Character Segmentation from Ancient Palm Leaf Manuscripts in Thailand

Rapeeporn Chamchong
School of Information Technology
Murdoch University
South Street, Murdoch, Western Australia
rapeeporn.c@gmail.com

Chun Che Fung
School of Information Technology
Murdoch University
South Street, Murdoch, Western Australia
l.fung@murdoch.edu.au

ABSTRACT
This paper presents a character segmentation system from ancient palm leaf manuscripts written in ancient Thai language. This aims to develop an automated system for the digitization and processing of ancient manuscripts. In this paper, the preprocessing stage of noise reduction is carried out. An optimal binarization is selected in order to reduce the unrelated noise and background information on the document. The proposed approach can improve the readability of the documents and enable selection of the optimal binarization technique. Text line segmentation is then applied to partial projection profiles, and the characters are separated by using the contour tracing algorithm and a trace of background skeleton. The experiment results have shown that this proposed system can be used to support subsequent steps such as automatic recognition of characters from Thai ancient palm leaf.

Categories and Subject Descriptors
I.4.6 [**Image Processing and Computer Vision**]: Segmentation – *Pixel classification*. I.7.5 [**Document and Text Processing**]: Document Capture—*Document analysis*;

General Terms
Algorithms, Experimentation

Keywords
Binarization, Image Segmentation, Text Segmentation, Character Segmentation, Document Image Analysis.

1. INTRODUCTION
Palm leaf manuscripts are collections of ancient documents in Thailand having great cultural value and containing invaluable knowledge about the history, culture, and local wisdom of Thai civilization. All these manuscripts have been used to record information in the past and they were mostly written by hand. These manuscripts are a heritage from past civilization passed down through many generations. There exists a huge collection of these manuscripts in libraries, museums and institutes in Thailand such as the National Library of Thailand [1], and Mahasarakham University [2]. These documents are deteriorating due to age and lack of preservation facilities at the place of collection.

Currently, computer technology has been used to record a large amount these documents in multimedia formats for future analysis and storage. Although current systems can store all these images, there is not yet available any specific system that is capable to retrieve relevant information efficiently and to extract knowledge from them. It is therefore a key objective of this project to develop an efficient image processing system that could be used to retrieve knowledge and information from the historical palm leaf documents in Thailand.

Palm leaf documents are different from other documents that were printed or produced by modern technology. Information on these physical media is harder to extract because formatting structure of documents are looser. In addition, these documents are of poor quality, due to their fragility and deterioration over age. The various problems are due to issues such as holes and spots on the media, blurriness, smearing, dirt and discoloration. These factors lead to poor contrast, and ghosting noise due to seeping ink from the other side of the manuscripts between the foreground text and the background. In addition, characters were handwritten in narrow spaced lines with overlapping and touching components. Moreover, characters have unusual, varying shapes, and different styles, which depend on the writer. Thai language by itself also imposes additional challenges due to lack of separation between words and the high number of consonants, vowels and tonal indicators, Digital image processing techniques are therefore necessary to improve the readability of the manuscripts.

In this paper, a proposal for the character segmentation from ancient palm leaf manuscripts is reported. These ancient manuscripts have been collected from projects reported in [1]. In the next section, a background of the study and related work are described. This is followed by section 3 in which image enhancement is presented. Then, section 4 describes the binarization and the optimal selection of binarization techniques by machine learning. Section 5 explains the process of text line and character segmentation. Experimental results and discussion are then shown in Section 6, and a conclusion and future work on future research are given in the last section.

2. BACKGROUND AND RELATED WORK
Prior to the stage of information extraction, characters or text on the images have to be recognized. There are four steps which need to be completed prior to the task of character recognition or extraction. The process of preprocessing system for information extraction [3] is given as follows

1. *Image acquisition:* acquire the scanned image in color format (RGB) then convert the RGB image to gray-scale image.

2. *Image enhancement:* remove noise and enhance the image. There are two categories of filtering technique. One is *linear filtering* such as low pass filter, enhancement filter, discrete Laplacian edge detector etc. Another is the *non-linear filtering* which includes median filter, morphological filtering, and others. Average filtering and Gaussian filtering are conventional methods of image enhancement [3].

3. *Binarization:* eliminate background and extract text. This is an essential part of the preprocessing step in image processing,

which is crucial to remove unrelated information, noise and background on the documents. If these steps are ineffective, the original characters from the image may be unrecognizable or more noise may be added. Therefore, these techniques are essential to improve the readability of the documents and the overall performance of the process. However, there are several available techniques and it is difficult to select the optimal algorithm and its associated parameters.

Binarization techniques can be separated into two main methods, they are *global* thresholding (such as Otsu, and Kapur [4]) and *local adaptive* thresholding techniques (such as Niblack [5], and Bernsen [6]).

Although there exist several binarization techniques, researchers (Trier and Taxt [7], Trier and Jain [6], and Leedham and et al.[8]) have proved that there is no single binarization technique that can be applied effectively to all kinds of digital documents. However, Trier and Taxt found Niblack's and Bernsen's methods have shown to have good performance in the binarization process. In the study, the global thresholding technique based on Otsu's method has demonstrated good performance with simple documents and bimodal histogram, while the performance may vary with different data sets.

4. *Text line and character segmentation:* separate text line and individual character. From past research, the process of character segmentation consists of three steps. Text line segmentation is the first. This is followed by word segmentation and then character segmentation. In ancient Thai writing system, there is no space to separate word like the English language, so character segmentation has to be done after text line segmentation. As a matter of fact, character segmentation could also be done without text line separation. However, in the optical character recognition process, flow of text component (character or alphabet) cannot be read unless they are in sequence. Consequently, line segmentation is used to form a horizontal script.

In the survey by Likforman-Sulem and et al.[9], text line segmentation of historical documents is separated into six categories, they are: projection-based, smearing, grouping, Hough-based, repulsive-attractive network and stochastic methods. They reported that piecewise projections which were proposed by Pal and Datta [10], and Zahour and et al.[11] are suitable for overlapping or touching lines, and another technique based on stochastic method [12] is also suitable for overlapping lines and more robust. Their summary stated that there is no single line segmentation technique that is suitable for all historical documents. The particular technique will depend on the characteristics of the writings such as script size, stroke width and average spacing

Character segmentation has been proposed many years ago. Casey and Leolinet [13] have reported a survey of the methods and strategies in character segmentation. The four strategies for segmentation are listed as the classical approach, recognition-based segmentation, holistic approach, and hybrid methods. Many techniques have also been proposed for segmentation of touching characters.

From the literature, character segmentation of Roman, Arabic, Indian, Chinese and Japanese scripts have been published but there are only few research on Thai language. Among the papers concerning Thai handwritten segmentation, they used the classical approaches to segment character [14], [15], [16] and the demonstrations are based on controlling the writing styles of the writers that is not practical. In addition, all those studies were based on modern Thai language and it is different from ancient Thai language.

Surinta and Chamchong [17] applied the recursive strip for line segmentation and connected component for character segmentation on Thai palm leaf manuscripts. They claimed their approach was able to segment their samples correctly at 71.81%. Their experiment found several components fell into wrong lines and several ancient Thai characters cannot be separated due to overlapped components and connected characters. While it was found that these techniques could be applicable, they however are not useful for practical documents.

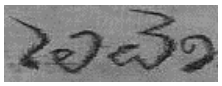
In terms of document analysis, it is desirable to have embedded tools for searching of blocks, lines and words, and the inclusion of a dedicated handwriting recognition system. Interactive tools are generally offered for segmentation and recognition correction purposes. Several projects in the past are concerned with printed materials. However, solutions to tackle Thai handwritten text perfectly are yet to be developed. Furthermore, there is no OCR system, tool, or development currently widely available for the processing of ancient Thai handwriting and documents.

3. IMAGE ENHANCEMENT

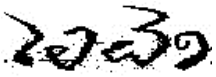
When a document is scanned and converted to gray scale image, noise may be generated and included in the image. Therefore, filtering techniques are used to reduce the noise and enhance data in image. In general, filtering techniques can be applied both before and after binarization. In this stage, filtering technique is applied before binarization by using linear smooth filter which is known as the *Gaussian filter* [3]. This filter can smooth data and enhance character. This filtering technique consisting of convolution between the gray-scale image, $g(x,y)$ and a Gaussian mask filter is given as follows

$$f(x,y) = \frac{1}{16}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix} \otimes g(x,y) \quad (1)$$

The comparison of the binary images with and without Gaussian filtering is shown in Fig.1. Gaussian filtering technique can filter salt and pepper noise, and enhance the edge of the character as shown in the figure.



a) Gray-scale image



(b) Binary image without enhancement

(c) Binary image with enhancement by Gaussian filtering

Figure 1. Binary image results illustrating no enhancement and enhancement by Gaussian filtering

4. BINARIZATION TECHNIQUES

Binarization is the technique of converting a gray-scale image to a binary image by using threshold selection techniques to categorize the pixels of an image into either one of the two classes.

From the study, there is no single technique that is suitable for all types of document. An automated selection of binarization technique could be implemented in order to assist the process and improve the system performance. This aims at selecting the most appropriate algorithm by using machine-learning techniques which has been explained and evaluated in [18]. There are six candidates of binarization techniques in this work. This approach consists of three modules: clustering, feature extraction, and selection.

4.1 Clustering Module

This module is to cluster the binary images by using k-means algorithm [18]. The objective is to group the images of similar characteristics together. When an optimal algorithm is selected, it could be applied to the other images within the same cluster.

4.2 Feature Extraction Module

Color histogram is the most commonly used feature for image characterization. It is found to be very effective. Beside, color moment, is used in this research in the forms of mean and standard deviation. They are used to convey the color distribution information. This forms a compact representation of the color feature used to characterize a particular color image.

4.3 Selection Module

This module is the main component of the optimal selection. The prediction model is established by learning from the training data set. In this study, Backpropagation [18] is used with a hidden layer. There are 258 input nodes (256 bins of histogram, mean, and average values), 132 hidden nodes, and 6 output nodes (binarization algorithms).

After this stage, the optimal binarization algorithm is selected for the image and then the background is eliminated from the gray-scale image by using the selected binarization algorithm. Consequently, text line and character will be separated from the binary image. This is explained in the next section.

5. TEXT LINE AND CHARACTER SEGMENTATION

This study developed a system for extracting scripts from palm leaf manuscripts that have Thai-Noi and Tham alphabets of Ancient Thai language which are different from modern Thai language. To this date, there is no effective system for automatic separation of the characters in these scripts. To extract lines and characters from the manuscript, partial projection profile [10, 11] is initially applied for text line extraction. The text line is then used to define the sequence of characters. After text line separation, character segmentation is then applied by integrating the Contour Tracing algorithm and a trace of the background skeleton [19].

5.1 Text Line segmentation

If the text is separated without text line segmentation, the sequence of the characters cannot be defined. Consequently, text line segmentation is purposed to define the sequence of the characters.

To separate text lines, the partial projection method is applied by dividing the text images into vertical columns. The width of the column is defined as one character which is calculated by manual from the mean value of the width of the characters that can be separated as individual character from the 43 palm leaf manuscript images. The approach to separate the lines is as follows:

- Crop the border of the image in order to reduce unnecessary information.
- Divide the image into vertical columns by using the width of a character.
- Calculate the horizontal projection profile in each column.

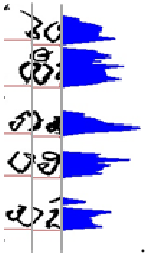


Figure 2. Sample of projections of a column

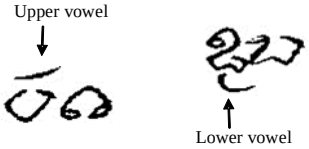


Figure 3. Consonants with a vowel

- Find the minimal value of the histogram by applying horizontal projection for all columns as shown in Fig.2. This minimal value of histogram indicates the top and bottom lines. A bottom line is chosen as a base line. If the height of the character of each line is less than the average height character which is calculated from the mean value of character heights from the data set, this based line is deleted. However, there are some vowels appear above or below the characters and they were drawn as isolated components as shown in Fig.3. The positions of these vowels occupy certain distance from the characters. This significantly affects the separating line. To calculate this value, two distances are calculated as shown in Fig.4. The value of *d1* defines the distance between the bottom of the upper line and the top of the vowel. *d2* defines the distance between the bottom of the vowel and top of the lower line. If (*d1>=d2*) then this vowel belongs to the lower line. If (*d2>d1*) then this vowel belongs to the upper line.
- Calculate the average value of a number of lines (*avg_num_line*) from all columns and this value is used to be a key of the number of lines in each document. If the number of lines in each column is less than the *avg_num_line*, a base line is added from the same line of the closest right/left column which has a lower base line. This process starts from the right column to the left column. If the characters of two lines are connected,

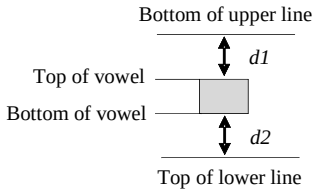


Figure 4. Analysing the distance of the vowels

they can be separated by checking against the height of the characters. Touching consecutive lines can be separated into two lines by setting the based line position of the upper line as in Equation (2).

$$line[i][j] = line[k][j] + \frac{|line[i][j] - line[k][j]|}{2} \quad (2)$$

where *j* is the current line position, *i* is the current column and *k* is the left line position.

- Join horizontal line by linking the position of horizontal line from right to left.
- Draw the joined line and then separate each line.

5.2 Character Segmentation

After the text line is extracted, isolated character is then segmented. There are two steps involved. First, the contour tracing algorithm is processed. This method is suitable for extracting over segmented character images and slant writing styles. Subsequently, the mean and standard deviation width of all the segmented components in each manuscript are calculated (μ and σ). The value is then used for the analysis of the width of the segmented components. If the width of the segmented component is more than the *criteria* value which is shown in Equation (3), it will be separated again by using a trace of the background skeleton.

$$criteria = \mu + \frac{1}{2}\sigma \quad (3)$$

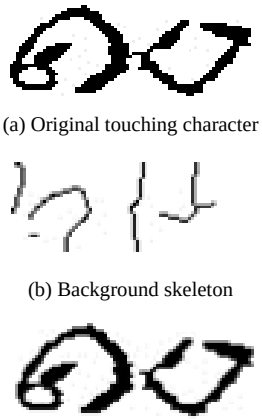
5.2.1 Contour Tracing Algorithm

Contour tracing is a technique that is applied to digital images for extracting the boundary of an object. This proposed system applies one of the recent contour tracing algorithms to separate character by using the *Theo Pavlidis*'s algorithm [20]. It works with 4-connected patterns. The width of the segmented components from this process is checked. If it is more than the *criteria* value of the average width of the components, it will be processed in the next stage. This means there are some touching characters are not separated.

5.2.2 Tracing of Background Skeleton

This approach is applied to separate some touching components. To segment touching characters, background skeleton is processed by using the *Zhang-Suen* thinning algorithm [21]. Then, contour tracing algorithm is applied to extract the skeleton of the background. An example result after this process is shown in Fig.5.

After this process, the characters in each line will be sorted by checking the column position in order to determine the sequence of the characters. This research has implemented the previously



(a) Original touching character

(b) Background skeleton

(c) After tracing background skeleton

Figure 5. An example result of tracing of background skeleton

described techniques and applied to practical data from ancient documents. The experimental results are presented in the next section.

6. EXPERIMENTAL RESULTS AND DISCUSSION

In this experiment, palm leaf manuscript images were collected and scanned by the Palm Leaf Manuscript Preservation Project in the Northeastern Region of Thailand, conducted at Mahasarakham University [2]. The proposed system has been implemented with Visual C++ and OpenCV library. The resolution of the input images is 200x200 dpi in RGB format. The input images were converted to gray-scale images and then noise is reduced by using Gaussians filtering technique. Then the filtered image is transformed to binary image by automatically selecting the optimal binarisation algorithm. This experiment has been reported in [18]. After this, line and character segmentation were applied. In this study, 43 binary images were considered for text line segmentation and 10 images from text line segmentation were investigated for character segmentation.

Example results of this system are shown in Fig.6. In the experiment, 195 text lines were considered from 43 palm leaf manuscripts. To check whether a text line is segmented correctly, a boundary is drawn between two lines as shown in Fig.6(b). The result of line extraction is measured by following the rules in [10]. Experimental results of the experiment are given in Table 1. The experiment shows that 114 lines of all lines (58.46%) were segmented correctly. 57 lines of all the lines (29.23%) have one component out of the correct lines. For the lines with 2 components out of the correct lines, the number is 10 (5.13%). The rest are more than two out of the correct lines.

The above technique can be used to separate some touching characters from consecutive lines. However, errors can be due to prolonged characters in each column and being adjacent between two consecutive columns. This has affected a few of the characters in the document.

The performance of the character segmentation processes are shown in Table 2. The data set consists of 2702 characters from ten palm leaf manuscripts. The correct character segmentation rate by using the contour tracing algorithm was 81.24% while the correct character segmentation after tracing background skeleton

was 82.57%. This demonstrates the tracing background skeleton can improve the performance of character segmentation. A main problem of segmentation is the touching characters. The techniques applied in this system have separated most characters correctly but in some cases, some touching characters are difficult to separate due to the writing style and smearing from ink. Moreover, some of binary images are unclear and they have a major effect on the accuracy of text line and character segmentation.

7. Conclusion and Future work

This research has applied image processing and intelligent approaches for isolating characters from practical ancient palm leaf manuscripts. This system can be applied as a preliminary stage of a fully automated system in the future. However, there are some problems caused by touching characters, dis-segmenting characters, noise surrounded characters, and incorrect line separation. For performance evaluations of character segmentation, such as using existing OCR system, this cannot be done as there is no present OCR system supporting the recognition of ancient Thai language. This is a challenge for Thai researchers to develop automated system for the recognition of ancient Thai language. Another challenge is to develop a Content Based Image Retrieval (CBIR) System for Thai manuscripts written in Thai language. Future development will aim at enhancing the performance of the proposed system. More data sets will be used to test the prototype in order to verify the fully automated knowledge and information extraction system for ancient Thai manuscripts. Furthermore, a framework of optimal selection of binarization techniques will have to be adopted and character segmentation has to be improved by considering the touching characters.

Table 1. Results from Text Line segmentation

Number of components out of the correct line	Number of lines correct	Percentage of lines correct	Percentage of components segmented within the correct line
0	114	58.46%	100%
1	57	29.23%	98.00-99.99%
2	10	5.13%	96.00-97.99%
3	11	5.64%	94.00-95.99%
≥4	3	1.54%	<=93.99%

8. ACKNOWLEDGMENTS

The authors wish to express their thanks and appreciation to members of the Preservation of Palm Leaf Manuscripts Project, Mahasarakham University, Thailand for their support and providing images from their database.

9. REFERENCES

[1] Ancient Languages. National Library of Thailand. Microfilm of palm leaf manuscripts with Thai alphabet.

[2] Palm Leaf Manuscripts in Northeastern Thailand: http://www.bl.msu.ac.th/2553/english_bl.htm

[3] Cheriet, M., Kharma, N., Liu, C.-L. and Suen, C. Y. Character recognition systems : a guide for students and practioners. John Wiley & Sons, Inc., New Jersey, 2007.

[4] Kapur, J. N., Sahoo, P. K. and Wong, A. K. C. A new method for gray-level picture thresholding using the entropy of the histogram. Graph. Models Image Process, 29(1985), 273-285.

[5] Niblack, W. An introduction to digital image processing. Prentice Hall, 1986.

[6] Trier, O. D. and Jain, A. K. Goal-directed evaluation of binarization methods. IEEE Trans. Pattern Analysis and Machine Intelligence, 17, 12 (December 12 1995), 1191-1201.

[7] Trier, O. D. and Taxt, T. Evaluation of binarization methods for document images. IEEE Trans. on Pattern Analysis and Machine Intelligence, , 17, 3 (1995), 312-315.

[8] Leedham, G., Chen, Y., Takru, K., Joie Hadi Nata, T. and Li, M. Comparison of some thresholding algorithms for text/background segmentation in difficult document images. 2003.

[9] Likforman-Sulem, L., Zahour, A. and Taconet, B. Text line segmentation of historical documents: a survey. International Journal Document Analysis and Recognition 9, 2 (April 2007), 123 - 138.

[10] Pal, U. and Datta, S. Segmentation of Bangla unconstrained handwritten text. 2003.

[11] Zahour, A., Taconet, B., Mercy, P. and Ramdane, S. Arabic hand-written text-line extraction. 2001.

[12] Tseng, Y. H. and Lee, H. J. Recognition-based handwritten Chinese character segmentation using a probabilistic Viterbi algorithm. Pattern Recognition Letters, 20, 8 (1999), 791-806.

[13] Casey, R. G. and Lecolinet, E. A Survey of Methods and Strategies in Character Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 18, No.7(Jul. 1996 1996), 690-706.

[14] Arunrungrusmi, S. and Chamnongthai, K. Adaptive blocks for skeleton segmentation in handwritten Thai character. 2000.

[15] Chatwiriya, W., Klinkhachorn, P. and Lass, N. Thai handwriting legal amounts recognition. 2003.

[16] Lohakan, M., Airphaiboon, S. and Sangworasil, M. Single-character segmentation for handprinted Thai word. 1999.

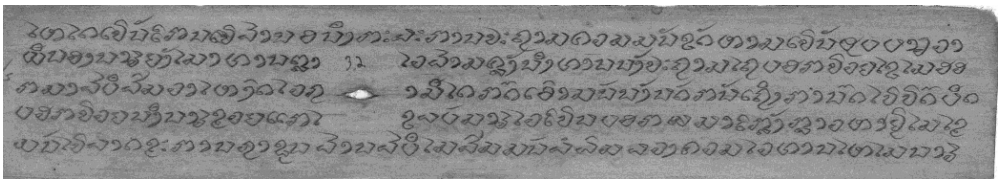
[17] Surinta, O. and Chamchong, R. Image segmentation of historical handwriting from palm leaf manuscripts. Springer, 2008.

[18] Chamchong, R. and Fung, C. C. Optimal selection of binarization techniques for the processing of ancient palm leaf manuscripts. 2010.

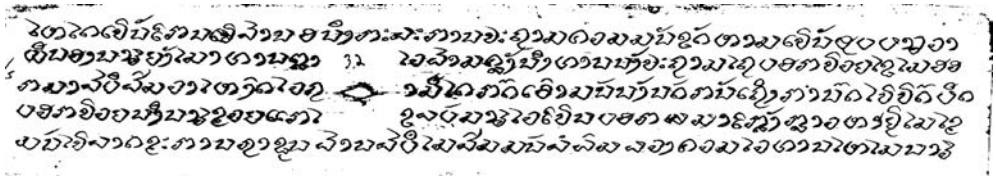
[19] Shuyan, Z., Zheru, C., Pengfei, S. and Qing, W. Handwritten Chinese character segmentation using a two-stage approach. 2001.

[20] Pavlidis, T. Algorithms for Graphics and Image Processing. Computer Science Press, Rockville, , Maryland, 1982.

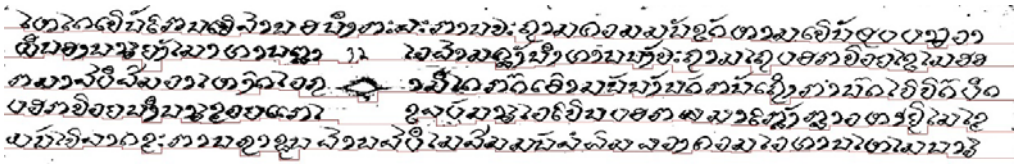
[21] Zhang, T. Y. and Suen, C. Y. A fast parallel algorithm for thinning digital patterns. Commun. ACM, 27, 3 (1984), 236-239.



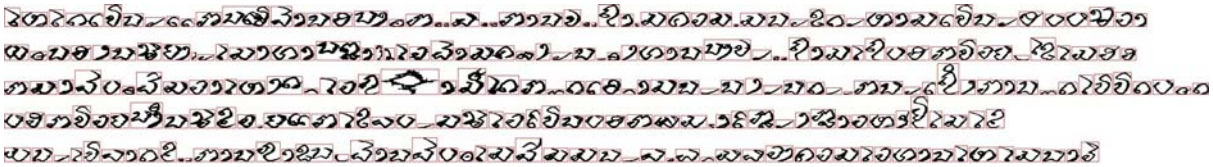
(a) Filtering image



(b) Binary image



(c) Line segmentation image



(d) Character segmentation image

(e) Some touching characters of segmentation

Figure 6. Example Results

Table 2. Accuracy of character segmentation

Manuscript	Accuracy						
	Total components	*Correct segmentation by using contour tracing algorithm*		*Correct segmentation after using tracing background skeleton*		*Difference between before and after using tracing background skeleton*	
		Components	*Rates*	*Components*	*Rates*	*Components*	*Rates*
1	288	234	81.25%	235	81.60%	1	0.35%
2	284	199	70.07%	204	71.83%	5	1.76%
3	298	258	86.58%	260	87.25%	2	0.67%
4	260	233	89.62%	240	92.31%	7	2.69%
5	269	202	75.09%	211	78.44%	9	3.35%
6	273	232	84.98%	234	85.71%	2	0.73%
7	305	248	81.31%	253	82.95%	5	1.64%
8	228	186	81.58%	186	81.58%	0	0.00%
9	270	216	80.00%	219	81.11%	3	1.11%
10	227	187	82.38%	189	83.26%	2	0.88%
Total	2702	2195	81.24%	2231	82.57%	36	1.33%