Panel Tracking for the Extraction and the Classification of Speech Balloons

Hadi S. Jomaa(✉), Mariette Awad, and Lina Ghaibeh

Electrical Engineering Department, American University of Beirut, Beirut, Lebanon
hsj04@mail.aub.edu, {mariette.awad,lg00}@aub.edu.lb

Abstract. Searching for texts inside a full comic strip may be exhaustive, and can be simplified by restricting the scope of the search to single panels, and better yet to within individual speech balloon. In this paper, a novel approach is devised where a tracking algorithm is employed for panel extraction, and speech balloons are identified using 'Roberts' edge detection operator as well as a classifier to find the number of balloons within every panel using a non-exhaustive projection method. Two main objectives in the field of comic strip understanding are achieved through our panel tracking for the extraction and classification of speech balloons (PaTEC). PaTEC may be incorporated as a precursor to text extraction and recognition reducing the computational time and effort of searching the whole image to the speech balloon area itself. PaTEC accuracy for panel extraction is *88.78%* while balloon classification accuracy is *81.49%* on a homegrown comic database.

Keywords: Comic strip · Panel · Classification · Speech balloon extraction · Comic page segmentation · Text detection

1 Introduction

Comic books represent a cultural expression and have been known as a cultural heritage for many countries. Most of the well-known comics are American, Japanese, and European from which animated movies have been made. They have been around for more than a century serving as means of entertainment and gathering an audience that ranges from kids to adults with content that expands from satirical caricatures to drama, mystery and erotica. With the growth of the mobile industry and digitization that is happening worldwide, comic books are falling behind in the race to be present in the digital world.

For user convenience, a significant number of newspapers can be now read online. Following in their footsteps, comic companies are starting to do the same. People like to view comics on their electronic devices, on the go, without the hassle of carrying around the hard copy. However, viewing the comic strip as a whole image on mobile electronic devices is not always convenient. The visual information present in the comic strip, along with the expressions of the characters in the frames and the text in balloons may be deteriorated if the comic strip page was to be fitted without proper processing. This problem can be solved through segmenting the comic strip to its

V. Murino and E. Puppo (Eds.): ICIAP 2015, Part II, LNCS 9280, pp. 394–405, 2015.
DOI: 10.1007/978-3-319-23234-8_37

individual panels, and extracting the text before fitting the panels to the electronic screen. This way, instead of viewing the comic strip as whole and having to zoom in, the user can scroll through the panels while reading the extracted texts typed below each one, or the text can pop out by tapping on the panels.

Recently, comic digitization field has gained growing attention, and presented here are some approaches that deal differently with the presented problem. Morphological operations were applied in [1] followed by region growing to highlight the background from the panels, whereas [2] employed a recursive algorithm that detects uniform color stripes and kept on segmenting the strip into sub regions until no more stripes were detected. The regions that can no longer be segmented were saved as panels. [3] proposed an ACS that also uses the X-Y recursive cut algorithm to segment the page into frames, which are then fitted to the size of the mobile screen. In [4], Rigaud et al. formulated the problem of speech balloon detection as fitting of a closed contour around text areas where the outline is not always confined to the image. However they assumed that the text location is already known. In [5] a watershed segmentation algorithm extracted the panels and the comic strips were preprocessed to classify the panels with uniform solid colors differently from that of the white background. Ishii et al in [6] detected separation lines through calculating the total gradient of the lines to get the basic shape of the panel. The corner candidates were then detected using Harrison's corner detection technique while Sobel filter calculated the image gradient. Frame lines were shifted in parallel until they reach the corner candidates. Evaluation values of all possible rectangular combinations were calculated based on the average density of gradient values of the quadrangle formed between each line and the intersection. Combinations with the highest evaluation values were kept as frames.

Using the pre-described XY recursive cut algorithm in [7], candidate points are detected for segmenting the frames and then these candidates are classified by a multilayer perceptron as corners of panels or not. In [8], screen division is applied using the density gradient after filling the quadrangle regions in each image with black. All possible lines that pass by every pixel are generated and an evaluation term of the density gradient is created. Based on an exhaustive criterion, the separation lines are detected.

In this paper, we propose a panel extraction technique followed by speech balloon extraction and classification, hereafter PaTEC, The mean-shift tracking algorithm is applied to a processed comic strip, after which every panel undergoes edge detection coupled with filtering and morphological operation to remove all but the speech balloons. The accuracy of panel extraction topped Burie's method [1], which showed better results than those in the literature, by more than 10%, while the speech balloon classification resulted in 81.49% accuracy. The differences between PaTEC and the existing approaches proposed in the literature are many: 1) it doesn't include recursions, such as the X-Y cut algorithm, 2) requires less computation and showed better results than region growing, and 3) doesn't require calculating the gradient of the image along different orientations. The remaining of this paper is as follows: section 2 deals with the methodology while experimental results are detailed in section 3. Section 4 concludes the paper along with possible future work.

2 Methodology

Throughout this paper, some comic book terminology described in Fig. 1 will be used. The comic strip corresponds to the page, the panel refers to the block which contains information on a specific event, and the gutter is the space separating the panels.

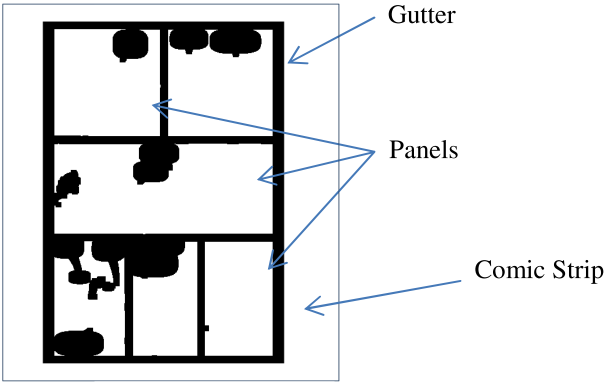


Fig. 1. Comic book terminology

An overview of PaTEC is represented in Fig. 2. As mentioned earlier, a tracking algorithm is applied to detect individual panels within the processed comic strip. Every panel is then treated individually. The speech balloons are extracted through ‘‘Roberts’’ edge detection operator coupled with a set of morphological operations and filtering, and then using histogram projections, they are classified. PaTEC workflow hence has four major steps, namely: Pre-Processing, Tracking, Extraction and Classification.

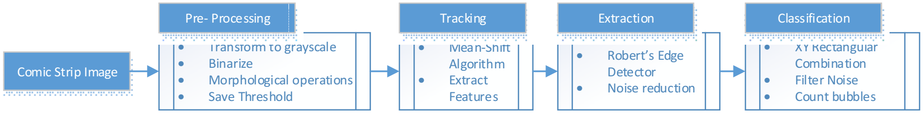
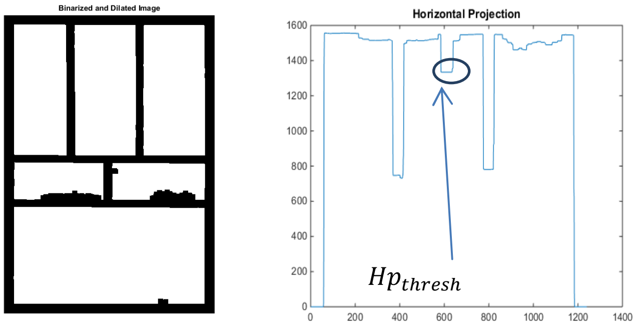


Fig. 2. PaTEC flowchart

2.1 Pre-processing

The input to the system is a colored image, size (m, n), of the comic strip that is first transformed to grayscale. The image is then binarized based on a grayscale threshold of *253*, chosen heuristically, since the white background has a value ranging between *253* and *255* A morphological operation, *N-dilations*, is applied to the image to widen the separation between the panels. The number “*N*” varies from one strip to another, and is chosen by user visual inspection to be the smallest number that ensures proper separation between the panels. Hence slight user interaction is required in setting the value of "*N*" which results in the optimal separation of the panels from the background. The user is also required to visually check the horizontal projection of the panel and save the value corresponding to the maximum number of pixels a gutter may have in the horizontal projection (Hp_{thresh}). In Fig. 3a, a binarized image is presented, with the horizontal projection in Fig. 3b. The value circled in Fig. 3b is the maximum number of pixels a gutter may have in the horizontal direction. The shorter the gutter in the comic, the larger the threshold.



(a) (b)

Fig. 3. (a) Binarized image (b) Horizontal projection with circled threshold

The horizontal gutters are considered as the spaces with zero vertical projections and their ordinates are saved automatically ($y_{horizontalgutter_i}$). They represent the separation lines between the panels.

2.2 Tracking

The tracker initiates from the top left part of the image, a couple of pixels before the beginning of the upper left panel at point P_1 (x_1, y_1). The mean shift algorithm (MSA) is initialized with center ($c_1 = P_1$) and horizontal direction ($d_1 = [1\ 0]$). The mean of the valid points within the specified semi-ellipse, i.e. the points that belong to the gutter and/or background, is calculated, and is saved as the new center,c_i of the semi-ellipse. The direction is also updated to be the difference between the old and the new center ($d_i = c_i - c_{i-1}$).After every iteration, the tracker checks whether or not it should initiate a change in direction, i.e. turn down. The decision is based on the value of the horizontal projection at the corresponding pixel coordinates (Hp_i). If the value is less than the pre-specified threshold,($Hp_i < Hp_{thresh}$), and no obstacles are present between the center and the nearest horizontal gutter ($y_{horizontal_{gutter,i}}$), this means that the tracker reached the edge of the panel, and should start rotating downwards. The horizontal gutter can be considered the first point whose vertical projection ($Vp_i = 0$) is zero. The turndown point is saved, $Td_i(xd_i, yd_i)$. After the change in direction is applied, i.e. turn downwards, the tracker proceeds until it reaches the nearest horizontal gutter separating the panel being tracked from the one below it, and then turns right. The turn right point is saved, $Tr_i(xr_i, yr_i)$. The tracker proceeds until it reaches the lower edge of the panel and when it goes past the abscissa (x_0) of the origin, rotates upwards. Notice that the rotation of the tracker is in a clockwise manner and it ends when the point is a couple of pixels near the initial point $(|y_i - y_0| < thresh)$. The order in which the tracker extracts panels is displayed in Fig. 4.

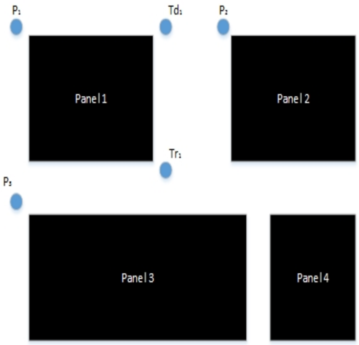


Fig. 4. Panel extraction order

The initial point for the second panel is considered a couple of pixels away from the turn-down point of the previous panel if it is still far from the end of comic strip. If not, the tracker jumps down to the second block of panels, and initiates from the same abscissa of the one above it. The vertical level of the new initial point is just a few pixels below the turn-right point of the panel above. $(if\ |n - xd_i| \gg 0\ , P_2\ (xd_i + x_1, yd_i); else\ P_2\ (x_i, yr_i - y)$). The pseudo code of extracting a single panel is presented in Fig. 5 followed by a pictorial depiction of panel extraction for a typical comic page present in the home-grown dataset in Fig. 6.

```
1   Initiate tracker at point $P_1$
    $d_1 = [1\ 0]$
    Set $done\ =\ false$
    **While** $\neg done$ **do**
2       Create semi ellipse
3       Get valid point coordinates $v_i$ for all $i \in gutter$
4       Calculate the mean $m_i$
5       Update center $c_{i+1} = m_i$
          **If** $Hp_i < thresh\ \wedge \sum_{j=i}^{y_{horizontal\ gutter,i}} I(x,y) = 0$ **then**
            $d_{i+1} = [0 - 1]$
            $(xd_i, yd_i) = c_{i+1}$
            **Repeat** steps 2,3,4,5 **until** $y_i\ = y_{horizontal\ gutter}$
            $d_i = [-1\ 0]$
            **Repeat** steps 2,3,4,5 **until** $|y_i - y_1| < thresh$
            Connect the final point $c_i$ and initial point $P_1$
            Set $done\ =\ true$
          **Else**
            Update the direction vector $d_{i+1} = c_{i+1} - c_i$
```

Fig. 5. Panel tracking algorithm

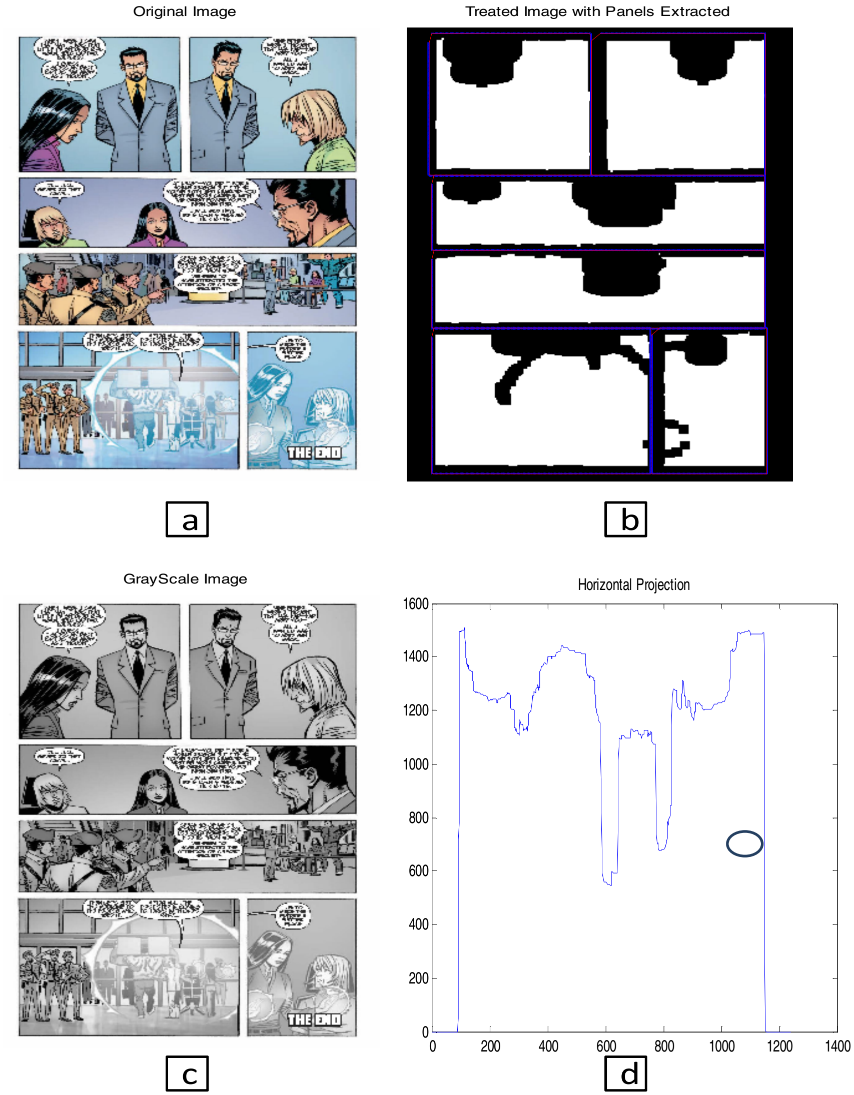


Fig. 6. (a) Original Image (b) Treated image with panels extracted (c) Grayscale image (d) Horizontal projection of treated image

2.3 Extraction

Speech balloons are extracted using the 'Roberts' operator for edge detection. The "Roberts" operator is effective in extracting speech balloons, since it calculates the gradient of the pixels through convolving them with a diagonal matrix, and keeping the ones with a gradient greater than an automatically set threshold. Other edge detection operators such as 'Prewitt' or 'Sobel' fail in detecting curved objects, and are mainly used to detect straight horizontal and/or vertical edges. Canny edge detector on the other side detects the edges of the speech balloon along with other unwanted edges making it difficult to locate the speech balloons.

Typically, after applying edge detection, random noises appear along with the speech balloon, thus further filtering is applied. The edge panel is projected vertically, and using connected components, labeled. We assume that speech balloons cover in length at least *3%* of the panel. Based on this empirical threshold, connected components falling below what is specified are discarded. The result is a panel with only speech balloons and some random unfiltered noise. A sample balloon extraction done on the same panel of Fig. 5 is presented in Fig 7.

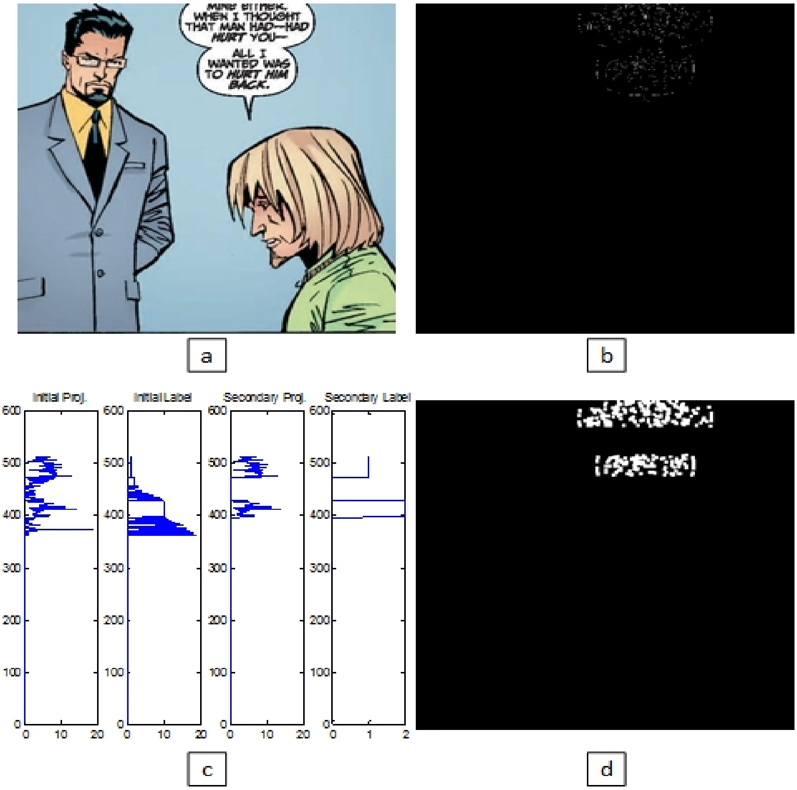


Fig. 7. (a) Original image (b) Edge image pre-filtering (c) Projection of edge image before and after filtering (d) Exracted balloons dilated

2.4 Classification

Classification is needed in order to get the real number of balloons within the panels. A non-exhaustive way to classify balloons is based on projections. Start by getting the horizontal and vertical limits of where balloons may be present based on the respective projections. Based on these limits, rectangular combinations are formed to explore the presence of bubble. Every rectangle hence may contain a bubble, noise, or nothing based on its content. The decision is based on the content of the rectangle. If the rectangle contains high density of pixels, above a certain threshold, it is considered as having a bubble, and added to the count; otherwise, the rectangle is considered to contain noise and consequently no balloons. The value of the threshold was taken to be 20% of the rectangular area, and was chosen heuristically based on the assumption that at least one fifth of the area covered by the balloon is filled with text.

3 Experimental Results

3.1 Experimental Setup

A data set consisting of 38 pages extracted from 5 different issues of 2 distinct comic books has been created. The pages were selected from the issues based on the following criteria: the comic strip should have white background, any comic strip with more than one irregular panel is discarded, and the panels should include round-shaped speech balloons. This data set contains *205* panels with various sizes and shapes. The result of panel extraction is expressed and compared using two methods, Page and Panel. A Page is considered well segmented when all the panels have been correctly extracted. The Panel section represents the accuracy of extracting any panel in the page. The script used was written and executed in *MATLAB R2014a*.

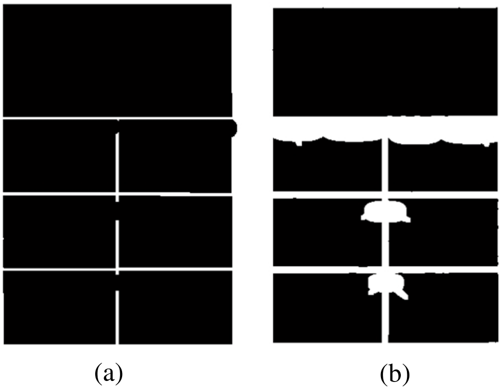
3.2 Panel Extraction

The results of the tracker are validated manually. [1]'s method was reproduced and tested on our dataset for the sake of comparison. In their paper, Burie *et al's* approach showed better results than others in the literature. Table 1 represents the accuracy of our method versus Burie *et al's* [1]. PaTEC failed to detect irregular panels, i.e. slanted panels and panels with more than 4 corners, present within the comic image that lead to this gap in accuracy.

Table 1. Accuracy of Panel Extracion

Method	PaTEC	Burie *et al.*
Page (%)	**89.05**	72.24
Panel (%)	**88.78**	75.63

PaTEC outperforms Burie's on both Panel and Page scale. The accuracy of extracting all the panels in the comic strip is considerably high. Looking also at the standard deviation of the results, our approach has a standard deviation *(18.88%)* less than that of Burie's *(32.11%)*. It means that in some cases Burie's method either extracted most of the panels in the page, or extracted none. The figure below displays a comic strip treated by the region growing method followed by *N-dilations* and *N-erosions* Fig. 8a compared to the same comic strip treated by PaTEC, Fig. 8b.



(a) (b)

Fig. 8. (a) Burie (b) PaTEC

Applying connected components to the left will result in extracting 4 panels, while on the right, applying the tracker yields 6 out of 6 panels. Since the balloons connected two separate panels, region growing failed in extracting each one individually. The balloons however where emphasized as part of the gutter after we applied our thresholding method, and broke the connection between joining panels.

3.3 Speech Balloon Classification

The panels successfully extracted contain a total of 335 speech balloons, where a balloon is considered to be a closed contour surrounding texts in the panel. The content-based classification process applied to the different rectangular combinations resulted in identifying 273 balloons (*81.49%* accuracy), which were visually inspected. The accuracy of the extraction is considerably high, and resulted in misclassification at certain instances when speech balloons overlapped, in which case more than one balloon are identified as one, or when there is significant unfiltered noise, adding a non-existent balloon to the count.

3.4 Discussion

PaTEC has certain limitations. Notably, the tracker can't extract all the panels in comic strips where panels exist on different horizontal levels *without* slight user intervention. The user must limit the scope of the image to that of the panel, before applying the tracker, because when projecting the whole image vertically, the small horizontal gutter is masked, and passed over by the tracker. Fig. 9a displays a comic strip that requires such user interference. Fig. 9b shows the result of extracting the panels automatically, while Fig. 9c displays the result with slight intervention. Notice how all three panels to the right of the Eiffel tower are extracted as one in Fig. 9b, while they are detected separately in Fig. 9c.

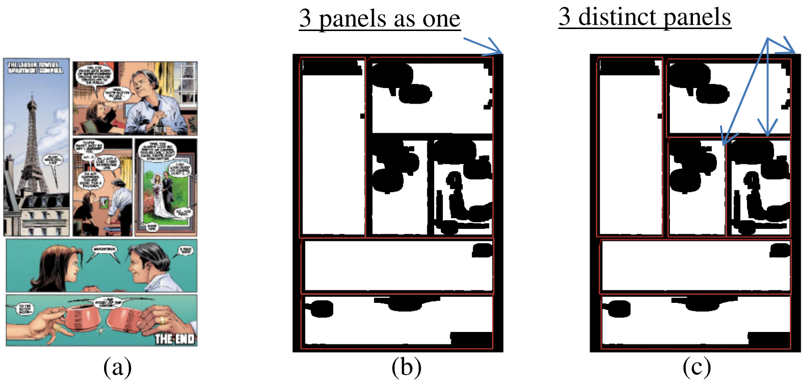


Fig. 9. (a) Original panel (b) No intervention (c) Slight user intervention

Slanted panels are not extracted since the vertical projection of the comic strip doesn't recognize the slanted gutters. In Fig 10b, the slanted gutter is masked by the pixels of the panel, preventing the tracker from identifying a horizontal gutter. Such failure results in extracting slanted panels as a part of a larger panel such as in Fig. 10c.

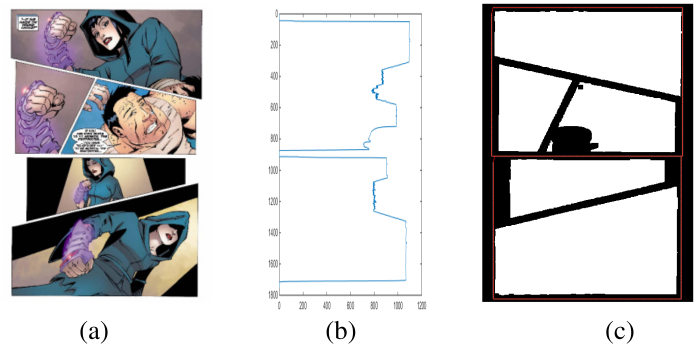


Fig. 10. (a) Slanted image (b) Vertical projection (c) Extracted panels not added to count

Other comic strips in which not all panels are extracted include panels that extend to the border of the image, panels with more than 4 corners and panels that are formed by objects rather than in blocks. Examples on such situations are found in Fig. 11.

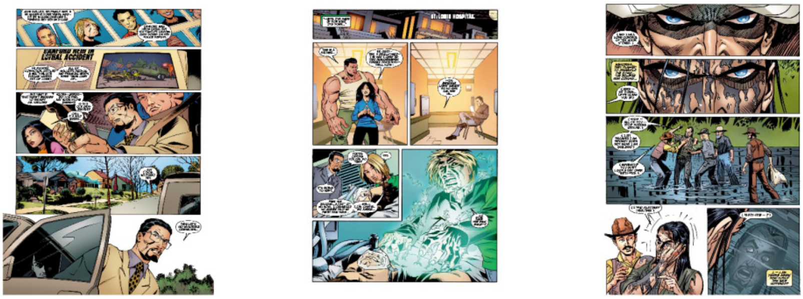
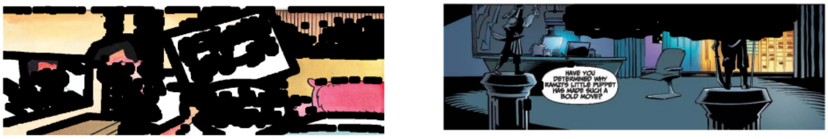


Fig. 11. Sample failures

Image processing of the comic strip varies from one to another since every image has different sized gutters. Hence the number of dilations which is directly related to the original size of the gutter between panels can't be fixed to constant value.

When speech balloons overlap, i.e. boundaries are almost connected to each other; the classifier fails to detect them as separate, and ends up classifying them as one. Some panels may end up with scattered noise that looks like a speech balloon and hence misclassified as one, Fig. 12a. In others, not all speech balloons survive the edge operator, and are overlooked by the classifier such as in the panel of Fig. 12b.


HAVE YOU DETERMINED WHY RAMZI'S LITTLE PUPPET HAS MADE SUCH A BOLD MOVE?

(a) (b)

Fig. 12. (a) Noise augmented with balloon (b) Missed balloon

4 Conclusion and Future Work

In this paper, we proposed panel tracking for speech balloon extraction. The comic strip is preprocessed through several image processing techniques. A tracker is employed to extract the panels of the image and then every panel undergoes other processing techniques for speech balloon extraction and classification. The results of panel extraction reached *88.78 %* and speech balloon classification accuracy was *81.49 %*.

PaTEC doesn't apply to irregular panels. Improved image processing techniques might be able to enhance the speech balloon preservation. Preprocessing, tracking and extraction, applied in PaTEC can be dealt with separately as any optimization in one step contributes to overall better results.

References

1. Ho, A.K.N., Burie, J.-C., Ogier, J.: Panel and speech balloon extraction from comic books. In: 2012 10th IAPR International Workshop on Document Analysis Systems (DAS). IEEE (2012)
2. Chan, C.H., Leung, H., Komura, T.: Automatic panel extraction of color comic images. In: Ip, H.H.-S., Au, O.C., Leung, H., Sun, M.-T., Ma, W.-Y., Hu, S.-M. (eds.) PCM 2007. LNCS, vol. 4810, pp. 775–784. Springer, Heidelberg (2007)
3. Han, E., Chun, S., Park, A., Jung, K.: Automatic conversion system for mobile cartoon contents. In: Fox, E.A., Neuhold, E.J., Premsmit, P., Wuwongse, V. (eds.) ICADL 2005. LNCS, vol. 3815, pp. 416–423. Springer, Heidelberg (2005)
4. Rigaud, C., et al.: An active contour model for speech balloon detection in comics. In: 2013 12th International Conference on Document Analysis and Recognition (ICDAR). IEEE (2013)
5. Ponsard, C., Fries, V.: An accessible viewer for digital comic books. Springer, Heidelberg (2008)
6. Ishii, D., Watanabe, H.: A study on frame position detection of digitized comics images. In: Proc. Workshop on Picture Coding and Image Processing, PCSJ2010/IMPS2010, Nagoya, Japan (2010)
7. Han, E., Kim, K., Yang, H.-K., Jung, K.: Frame segmentation used MLP-based X-Y recursive for mobile cartoon content. In: Jacko, J.A. (ed.) HCI 2007. LNCS, vol. 4552, pp. 872–881. Springer, Heidelberg (2007)
8. Tanaka, T., et al.: Layout Analysis of Tree-Structured Scene Frames in Comic Images. In: IJCAI, vol. 7 (2007)