideration, it is limited to only a few characters when building the audio story within our scope.

5.1 Future Improvements

For the future improvements recognition of speech tone and emotions of the characters, Extending speech balloon extraction to extraction of balloons with partial contours and Extending character and object recognition models to identify more classes can be taken into consideration.

Since the recognition of selected characters and objects was able to do with a successful rate, successful recognition for rest of the objects and characters by training the algorithm with proper expanded data sets can be assumed safely. Text recognition also will be able to recognize for font styles with the increased data sets and text to speech component can be expanded with more comic voices with the expanded audio training streams of more comic characters.

6 CONCLUSION

There are two main problems that trying to address by the prototype proposed by this novelty. Providing the comic experience to children/adults who have literacy level issues, and helping busy parents to narrate stories to their children are those main problems which have tried to solve here.

The proposed web application solution is composed with content extraction processes of an uploaded image and the integration of the relationships between those contents to build a storyline and deliver it as an audio story to the audience. Through that, it was possible to address the above mentioned problems and to let the targeted audience to hear the generated audio and to have the new experience of digitised comic books.

It is possible to scale up the scope of Snap & Hear project and consider the scenarios mentioned above in the discussion to the analysis parts to give a better experience to the end user by integrating those missing inputs to the story as well.

REFERENCES

Duc-Minh Pham, Trong-Nhan Dam-Nguyen, Phuc-Thinh Nguyen-Vo and Minh-Triet Tran, "Smart Teddy Bear a vision-based story teller", *2013 International Conference on Control, Automation and Information Sciences (ICCAIS)*, 2013. Available: 10.1109/iccais. 2013.6720564

C. Rigaud, "Segmentation and indexation of complex objects in comic book", *ELCVIA Electronic Letters on Computer Vision and Image Analysis*, vol. 14, no. 3, 2016. Available: 10.5565/rev/elcvia.833

C. Ponsard and V. Fries, "Enhancing the Accessibility for All of Digital Comic Books", vol. I, no. 5, 2009. Available: http://www.eminds.hci-rg.com.

Phaisarn Sutheebanjard & Wichian Premchaiswadi, "A Modified Recursive X-Y Cut Algorithm for Solving Block Ordering Problems", *2010 2nd International Conference on Computer Engineering and Technology*, Available: v3-307.

Xufang Pang, Ying Cao, Rynson W.H. Lau, and Antoni B. Chan, "A Robust Panel Extraction Method for Manga", 2014.

N. Nguyen, C. Rigaud and J. Burie, "Multi-task Model for Comic Book Image Analysis", *MultiMedia Modeling*, pp. 637-649, 2018.

N. Nguyen, C. Rigaud and J. Burie, "Digital Comics Image Indexing Based on Deep Learning", *Journal of Imaging*, vol. 4, no. 7, p. 89, 2018.

T. Ogawa, A. Otsubo, R. Narita, Y. Matsui, T. Yamasaki and K. Aizawa, "Object Detection for Comics using Manga109 Annotations", *Research Gate*, 2018. Available: https://www.researchgate.net/publication/324005785_Object_Detection_for_Comics_using_Manga109_Annotations/citations.

K. Ahmed Siddiqui, "Skin Detection Of Animation Characters", *International Journal on Soft Computing*, vol. 6, no. 1, pp. 37-52, 2015.

M. Shejwal and S. Bharkad, "Segmentation and extraction of text from curved text lines using image processing approach", *2017 International Conference on Information, Communication, Instrumentation and Control (ICICIC)*, 2017.

S. Muhammad Arsalan Bashir, "Font Acknowledgment and Character Extraction of Digital and Scanned Images", *International Journal of Computer Applications*, vol. 70, no. 8, pp. 1-5, 2013.

H. Tolle and K. Arai, "Method for Real Time Text Extraction of Digital Manga Comic," *International Journal of Image Processing (IJIP)*, vol. 4, no. 6, pp. 669–676, 2011.

A. K. N. Ho, J. C. Burie and J.M. Ogier, "Panel and Speech Balloon Extraction from Comic Books," presented at *Tenth IAPR International Workshop on Document Analysis Systems*, pp. 424-428, Mar. 2012.

C. Rigaud, J. C. Burie and J.M. Ogier, "Text-Independent Speech Balloon Segmentation for Comics and Manga," 2017, pp. 133-147.

C. Rigaud, J. C. Burie, J.M. Ogier, D. Karatzas and Jo, "An Active Contour Model for Speech Balloon Detection in Comics," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR 2013*

Rigaud, C., Thanh, N.L.; Burie, J.C.; Ogier, J.M.; Iwata, M.; Imazu, E.; Kise, K. "Speech balloon and speaker association for comics and manga understanding," in *Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, Tunis, Tunisia, Aug. 23–26, 2015; pp. 351–355.

Xin Wang, Wenhu Chen, Yuan-Fang Wang and William Yang Wang, "No Metrics Are Perfect: Adversarial Reward Learning for Visual Storytelling." *ACL 2018.*

Wang, Jing, Jianlong Fu, Jinhui Tang, Zechao Li and Tao Mei, "Show, Reward and Tell: Automatic Generation of Narrative Paragraph From Photo Stream by Adversarial Training," *AAAI 2018.*

Y. Liu, J. Fu, C. W. Chen, "Let Your Photos Talk: Generating Narrative Paragraph for Photo Stream via Bidirectional Attention Recurrent Neural Networks," *AAAI Conference on Artificial Intelligence*.

Ajinkya Domale, Bhimsen Padalkar, Raj Parekh, M.A. Joshi, "Printed book to audio book converter for visually impaired",*2013 Texas Instruments India Educators' Conference.*

Mishra, Taniya & Greene, Erica & Conkie, Alistair. (2012). Predicting Character-Appropriate Voices for a TTS-based Storyteller System. *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*. 3.