

Reihe 8

Mess-,
Steuerungs- und
Regelungstechnik

Nr. 1251

Dipl.-Ing. Thomas Guthier,
Frankfurt am Main

Visual Motion Processing

Berichte aus dem

Institut für
Automatisierungstechnik
und Mechatronik
der TU Darmstadt



Fortschritt-Berichte VDI

Reihe 8

Mess-, Steuerungs-
und Regelungstechnik

Dipl.-Ing. Thomas Guthier,
Frankfurt am Main

Nr. 1251

Visual Motion Processing

Berichte aus dem

Institut für
Automatisierungstechnik
und Mechatronik
der TU Darmstadt



Guthier, Thomas

Visual Motion Processing

Fortschr.-Ber. VDI Reihe 8 Nr. 1251. Düsseldorf: VDI Verlag 2016.

166 Seiten, 61 Bilder, 12 Tabellen.

ISBN 978-3-18-525108-5, ISSN 0178-9546,

€ 62,00/VDI-Mitgliederpreis € 55,80.

Keywords: Human Action Recognition – Computational Neuro-science – Computer Vision – Machine Learning – Deep Learning

The capability to recognize biological motion, i.e. gestures, human actions or face movements is crucial for social interactions, for predators, prey or artificial systems interacting in a dynamic environment.

In this thesis an artificial feed-forward neural network for biological motion recognition is proposed. Like its natural counterpart, it consists of multiple layers organized in two streams, one for processing static and one for processing dynamic form information. The key component of the proposed system is a novel unsupervised learning algorithm, called VNMF, that is based on sparsity, non-negativity, inhibition and direction selectivity.

In the first layer of the dorsal stream, the VNMF is modified to solve the optical flow estimation problem. In the subsequent layer the VNMF algorithm extracts prototypical patterns, such as optical flow patterns shaped e.g. as moving heads or limb parts. For the ventral stream the VNMF algorithm learns distinct gradient structures, resembling edges and corners. All these patterns represent simple cells of the feed-forward hierarchy. The classification performance of the feed forward neural network is analyzed on three real world datasets for human action recognition and one face expression recognition dataset, achieving results comparable to current computer vision approaches.

Bibliographische Information der Deutschen Bibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet unter <http://dnb.ddb.de> abrufbar.

Bibliographic information published by the Deutsche Bibliothek

(German National Library)

The Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliographie (German National Bibliography); detailed bibliographic data is available via Internet at <http://dnb.ddb.de>.

D 17

© VDI Verlag GmbH · Düsseldorf 2016

Alle Rechte, auch das des auszugsweisen Nachdruckes, der auszugsweisen oder vollständigen Wiedergabe (Fotokopie, Mikrokopie), der Speicherung in Datenverarbeitungsanlagen, im Internet und das der Übersetzung, vorbehalten.

Als Manuskript gedruckt. Printed in Germany.

ISSN 0178-9546

ISBN 978-3-18-525108-5

Visual Motion Processing

Vom Fachbereich
Elektrotechnik und Informationstechnik
der Technischen Universität Darmstadt
zur Erlangung des akademischen Grades eines
Doktor-Ingenieurs (Dr.-Ing.)
genehmigte Dissertation

von

Dipl.-Ing. Thomas Guthier

geboren am 16. Juni 1983 in Heppenheim/Hessen

Referent: Prof. Dr.-Ing. J. Adamy
Korreferent: Prof. Dr.rer.nat. B. Sendhoff
Tag der Einreichung: 15. April 2015
Tag der mündlichen Prüfung: 14. October 2015

D17
Darmstadt 2015

Danksagung

Mein größter Dank gilt meiner Familie, welche bei allen meinen Entscheidungen stets bedingungslos hinter mir steht. Diese Sicher- und Geborgenheit hat mir die angemessene Zeit gegeben, welche eine Dissertation von jedem einfordert.

Mein nächster Dank gilt Julian Eggert und Volker Willert, von denen ich alles wichtige gelernt habe was ich über Bildverarbeitung und Lernalgorithmen weiß. Ihr beider Anteil an dieser Arbeit ist in gewisser Weise größer als der Meinige.

Prof. Adamy, Prof. Sendhoff und Prof. Körner danke ich für ihr Vertrauen und das sie die vorliegende Arbeit überhaupt erst ermöglicht haben. Ich danke Prof. Adamy dafür, dass er meine Ausbildung, fachlich und darüber hinaus, gefördert hat.

Den Kaffeepausen und den daran teilnehmenden Kollegen danke ich für unbezahlbare Unterhaltungen. Letztendlich dafür, dass ich mich nahezu an jedem Morgen auf die Arbeit freuen durfte.

Dann wären da noch meine Studenten, welche mit ihren Ideen und ihrer investierten Zeit wesentlich zum Gelingen dieser Dissertation beigetragen haben. Vielen Dank dafür, dass ihr mir die Betreuung eurer Abschlussarbeit anvertraut habt.

Bei der Natur bedanke ich mich dafür, dass sie so etwas komplexes wie das menschliche Gehirn hat entstehen lassen. Ein spannenderes Themenfeld als dieser schwer definierbare Raum zwischen Biologie und vereinfachter, mathematisch-technischer Algorithmik hätte ich mir nicht erträumen können.

Es ist schon erstaunlich wieviele Menschen letztendlich solch eine Arbeit beeinflussen. Mein letzter Dank gilt all denen welche sich oben eventuell nicht wiederfinden, aber die auf ihre eigene Art meine Dissertation beeinflusst haben. Im speziellen...

Isabell, Marlene, Dominik Haumann, Erich Lenhardt, Jochen Grieser, Andreas Lutz, Thorsten Graber, Valentina Ansel, Adrian Šošić, Nikola Aulig, Andrea Schnall, Matthias Platho, Moritz Schneider, Kim Listmann, Dieter Lens und viele andere ...

Contents

Abbreviations	IX
Abstract	X
1 Introduction	1
1.1 Biological Motion Recognition	4
1.1.1 Temporal and View-Point Variations	7
1.1.2 Discriminative Features	8
1.2 Computational Models for Biological Motion Recognition	8
1.2.1 Computational Neuroscience	9
1.3 Summary & Thesis Structure	10
2 Computational Model	12
2.1 Biological Motion Recognition in the Brain	12
2.1.1 Different Motion Representations	14
2.1.2 Static and Dynamic Form Description	15
2.1.3 Neurophysiological Experiments	16
2.1.4 Brain Areas based on Neurophysiological Experiments	16
2.1.5 Dorsal Stream Areas In Biological Motion Recognition	17
2.1.6 Mid-Level Motion Patterns	18
2.1.7 Ventral Stream Areas In Biological Motion Recognition	19
2.1.8 Posterior Superior Temporal Sulcus (STSp)	20
2.2 Proposed Computational Model	21
2.2.1 Feed-Forward Neural Networks	21
2.2.2 Invariance Properties of Feed-Forward Neural Networks	23
2.2.3 Related Work	23
2.2.4 Proposed Computational Model	24
3 Unsupervised Pattern Learning	26
3.1 Related Work	28
3.1.1 Principal Component Analysis	28
3.1.2 Independent Component Analysis	29

3.1.3	Extensions of NMF	29
3.2	Properties of Parts-based Representations	30
3.2.1	Basic Constraints	31
3.2.2	Non-negativity	32
3.2.3	Sparsity	33
3.2.4	Local and Lateral Inhibition	34
3.2.5	Resulting Energy Function and Notations	35
3.3	Sparse Non-negative Matrix Factorization	36
3.3.1	Sparse Activations	37
3.3.2	Normalized Basis Vectors	37
3.3.3	Sparse Basis Vectors	38
3.3.4	Reconstruction Energy	38
3.3.5	sNMF Learning Algorithm	39
3.3.6	Orthogonality and Enforced Parts-Basedness	40
3.4	Non-negative Representations of Real-valued Data	42
3.4.1	Multidimensional Input	42
3.4.2	Multidimensional Basis Vectors	43
3.4.3	Multidimensional Activations	44
3.4.4	Sparse Activation Amplitudes	45
3.4.5	Positive and Negative Input	45
3.4.6	Strict Non-negativity	46
3.4.7	Weak Non-negativity	46
3.4.8	Orthogonality between Positive and Negative Reconstructions	48
3.5	Translation-invariant NMF	49
3.5.1	Reconstruction Energy	50
3.5.2	Sparse Activations	52
3.5.3	Orthogonality between Positive and Negative Representation	53
3.5.4	Enforced Topological Sparsity	54
3.5.5	VNMF Learning Algorithm	55
3.6	Algorithm Summary	57
4	Optical Flow Estimation	58
4.1	Problem Formulation	59
4.1.1	General Algorithmic Approaches	61
4.1.2	Correlation Methods	62
4.1.3	Differential Methods	63
4.1.4	Method Comparison	65

4.2	Related Work	66
4.2.1	Horn and Schunk	66
4.2.2	Lukas and Kanade	67
4.2.3	Extensions of the Classical Methods	67
4.2.4	Multi-Scale Methods	68
4.2.5	Other OFE-algorithms	68
4.3	VNMF-OFE Approach	69
4.3.1	Restrict Optical Flow Field to Model	69
4.3.2	Enforced Non-Negativity	70
4.3.3	Penalize Opposing Directions	71
4.3.4	Sparse Activity Amplitudes	72
4.3.5	Lateral Competition	72
4.3.6	VNMF-OFE Learning Algorithm	73
4.3.7	VNMF-OFE Algorithm for Activation Inference	75
4.4	Learning the Basis Vectors	76
4.4.1	Varying Model Parameters	76
4.4.2	Varying Energy Parameters	77
4.4.3	Learned vs Designed Basis Vectors	80
4.4.4	Discussion of the Parameter Settings	81
4.5	Comparison & Results	83
4.5.1	Comparison to Related Work	84
4.5.2	VNMF-OFE for Human Actions	86
4.6	Summary & Discussion	86
5	Feature Extraction	89
5.1	Optical Flow Patterns	90
5.1.1	Preprocessing	90
5.1.2	Varying Energy Parameters	91
5.1.3	Varying Basis Vector Parameters	94
5.1.4	Detailed Analysis of the Learning Process	95
5.1.5	Comparison to PCA and sNMF	97
5.1.6	Basis Vectors learned on Face Data	99
5.2	Gradient Patterns	100
5.2.1	Preprocessing	101
5.2.2	Varying Energy Parameters	102
5.2.3	Varying Basis Vector Parameters	103
5.2.4	Detailed Analysis of the Learning Process	104
5.2.5	Comparison to PCA and sNMF	105
5.2.6	Basis Vectors learned on Face Data	105

5.3	VNMF as Feature Descriptor	106
5.3.1	Simple Cell Response	108
5.3.2	Complex Cell Response	111
5.3.3	Relation to HOG/HOF Descriptor	113
6	Human Action Recognition	116
6.1	Support Vector Machine (SVM)	116
6.2	Results for Different Basis Vector Sets	117
6.2.1	Varying Basis Vector Parameters	118
6.2.2	Varying Energy Parameters	119
6.2.3	Comparison to PCA and sNMF Patterns	119
6.2.4	Varying Simple Cell Response	120
6.3	Facial Expression Recognition	121
6.4	Comparison to Related Work	122
6.4.1	HOG/HOF Results	122
6.4.2	Benchmark Results	123
7	Conclusion	126
7.1	Summary & Discussion	126
7.1.1	Optical Flow Estimation (VNMF-OFE)	127
7.1.2	Feature Extraction (VNMF)	127
7.1.3	Biological Motion Recognition Model (FFNN)	129
7.2	Outlook	129
A	Bag of Words	132
B	Visual Cortex	133
C	Gradient Derivations	135
C.1	Translation Invariant Learning	135
C.2	Topological Sparsity	136
D	Sparse Non-Negative Linear Dynamic Systems	137
D.1	Temporal Extension of sNMF	137
D.2	Related Work	138
D.3	Transition Energy	139
D.3.1	Sparsity in the Transitions	140
D.3.2	sNN-LDS Learning Algorithm	140
D.4	Results	141
	Bibliography	143

Abbreviations

BCA	Brightness Consistency Assumption
BCE	Brightness Consistency Equation
BOW	Bag of Words
EBA	Extratriate Body Area
FFNN	Feed Forward Neural Network
FER	Facial Expression Recognition
HAR	Human Action Recognition
HOG	Histogram of Oriented Gradients
HOF	Histogram of Optical Flow
HS	Horn and Schunk
ICA	Independent Component Analysis
ISA	Independent Subspace Analysis
IT	Inferior Temporal
LK	Lukas and Kanade
MT	Middle Temporal
MST	Medial Superior Temporal
NMF	Non-negative Matrix Factorization
OF	Optical Flow
OFE	Optical Flow Estimation
PCA	Principal Component Analysis
SC	Sparse Coding
sNMF	Sparse Non-negative Matrix Factorization
sNN-LDS	Sparse Non-negative Linear Dynamic Systems
STSp	Posterior Superior Temporal Sulcus
SVM	Support Vector Machine
V1	Primary Visual Cortex
VNMF	Vector Non-negative Matrix Factorization

Abstract

The capability to recognize biological motion, *i.e.* gestures, human actions or face movements is crucial for social interactions, for predators, prey or artificial systems interacting in a dynamic environment. The famous point-light-walker experiments [58] reveal that humans have a highly skilled mechanism dedicated to the analysis of motion information, however the exact details of this mechanism remain largely unclear. A popular theory is, that visual recognition is performed in a hierarchical feed-forward process, consisting of multiple learned simple cell/complex cell layers [53]. In the case of biological motion recognition these layers are spread throughout the ventral and dorsal stream of the visual cortex, the ventral stream being dedicated to static visual information, such as spatial gradient structures and the dorsal stream is related to dynamic visual information, such as the motion for each pixel in the input, also known as the optical flow.

In this thesis an artificial feed-forward neural network for biological motion recognition is proposed. Like its natural counterpart, it consists of multiple layers organized in two streams, one for processing static and one for processing dynamic form information. The key component of the proposed system is a novel unsupervised learning algorithm, called VNMF, that is based on sparsity, non-negativity, inhibition and direction selectivity.

In the first layer of the dorsal stream, the VNMF is modified to solve the optical flow estimation problem. In the subsequent layer the VNMF algorithm extracts prototypical patterns, such as optical flow patterns shaped *e.g.* as moving heads or limb parts. For the ventral stream the VNMF algorithm learns distinct gradient structures, resembling edges and corners. All these patterns represent the simple cells of the feed-forward hierarchy, while the complex cells are modeled by a non-linear maximum pooling operation.

The classification performance of the feed forward neural network is analyzed on three real world datasets for human action recognition and one face expression recognition dataset, outperforming other biological inspired models while being competitive with current computer vision approaches.

Kurzfassung

Gesten, Mimiken und andere natürliche Bewegungen sind ein wesentlicher Bestandteil zwischenmenschlicher Kommunikation. Darüber hinaus ist die visuelle Wahrnehmung von Bewegungen notwendig um sich in einer sich stetig verändernden Umgebung zurechtzufinden. Die berühmten *Point-Light-Walker* Experimente von Johansson [58] zeigen, dass Menschen Bewegungen auch ohne klar definierte Formen wahrnehmen können. Allerdings ist es nach wie vor unklar wie die Bewegungsinformationen im menschlichen Gehirn verarbeitet werden. Eine populäre Theorie [53] besagt, dass visuelle Informationen in aufeinander folgenden, gelernten Neuronen-schichten verarbeitet werden. Im Fall der visuellen Bewegungsanalyse sind die Schichten im ventralen und dorsalen Pfad des visuellen Kortex verteilt. Der ventrale Pfad verarbeitet statische, z.B. Kanten, Informationen, während der dorsale Pfad eher dynamische Informationen, z.B. Punktbewegungen, auch optischer Fluss genannt, verarbeitet.

In der vorliegenden Dissertation wird ein künstliches neuronales Netzwerk zur Erkennung von natürlichen Bewegungen vorgestellt, welches, dem biologischen Vorbild gleich, aus zwei parallelen Pfaden besteht. Die Schlüsselkomponente des vorgestellten Systems ist ein neuer Lernalgorithmus, welcher die neuronalen Verbindungen der verschiedenen Schichten ausschließlich anhand von Beobachtungen lernt. Die Kodierung der Bewegungsinformation erfolgt richtungsspezifisch anhand von spärlichen, nicht-negativen Aktivitäten, welche mit anderen Aktivitäten in ihrer lokalen Nachbarschaft konkurrieren. In der ersten Schicht des dorsalen Pfades wird das optische Flussfeld mit Hilfe des neuen Lernalgorithmus geschätzt. In der darauf folgenden Schicht werden prototypische Muster gelernt, deren Formen bewegliche Körperteile beschreiben. Im ventralen Pfad wird der VNMF Algorithmus verwendet um Kantenstrukturen zu lernen.

Die Klassifikationseigenschaften des neuronalen Netzes werden anhand von drei Datensätzen für Körper- und Gesichtsbewegungen evaluiert. Die Klassifikationsergebnisse des vorgestellten Systems sind genauer als die anderer biologisch inspirierter Modelle und vergleichbar mit aktuellen Modellen der Bildverarbeitung.

