Removing Back-to-Front Interference in Documents with Mirror Filtering

R. D. Lins and I.G. da Silva Netto

***Abstract*—Very often documents are written on both sides on translucent paper making visible the ink from one side on the other. This artifact was called "back-to-front interference". The direct binarization of documents with such interference yields unreadable documents. A mirror transformation to remove such noise was suggested over a decade ago, but there is no record in the literature either on how to implement it or of its effectiveness. This paper proves viable the mirror transformation method for generating high-quality monochromatic images of documents with back-to-front interference.**

***Index Terms*—Back-to-front interference, Bleeding.**

I. Introduction

THE method presented herein is part of a larger project for processing and automatic transcription of historical documents from before the nineteenth century belonging to Joaquim Nabuco's bequest held by Joaquim Nabuco Foundation [1] (a social science research institute in Recife-Brazil), aiming at preserving their content to future generations and granting access to them in a number of ways including the possibility of making them available through networks. Processing this kind of document is more difficult than more recent ones because while the paper darkens with age, the printed part, either handwritten or typed, tends to fade. These two factors acting simultaneously narrow the discrimination gap between the two predominant color clusters within documents. If a document is typed or written on both sides and the opacity of the paper is such as to allow the back printing to be visualized on the front side, the degree of difficulty of good segmentation increases enormously. A new set of hues of paper and printing colors appears. This phenomenon, first addressed in the literature by reference [2] was called "back-to-front interference". Whenever the document is either in true-color or gray-scale the human eye is able to filter out that sort of noise keeping document readability. This is not the case with automatic tools such as OCRs. Thus, it is important to find better segmentation techniques to suitably solve that problem.

For purposes of image preservation, those documents are scanned with a resolution of 300 dots per inch in true-color (16 million of colors), in general.

Monochromatic images claim less storage space, allow for faster network transmission, and are suitable to be processed by most commercial OCR tools. Thus, documents are binarized into monochromatic. Image processing environments (such as Jasc Paint Shop Pro™) offer a great variety of binarization filters. However, the use of such software requires a specialized operator and that is not feasible to handle large quantities of documents. Besides that, the palette reduction algorithms provided by standard commercial tools whenever applied to documents with back-to-front interference yield unreadable images, even for humans. Fig. 1 presents an example of a letter with back-to-front interference, obtained from the Nabuco's file. The monochromatic version of the same document generated by the direct application of the binarization algorithm by using Jasc Paint Shop Pro™ version 8 (Palette component: Grey values, Reduction component: nearest color, Palette weight: non-weighted) is completely unreadable, as one may observe in Fig. 2.

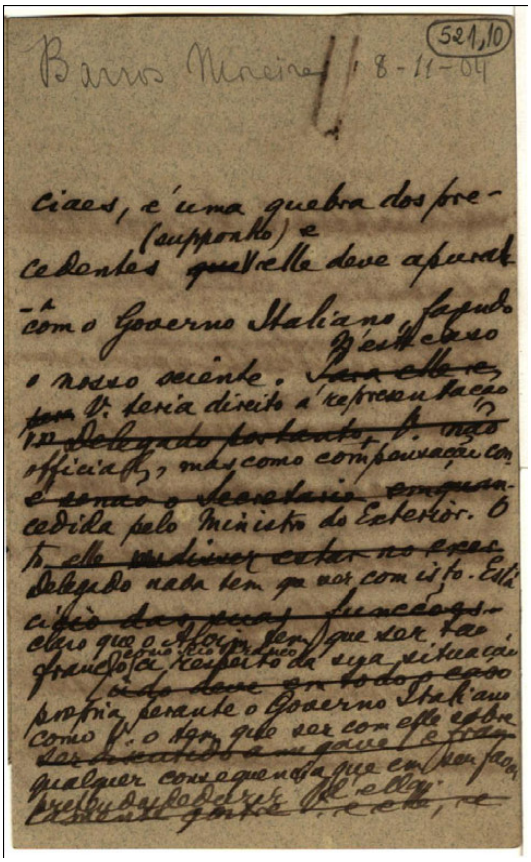


Fig. 1. Document from Nabuco's bequest.

In a document such as the one presented in Fig. 1, one expects to find three color clusters corresponding to the ink

Rafael Dueire Lins and Ismael Gomes Netto, Department of Electronic and Systems, Telecommunication Group, Center of Technology and Geosciences, Federal University of Pernambuco, Recife, Brazil, e-mails: rdl@ufpe.br, rdl.ufpe@gmail.com .

in the foreground, the paper background and the trespassed ink (the back-to-front interference). Unfortunately, no image representation provided such clustering to allow the easy filtering out of the back-to-front interference.

Several papers in the literature addressed the back-to-front interference problem. Some authors use waterflow models [3], other researchers have used wavelet filtering [4], but the technique of most widespread use is thresholding [5]-[6]-[7]. The most successful techniques for filtering out back-to-front interference are based on the entropy [8] of the grey-scale document [9]-[10]. Although recent advances were made in finding efficient algorithms that yield good quality images [11], a final solution to the filtering of back-to-front interference is still sought off.

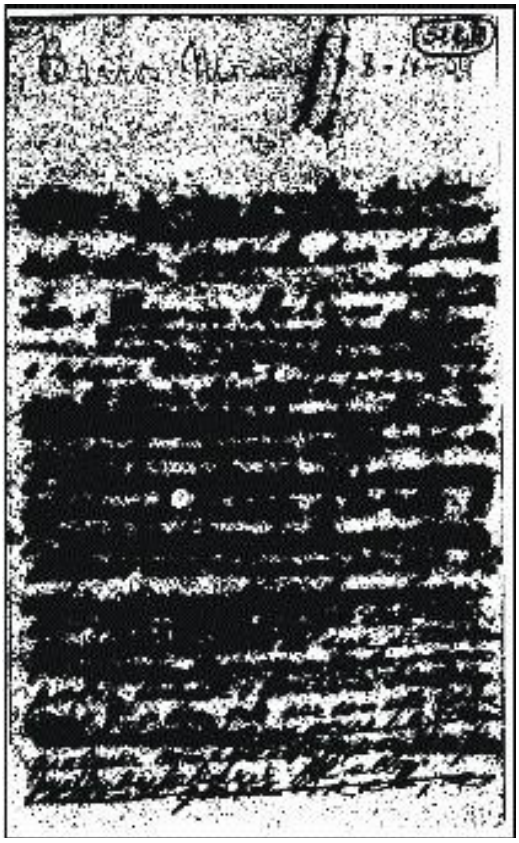


Fig. 2. Binarized document of Fig. 1.

Visual inspection of the filtered images provides a weak qualitative assessment of the performance of the algorithms under comparison. Analyzing the quality of images produced by filtering algorithms is far from being a trivial task. Subjectivity must be avoided by every means. Thus, a quantitative method to measure the quality of algorithms for binarizing documents with back-to-front interference is introduced here.

II. GRAY-SCALE CONVERSION

The conversion from true-color image into 256 levels grey scale images as an intermediate step towards image binarization has shown to be a valuable simplification, adopted by all the binarization algorithms either used or developed for removing back-to-front interference.

The first processing step is generating grey-scale documents from the true-color ones by using the standard equation to calculate the value of the new pixel:

$$gray\_level = 0.3R + 0.59G + 0.11B$$

where *R*, *G*, and *B* are the Red, Green and Blue values of the original pixel.

As one may observe from the image of the document exhibited in Fig. 3, the grayscale conversion of documents tends to preserve their readability.

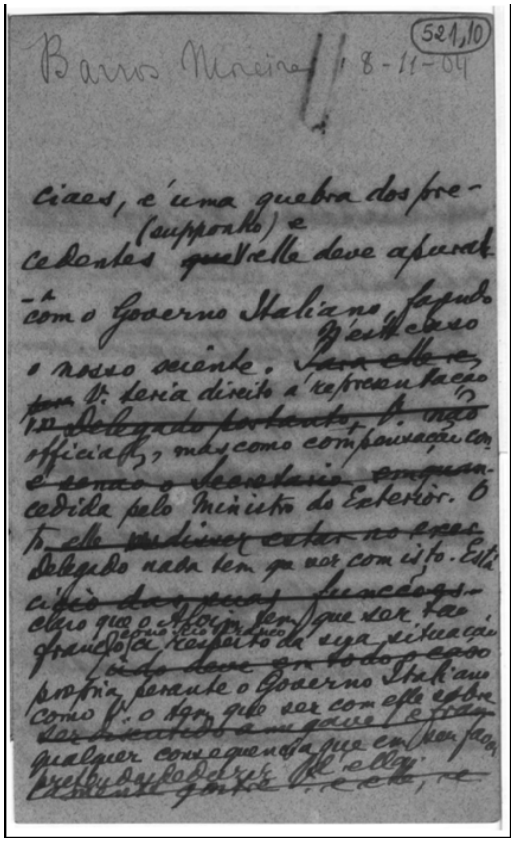


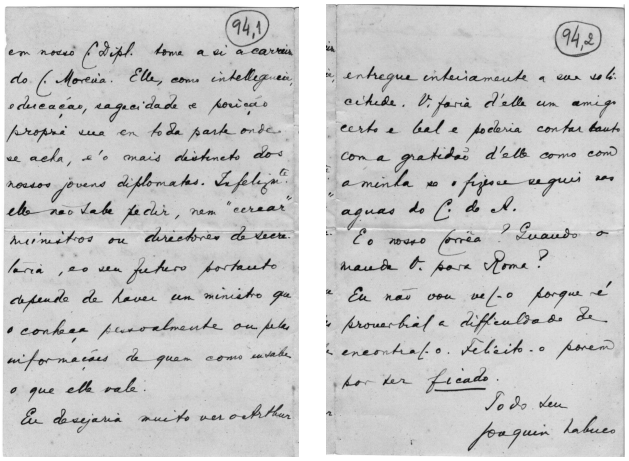
Fig. 3. Grayscale image of document of Fig. 1.

III. SYNTHESIS OF IMAGES WITH INTERFERENCE

This section presents how test images with back-to-front interference are generated. The basic idea is to introduce such interference in a well controlled way, thus one is able to really know which pixels ought to be removed and with should not be removed in the filtered image. The number of mismatching pixels from the reference and filtered images provides a quality factor of the algorithm, allowing a fair comparison between the results obtained. In what follows test image generation is detailed.

1) The first step is take two 256-grayscale images without back-to-font interference, such as the ones presented in Fig. 4.
   - S – a document that plays the role of foreground information (signal image); and
   - I – the image that plays the role of back-to-front noise (interfering image).

2) The second step is to synthesize a third image, called G_{fade}, by overlapping image S with a faded mirrored



Front (S) Back (I)

Fig. 4. Images of documents without back-to-front interference from Nabuco's bequest.

version of image I, as follows:

- One fades image I, producing the I_{fade} image given by

$$i_{fade}(m,n) = \lceil i\ (m,n) + fade \rceil^{255}$$

where the $i\ (m,n)$ and $i_{fade}(m,n)$ are the intensity values in the pixel (m,n) from the images I and I_{fade}, respectively, and *fade* is the brightness offset applied. Notice that the maximum value of the sum is 255.

- Finally, the overlapping process merges images S and G_{fade}, selecting the darker pixel between $s(m,n)$ and $g_{fade}(m,n)$, then,

$$g_{fade}(m,n) = \begin{cases} s(m,n) & \text{, if } s(m,n) \le g_{fade}(m,n) \\ g_{fade}(m,n) & \text{, if } s(m,n) > g_{fade}(m,n) \end{cases},$$

where $s(m,n)$ and $g_{fade}(m,n)$ are the intensity values in the pixel (m,n) from the images S and G_{fade}, respectively.

To assess the filtering capability of algorithms *fade* assumes values from 0 to 255. The effect of fade variation on the final synthesized document generated from the documents presented in Fig. 4 with I mirror-reflected is presented in Fig. 5.

IV. Assessment Method

This section introduces a new method to qualitatively assess the quality of images with back-to-front interference that have been filtered by binarization algorithms. After the image synthesis process shown above produced a series of 256 images with different interference levels two quality factors are calculated:

After the synthesize process one goes to the third step.

- The first takes as reference the image $S^{(manual)}$, which is obtained from S by manually searching a threshold that yields a good quality binary image. Now, one compares image $S^{(manual)}$ with each of the 256 $G_{fade}^{(k)}$ images (yielded by the application of the algorithm k to the images G_{fade}) and calculates the number of mismatching pixels:

$$q_{fade}^{(\text{absolute reference})} = \sum_{n=1}^{N}\sum_{m=1}^{M} | s^{(\text{manual})}(m,n) - g_{fade}^{(k)}(m,n) |$$

This is the *absolute reference mismatching factor*, because the reference image is the same for all algorithms.

- The second quality factor, called *self-referent mismatching factor*, takes as reference image $S^{(k)}$, which is obtained by the application of the algorithm k onto image S, for its binarization. The number mismatching pixels between the $S^{(k)}$ and $G_{fade}^{(k)}$ is calculated as:

$$q_{fade}^{(\text{self-referent})} = \sum_{n=1}^{N}\sum_{m=1}^{M} | s^{(k)}(m,n) - g_{fade}^{(k)}(m,n) |$$

fade = 0 *fade* = 30 *fade* = 50

fade = 70 *fade* = 100 *fade* = 130

fade = 150 *fade* = 170

Fig. 5. Pieces of synthesized images for different *fade* values.

V. Results and Analysis

The proposed assessment method was applied to eight binarization algorithms having as input 20 pairs of images selected from the Joaquim Nabuco's bequest:

- Algorithm 1 – da Silva-Lins-Rocha [11];
- Algorithm 2 – Mello-Lins [9]-[10];
- Algorithm 3 – Pun [12];
- Algorithm 4 – Kapur-Sahoo-Wong [13];
- Algorithm 5 – Wu-Songde-Hanqing [14];
- Algorithm 6 – Otsu [15];
- Algorithm 7 – Yen-Chang-Chang [16];
- Algorithm 8 – Johannsen-Bille [17].

For each pair of images and each algorithm the two quality factors introduced were measured and a graph was plotted. The 20 experiments performed exhibited similar resulting curves.

Fig. 14 presents the assessment plot with the *absolute reference mismatch factor* produced for the pair of images shown in Fig. 04, while Fig. 15 plots the *self-referent mismatch factor* for the same pair of figures. Fig. 16 and 17

present a zoom into the critical areas of the graphs presented in Fig. 14 and 15, respectively.

The analysis of the four graphs in Fig. 14-17 allows one to observe that the Johanssen-Bille algorithm produced highly unstable results for $0 \leq fade \leq 70$. That means that whenever the back-to-front interference is strong (see Fig. 05) the Johanssen-Bille algorithm [17] does not filter the noise satisfactorily. This fact is corroborated by the visual inspection of the resulting images, as shown in Fig. 06, taking *fade* = 90, one of the values most frequent of back-to-front interference.

Fig. 6. Filtering with algorithm by Johannsen-Bille (*fade*=90)

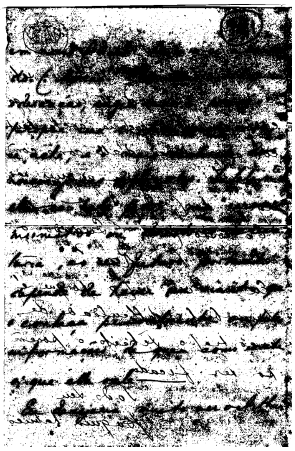


Fig. 7. Binarized by Pun´s algorithm (*fade*=90)

The performance of Pun´s algorithm, although far superior than Johanssen-Bille´s, as may be observed from the plots in Fig. 14-17, yields unsatisfactory images (see Fig. 7).

The graphs in Fig. 17-18 show that the algorithms by Yen-Chang-Chang and Kapur-Sahoo-Wong only produce reasonable filtering for images with medium-to-weak back-to-front interference (*fade*≥120). In the most frequent noise region (*fade*=90) these algorithms are unable to filter out significant amount of the back-to-front interference, as shown in Fig. 8 and 9.

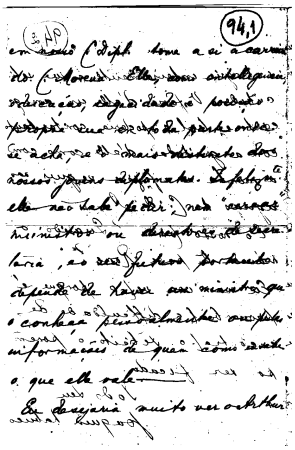


Fig. 8. .Filtering with algorithm by Yen-Chang-Cheng (*fade*=90)

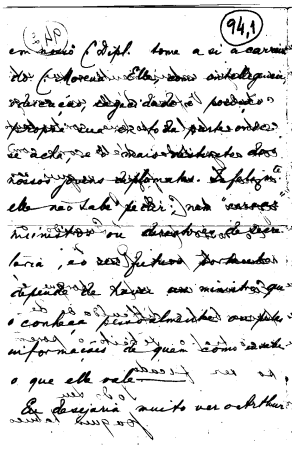


Fig. 9. Filtering with algorithm by Kapur-Sahoo-Wong (*fade*=90)

According to the assessment method proposed herein only four of the eight algorithms analyzed are suitable to remove the bleeding noise in documents. Their performances vary according to the strength of back-to-front interference. For images with strong noise (fade ≈ 50), the algorithm proposed by Wu-Songde-Hanqing has good chances of performing well in back-to-front noise removal, however it tends to be greedy and remove part of the foreground information. Fig. 10 presents the result of applying that algorithm with *fade* = 90.

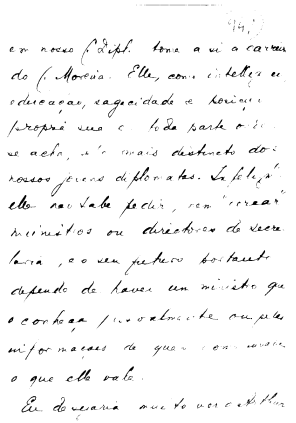


Fig. 10. Result with Wu-Songde-Hanqing (*fade*=90)

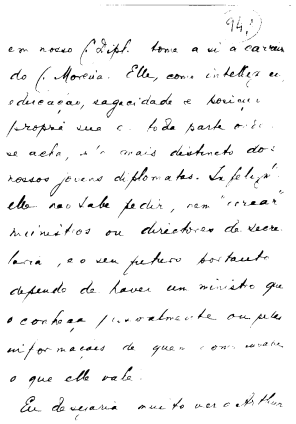


Fig. 11. Filtering with algorithm by Mello and Lins (*fade*=90)

According to the graphs shown Mello-Lins algorithm is suitable to filter-out strong back-to-front inference (*fade*≈40) degrading its performance when the noise weakens (*fade*≥70). Fig. 11 exhibits the image produced by Mello-Lins algorithm with *fade* = 90.

The steadiest good performance in filtering out the bleeding noise is provided by da Silva-Lins-Rocha algorithm, whose may be seen in Fig. 12.

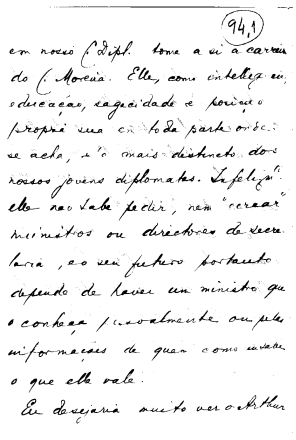


Fig. 12. Result with *fade*=90 for da Silva-Lins-Rocha algorithm

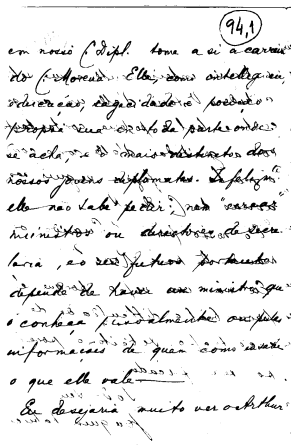


Fig.13 Filtering with algorithm Otsu algorithm at *fade*=90

In the case of images with very low back-to-front noise (*fade*>120), Otsu's algorithm surpasses all other algorithms. Its performance at the most frequent (and commonly found in documents) interference range (fade≈90), runs behind Mello-Lins and da Silva-Lins-Rocha algorithms, as depicted in Fig. 13.

VI. Conclusions and Lines for Further Work

A quantitative method to assess the quality of binarization algorithms for images with back-to-front interference is introduced. The results obtained with this assessment

method are consistent with the results obtained by visual inspection of filtered documents.

An important point of the proposed method is that it is able to spot which algorithm is more likely to perform better at filtering out the bleeding noise by analyzing the features of the document. This feature may allow the automatic choice of the best suitable algorithm to filter an specific document, thus permitting to be incorporated into a document processing environment such as BigBatch [18].

The assessment method proposed here did not take into account the color of the background as a controlled parameter. Work on progress is widening the scope of this work to model aged background, giving complete control of all document parameters.

REFERENCES

[1] FUNDAJ: http://www.fundaj.gov.br

[2] R. D. Lins, et al. *An Environment for Processing Images of Historical Documents. Microprocessing & Microprogramming*, pp. 111-121, North-Holland, 1995.

[3] Hyun-Hwa Oha, Kil-Taek Limb, Sung-Il Chienc. *An improved binarization algorithm based on a water flowmodel for document image with inhomogeneous backgrounds*. Pattern Recognition 38 (2005) 2612 – 2625, 2005.

[4] C. L. Tan, R. Cao, P. Shen, *Restoration of archival documents using a wavelet technique*, IEEE Tansactions on Pattern Analysis and Machine Intelligence, Vol.24, No. 10, pp. 1399-1404, October 2002.

[5] E. Kavallieratou and H. Antonopoulou, *Cleaning and Enhancing Historical Document Images*, Intelligent Vision Systems, Springer-Verlag 3708, pp. 681-688, 2005.

[6] G. Leedham, S. Varma, A. Patankar, V. Govindaraju, *Separating text and background in degraded document images—a comparison of global thresholding techniques for multi-stage thresholding*, Proceedings of the Eighth International Workshop on Frontiers in Handwritten Recognition, pp. 244–249, 2002.

[7] Q. Wang, C. L. Tan, *Matching of double-sided document images to remove interference*, IEEE CVPR2001, Hawaii, USA, 8-14 Dec 2001.

[8] N. Abramson, *Information Theory and Coding*. McGraw-Hill Book Co, 1963.

[9] C. A. B. Mello and R. D. Lins. *Image segmentation of historical documents*, Visual 2000, Mexico City, Mexico.

[10] C. A. B. Mello and R. D. Lins. *Generation of images of historical documents by composition*. ACM Document Engineering 2002, McLean, VA, USA.

[11] J. M. M. da Silva, R.D.Lins and V.C.da Rocha Jr. *Binarizing and Filtering Historical Documents with Back-to-Front Interference*, ACM-SAC 2006, Nancy, April 2006.

[12] T. Pun, *Entropic Thresholding, A New Approach*, C. Graphics and Image Processing, 16(3), 1981.

[13] J. N. Kapur, P. K. Sahoo and A. K. C. Wong. *A New Method for Gray-Level Picture Thresholding using the Entropy of the Histogram*, Computer Vision, Graphics and Image Processing, 29(3), 1985.

[14] L. U. Wu, M. A. Songde, and L. U. Hanqing, *An effective entropic thresholding for ultrasonic imaging*, ICPR'98: Intl. Conf. Patt. Recog., pp. 1522–1524 (1998).

[15] N. Otsu. *A threshold selection method from gray level histograms*. IEEE Trans. Syst. Man Cybern. SMC-9, 62–66 (1979).

[16] J. C. Yen, F. J. Chang, and S. Chang. *A new criterion for automatic multilevel thresholding*. IEEE Trans. Image Process. IP-4, 370–378 (1995).

[17] G. Johannsen and J. Bille. *A threshold selection method using information measures*. ICPR'82: Proc. 6th Intl. Conf. Patt. Recog., pp. 140–143 (1982).

[18] R.D.Lins and B.T.Ávila. *Tool Demo – BigBatch: A Toolbox for Monochromatic Documents*. ACM Symp. on Document Engineering 2005, Bristol, Sep. 2005.

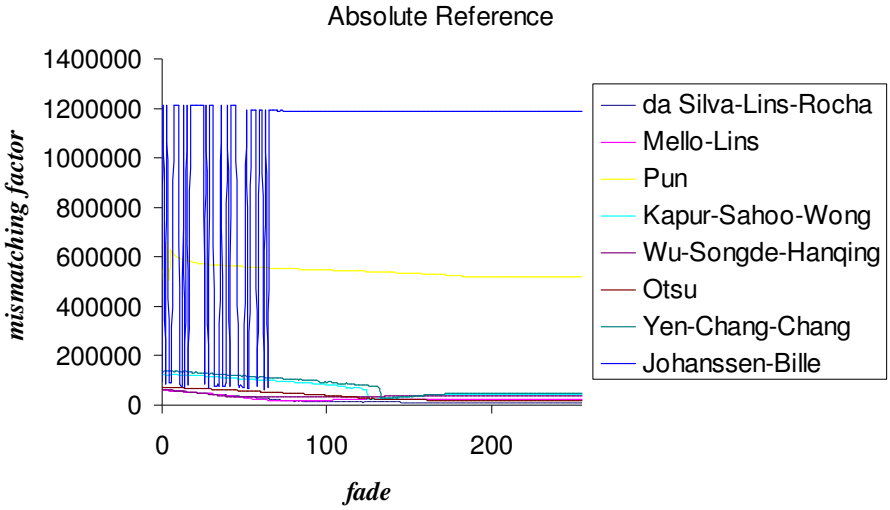


Figure 14. Absolute reference quality factor

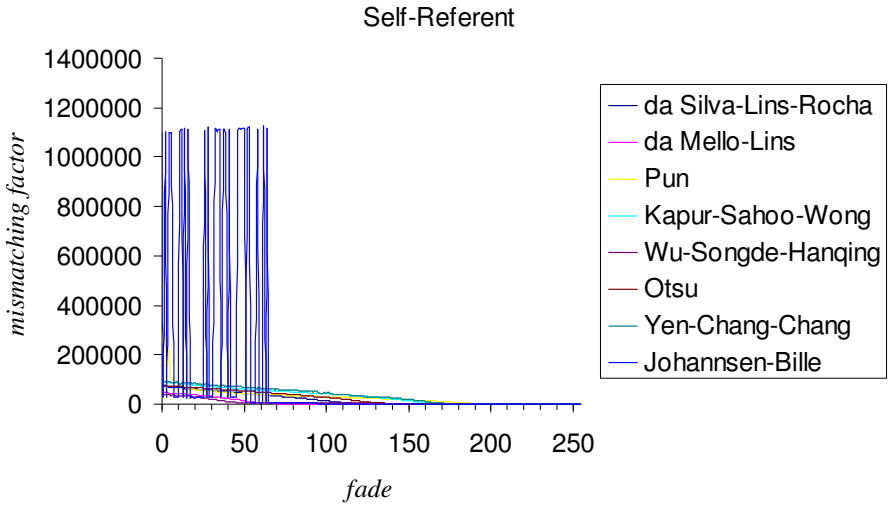


Figure 15. Self-referent quality factor

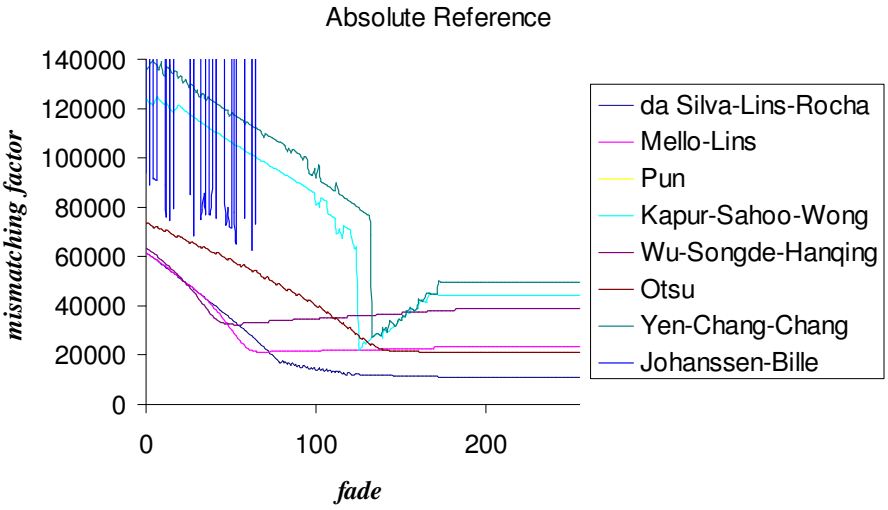


Figure 16. Zoom on critical part of Fig. 14 (Absolute reference)

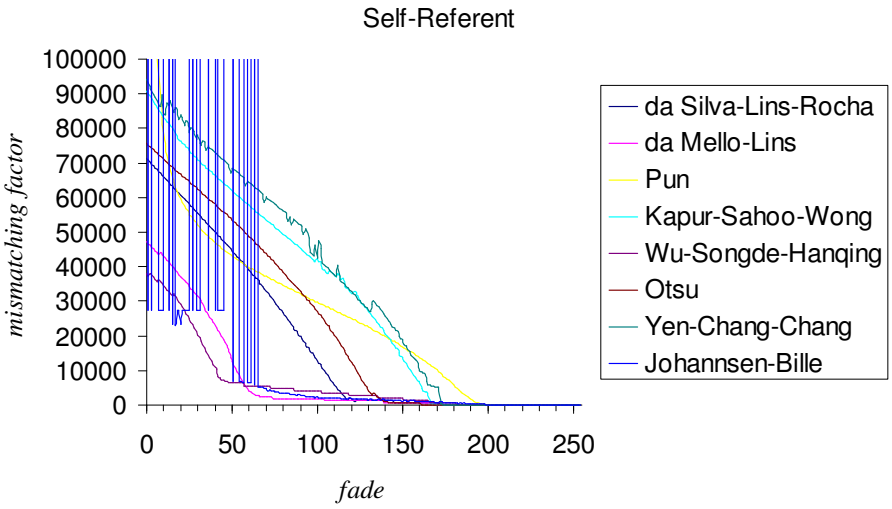


Figure 17. Zoom on critical part of Fig. 15 (Self-referent)