University of Groningen

The snowball principle for handwritten word-image retrieval

van Oosten, Jean-Paul

DOI:
[10.33612/diss.160750597](https://doi.org/10.33612/diss.160750597)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2021

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):
van Oosten, J-P. (2021). *The snowball principle for handwritten word-image retrieval: The importance of labelled data and humans in the loop*. [Thesis fully internal (DIV), University of Groningen]. University of Groningen. <https://doi.org/10.33612/diss.160750597>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

The snowball principle for handwritten word-image retrieval

The importance of labelled data and humans in the loop

Jean-Paul van Oosten

Cover illustration by Bente van der Graaf

Printed by Proefschriftmaken (www.proefschriftmaken.nl)

This research was made possible thanks to the SNN project
Target

This research was also in part made possible by the University of
Groningen, Slimmer AI and the Human Interface Laboratory at
Kyushu University

Special thanks to the National Archive for the use of the collec-
tion *Kabinet der Koningin*

© 2021, Jean-Paul van Oosten

rijksuniversiteit
 groningen

The snowball principle for handwritten word-image retrieval

The importance of labelled data and humans in the loop

Proefschrift

ter verkrijging van de graad van doctor aan de
Rijksuniversiteit Groningen
op gezag van de
rector magnificus prof. dr. C. Wijmenga
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

vrijdag 26 maart 2021 om 16.15 uur

door

Jean-Paul van Oosten

geboren op 2 september 1984
te Westvoorne

Promotor

Prof. dr. L.R.B. Schomaker

Beoordelingscommissie

Prof. dr. A.P.J. van den Bosch

Prof. dr. M. Biehl

Prof. dr. R.I. Ingold

CONTENTS

1	INTRODUCTION	1
1	From monks to the Monk system	1
2	Human involvement in the handwriting recognition pipeline and misleading assumptions on . . .	5
2.1	Machine learning algorithms	8
2.2	Features	8
2.3	The origin and availability of labels	9
3	Research questions	11
4	Outline of the thesis	15
2	EXAMINING COMMON ASSUMPTIONS ABOUT THE CONVERGENCE OF THE BAUM-WELCH TRAINING ALGORITHM FOR HIDDEN MARKOV MODELS	17
1	Introduction	17
2	Method	21
3	Relation between distance to the global optimum and performance in terms of Maximum Likelihood	24
4	Relation between initial and trained distance . . .	27
5	Training with partially known models	31
6	Implications	32
6.1	How much do we need to 'push' a model in the right direction?	33
6.2	Is meta-learning necessary?	35
7	Discussion and conclusion	36
3	A REEVALUATION AND BENCHMARK OF HIDDEN MARKOV MODELS	41
1	Introduction	41
2	Benchmark	43
3	Learning the topology of a transition matrix . . .	46
4	The importance of temporal modelling	49
5	Discussion	53
4	SEPARABILITY VERSUS PROTOTYPICALITY IN HANDWRITTEN WORD-IMAGE RETRIEVAL	57
1	Introduction	57

2	Separability versus Prototypicality	61
3	Methods	65
4	Results	70
5	Conclusions	73
5	GENERAL DISCUSSION	79
1	Machine learning and representation	79
1.1	Baum-Welch training of HMMs	80
1.2	The relative importance of transition vs. observation probabilities	81
2	Labels	82
3	Loops and snowballs, not pipelines	85
4	Deep learning	86
5	Conclusion	88
	APPENDIX	93
	SUMMARY	97
	SAMENVATTING	103
	PUBLICATIONS BY THE AUTHOR	109
	BIBLIOGRAPHY	111
	ACKNOWLEDGEMENTS	123