EFFICIENT 3D POSE ESTIMATION AND TRACKING OF ARTICULATED BODY FROM UNCALIBRATED MONOCULAR IMAGE SEQUENCE

MRS. KITTIYA KHONGKRAPHAN

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
(ELECTRICAL AND COMPUTER ENGINEERING)
FACULTY OF ENGINEERING
KING MONGKUT'S UNIVERSITY OF TECHNOLOGY THONBURI
2010

Efficient 3D Pose Estimation and Tracking of Articulated Body from Uncalibrated Monocular Image Sequence

Mrs. Kittiya Khongkraphan M.Sc. (Computer Science)

A Dissertation Submitted in Partial Fulfillment
of the Requirements for
the Degree of Doctor of Philosophy (Electrical and Computer Engineering)
Faculty of Engineering
King Mongkut's University of Technology Thonburi
2010

Dissertation Committee

(Assoc. Prof. Suthep Madarasmi, Ph.D.)	Chairman of Dissertation Committee
(Assoc. Prof. Pakorn Kaewtrakulpong, Ph.D.)	Member and Dissertation Advisor
(Asst. Prof. Natasha Dejdumrong, D.Tech.Sci.)	Member
(Asst. Prof. Thanarat Chalidabhongse, Ph.D.)	Member
(Pradit Mittrapiyanuruk, Ph.D.)	Member

Dissertation Title	Efficient 3D Pose Estimation and Tracking of Articulated Body from Uncalibrated Monocular Image Sequence
Dissertation Credits	36
Candidate	Mrs. Kittiya Khongkraphan
Dissertation Advisor	Assoc. Prof. Dr. Pakorn Kaewtrakulpong
Program	Doctor of Philosophy
Field of Study	Electrical and Computer Engineering
Department	Computer Engineering
Faculty	Engineering
B.E.	2553

Abstract

This dissertation presents a marker-less based approach for estimating and tracking 3D articulated human pose from an uncalibrated monocular image sequence. Unlike previous approaches, our proposed method does not require training or data set of 3D known pose exemplars, nor does it require the assumption that at least one predefined segment is parallel to the image plane or that most human parts are close to a plane parallel to the image plane in every frame. Our work assumes a simpler assumption, for example, the actor stands vertically parallel to the image plane and not all of his/her joints lie on a plane parallel to the image plane in the first frame. The basic idea of our approach is to reconstruct 3D relative human pose from 2D point correspondences. Firstly, 2D joint points are accurately tracked by a new Quick Shift Belief Propagation (QSBP) based approach which benefits from Quick Shift mode seeking. The joint points are then used to reconstruct a set of 3D possible human poses using perspective concept (due to non-uniqueness of solutions). From this set, the optimal solution is selected by an efficient technique based on the concept of Multiple Hypothesis Tracking (MHT) with a motion-smoothness function between consecutive frames. Moreover, feedback information from 3D human pose is applied to alleviate several problems inherit in the monocular approach e.g. self-occlusion and observation ambiguity problems. Additionally, a motion model based on feedback information and geometric constraint is introduced for initializing state and (re)initializing state in case of tracking loss. The performance of our proposed approach is characterized using both real and synthesized image sequences and shows very good results. The accuracy, measured as an overall distance from the ground truth, is compared with previous approaches.

Keywords: 3D Human Body Reconstruction / 2D Human Body Tracking / Monocular Image Sequence / Perspective Model / Belief Propagation / Quick Shift / Multiple Hypothesis Tracking

หัวข้อวิทยานิพนธ์	การประมาณและติดตามท่าทางแบบสามมิติของร่างกายที่มีรูปแบบเป็นข้อต่ออย่างมีประสิทธิภาพจากภาพที่ได้จากกล้องเดี่ยวที่ไม่ได้สอบเทียบ
หน่วยกิต	36
ผู้เขียน	นางกิตติยา คงกระพันธ์
อาจารย์ที่ปรึกษา	รศ. ดร. ปกรณ์ แก้วตระกูลพงษ์
หลักสูตร	ปรัชญาดุษฎีบัณฑิต
สาขาวิชา	วิศวกรรมไฟฟ้าและคอมพิวเตอร์
ภาควิชา	วิศวกรรมคอมพิวเตอร์
คณะ	วิศวกรรมศาสตร์
พ.ศ.	2553

บทคัดย่อ

วิทยานิพนธ์นี้นำเสนอวิธีการประมาณและติดตามท่าทางแบบสามมิติของร่างกายที่มีรูปแบบเป็นข้อต่อจากชุดภาพที่ได้จากกล้องเดี่ยวที่ไม่ได้สอบเทียบ วิธีการนี้แตกต่างกับวิธีการที่มีอยู่เดิม คือ ไม่ต้องการการสอนหรือการพิจารณาจากฐานข้อมูลที่ทราบตำแหน่งสามมิติของร่างกายมาก่อน นอกจากนี้ยังไม่ต้องการข้อสมมุติต่างๆ เกี่ยวกับร่างกาย เช่น มีอย่างน้อยหนึ่งท่อนของร่างกาย ขนานกับฉากภาพในทุกๆ ภาพ หรือเกือบทุกท่อนของร่างกายอยู่ใกล้กับฉากที่ขนานกับฉากภาพในทุกๆ ภาพ โดยมีข้อสมมุติอย่างง่ายเพียงข้อเดียวในภาพแรก คือ ผู้แสดงต้องยืนตรงขนานกับฉากภาพและทุกท่อนของร่างกายจะต้องไม่อยู่บนฉากเดียวกัน แนวคิดหลักเบื้องต้นของวิทยานิพนธ์นี้ คือ การสร้างแบบสามมิติของร่างกายที่มีรูปแบบเป็นข้อจากข้อมูลสองมิติของตำแหน่งข้อต่อที่ตรงกัน โดยเริ่มจากการหาตำแหน่งสองมิติของข้อต่อของร่างกายด้วยวิธีการสืบทอดความเชื่ออย่างรวดเร็ว (Quick Shift Belief Propagation) หลังจากนั้นตำแหน่งสองมิติของข้อต่อจะถูกนำไปใช้สร้างกลุ่มท่าทางแบบสามมิติของร่างกายที่เป็นไปได้ทั้งหมด ซึ่งเกิดจากปัญหาความเป็นไปได้มากกว่าหนึ่งคำตอบ (non-uniqueness of solutions) โดยในการสร้างท่าทางแบบสามมิติของร่างกายจะใช้วิธีการเทคนิคการสร้างภาพให้ได้สัดส่วนความลึกฉายจากกล้อง (perspective concept) หลังจากนั้นคำตอบที่ถูกต้องจะถูกเลือกจากกลุ่มท่าทางแบบสามมิติของร่างกายที่เป็นไปได้ โดยในวิทยานิพจน์ได้เสนอวิธีการเลือกด้วยการประยุกต์ใช้การสร้างหลายสมมุติฐานในการติดตาม (Multiple hypothesis Tracking) ร่วมกับข้อมูลความราบรื่นของการเคลื่อนที่ระหว่างภาพ (smoothness function) นอกจากนี้วิทยานิพนธ์นี้ได้นำข้อมูลย้อนกลับมาใช้แก้ปัญหาที่มักจะเกิดกับการหาท่าทางแบบสามมิติของร่างกายด้วยกล้องเดี่ยว เช่น การบดบังตัวเองของร่างกาย หรือความกำกวมของข้อมูลภาพ และได้นำข้อมูลย้อนกลับร่วมกับข้อบังคับทางเรขาคณิต (geometric constraint) มาใช้เป็นรูปแบบของการเคลื่อนที่ของร่างกาย เพื่อใช้ในการกำหนดตำแหน่งเริ่มต้นหรือทำการกำหนด

ตำแหน่งเริ่มต้นซ้ำกรณีที่มีการสูญหายในการติดตามของร่างกาย การวัดประสิทธิภาพของวิธีการที่นำเสนอได้แบ่งออกเป็นสองส่วน คือ ด้วยภาพที่สร้างขึ้นมาและภาพที่ได้จากกล้องถ่ายภาพจริง ซึ่งทั้งสองส่วนได้ผลลัพธ์ที่ดี โดยความถูกต้องจะถูกเปรียบเทียบกับคำตอบจริง (ground truth) และเปรียบเทียบกับวิธีอื่นๆ ที่มีอยู่เดิม

คำสำคัญ : การสร้างแบบสามมิติของร่างกาย / การติดตามร่างกายแบบสองมิติ / ชุดภาพจากกล้องเดี่ยว / เทคนิคการสร้างภาพให้ได้สัดส่วนความลึกฉายจากกล้อง / วิธีการสืบทอดความเชื่อ / การเลื่อนอย่างรวดเร็ว / การติดตามแบบหลายสมมุติฐาน

ACKNOWLEDGMENTS

First and foremost, I am extremely grateful to my advisor Assoc. Prof. Dr. Pakorn Kaewtrakulpong. He has guided me into computer vision field and has always provided me with invaluable feedback through many stimulating discussions. He has been everything that one would desire as an advisor.

I would like to acknowledge Prof. Dr. Mubarak Shah and Dr. Arslan Basharat for spending their valuable time to fruitful discussion on various ideas in research during my six months visit at computer vision lab in University of Central Florida.

I would like to thank my fellow PhD students as well as members of Machine Vision and Computer Intelligent laboratory (MVCI Lab) for being great friends, providing their help and keeping me company throughout my stay at King Mongkut's University of Technology Thonburi.

Finally, I would like to thank my parents Yindee and Lop Yoisertsut for all their love and encouragement. I also would like to thank my husband Phanpong for understanding and supporting me spiritually. The last gratefully special thanks to my lovely sons Napas and Burith for giving me happiness and making me relax during the hard times of my PhD study.

CONTENTS

	PAGE
ENGLISH ABSTRACT	ii
THAI ABSTRACT	iii
ACKNOWLEDGEMENTS	v
CONTENTS	vi
LIST OF TABLES	ix
LIST OF FIGURES	x
LIST OF SYMBOLS	xii
LIST OF TECHNICAL VOCABULARY AND ABBREVIATIONS	xiii
CHAPTER	
1. INTRODUCTION	**1**
1.1 Challenges	1
1.2 Objectives	5
1.3 Contributions	5
1.4 Dissertation Outline	6
2. LITERATURE REVIEW	**8**
2.1 Human Model Representation	8
2.1.1 Kinematic Model	8
2.1.2 Part-based Model	9
2.1 Image Features	9
2.2.1 Edge	9
2.2.2 Silhouette	9
2.2.3 Color	9
2.2.4 Contour	10
2.2.5 Optical Flow	10
2.3 3D Human Pose Inference	10
2.3.1 A Single View	10
2.3.2 Multiple Views	11
2.4 3D Human Motion Estimation	11
2.4.1 Top-down Approach	11
2.4.2 Bottom-up Approach	12
2.5 Quantitative Evaluation	13
3. MATHEMATICAL BACKGROUND	**14**
3.1 Graphical Model	14
3.2 Belief Propagation	14
3.3 Mode Seeking Technique	17
3.3.1 Mean Shift Concept	19

3.3.2 Quick Shift Concept 21

4. OVERVIEW OF THE PROPOSED METHOD 23

4.1 A System Overview 24

4.2 Human Model 25

4.3 Assumptions 26

4.4 Initialization 26

5. 2D HUMAN BODY ESTIMATING AND TRACKING 28

5.1 Related Work 28

5.2 Mean Shift Belief Propagation [29] 30

5.3 Proposed Method 31

5.3.1 Human Representation 32

5.3.2 Motion Model 33

5.3.3 Quick Shift Belief Propagation 34

5.3.4 Mode Seeking in Belief Propagation 35

5.3.5 Observation Function 37

5.3.6 Potential Function 38

5.3.7 A Binary Occlusion Mask Determination 39

5.4 Experiments 39

5.4.1 Performance of 2D Tracking with Motion Model 39

5.4.2 Performance Comparison with MSBP [29] in 1D 41

5.4.3 Performance Comparison with Other Approaches 42

5.5 Conclusion 48

6. RECONSTRUCTION AND TRACKING OF 3D-ARTICULATED HUMAN BODY FROM 2D POINT CORRESPONDENCES OF A MONOCULAR IMAGE SEQUENCE 49

6.1 Previous Work 49

6.2 3D Articulated Body Reconstruction 52

6.2.1 Problem Formulation 53

6.2.2 Determining Reference Distance 54

6.2.3 Determining Scaling Factor of Root-node 55

6.2.4 Determining Scaling Factor of other Nodes 56

6.3 3D Human Body Tracking 57

6.3.1 The Joint Angle Limitation Constraint 57

6.3.2 Optimal Solution Selection by MHT 58

6.4 Experiments and Evaluations 60

6.4.1 Performance Comparison with other Approaches 60

6.4.2 Results on Real-world Image Sequences 63

6.4.3 Performance of 3D Tracking with Motion-smoothness Function 70

6.4.4 Robustness against Errors in Reference Distance Determination 74

6.4.5 Robustness against Noise in 2D Data 74

6.5 Conclusions 75

7. DISCUSSIONS AND CONCLUSIONS **76**
7.1 Discussions and Conclusions 76
7.2 Suggestions 77
7.3 Publications 77
7.3.1 International Conference 77
7.3.2 International Journal 77

REFERENCES **78**

APPENDICES **87**
A Big-O Analysis 88
B Publications (International Conference) 90
C Publications (International Journal) 96

CURRICULUM VITAE **120**

LIST OF TABLES

TABLE		PAGE
6.1	The relative length of each segment in the human model	53
6.2	Attributes of four test image sequences	60
6.3	Averaged per-joint error distances (in centimeters) from different approaches	61
6.4	The error distances (in centimeters) of our smoothness function	74

LIST OF FIGURES

FIGURE		PAGE
1.1	Reconstruction of a 3D human pose from 2D point correspondences	2
1.2	Problem of non-uniqueness of solutions in 3D human body pose tracking	3
1.3	High dimensional space of human representation	3
1.4	Problems inherit the monocular approach	4
1.5	An example of the lost tracking	5
2.1	Human model representation	8
3.1	Graphical model	15
3.2	Belief propagation on pair-wise Markov Random Fields	16
3.3	Belief propagation	17
3.4	Kernel density estimation	19
3.5	Mode seeking by Mean Shift concept	20
3.6	Probability surface and motion of data points toward mode value of Quick Shift	21
3.7	Mode seeking by Quick Shift	22
4.1	Overview of the proposed method.	25
4.2	The 3D skeleton human model in this dissertation	26
4.3	Initial posture in the first frame	27
5.1	Overview of part-based approach	29
5.2	The graphical model in our approach	32
5.3	Our motion model process	34
5.4	A comparison of moving toward the optimal solution	36
5.5	Euclidean distance between the connected points of body parts	38
5.6	Different binary occlusion masks	39
5.7	Sample of results in 2D human body tracking	40
5.8	Mode seeking by MSBP	41
5.9	Mode seeking by QSBP	42
5.10	Prediction by model video	43
5.11	Some results of human body tracking by using model video in sample prediction	46
5.12	A performance comparison between the proposed method and BP and MSBP	47
6.1	The scaled-orthographic camera model of two pairs of corresponding points	50
6.2	The relationship of real-word, reference and image coordinate system	52
6.3	The perspective camera model of two pairs of corresponding points	53
6.4	Initial posture in the first frame that two joints are not in the same plane	55
6.5	The intersection points between a sphere and a line	57
6.6	Scaling factor	57
6.7	Joint angle limitation	58
6.8	Flow chart of the multiple hypothesis algorithm	59
6.9	Samples of ground truth and results from our approach in synthesized data	61
6.10	Samples of results from Taylor'method and Remondino and Roditakis'method	62

6.11 Some results for the walking scene 63
6.12 Some results from the ball throwing scene 64
6.13 Some results on the back flip scene 65
6.14 The error distances of walking, throwing and back-flipping 66
6.15 Some results on the first image sequence of aerobics style activities 67
6.16 Some results on the second image sequence of aerobics style activities 68
6.17 Some results on walking image sequence 69
6.18 Trajectories of the left wrist in the walking 71
6.19 Trajectories of the left wrist in the ball-throwing 72
6.20 Trajectories of the left wrist in the back-flipping sequences 73
6.21 Error distance due to errors in difference reference distance determination 74
6.22 Average error distance due to errors in 2D joint locations 75

LIST OF SYMBOLS

SYMBOL		UNIT
$\mathbf{x}$	Hidden node	-
$\mathbf{z}$	image observation node	-
$\mathbf{E}$	Set of edges in graphical model	-
$\mathbf{G}$	Graph set	-
$\mathbf{V}$	Set of nodes in graphical model	-
$\mathbf{X}$	Hidden node set	-
$\mathbf{Z}$	Corresponding observation set	-
$\phi(\mathbf{x}_i, \mathbf{z}_i)$	Observation function	-
$\psi(\mathbf{x}_i, \mathbf{x}_j)$	Potential function	-
$p^n(\mathbf{x}_i\|\mathbf{Z})$	Marginal probability	-
$m_{ji}^n(\mathbf{x}_i)$	Message	-
$\Gamma(j)\backslash i$	Neighboring nodes of j expect node i	-
$D(\mathbf{x}_i, \mathbf{x}_j)$	Distance	-
y_i^t	Updated position	-
w_i	Region overlapping	-
$n_{i,o}$	Number of pixels in overlapping region	pixels
$n_{i,p}$	Number of pixels in projected region	pixels
υ	Standard deviation of observation function	-
σ	Standard deviation of potential function	-
R	Reference distance	-
s_a	Scaling factor	-
(x_a, y_a, z_a)	Coordinate point around reference plane	-
(X_a, Y_a, Z_a)	real-world Coordinate point	-
(x'_a, y'_a)	Coordinate point on image	-
ε_{ab}	Human body segment	-
l_{ab}	Segment length around image plane	-
L_{ab}	Segment length in real-world Coordinate	-
$\theta_i(t)$	Possible configurations	-
Θ_i^t	Pose sequence	-
E_i^t	Fitness function	-
$\Delta(\theta_i(t))$	Sum of distance to the origin	-

LIST OF TECHNICAL VOCABULARY AND ABBREVIATIONS

2D	=	Two-Dimensional
3D	=	Three-Dimensional
BNs	=	Bayesian networks
DOFs	=	Degrees of Freedom
EM	=	Expectation-maximization
HSI	=	Hue-Saturation-Lightness
MHT	=	Multiple Hypothesis Tracking
MRFs	=	Markov Random Fields
MS	=	Mean Shift
MSBP	=	Mean Shift Belief Propagation
NBP	=	Non-parametric Belief Propagation
QSBP	=	Quick Shift Belief Propagation
RGB	=	Red-Green-Blue