

Schriftenreihe des Fachbereichs Elektrotechnik und Informatik

herausgegeben von

Prof. Dr.-Ing. Hans Dieter Beims  
Fachbereich Elektrotechnik und Informatik  
Hochschule Niederrhein

Band 8/2009

**C. Dalitz (Ed.)**

**Document Image Analysis  
with the Gamera Framework**

Shaker Verlag  
Aachen 2009

**Bibliographic information published by the Deutsche Nationalbibliothek**

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at <http://dnb.d-nb.de>.

Copyright Shaker Verlag 2009

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the publishers.

Printed in Germany.

ISBN 978-3-8322-8594-4

ISSN 1610-9392

Shaker Verlag GmbH • P.O. BOX 101818 • D-52018 Aachen

Phone: 0049/2407/9596-0 • Telefax: 0049/2407/9596-9

Internet: [www.shaker.de](http://www.shaker.de) • e-mail: [info@shaker.de](mailto:info@shaker.de)

## Herausgeber der Beiträge:

Christoph Dalitz      Fachbereich Elektrotechnik und Informatik,  
Hochschule Niederrhein (Krefeld, Deutschland)

## Autoren:

René Baston      Fachbereich Elektrotechnik und Informatik,  
Hochschule Niederrhein (Krefeld, Deutschland)

Christoph Dalitz      Fachbereich Elektrotechnik und Informatik,  
Hochschule Niederrhein (Krefeld, Deutschland)

Michael Droettboom      Space Telescope Science Institute (Baltimore, USA)

Vanni G. Rizzo      Junior Enterprise Catania – JECT (S. Gregorio, Italien)

Andrea Spadaccini      Dipartimento di Ingegneria Informatica e delle  
Telecomunicazioni, University of Catania (Catania, Italien)



# Vorwort

Es ist mir eine Freude, im diesjährigen Band unserer Schriftenreihe eine Reihe von Arbeiten zu präsentieren, die alle im Zusammenhang mit dem internationalen Softwareprojekt “Gamera” stehen, das ursprünglich an der Johns Hopkins University (USA) begonnen wurde und jetzt an der Hochschule Niederrhein unter meiner Leitung weiterentwickelt wird. Diese Software ist eine Art Baukasten, mit dem sowohl komplette Dokumenterkennungssysteme erstellt werden können, als auch neue Verfahren im Bereich der Dokumenterkennung entwickelt und untersucht werden können.

Besonders freut mich, dass neben eigenen Beiträgen aus unserer Hochschule auch zwei externe Beiträge aus den USA und Italien dabei sind. Im Einen gibt *Michael Droettboom*, der langjährige Hauptentwickler von Gamera, einen Überblick über die Entstehungsgeschichte von Gamera. Im Anderen beschreiben *Andrea Spadaccini* und *Vanni Rizzo* ein automatisches Erkennungssystem für Multiple-Choice Tests, das sie auf der Basis von Gamera entwickelt haben.

Die drei anderen Artikel beschreiben von mir an der Hochschule Niederrhein entstandene Weiterentwicklungen, sowie eine Zusammenarbeit mit *René Baston*, einem Absolventen unseres Fachbereichs, der sich im Rahmen seiner Bachelorarbeit mit der Anwendung von Gamera auf gewöhnliche Textdokumente beschäftigt hat. Da Gamera weltweit in zahlreichen Forschungsprojekten eingesetzt wird, sind die Artikel dieses Bandes auf englisch verfasst, so dass sie einem internationalen Leserkreis zugänglich sind.

Herzlichen Dank an Prof. Dr. Hans Dieter Beims für sein Engagement um die Herausgabe dieser Schriftenreihe und an den Fachbereich für die Finanzierung des vorliegenden Bandes.

Prof. Dr. Christoph Dalitz



## Inhaltsverzeichnis

|   |    |
|---|----|
| A brief History of the Gamera Document Image Analysis System            | 1  |
| <i>M. Droettboom</i>  |    |
| A Multiple-Choice Test Recognition System based on the Gamera Framework | 5  |
| <i>A. Spadaccini, V.G. Rizzo</i>  |    |
| Reject Options and Confidence Measures for kNN Classifiers              | 16 |
| <i>C. Dalitz</i>  |    |
| Kd-Trees for Document Layout Analysis                                   | 39 |
| <i>C. Dalitz</i>  |    |
| Optical Character Recognition with the Gamera Framework                 | 53 |
| <i>C. Dalitz, R. Baston</i>   |    |