Segmentation-free speech text recognition for comic books

Christophe Rigaud, Jean-Christophe Burie, Jean-Marc Ogier
Laboratoire L3i
Avenue Michel Crépeau, 17042 La Rochelle
Université de La Rochelle
La Rochelle, France
{christophe.rigaud,jean-christophe.burie,jean-marc.ogier}@univ-lr.fr

ABSTRACT

Speech text in comic books is written in a particular manner by the scriptwriter which raises unusual challenges for text recognition. We first detail these challenges and present different approaches to solve them. We compare the performances of pre-trained OCR and segmentation-free approach for speech text of comic books written in Latin script. We demonstrate that few good quality pre-trained OCR output samples, associated with other unlabeled data with the same writing style, can feed a segmentation-free OCR and improve text recognition. Thanks to the help of the *lexicality* measure that automatically accept or reject the pre-trained OCR output as pseudo ground truth for a subsequent segmentation-free OCR training and recognition.

Keywords

Text recognition; pseudo ground truth; segmentation-free OCR; comic book image analysis.

1. INTRODUCTION

Comic books are part of the cultural heritage of many countries and their massive digitization allows information retrieval. Text is one of the crucial information to retrieve in order to index comic book content. There are different types of text present in comics. We focus on the most frequent one which is speech text. There are other types of text such as title, caption, illustrative and drawing text. This paper highlights previous work about text recognition applied to comics and propose a new segmentation-free text line recognition system. Figure 1 shows some results of segmentation-free OCR output. In the rest of the paper, we use the expression "segmentation-free" to designate a segmentation that is fully automatic with no human action.

Text recognition in comics is really challenging because it includes most of the difficulties from text recognition in document analysis domain. Especially, if we consider high variability of the text types that compose a comics: from typewritten to handwritten, free-form text in uniform to complex background including noisy background, text deformation and overlap.

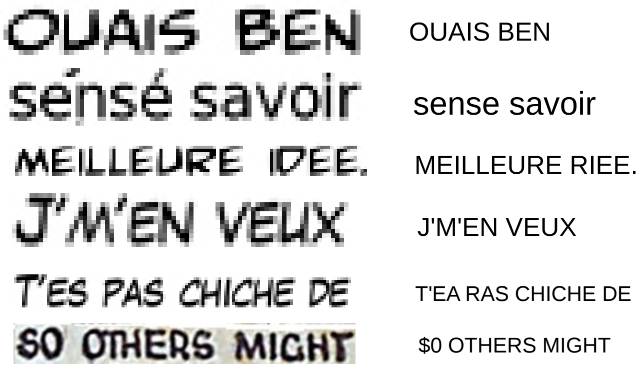


Figure 1: Examples of text line images and associated segmentation-free transcriptions. Considered text lines contain few words and have been digitized between 75 and 300DPI. Image credits: eBDtheque dataset [6].

A desirable research contribution has been made in the field of text recognition considering scanned document pages from printed and handwritten text books. The main published studies on text recognition and identification are discussed below along with recent studies about segmentation-free text recognition techniques.

Automatic text extraction and recognition of speech balloon considering digital English comics was investigated in [11]. In their investigation, a region based text extraction method was applied initially, and furthermore, two sub-approaches: connected component and edge-based techniques were introduced. Connected component labeling-based algorithm was applied to remove noises from color images for detecting the connected components in black and white images. Connected component-based methods were applied by grouping small components into successively larger ones until all regions were identified in the image. In their research, digital color English comic images were taken as input and color band selection was performed. During recognition phase, the process of Optical Character Recognition (OCR) was divided into segmentation, correlation and classification steps. In segmentation step, each text character was cropped and

MANPU '17 Kyoto, Japan
© 2017 ACM. ISBN ???…$???
DOI: ???

in the correlation step, cropped characters were matched with the datasets. In the classification process, text images were recognized considering the matching process of cropped characters with the datasets. Another method using OCR output have been proposed by Ponsard [10]. It is part of an end-to-end process and focuses on speech text of a single French typewritten font, for which a specific OCR system have been trained for.

Recently, a comic text recognition method was investigated [4]. In this work, manga images (Japanese comics) were used for text extraction and text recognition process. A median filtering-based technique was considered in the preprocessing stage for noise removal. A connected component labeling-based algorithm was taken into account for speech balloon detection and subsequently the OCR was used for text recognition within the balloons. The OCR process, described in their study was also divided into: segmentation, correlation and classification steps. Text lines, words and characters were segmented manually before feeding characters into the OCR system. A desirable recognition rate at character level was achieved in the experimentation.

Segmentation-free OCR training has been very well detailed in Ul-Hasan *et* al. [17]. It replaces manual ground truth production by first training a standard OCR system (Tesseract) on a historically reconstructed typeface with subsequent OCR training (OCRopus) on the actual book using Tesseract's output as pseudo ground truth. It has also achieved accuracies above 95% but shifts the transcription effort to the manual (re)construction of the typeface. In this paper, we will have a similar approach but instead of shifting the manual work we will show how to remove it totally.

In the next section, we detail the proposed approach. Section 3 details the experiments we carried out on a public dataset and Section 4 concludes this work.

2. PROPOSED APPROACH

In this section, we detail the proposed segmentation-free approach which means annotation and training are fully automatic (no human work needed). Note that the proposed approach is designed to learn how to recognize text of a specific writing style (handwritten font) that's why it has to be applied to each album (or scriptwriter) separately. The approach consists in three steps after having extracted text lines crops from comic book images. First we apply one or several standard OCR systems to try to recognize all extracted text lines. Then, we check the quality of each recognized text lines for each OCR system using the lexicality measure (see Section 2.4). Note that it requires a lexicon corresponding to the language of the analyzed text. In general, the quality of the recognized text is quite low at this stage because comic book writing styles are quite different from the generic fonts that they have been trained for. In a third step, the recognized text lines are used as input for training another OCR system from scratch. Finally, we end up with a newly trained OCR that we use for recognizing (again) all the extracted text lines from the related album (taking the OCR result of mixed models as pseudo ground truth for the subsequent training), see Figure 2. This new model being specifically trained for a specific writing style, it does provide better automatic text transcription (see Section 3).

2.1 Text localization

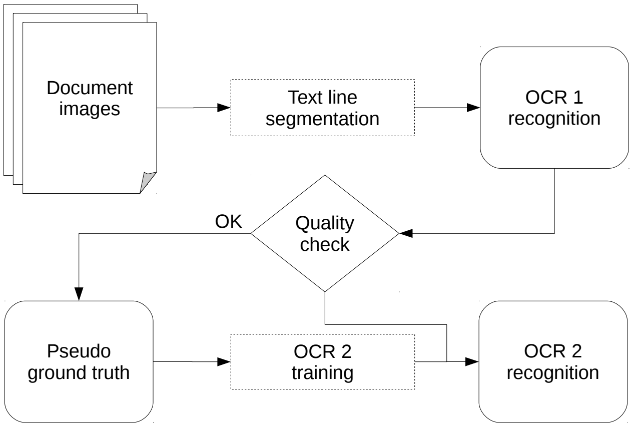


Figure 2: The complete pipeline of the proposed segmentation-free OCR system. First row represents traditional text recognition sequence (single or a bunch of OCRs). In the middle block, text lines are quality checked and the good ones (text line images and associated transcriptions) are used as pseudo ground truth for training the second OCR (OCR 2). Final OCR output are produced by OCR 2 using its automatically trained model.

In this study, we propose to rely on text localization to find speech text from speech balloons. At this stage, any speech text extractor can be used but text recognition performance highly relies on its performance. In this paper, we use a state of the art algorithm that reaches 75.8% recall and 76.2% precision for text line localization on eBDtheque dataset [12].

Note that no post-processing of the extracted text line images has been performed in order to be as close as possible of a real case study for text recognition (e.g. low definition, degraded image, inexact cropping, touching characters).

2.2 Pre-trained OCR systems

In this study, an attempt has been made to explore a segmentation-free recognition technique of comic text images. The proposed approach requires at least one pretrained OCR system able to recognize at least some text lines with a good accuracy in order to feed the learning stage of a subsequent OCR system. The second OCR is expected to recognize much more text lines than the first OCR.

Tesseract and FineReader are the two most popular OCR systems presently available. Tesseract, is considered to be one of the most accurate free open source OCR engines [15]. It is an open-source software that can be easily integrated into research experiments, that's why we chose to use it. FineReader OCR engine is a well-known commercial OCR system. We did not use it in this experiment but it can be added easily as any other pre-trained OCR. Traditional OCR systems require annotated data at the level of letter which is inappropriate for cursive fonts [7]. A more recent alternative has been proposed to cope with this issue [1]. This efficient algorithm called OCRopus is based on Long Short Term Memory neural networks (LSTM) that has proven its efficiency for handwritten text recognition [2, 9, 16, 14].

Note that the number of pre-trained OCR is not limited, a bunch of pre-trained OCR systems can be used and the one which gives the best result has to be chosen for each text line in order to feed the second OCR system with as much as correct text lines as possible (improves training and recognition quality).

2.3 Segmentation-free training

As introduced in Section 1, handwritten text is very challenging for OCR systems. They require a lot of annotated single letters of each font to train their optical model in order to be able to recognize text. In fact, it is not really feasible to annotate all scriptwriter styles as they are continuously trying to publish comics that are different from others (new authors are making comics everyday). Instead of annotating a huge amount of handwritten styles and try to build a generic handwritten OCR system, we propose to automatically train a specific OCR for each writing style from a single scriptwriter (person who writes text in speech balloons). This approach has the advantage of minimizing confusions between visual similarities (e.g. letter “i” from scriptwriter A may be similar to letter “l” from another scriptwriter B).

The idea is to use, for each writing style, only good pre-trained OCR output to train OCRopus algorithm and then recognize all text with same writing style using OCRopus with its newly trained LSTM model.

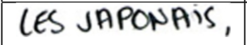
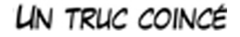
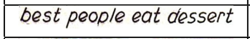
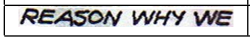
This approach removes image annotation time (ground-truthing) but may introduce false positive and false negative text lines in the recognition step. False negatives are not important in our study because they will not decrease the quality of the ground truth, they just ignore some text lines from the story. However, false positives may bias the ground truth so they must not be used as pseudo ground truth. In order to detect and ignore false positives, we evaluate the recognition quality using a lexicality measure.

2.4 The lexicality measure

In order to evaluate the recognition quality of each pre-trained OCR output in the absence of any ground truth, we use a readily observable quantity that correlates well with true accuracy. It gives us a ranking scale which allows to automatically select the transcription given by the (nearly) best performing model. We chose to use the *mean token lexicality L*, a distance measure of OCR tokens (words) which can be calculated for any OCR engine [16]. The original paper also mention the *mean character confidence C* but because it is OCR specifics we do not use it. The lexicality measure calculates, for each OCR token, the minimum edit distance (Levenshtein distance) to its most probable lexical equivalent (from a lexicon of the corresponding language). The sum of these Levenshtein distances over all tokens is therefore a (statistical) measure for the OCR errors of the text, and the lexicality defined as *L = (1 - mean Levenshtein distance per character)* is a measure for accuracy. Problems with this measure arise from lexical gaps (mostly proper names) and very garbled tokens (i.e. short text lines such as sequences of single letters, or too long because of merged tokens with unrecognized whitespace). These issues are not restrictive in our case because we don't need such garbled tokens to build a good pseudo ground truth, they could even bias it if they are not so frequent.

Examples of lexicality measure are given in Table 1. In the first row, the image where it is written “LES JAPONAIS” has been mistakenly transcribed with the 9 following characters “Les mmmsc,” (ignoring spaces). The first 3 characters composing the word “Les” is part of the French lexicon (Levenshtein distance equal to 0) but the second word is not a French word. The “closest” word from the lexicon is “immiscé” which is at a Levenshtein distance of 3 from “mmmsc,” (number of characters that are different or at a different place). The lexicality $L = 1 - (0 + 3/9) = 0.67$.

Table 1: Examples of different lexicality measure related to the percentage of recognized words that are part of the lexicon. Image credits: eBDtheque dataset [6] and public domain.

Image	OCR output	Lex.
 LES JAPONAIS,	Les mmmsc,	67%
 UN TRUC COINCÉ	UN TRUC COINŒ	91%
 best people eat dessert	best” people eat dessert	95%
 REASON WHY WE	REASON WHY WE	100%

3. PERFORMANCE EVALUATION

In this section we compare speech text recognition quality between pre-trained and segmentation-free OCR systems. The idea is to highlight the limitations of standard pre-trained OCR systems at recognizing speech text in comics. Also to measure the benefit of feeding another OCR with such output as pseudo ground truth for training.

We evaluate the proposed method using human-made ground truth for the testing set and automatically generated ground truth for the training set. This is similar to a real case study where only few human power (or annotated images) is available and a huge amount of training data is necessary. It also demonstrates the capability of the method to learn and generalize from inexact ground truth.

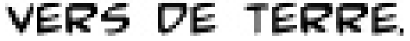
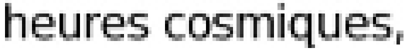
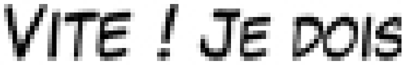
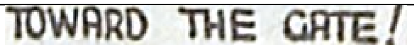
3.1 Metrics

We rely on standard metrics of speech recognition such as Character Error Rate (CER), Word Error Rate (WER) and Sentence Error Rate (SER) for determining the recognition accuracy of the segmentation free OCR output [8, 5]. Note that we measure the ability of pre-trained OCR systems to generate pseudo ground truth only by counting the number of validated text lines they produce. Because in our case, the training set is not annotated.

3.2 Dataset

We selected some manually annotated comic book images from the eBDtheque dataset [6] to build the test set and other collected images from the same albums to build the training set (images were collected from other sources like Internet, public library, private collection). For example, given an album A, we use all the annotated images of this album from the eBDtheque dataset as testing set and collect other unannotated images of this album from other sources to build the training set. We selected the eBDtheque dataset because it provides text transcription for all images and from diverse writing styles (more representative than Manga109 dataset [3]). This dataset is composed by one hundred images containing 4691 annotated text lines. It has images scanned from French comic books (46%), French webcomics (37%) with various formats and definitions, public domain

Table 2: Image examples for each tested album. Transcriptions are written between quotes below text line images for the ground truth (GT) and the two OCR system outputs (Tesseract and OCRopus), with corresponding Character Error Rate (CER). Original image sizes in pixels from top to bottom are: 113x10, 118x14, 92x15 and 237x23. Image credits: eBDtheque dataset [6].

OCR/im.	Image/transcription	CER
Album 1		
GT	“VERS DE TERRE”	
Tess.	“vgrs as r”	0.84
OCRop.	“VERX DE TERRE”	0.07
Album 2		
GT	“heures cosmiques,”	
Tess.	“heures cusmuques,”	0.12
OCRop.	“neures cosmlques.”	0.18
Album 3		
GT	“VITE ! JE DOIS”	
Tess.	“VITE ! Je POIS”	0.14
OCRop.	“VITE ! E DOIS”	0.07
Album 4		
GT	“TOWARD THE GATE!”	
Tess.	“TDWRRD THE GRTE.âĂŹ”	0.31
OCRop.	“TOWTRD THE CUT !”	0.25

American comics (11%) and unpublished artwork of manga (6%).

From eBDtheque dataset, we selected all available pages for three albums in French and one album in English to illustrate the performance and the language independence of the proposed approach. Album 1 (CYB BUBBLEGOM T01) is a gray scale web comics and is written with well separated uppercase letters. Album 2 (CYB COSMOZONE) is colorful and also typewritten and lowercase with some touching letters due to low definition. Album 3 (LAMISSEB ETPISTAF) is typewritten with a comic-like font in uppercase and slightly tilted letters. Album 4 (MCCALL ROBINHOOD T31) is handwritten and uppercase with a degraded yellowish background (golden age American comics). Note that album 1 to 3 have been digitized between 75 and 150DPI and album 4 in 300DPI. An example of text line from each album is given Table 2.

The exhaustive list of album, page number, generated pseudo ground truth and learned models are available here[1].

3.3 Pre-trained OCR

In order to generate the pseudo ground truth for the segmentation-free OCR, we used only one pre-trained OCR (Tesseract version 3.04) to ease the comparison. However, several OCR systems can be used in parallel and only the best output should be used as pseudo ground truth, as detailed in Section 2.

3.4 Lexicons

We selected complete lexicons for each language containing the list of flexed forms for measuring the *lexicality* of each text line. For French, we used the “Dicollecte lexicon” version 6.1 which is also used in LibreOffice and Firefox. For English, we used the “Dallas lexicon” from the SIL International Linguistics Department which contain inflected forms, such as plural nouns and the -s, -ed and -ing forms of verbs. Both lexicons contain about 500,000 and 110,000 entries respectively.

3.5 Results

In this section we compare the performance of the pre-trained OCR system itself and when its output is redirected to a segmentation-free OCR as presented in Section 2.

For the first OCR, we did not manually re-train Tesseract on the font (writing style) used in the image but instead, we used its pre-trained data for French and English. For the second OCR, a new model for OCRopus was automatically trained from scratch with valid text lines from the first OCR (*L = 100%*).

For training an individual model, we used the set of valid text lines as training set. The resulting model was saved every 10,000 learning steps until 50,000. Each step consists in comparing one text line image and its associated ground truth from the test set[2]. We recommend to start by training OCRopus only if at least 100 valid text lines have been recognized by the pre-trained OCRs. This minimum value has been experimented in paper [13].

After training was completed, the model with the best accuracy was chosen for later recognition tasks. We observed that 10,000 iterations were usually sufficient to train such system but we suggest to iterate over 50,000 or even 100,000 iterations in order to be sure do not be affected by the lack of training issue.

Table 3 shows the results of the experiments with the pre-trained OCR Tesseract alone and then post processed by OCRopus (segmentation-free). The number of pages for testing and training depends on one hand, of the number of pages available in eBDtheque dataset (test set) and, on the other hand, how many pages have been found from different sources about the same album (train set). Note that the train sets is 4 to 10 times bigger than the test set and can be easily extended because it does not require any manual annotation.

The number of extracted text lines is the output of the text line extraction algorithm presented in Section 2.1. The number of validated text lines is a subset of the extracted text lines that have a lexicality measure *L= 100%* (see Section 2.4).

The results of CER, WER and SER of the both OCR systems (pre-trained and segmentation-free) are computed on the annotated images from the test set. The segmentation-free OCR is automatically and only trained on validated text line images.

The results show that the segmentation-free OCR always improves the results from pre-trained OCR at character level (CER) and most of the time also at word (WER) and sentence level (SER). This means that the quality of the validated text line from pre-trained OCR are good enough to

[1] https://github.com/crigaud/publication/tree/master/2017/MANPU/segmentation-free_speech_text_recognition_for_comic_books

[2] http://www.danvk.org/2015/01/11/training-an-ocropus-ocr-model.html

be used as pseudo ground truth for training automatically a new model that outperforms the initial OCR results.

The error rate of CER, WER and SER are quite high compared to usual results from the literature. This is partially due to some pseudo ground truth errors introduced in the automatic text line validation process. The error rates can also be reduced by using images with higher definition (OCR systems usually recommend 300DPI). Post processing text extraction output (i.e. remove surrounding letter parts, separate letters from image border) could also generate better quality pseudo ground truth.

4. CONCLUSIONS

In this paper we measured the ability of standard pre-trained OCR systems to generate pseudo ground truth in order to automatically train a second OCR from scratch. Both systems have been tested on several Latin script albums from the public eBDtheque dataset. The result analysis shows that pre-trained OCR systems can boost specific OCR training if the quality of the generated pseudo ground truth is very high. We measured the influence of the amount of training data to train such OCR system. This amount is related to the number of writing styles to recognize and to its level of difficulty (e.g. uppercase only, mixed, cursive). In the future, we would like to extend this approach to all albums from eBDtheque dataset. This means we need to collect a lot of unannotated images for each album of this dataset. Then, we would like to reuse previously learned models as a bunch of pre-trained OCRs in order to increase the number of valid text lines, increase the quality of the pseudo ground truth and, hopefully, reduce the error rates.

5. ACKNOWLEDGMENTS

This work is supported by the University of La Rochelle (France), the town of La Rochelle and the PIA-iiBD ("Programme d'Investissements d'Avenir"). We are grateful to all authors and publishers of comics images from eBDtheque dataset for allowing us to use and share their works.

6. REFERENCES

[1] T. M. Breuel. The ocropus open source ocr system. In *Proc. SPIE 6815, Document Recognition and Retrieval XV*, pages 68150F–15, 2008.

[2] T. M. Breuel, A. Ul-Hasan, M. A. Al-Azawi, and F. Shafait. High-performance ocr for printed english and fraktur using lstm networks. In *2013 12th International Conference on Document Analysis and Recognition*, pages 683–687. IEEE, 2013.

[3] A. Fujimoto, T. Ogawa, K. Yamamoto, Y. Matsui, T. Yamasaki, and K. Aizawa. Manga109 dataset and creation of metadata. In *Proceedings of the 1st International Workshop on coMics ANalysis, Processing and Understanding*, MANPU '16, pages 2:1–2:5, New York, NY, USA, 2016. ACM.

[4] M. R. Gaikwad and N. Pardeshi. Text extraction and recognition using median filter. *International Research Journal of Engineering and Technology*, 3(1):717–721, 2016.

[5] M. J. F. Gales, X. Liu, R. Sinha, P. C. Woodland, K. Yu, S. Matsoukas, T. Ng, K. Nguyen, L. Nguyen, J. L. Gauvain, L. Lamel, and A. Messaoudi. Speech recognition system combination for machine translation. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, volume 4, pages IV–1277–IV–1280, April 2007.

[6] C. Guérin, C. Rigaud, A. Mercier, and al. ebdtheque: a representative database of comics. In *Proceedings of International Conference on Document Analysis and Recognition (ICDAR)*, pages 1145–1149, Washington DC, 2013.

[7] M. Heliński, M. Kmieciak, and T. Parkoła. Report on the comparison of tesseract and abbyy finereader ocr engines. 2012.

[8] H. Helmke, H. Ehr, M. Kleinert, F. Faubel, and D. Klakow. Increased acceptance of controller assistance by automatic speech recognition. In *Tenth USA/Europe Air Traffic Management Research and Development Seminar (ATM2013)*, pages 1–10, June 2013.

[9] M. Jenckel, S. S. Bukhari, and A. Dengel. anyocr: A sequence learning based ocr system for unlabeled historical documents. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 4035–4040, Dec 2016.

[10] C. Ponsard, R. Ramdoyal, and D. Dziamski. An ocr-enabled digital comic books viewer. In *Computers Helping People with Special Needs*, pages 471–478. Springer, 2012.

[11] S. Ranjini and M.Sundaresan. Extraction and recognition of text from digital english comic image using median filter. *International Journal on Computer Science and Engineering (IJCSE)*, 5(4):238–244, April 2013.

[12] C. Rigaud, D. Karatzas, J. Van de Weijer, J.-C. Burie, and J.-M. Ogier. Automatic text localisation in scanned comic books. In *9th International Conference on Computer Vision Theory and Applications*, 2013.

[13] C. Rigaud, S. Pal, J.-C. Burie, and J.-M. Ogier. Toward speech text recognition for comic books. In *Proceedings of the 1st International Workshop on coMics ANalysis, Processing and Understanding*, MANPU '16, pages 8:1–8:6, New York, NY, USA, 2016. ACM.

[14] F. Simistira, A. Ul-Hassan, V. Papavassiliou, B. Gatos, V. Katsouros, and M. Liwicki. Recognition of historical greek polytonic scripts using lstm networks. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 766–770, Aug 2015.

[15] R. Smith. An overview of the tesseract ocr engine. In *Proc. Ninth Int. Conference on Document Analysis and Recognition (ICDAR)*, pages 629–633, 2007.

[16] U. Springmann, F. Fink, and K. U. Schulz. Automatic quality evaluation and (semi-) automatic improvement of mixed models for OCR on historical documents. *CoRR*, abs/1606.05157, 2016.

[17] A. Ul-Hasan, S. S. Bukhari, and A. Dengel. Ocroract: A sequence learning ocr system trained on isolated characters. In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 174–179, April 2016.

Table 3: Pre-trained and segmentation-free OCR results.

Album	Number of pages		Number of text lines		Pre-trained OCR (Tesseract)			Segmentation-free OCR (OCRopus)		
	Test	Train	Extracted	Validated	CER	WER	SER	CER	WER	SER
1	10	41	432	230	82.13	94.97	99.57	**35.61**	**65.27**	**82.77**
2	5	89	804	428	33.26	70.82	92.92	**28.34**	**63.16**	**91.2**
3	5	51	183	123	59.19	80.18	94.44	**23.48**	**41.79**	**76.92**
4	4	42	354	182	51.40	**78.77**	**97.52**	**46.42**	83.63	98.34