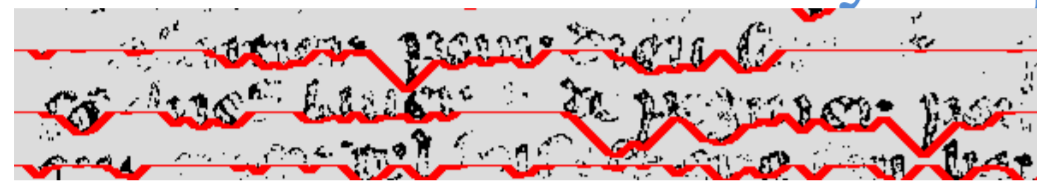


Seam Carving for Text Line Extraction on Grayscale Historical Manuscripts

Abstract

We propose a novel algorithm for automatic text line extraction on grayscale manuscripts without prior binarization. Our method is based on constrained seam carving to compute separating seams between consecutive text lines. Extensive experimental evaluations on diverse manuscripts show that we improve upon the state-of-the-art for grayscale text line extraction.

Why no prior binarization?



With binarization



Without binarization

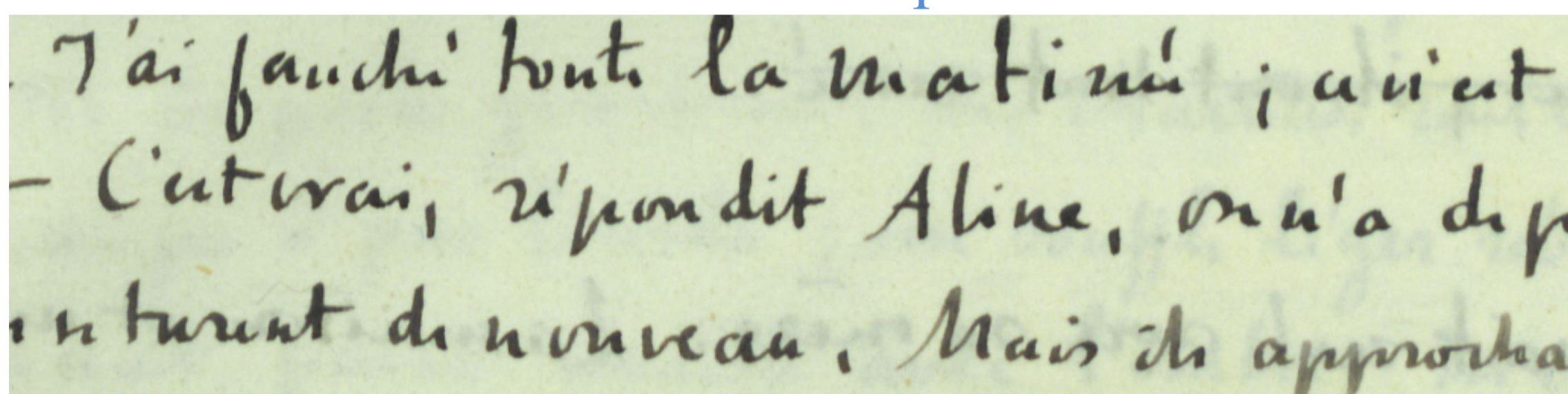
Due to the low quality of the manuscript, the binarization method results in extensive information loss. Any algorithm based on binary input is non-applicable.

Our Approach

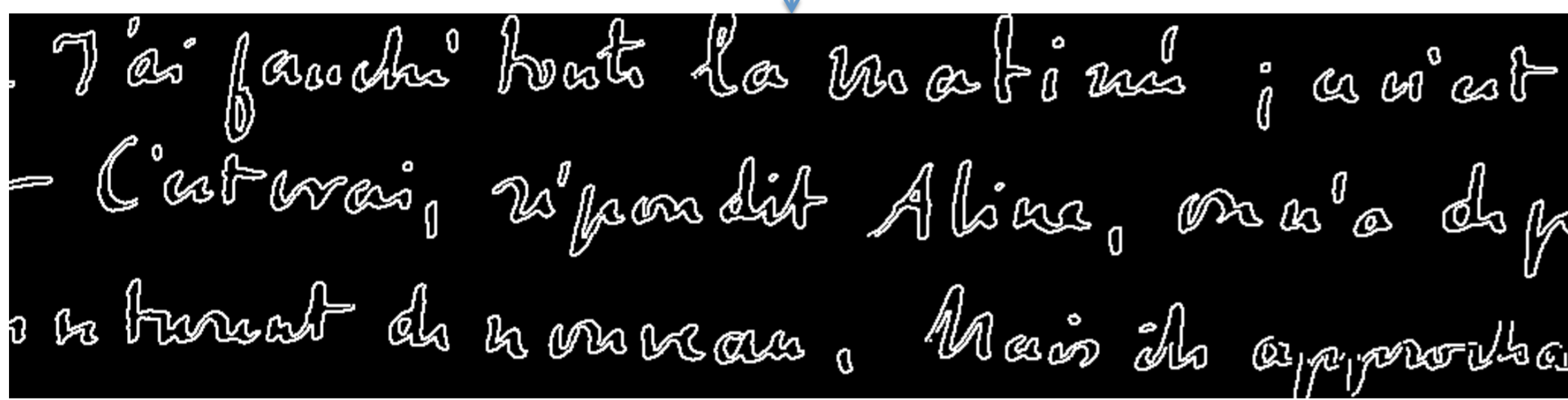
Our algorithm takes as input a grayscale image $I \in \mathbb{R}^{n \times m}$. It consists of two steps:

- 1) Medial seam computation using a projection profile matching approach similar to [3].
- 2) Separating seam computation using a modification of the seam carving procedure [4].

Medial Seam Computation



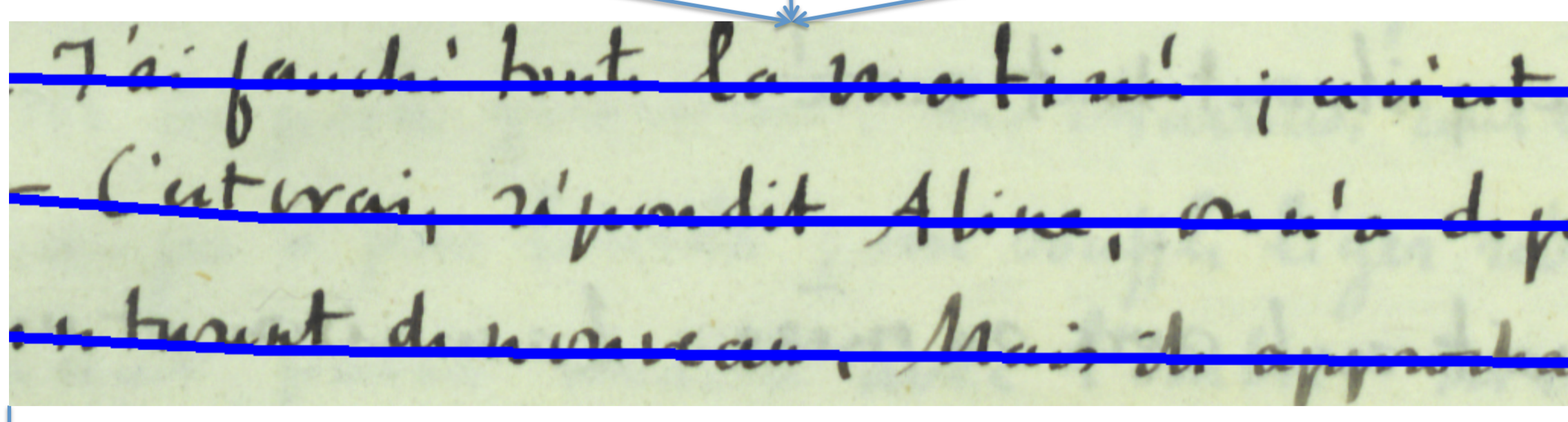
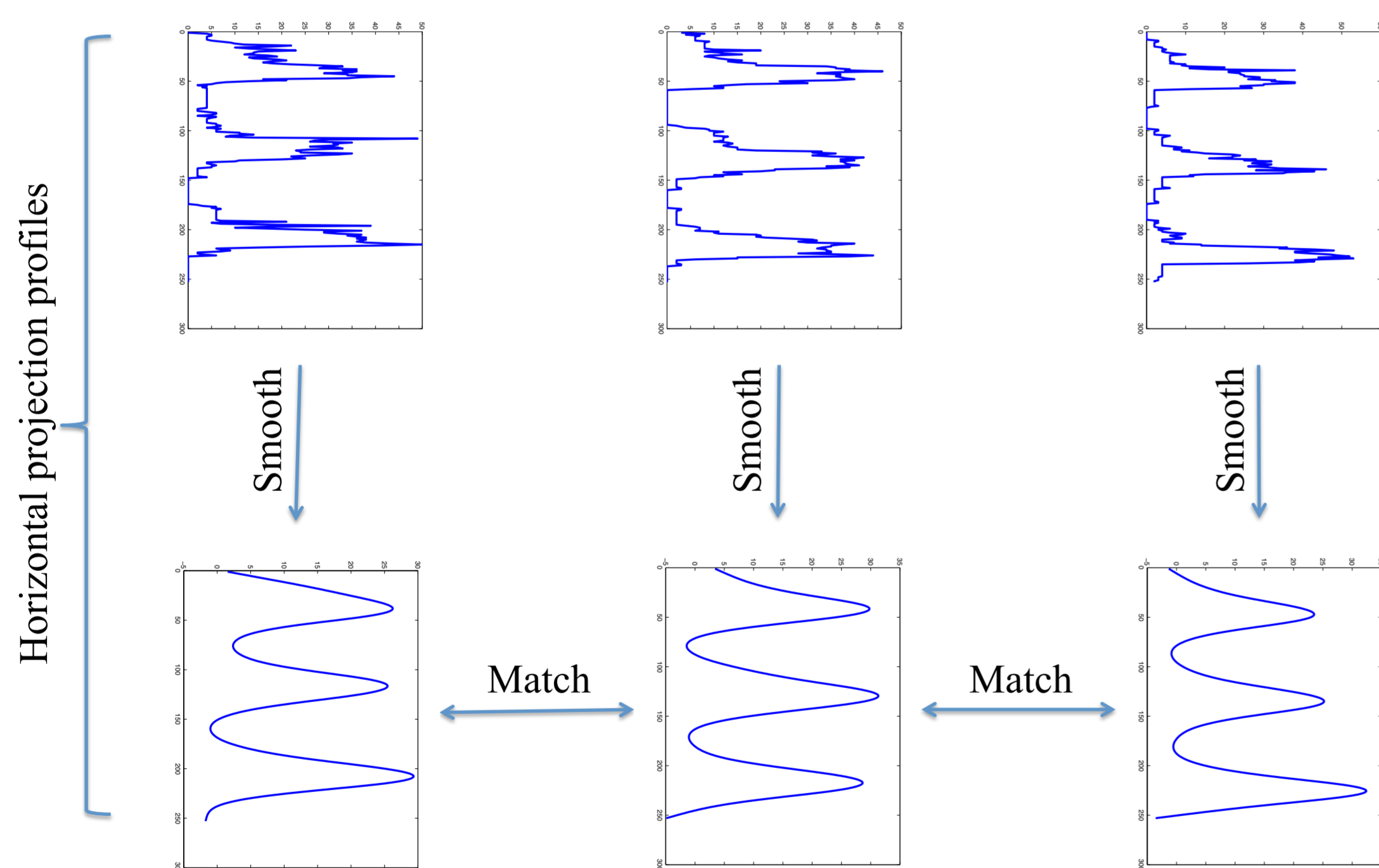
Edge filter



1st slice

2nd slice

3rd slice



Medial seams $L_{h,j}$, $h = 1, \dots, l$, $j = 1, \dots, m$

Acknowledgments

We thank Abdelkadir Asi for providing us the results of [2]. We also thank the Bibliothèque Cantonale et Universitaire de Lausanne for providing us with the manuscript pages of Aline.

Separating Seam Computation

Compute minimum-energy horizontal seams on the gradient image E :

- Horizontal seam definition:

$$s_h = \{s_{h,j}\}_{j=1}^m = \{(y_h(j), j)\}_{j=1}^m, \quad |y_h(j) - y_h(j-1)| \leq 1, \\ y_h(j) = L_{h,j}, \dots, L_{h+1,j}, \quad y_h : [1, \dots, m] \rightarrow [L_{h,j}, \dots, L_{h+1,j}]$$

- Seam computation with DP [4]:

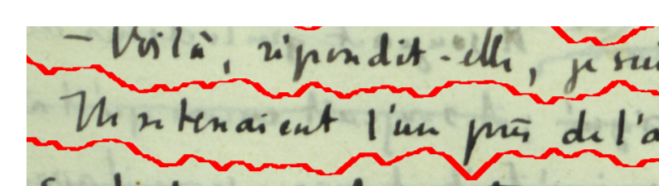
1. Cumulative minimum energy computation:

$$M_{y_h(j),1} = E_{y_h(j),1}, \quad M_{y_h(j),j} = E_{y_h(j),j} + \min \begin{cases} M_{y_h(j)-1,j-1} \\ M_{y_h(j),j-1} \\ M_{y_h(j)+1,j-1} \end{cases}$$

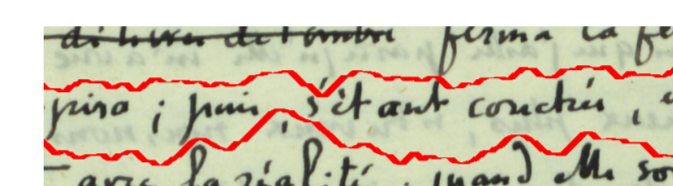
2. Traverse M backwards for optimal seam generation.

Results

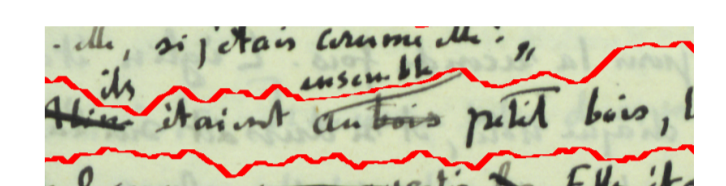
We distinguish three seam types:



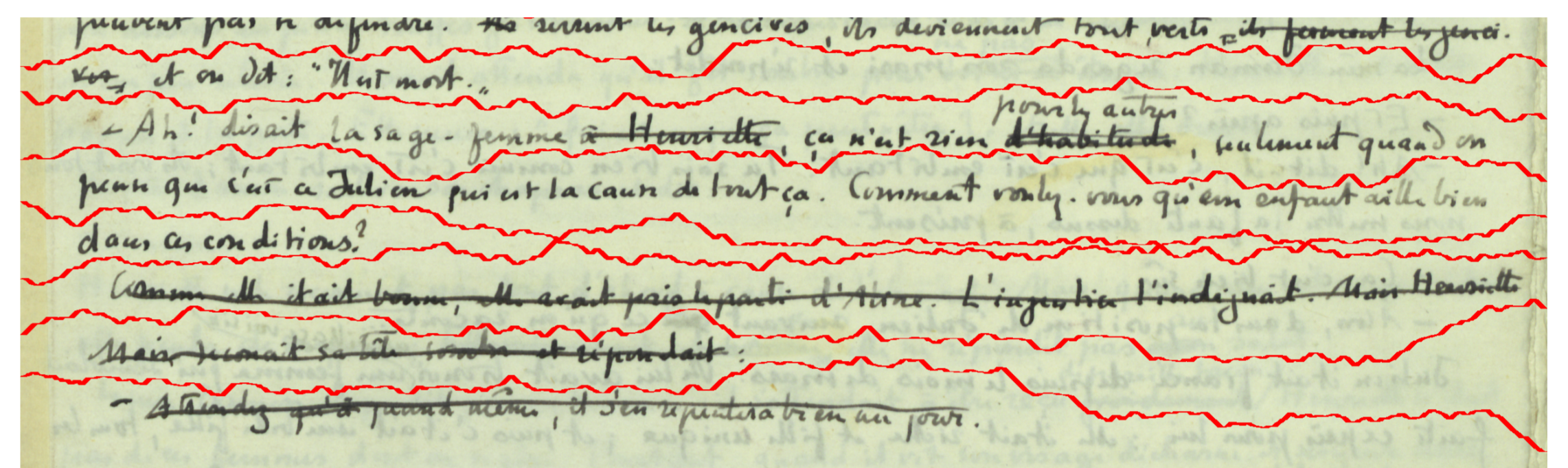
Type I seams, perfect.



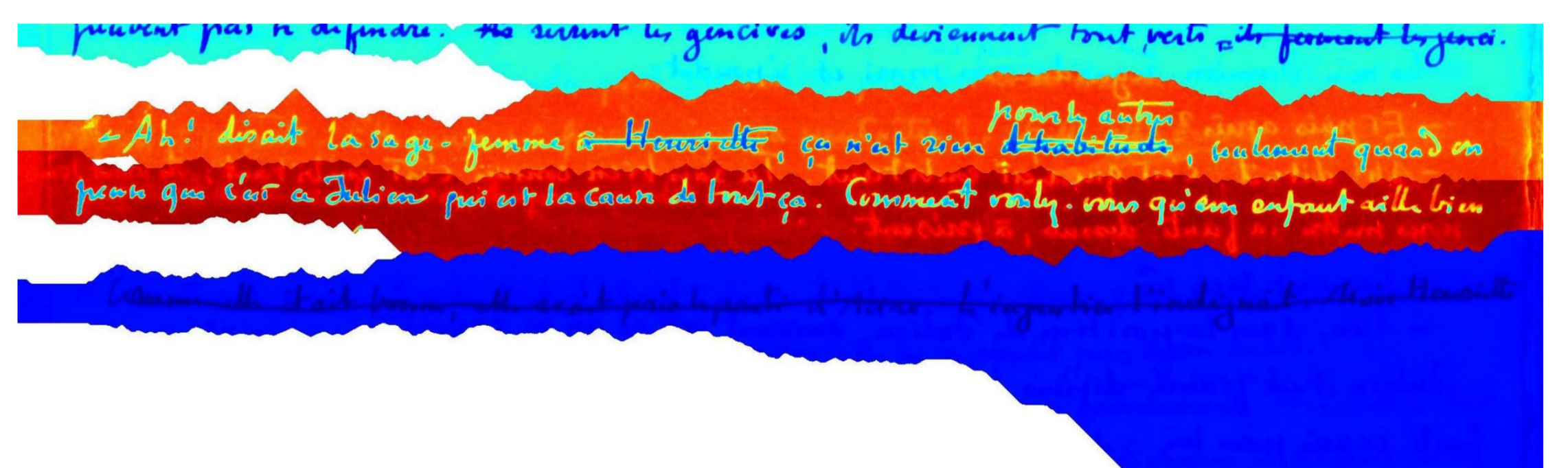
Type II seams, letter cuts.



Type III seams, wrong words.



Seams with our approach.



Seams with the state-of-art [2].

Collection	Seam Type					
	I		II		III	
	Ours	[2]	Ours	[2]	Ours	[2]
Al-Majid-2	98.2%	69.9%	1.8%	26.4%	0%	3.7%
Wadod-2	78%	21.5%	21.5%	45.6%	0.5%	0.5%
AUB-2	92%	53.9%	8%	43.4%	0%	2.7%

Table I. Seam type comparison with the state-of-the-art [2].

Collection	Accuracy	
	Ours	[2]
Al-Majid-1	99.30%	97.59%
Al-Majid-2	99.97%	98.19%
Wadod-1	99.04%	98.35%
Wadod-2	99.87%	97.53%
AUB-1	99.75%	98.05%
AUB-2	99.97%	96.15%
Thomas Jefferson	97.75%	95.21%

Table II. Automatic comparison with the state-of-the-art [2].

References

- [1] N. Arvanitopoulos and S. Süsstrunk, "Seam Carving for Text Line Extraction on Color and Grayscale Historical Manuscripts," to appear in ICFHR 2014.
- [2] R. Saabni, A. Asi, and J. El-Sana, "Text Line Extraction for Historical Document Images," PRL 2014.
- [3] M. Liwicki, E. Indermühle, and H. Bunke, "On-line Handwritten Text Line Detection Using Dynamic Programming," ICDAR 2007.
- [4] S. Avidan and A. Shamir, "Seam Carving for Content-Aware Image Resizing," SIGGRAPH 2007.