A Bilingual Gurmukhi-English OCR Based on Multiple Script Identifiers and Language Models

Gurpreet Singh Lehal
Department of Computer Science,
Punjabi University Patiala, Punjab, India
gslehal@gmail.com

ABSTRACT

English words are frequently encountered in Gurmukhi texts. A monolingual Gurmukhi OCR will recognize such words as garbage. It becomes necessary to add bilingual capability to the Gurmukhi OCR to recognize English text too. But adding bilingual capability reduces the recognition accuracy for monolingual texts due to errors in script identification. Even a system with 99% script identification accuracy results in reduction of 1% recognition accuracy on monolingual text. In this paper, we present a bilingual OCR, which recognizes both English and Gurmukhi scripts without any significant reduction in recognition accuracy as compared to the monolingual Gurmukhi OCR when recognizing monolingual Gurmukhi text. This is achieved by using multiple script identification engines and language models for both English and Gurmukhi scripts. For the first time, such a system has been developed, which recognizes with high accuracy document images containing mixed Gurmukhi and English text or only Gurmukhi/English text.

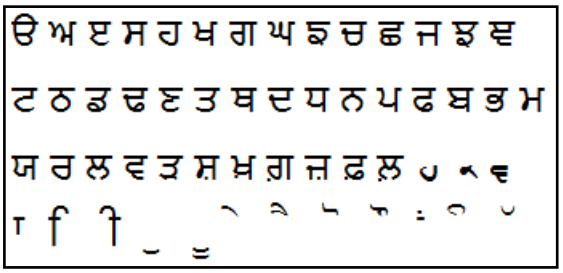
1. INTRODUCTION

English words are frequently encountered in Gurmukhi texts. A monolingual Gurmukhi OCR will recognize such words as garbage. It becomes necessary to add bilingual capability to the Gurmukhi OCR to recognize English text too. However, the performance of multilingual OCR system is lower than that of single language OCR system, as word segmentation and script discrimination errors are introduced when multilingual mixed document are processed. So special care has to be taken that adding the bilingual capability does not degrade the performance of OCR for monolingual text. In this paper, we present a bilingual Gurmukhi-English OCR system, which recognizes bilingual Gurmukhi/English texts as well as monolingual Gurmukhi and English with a fairly high accuracy.

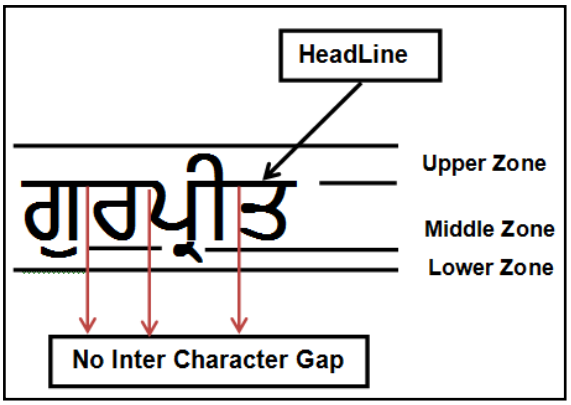
Two types of approaches are followed in the development of a multi-script OCR. In the first approach, script identification is done at word level and this information is used to invoke the corresponding OCR developed for that particular script[1-11]. In the other combined database approach, characters from all the participating scripts are treated identically irrespective of their scripts. But in this method the search space in the database increases as it contains alphabets from all the scripts and the overall accuracy of the system goes down. In our current work, we have followed the former approach.

2. PROPERTIES OF GURMUKHI SCRIPT

Gurmukhi script is used primarily for Punjabi language, which is the world's 12th most widely spoken language. The populace speaking Punjabi is not only confined to North Indian states but is spread all over the world. Gurmukhi script is cursive and the Gurmukhi script alphabet consists of 41 consonants/vowel carriers, 3 half characters, 10 vowels, and 3 special symbols (Fig. 1a). Most of the characters have a horizontal line at the upper part. The characters of words are connected mostly by this line called headline and there is usually no vertical inter-character gap in the letters of a word. A word in Gurmukhi script can be partitioned into three horizontal zones (Fig 1b). The upper zone denotes the region above the head line, where vowels reside, while the middle zone represents the area below the head line where the consonants and some sub-parts of vowels are present. The lower zone represents the area below middle zone where some vowels and certain half characters lie in the foot of consonants. The bounding boxes of 2 or more characters in a word may intersect or overlap vertically.



a)

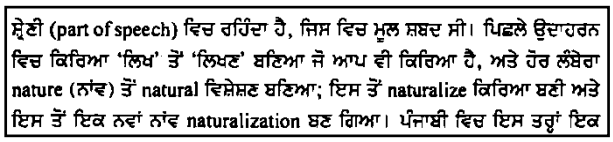


b)

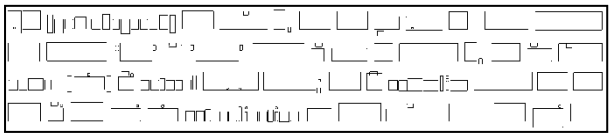
Figure 1. a) Basic Gurmukhi character set b)A Gurmukhi word image

3. SYSTEM ARCHITECTURE

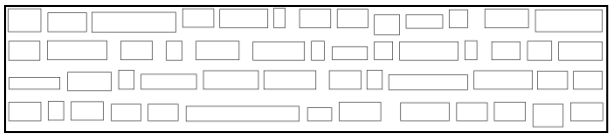
The document image (Fig 2a) is first subjected to pre-processing to remove the background noise and skewness. The image is then segmented into individual lines using horizontal projection profile and the connected components are extracted from each line. The bounding boxes of the connected components could represent Gurmukhi or English characters/words and punctuation marks (Fig 2b). A statistical analysis of vertical gaps in connected components is performed to determine inter word and inter character gaps. Then in each line, all the adjacent connected components, whose vertical gaps are lesser than the inter word gap are merged (Fig. 2c). The merged connected components represent word images and these images are sent to script identification and character recognition engines for conversion to text. For recognizing the Gurmukhi word images, we have used the multi classifier based Gurmukhi recognition engine[12], while the Tesseract OCR engine[13] is used for recognition of English word images.



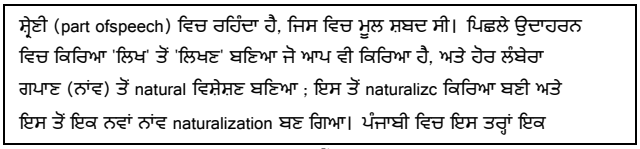
a)



b)



c)



d)

Figure 2. a) A bilingual document image b) Bounding boxes of connected components c) Bounding boxes of connected components after merging d) Recognized text by Approach 3 (discussed later)

4. SCRIPT IDENTIFICATION ENGINES

The script identification module has to be highly accurate, as words with wrongly identified scripts will be recognized as junk. To increase the script identification accuracy, we have experimented with two script identification engines. The first engine (S1) uses statistical features while the second engine (S2) uses structural features to identify the script. The S1 engine uses Gabor Filters and Support Vector Machine for classification. The word image is normalized to 32x32 pixels and partitioned into four equal non overlapping subregions of size 16x16. These subregions are further partitioned into four equal non overlapping sub-subregions of size 8x8 and thus we obtain 16 small regions in different parts of the image. These 21 images are convolved with odd symmetric and even symmetric Gabor filters in nine different angles of orientation of 20 degrees, to obtain a feature vector of 189 values. In earlier experiments it has been reported that similar features and classifiers identify the script with more than 99% accuracy[10]. But those experiments were carried on isolated words. When we applied the similar features on real life data, then due to word segmentation errors, noise and poor quality of text, the actual script identification accuracy dropped to 98.06% for Gurmukhi words and 97.08% for English words. As majority or all of the text will be in Gurmukhi script and our aim is to add the bilingual capability without reducing the recognition accuracy for Gurmukhi text, so one disadvantage of this approach is that if the text contains only Gurmukhi words or only few English words then the overall Gurmukhi OCR accuracy is reduced by 2% due to script identification errors. To reduce the script identification error, we added an additional script identification engine to the system.

The second script identification engine (S2) uses the following structural features to discriminate between Gurmukhi and English characters/words:

a) *Headline pixel count:* Most of the Gurmukhi characters and words have a headline in upper half of the image which is usually missing in English words (Fig. 1).

b) *Inter character gap:* In a Gurmukhi word, the characters are glued along the headline (Fig. 2). Thus if there is no inter character gap then it is highly probable that the word is in Gurmukhi.

c) *Right Vertical Bar:* Many Gurmukhi characters (ਅ ਸ ਹ ਖ ਗ ਘ ਚ ਜ ਥ ਧ ਪ ਬ ਮ ਰ etc.) have a vertical bar at the right extreme, while few Roman alphabets have such bar.

d) *Regions Beyond Headline Projection:* In Gurmukhi there are very few characters/words, which have any regions beyond the headline projection, while many English characters (q u i o a d g h l b n m etc.) have this property.

e) *Black-White Transactions:* Count of columns which contain exactly one black-white transaction, along the row corresponding to the headline is also helpful in distinguishing between Gurmukhi and Roman words.

It was found that the script identification accuracy of S2 is significantly lesser than S1. From experiments we found that S2 identifies the script of Gurmukhi words with 97.14% accuracy but recognition accuracy of Roman words is only 90.25%. Even though the recognition accuracy of S2 is lesser than S1, but if we combine the results of the two classifiers we get better accuracy for Gurmukhi script, which is our main concern.

We observed that even though 1.94% of Gurmukhi words were wrongly recognized as English words by S1 engine and 2.86% of Gurmukhi words were wrongly recognized as English words by S2 engine, but there were 0.74% of cases where both the engines wrongly recognized the Gurmukhi script. So if both S1 and S2 determine the script to be Gurmukhi or Roman then we

abide by their decision. But in case of split decision we take the script as Gurmukhi. This reduces the script identification error for Gurmukhi. If by mistake a Gurmukhi word is recognized as English word by S1, but if S2 identifies it as Gurmukhi then the script is taken as Gurmukhi. In Fig. 3, we have some samples of Gurmukhi words whose script was wrongly identified as Roman by S1 but correctly identified as Gurmukhi by S2 and we finalise the script as Gurmukhi. One disadvantage of this approach is that it is heavily biased against Roman script and if by mistake one of the classifier classifies the script of an English word as Gurmukhi, it is taken as Gurmukhi (Fig. 4), but practically that does not make much difference as majority of text is in Gurmukhi.

Even though the addition of another script identification engine improves the script identification, but still we see that in 0.74% of cases the script of Gurmukhi word is wrongly recognized by both the engines. This is usually in case of small sized words, broken words, words with punctuation marks such as hyphen or words containing characters with missing headlines. So still the overall recognition accuracy goes down by 0.74% on texts containing only Gurmukhi words or a few English words. So it is important to further reduce the script identification errors especially for Gurmukhi as most of the text will be in Gurmukhi only. For this purpose, we make use of the linguistic resources along with recognition engines for Gurmukhi and English to develop a rule based bilingual recognition engine, as discussed in next section.

Figure 3. Samples of words wrongly recognized as Roman by S1 but correctly recognized by S2

Figure 4. Samples of words wrongly recognized as Gurmukhi by either S1 or S2

5. RULE BASED BILINGUAL RECOGNITION ENGINE

The rule based bilingual recognition engine uses the results produced by script identification engines and recognition engines and validates them with the language models.

The following language models and linguistic resources have been used by the engine:

1. *Character level trigram language model for English:* A character trigram is a set of 3 consecutive characters extracted from a word. We have analysed an English corpus of 10 million words and generated the probabilities of all the character trigrams occurring in the text. Trigrams such as 'xxz', 'yzz', 'qvk' etc. not found in the text are assigned zero probability.
2. *Character level trigram language model for Gurmukhi:* Similarly a Gurmukhi corpus of 12 million words was analysed to generate the probabilities of Gurmukhi character trigrams.
3. Ten thousand most frequent words of English (*Elist*) **:** The ten thousand most frequently occurring words are collected from English corpus.
4. Ten thousand most frequent words of Gurmukhi (*Glist*) : The ten thousand most frequently occurring words are collected from Gurmukhi corpus.

The traditional approach followed by most of the researchers is to first determine the script of a word and then recognize the text. We have experimented with a slightly different approach and run both the script identification and word recognition engines in parallel. The word image is fed to the four engines in parallel and their outputs are combined with the language models to finally recognize the text. The system architecture of our bilingual recognition system is shown in Fig. 5. The bilingual recognition engine gets inputs from the four engines and uses the language models and word frequency lists to output the final recognized word. s1 and s2 are the scripts identified by the two script identification engines, while w1 and w2 are the recognized English and Gurmukhi words. The main purpose of the language models is check if the recognized word is valid by using the trigram language models. The recognized word is split into trigrams and if any trigram with zero probability is found the word is considered as invalid. As for example, consider the word 'hajx', the word has six trigrams **h, *ha, haj, ajx, jx* and x***. We find that the trigram *ajx* has zero probability of occurrence and so the word is considered to be invalid. In addition, the word frequency lists, *Elist* and *Glist*, are used to check for high frequency words.

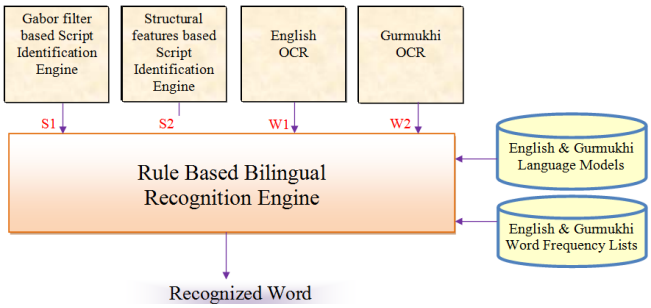


Figure 5. System architecture of Bilingual Gurmukhi/ English OCR

The rule based bilingual recognition engine recognizes the Gurmukhi/English word image using following rules:

Rule 1 : If both s1 and s2 are Gurmukhi, then output w2.

Rule 2 : If both s1 and s2 are Roman, then if w2 $\in$ Glist and w1 $\notin$ EList, then output w2 else output w1. This rule is needed to take care of the condition, where both the script identification engines wrongly recognize Gurmukhi word image as English. As majority of text is in Gurmukhi, so to make sure that Gurmukhi image is not identified as English, we use the frequency lists to check the recognized word in English and Gurmukhi word frequency lists.

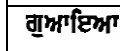
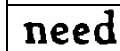
Rule 3 : If s1 and s2 are distinct, then use the language models sequentially as follows.

- If w2 is a valid Gurmukhi word (all the trigrams have non zero probability), output w2.
- If w1 is a valid English word (all the trigrams have non zero probability), output w1.
- If the recognized word is invalid in both scripts, then as Gurmukhi as majority of words are in Gurmukhi so output w2.

We see in Table 1, some sample images and how they are recognized by our rule based bilingual recognition engine. The script identification results returned by script identification engines are shown under columns s1 and s2, while w1 and w2 are the words recognized by English and Gurmukhi recognition

engines respectively. As seen in Table 1, the script of the first word is recognized as Gurmukhi by both engines and so the word is recognized as ਮੂਲ. The second word's script is identified as Roman by both engine, but as w2(ਪੱਖ) ∈ Glist and w1(aiu) ∉ EList, so the final recognized word is script is ਪੱਖ. There is a split verdict for third and fourth words by the script identification engines. As w2 is valid Gurmukhi word so the final recognized word is w2. For the fourth word, the trigram combination 'ਦਦਧ' in w2 has zero probability and so w1 is selected as recognized word, since w1 is a valid word. For the fifth word, the opening bracket is lying close to the word image and it is considered as part of word image and sent for script identification. Both the engines recognize the script as Roman. The Gurmukhi word, ਇਹਗਂ, recognized by Gurmukhi recognition engine ∉ GList and so the script is finalised as Roman. The last word is an example of failure case. The two script identification engines disagree on the script and the Gurmukhi recognized word turns out to be a valid word and so the script is taken as Gurmukhi. It is worth noting that the English word is also valid word, but as mentioned earlier, we are giving more weightage to Gurmukhi words so the English word is ignored.

Table I. Word Recognition by Rule based bilingual recognition engine

Image	S1	S2	W1	W2	Recognised word
	Gurmukhi	Gurmukhi	Ml)	ਮੂਲ	ਮੂਲ
	Roman	Roman	aiu	ਪੱਖ	ਪੱਖ
	Roman	Gurmukhi	aMrfaMr	ਗੁਆਇਆ	ਗੁਆਇਆ
	Roman	Gurmukhi	need	ਦਦਧਇ	need
	Roman	Roman	(part	ਇਹਗਂ	(part
	Roman	Gurmukhi	nature	ਗਪਾਣ	ਗਪਾਣ

The image of Fig 2a was recognized by the above bilingual recognition engine and the output is in Fig 2d. The last two sample images of Table 1 have been taken from image of Fig. 2a. We can see in Fig. 2d, the Gurmukhi text is recognized without any errors, but for English text we have two errors. In first sentence there is a word segmentation error and two English words are joined together, while the script of the English word in third sentence is wrongly recognized as discussed above.

6. Experimental Results

We performed four sets of experiments. For the first set(Set1), 105 images randomly taken from 7 Gurmukhi books were used. The pages contain 145, 373 characters. Out of these 105 pages, 76 pages contained only Gurmukhi words while 29 pages had a mixture of Gurmukhi and English words, though English words were in minority. Overall in these 105 pages 1.16% of characters are in Roman script while rest are in Gurmukhi script. The second set (Set2) is made up of 29 images from Set1 containing at least one English word. The purpose of this set is to see the performance in the presence of English words. In these 29 pages 96.04% percent of text is in Gurmukhi and rest in English. The third set (Set3) is made up of images containing only Gurmukhi text. These are the 76 images taken from Set1, which do not contain any English word. The fourth set(Set4), comprises of 10 images containing only English words. The fourth set is to test the performance of the bilingual OCR on English images, though practically, we shall not be using the bilingual OCR for recognizing image containing only English text.

We first executed the monolingual Gurmukhi OCR on these sets and then added the bilingual capability to it. We also tested with the traditional approach being followed by researchers, where first the script is identified and then the word is sent to appropriate OCR engine. In the first approach we have used the Gabor filter based script identification engine. In second approach we have used both Gabor filter based and structural feature based script identification engines as discussed in previous section. Finally the third approach, which uses both script identification engines and language models has been tested. The results are displayed in Table 2.

Table II. Character level Recognition Accuracy for different approaches

Experiment	Monolingual Gurmukhi OCR	Approach 1	Approach 2	Approach 3
Set1	95.80%	95.01%	96.05%	**96.86%**
Set2	93.71%	94.70%	95.58%	**96.70%**
Set3	**97.68%**	95.86%	96.96%	97.64%
Set4	--	**93.55%**	82.34%	88.11%

As we can see for Set1, which is a representative collection of Gurmukhi pages taken from 7 different books published over different periods of time, the recognition accuracy for approach 1 actually goes down from 95.80% for monolingual Gurmukhi OCR to 95.01%. The reason is obvious as the error rate of S1, the script identification module used, is 2% and share of English text is only 1.16%, so the error rate increases. Thus the usual methodology of determining the script of word image using a single script classifier such as Gabor filter and passing the image to appropriate recognition engine, does not work well in our case. The addition of second script classifier helps in raising the accuracy by 1.04% and the overall recognition accuracy is 0.25% more than monolingual OCR. But the only substantial gain comes in approach 3, where the accuracy increases by 1.06%.

Set2 is collection of images containing at least one English word, and with 3.96% of text being in English we have an increase in accuracy in all the approaches as compared to the monolingual OCR. But again as we see, approach 3 outperforms other approaches and we have a net gain of around 3% as compared to the monolingual OCR.

The next experiment on Set3 was done to see how much the performance of monolingual Gurmukhi OCR degrades when we add the bilingual capability to it as Set3 contains

images with only Gurmukhi text. As we can see there is only a marginal reduction of 0.04% in approach 3, while for other approaches the accuracy reduction is substantial. Thus we can observe from the experiments on these three sets that we can use approach 3 both for monolingual and bilingual texts without any performance degradation.

Experiments on Set4 are conducted to see how the OCR performs for different approaches when the text contains only English words. By default, our system treats English document as Gurmukhi document only and applies same line and word segmentation techniques. Usually in Gurmukhi there is no inter character gap, while in English there is vertical gap both between characters and words. This leads to word segmentation errors for English words. The recognition accuracy is best for approach 1, as the script identification engine S1 identifies English script with 97% accuracy. The recognition errors from English OCR and word segmentation errors contribute further and the text is recognized with 92.55% accuracy. Approach 2 is heavily based against English script and identifies word image as English only when both classifiers identify it as English. From experiments we found that in 87.5% of cases both the classifiers identified the text as English and rest were treated as Gurmukhi and thus the overall recognition is 82.34% for approach2. For approach 3, we have slight improvement in recognition (87.11%) as we use the language models to filter out invalid Gurmukhi words.

7. CONCLUSION

In this paper, we have presented a scheme to add the bilingual capability to Gurmukhi OCR, without compromising on the recognition accuracy for monolingual text, so that the same OCR can be used to recognize both monolingual and bilingual images without any significant loss in recognition accuracy. We have experimented with three different approaches for script classification. The traditional approach of first identifying the script of the word image and then sending it for recognition to appropriate OCR actually results in increasing the error rate in recognition of randomly taken monolingual and bilingual pages, if there are only few English words. The addition of second script identification engine to the system slightly increased the overall recognition accuracy for bilingual documents but for monolingual Gurmukhi documents the recognition accuracy was still 0.72% lower than monolingual Gurmukhi OCR. The addition of linguistics resources such as trigram language models and word frequency lists helped in further development of a robust and high accuracy bilingual recognition system. The system performs very well for all type of documents. For bilingual documents, there is a gain of 2.99% character recognition accuracy in our experiments. Even for monolingual Gurmukhi documents the recognition accuracy is only 0.04% lower than the monolingual Gurmukhi OCR. The English text images were recognized with 88.11% accuracy. This is the first time, such a system has been developed, which recognizes all types of bilingual Gurmukhi and English documents as well as monolingual documents with decent accuracy.

8. ACKNOWLEDGMENT

This research work is sponsored by Ministry of Communications and Information Technology under the project : Development of Robust Document Analysis and Recognition System for Printed Indian Scripts**.**

9. REFERENCES

[1] Kunte, R. S., and Samuel, R. S., "A Bilingual Machine-Interface OCR for Printed Kannada and English Text Employing Wavelet Features," In Proceedings of 10th International Conference on Information Technology, (ICIT 2007). pp. 202-207 2007.

[2] Rezaee, H., Geravanchizadeh, M. and Razzazi, F., "Automatic language identification of bilingual English and Farsi scripts," In Application of Information and Communication Technologies, 2009. AICT 2009. International Conference on (pp. 1-4). 2009..

[3] Chanda,S., Terrades, O. R. and Pal, U., "SVM based scheme for Thai and English script identification," In Ninth International Conference on Document Analysis and Recognition, 2007. ICDAR 2007. pp. 551-555. 2007

[4] Haboubi, S. Maddouri, S.S. and Amiri, H., "Discrimination between Arabic and Latin from bilingual documents," In International Conference on Communications, Computing and Control Applications (CCCA), 2011, pp. 1-6. 2011.

[5] Joshi,G.. Garg,S. and Sivaswamy, J., "Script identification from Indian documents," Document Analysis Systems VII, pp. 255-267. 2006

[6] Dhandra, B. V., Hangarge, M., Hegadi, R. and Malemath, V. S. "Word level script identification in bilingual documents through discriminating features," *In* International Conference on Signal Processing, Communications and Networking, 2007. ICSCN'07. pp. 630-635. 2007

[7] Pati, P. B. and Ramakrishnan, A. G. "Word level multi-script identification," *Pattern Recognition Letters*, *29*(9), pp. 1218-1229, 2008.

[8] Chanda, S. Sinha, S., and U. Pal. "Word-wise English devanagari and oriya script identification, Speech and Language Systems for Human Communication, pp. 244–248, 2004.

[9] Sinha, S. Pal, U. and Chaudhri, B.B. "Word-wise script identification from Indian documents," In Proc. IAPR Int'l Workshop Document Analysis Systems, pp. 310–321, 2004.

[10] Rani, R., Dhir, R. and Lehal, G. S. "Performance analysis of feature extractors and classifiers for script recognition of English and Gurumukhi words," Proceedings of the Workshop on Document Analysis and Recognition (DAR 2012), Mumbai, Publisher ACM, USA. pp. 30-36. 2012

[11] Ghosh, D., Dube, T. and Shivaprasad, A. P. "Script recognition—A review," Pattern Analysis and Machine Intelligence, IEEE Transactions on, 32(12), pp. 2142-2161, 2010

[12] Lehal, G.S. "Optical Character Recognition of Gurmukhi Script using Multiple Classifiers", Proceedings of International Workshop of Multilingual OCR, Article No.7, . 2009.

[13] http://code.google.com/p/tesseract-ocr/, last accessed 12 January 2013.