



University of Technology, Sydney

Faculty of Engineering and Information Technology  
SCHOOL OF COMPUTING AND COMMUNICATIONS

**PhD Thesis**

---

---

*Tracking and Fine-Grained Activity  
Recognition in Depth Videos*

---

---

Prepared By: SARI AWWAD

Principal Supervisor: Prof. Massimo Piccardi

Co-Supervisor: Dr. Richard Xu

NOVEMBER, 2016

## **Certificate of Authorship and Originality**

Title: **Tracking and Fine-Grained Activity Recognition in Depth Videos**

Author: **Sari Awwad**

Date: **November , 2016**

Degree: **PhD**

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of author

## Acknowledgements

Foremost, I would like to express my sincere gratitude to my principal supervisor, Professor Massimo Piccardi for the continuous support in my PhD study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better supervisor and mentor for my PhD study.

My sincere thanks also go to my home country “Jordan”, my sponsor “The Hashemite University” and Dr. Ahmad Ootom for giving me the greatest opportunity to complete my PhD with a greatest supervisor, and my thanks also goes to Dr. Bashar and Dawacom company.

I would like to thank my mother for her continuous prayer for me, and thanks for my brothers: Osama, Samer, Saed, Sameh, and my lovely sister Joman for their continuous supports.

I thank my colleagues: Dr. Shaukat Abidi, Mrs. Fairouz Hussein, Dr. Ava Bargi and Dr. Ehsan Zare Borzeshi, they are a great group, thank you all.

To a wonderful companion, great gratitude to my wife, Lina. Her support, encouragement, quiet patience and unwavering love were undeniably the bedrock upon which the past fourteen years of my life have been built. Her tolerance of my occasional vulgar moods is a testament in itself of her unyielding devotion and love.

Also, I thank Lina’s parents, Mr. Asad and Mrs. Somayya, for their unending encouragement and support.

To the lovely kids, my daughters Farah and Aseel, and my son Osama who have decorated my life and made it full of happiness and joy.

Finally, this thesis is dedicated to the soul of my father, may Allah forgive him and grant him his highest paradise (Ameen).

Sari Awwad.

November, 2016, Sydney

# *Abstract*

Tracking and activity recognition in video are arguably two of the most active topics within the field of computer vision and pattern recognition. Historically, tracking and activity recognition have been performed over conventional video such as color or grey-level frames, either of which contains significant clues for the identification of targets. While this is often a desirable feature within the context of video surveillance, the use of video for activity recognition or for tracking in privacy-sensitive environments such as hospitals and care facilities is often perceived as intrusive.

For this reason, this PhD research has focused on providing tracking and activity recognition solely from *depth* videos which offer a naturally privacy-preserving visual representation of the scene at hand. Depth videos can nowadays be acquired with inexpensive and highly-available commercial sensors such as Microsoft Kinect and Asus Xtion. The two main contributions of this research have been the design of a specialised tracking algorithm for tracking in depth data, and a fine-grained activity recognition approach for recognising activities in depth video. The proposed tracker is an extension of the popular Struck algorithm, an approach that leverages a structural support vector machine (SVM) for tracking. The main contributions of the proposed tracker include a dedicated depth feature based on local depth patterns, a heuristic for handling view occlusions in depth frames, and a technique for keeping the number of support vectors within a given budget, so as to limit computational costs. Conversely, the proposed fine-grained activity recognition approach leverages multi-scale depth measurements and a Fisher-consistent multi-class SVM. In addition to the novel approaches for tracking and activity recognition, in this thesis we have canvassed and developed a practical computer vision application for the detection of hand hygiene at a hospital. This application was developed in collaboration with clinical researchers from the Intensive Care Unit of Sydney's Royal Prince Alfred Hospital. Experiments presented through the thesis confirm that the proposed approaches are effective, and either outperform the state of the art or significantly

reduce the need for sensor instrumentation. The outcomes of the hand-hygiene detection were also positively received and assessed by the clinical research unit.

# Contents

<b>Abstract</b>	<b>i</b>
<b>Contents</b>	<b>iii</b>
<b>List of Figures</b>	<b>vi</b>
<b>List of Tables</b>	<b>ix</b>
<b>Abbreviations</b>	<b>x</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background and Motivation . . . . .	1
1.2 Research Questions and Main Contributions . . . . .	6
1.3 Significance . . . . .	8
1.4 Thesis Structure . . . . .	9
<b>2 Literature Review and Background</b>	<b>12</b>
2.1 Object Tracking . . . . .	13
2.1.1 Tracking Algorithms . . . . .	15
2.1.1.1 Detection-Based Tracking and Kernels . . . . .	15
2.1.1.2 Blob Tracking . . . . .	18
2.1.1.3 Kalman Filter . . . . .	20
2.1.1.4 Particle Filters . . . . .	22
2.2 Feature Descriptors . . . . .	26
2.2.1 Global descriptors . . . . .	27
2.2.1.1 Grid-based global descriptors . . . . .	27
2.2.1.2 Shape-contour based global descriptor . . . . .	28
2.2.1.3 Spatio-temporal based global descriptors . . . . .	29
2.2.2 Local descriptors . . . . .	31
2.2.2.1 Grid-based local descriptors . . . . .	33
2.2.2.2 Texture-based local descriptors . . . . .	35

2.2.2.3	Space-time interest points local descriptors . . . .	36
2.2.3	Depth Features . . . . .	38
2.3	Detecting events and activities in a video . . . . .	40
2.3.1	Human-centred action detection . . . . .	41
2.3.2	The “blind” approach . . . . .	41
2.3.3	Fine-grained activity recognition . . . . .	42
2.3.3.1	Object localization . . . . .	43
2.3.3.2	Classification approaches in fine-grained activity recognition . . . . .	44
2.3.4	Role of computer vision in various fields of medicine . . . .	45
2.4	Support vector machines . . . . .	45
2.4.1	Binary SVM . . . . .	46
2.4.2	Kernels: SVM from Linear to Nonlinear Classifiers . . . .	48
2.4.3	Multi-Class SVM . . . . .	50
2.4.4	Structural SVM . . . . .	51
<b>3</b>	<b>Tracking in depth videos</b>	<b>56</b>
3.1	Introduction and Background . . . . .	56
3.2	Related work . . . . .	57
3.3	The Struck tracker: overview . . . . .	59
3.4	Extensions for depth tracking . . . . .	63
3.4.1	Local depth features for tracking . . . . .	63
3.4.2	Support vector removal based on prototype selection . . . .	64
3.4.3	Occlusion handling . . . . .	68
3.5	Experiments . . . . .	69
3.5.1	Datasets . . . . .	69
3.5.2	Experimental results . . . . .	70
<b>4</b>	<b>Fine-Grained Activity Recognition in Depth Videos</b>	<b>76</b>
4.1	Introduction . . . . .	76
4.2	Background and related work . . . . .	77
4.3	Proposed Approach . . . . .	79
4.3.1	The Local Depth Feature: LDPT . . . . .	80
4.3.2	Feature Encoding . . . . .	82
4.3.3	Multi-Class Classification by M-SVM <sup>2</sup> . . . . .	83
4.4	Experiments . . . . .	85
4.4.1	Dataset . . . . .	85
4.4.2	Features extraction and classification . . . . .	86
4.4.3	State of the Art on the Dataset . . . . .	87
4.4.4	Experimental Results and Discussion . . . . .	88

---

<b>5</b>	<b>Automated Hand Hygiene Detection</b>	<b>91</b>
5.1	Introduction and background . . . . .	91
5.1.1	Study Objectives . . . . .	94
5.2	Methods . . . . .	94
5.2.1	Simulation of the clinical environment and the first moment of hand hygiene . . . . .	94
5.2.2	Capture and processing of RGB and depth images . . . . .	95
5.2.3	Maintenance of privacy during development of the image analysis . . . . .	95
5.2.4	Outcome measures and diagnostic accuracy . . . . .	97
5.3	Hand Hygiene approach and experiments . . . . .	98
5.3.1	Dataset . . . . .	98
5.3.2	Hand Hygiene Events . . . . .	98
5.3.3	Computer vision techniques for detection of dispensing alcohol- based hand rub (Event 1A) . . . . .	101
5.3.4	Computer vision techniques for detection hand rubbing (Event 1B) . . . . .	101
5.3.5	Computer vision techniques for detection of touching the patient (Event 2) . . . . .	102
5.3.6	Hand Hygiene Detection Experimental results . . . . .	103
5.4	Discussion . . . . .	104
<b>6</b>	<b>Conclusion and Future Work</b>	<b>107</b>
	<b>Bibliography</b>	<b>110</b>



# List of Figures

1.1	Depth frame example from a simulated hospital scenario. . . . .	2
1.2	Depth frame examples from a simulated hospital scenario, the Princeton Tracking Benchmark, and "50 salads" datasets. . . . .	5
1.3	Thesis Structure . . . . .	11
2.1	Tracking Stages in Computer Vision (Yang et al., 2011) . . . . .	14
2.2	Tracking by detection example; using super pixel-based discriminative appearance where (a) a new frame at time $t$ . (b) surrounding region of the target in the last frame, i.e., at state $X_t$ . (c) segmentation result of (b). (d) the computed confidence map of super pixels The super pixels coloured with red indicate a strong connection to the target, and those coloured with dark blue indicate a strong connection to the background. (e) the confidence map of the entire frame. (f), (g) and (h), (i) show two target candidates with high and low confidence, respectively. Model) (Wang et al., 2011) . . . . .	18
2.3	Blob tracking example: Two-dimensional blob tracking by applying the mean-shift algorithm to an image where pixel values represent likelihood of being on the tracked object (Collins, 2003) . . . . .	19
2.4	Particle tracking example: by using appearance-adaptive models (top row is the adaptive velocity model and the bottom row is the zero-velocity model) (Zhou et al., 2004) . . . . .	25
2.5	Bounding box and its partitions by dividing the ROI into horizontal slices (Danafar and Gheissari, 2007) . . . . .	28
2.6	Running shape: the kinematic constrains for knee and elbows angles are used for feature detection . . . . .	29
2.7	Movement recognition based on a sequence of space-time shapes in a video (Deng et al., 2010) . . . . .	30
2.8	(a) Space-time volume of stacked silhouettes, (b) Motion history volumes. Figures reprinted from (Poppe, 2010) . . . . .	31
2.9	Examples of local descriptors within a still image (Mikolajczyk and Tuytelaars, 2015) . . . . .	32

2.10	An example on space-time features represented by three frames of a drinking action. Three types of features with different arrangement of histogram blocks. Histograms for composed blocks (Temp-2, Spat-4) are concatenated into a single feature vector. (Ikizler and Duygulu, 2009) . . . . .	34
2.11	An example on texture features represented by gradients, where (a to c) are represent training stages and (d to g) represent testing stages . (Dalal and Triggs, 2005) . . . . .	36
2.12	Examples of spatio-temporal interest points for a “walking” action: (a) 3D plot of leg pattern shown upside down to simplify interpretation; (b) spatio-temporal interest points detection overlayed on walking legs (Laptev, 2005) . . . . .	37
2.13	Space-time cells of a depth sequence of the forward kick action. For each time segment, the frames are placed together in the same space. (Vieira et al., 2012) . . . . .	39
2.14	Example of HON4D descriptors for each cell and their concatenation. (Oreifej and Liu, 2013) . . . . .	40
2.15	Binary SVM(Meyer and Wien, 2015) . . . . .	47
2.16	An example of non-linear classifiers using the Gaussian kernel (Stanevski and Tsvetkov, 2005) . . . . .	50
2.17	An example of natural language parsing by structural SVM from (Le Nguyen et al., 2005) . . . . .	53
3.1	The main steps of Struck: a) the estimated ground-truth bounding box at frame $i$ (a positive support vector); b) other bounding boxes around the ground truth (negative support vectors); c) the score, $w^\top \phi(x, y)$ , of all bounding boxes is computed; d) the constraints in equation (3.2) impose that the score of the true displacement, $y_i$ , is greater than that of any other displacement, $y \neq y_i$ , by an amount set by the chosen loss function, $\Delta(y_i, y)$ . At its turn, $\Delta(y_i, y)$ is chosen to be complementary to the overlap between bounding boxes $y_i$ and $y$ . . . . .	61
3.2	Examples of occlusion handling in A) the hospital simulation and B) PTB datasets. . . . .	67
3.3	Cases of success and failure for the proposed tracker and the original Struck tracker. . . . .	73
3.4	Comparison between the proposed tracker and the original Struck tracker with various features; A) by varying the search radius; B) by varying the budget size. . . . .	74
4.1	Examples of depth frames from the “50 Salad” dataset. . . . .	79
4.2	Overview of the proposed approach. . . . .	80

4.3	The hierarchy of cells (smallest), depth patterns (intermediate; numbered from 1 to 12) and LDPTs (largest). This figure should be viewed in color. . . . .	81
4.4	Confusion matrix for the proposed method. Rows and columns represent ground-truth and predicted class labels, respectively. Numbers represent frequencies in percentages and the cells' gray-levels visually encode the frequencies from 0% = black to 100% = white. .	90
5.1	The Five Moments of Hand Hygiene . . . . .	93
5.2	An example of acquisition software that includes depth, RGB, and skeleton . . . . .	96
5.3	An example of the images that were used in this work. For the sake of visualization, the small RGB patch is superimposed to the bottle area . . . . .	96
5.4	Event 1 of Hand Hygiene detection approach . . . . .	99
5.5	Event 2 of Hand Hygiene detection approach . . . . .	100
5.6	Background removal procedure . . . . .	102

# List of Tables

2.1	Most used kernels functions (Stanevski and Tsvetkov, 2005) . . . .	49
3.1	Accuracy comparison for the proposed tracker and other trackers on the Princeton Tracking Benchmark. . . . .	71
3.2	Comparison of average accuracy with different prototype selection techniques and distances. . . . .	71
4.1	Dataset activities and video frame counts . . . . .	86
4.2	Recall, precision and F1 score for each activity class with the proposed approach. . . . .	89
4.3	Comparison of recognition performance. . . . .	89
4.4	Recall and precision for the proposed method with and without PCA. . . . .	90
5.1	Accuracy of Moment 1 monitoring. . . . .	103

# Abbreviations

<b>DP</b>	<b>Depth Patterns</b>
<b>EM</b>	<b>Expectation Maximization</b>
<b>FV</b>	<b>Feature Vector</b>
<b>HAI</b>	<b>Healthcare Associated Infections</b>
<b>KPF</b>	<b>Kalman Particle Filter</b>
<b>LDP</b>	<b>Local Depth Patterns</b>
<b>LDPT</b>	<b>Local Depth Patterns for Tracking</b>
<b>MEI</b>	<b>Motion Energy Images</b>
<b>MHT</b>	<b>Multilple Hypothesis Tracking</b>
<b>PCA</b>	<b>Principal Component Analysis</b>
<b>PTB</b>	<b>Princeton Tracking Benchmark</b>
<b>RFID</b>	<b>Radio Frequency Identification Device</b>
<b>ROI</b>	<b>Region Of Interest</b>
<b>RPAH</b>	<b>Royal Prince Alfred Hospital</b>
<b>SMO</b>	<b>Sequential Minimal Optimization</b>
<b>SVM</b>	<b>Support Vector Machine</b>
<b>WHO</b>	<b>World Health Organisation</b>