

Towards Practical Automated Human Action Recognition

A Thesis Submitted for the Degree of Doctor of Philosophy

By

Zia Moghaddam

in

School of Computing and Communications

Faculty of Engineering and Information Technology

UNIVERSITY OF TECHNOLOGY, SYDNEY (UTS)

AUSTRALIA

March 2012

CERTIFICATE

Date: **1st March 2012**

Author: **Zia Moghaddam**

Title: **Towards Practical Automated Human Action Recognition**

Degree: **Ph.D.**

I certify that this thesis has not already been submitted for any degree and is not being submitted as part of candidature for any other degree.

I also certify that the thesis has been written by me and that any help that I have received in preparing this thesis, and all sources used, have been acknowledged in this thesis.

Production Note:
Signature removed
prior to publication.

Signature of Author

*This thesis is dedicated to my wife, Hamideh,
and my two sons, Mohammad Hossein and Mohammad Sadra*

Acknowledgments

First of all, I am very appreciative of my supervisor, Prof. Massimo Piccardi, for his complete guidance and direction through my PhD, for being always willing to answer questions, for his understanding of my family situation, and for his full support and encouragement throughout all years of my study.

I also thank my friends and colleagues at the School and our research group, iNEXT, for their kindness and support.

Furthermore, this author is grateful to SenSen Pty Ltd for partially supporting the work presented in this thesis.

Finally, thanks to my wife, Hamideh, for her tremendous support and encouragement which allowed me to pursue my research.

Author's Publications

The following publications have been produced during the course of this thesis:

Conference papers:

- **Z. Moghaddam** and M. Piccardi, "Deterministic Initialization of Hidden Markov Models for Human Action Recognition", *Proc. of the Digital Image Computing: Techniques and Applications (DICTA)*, 2009, pp. 188-195.
- **Z. Moghaddam** and M. Piccardi, "Histogram-Based Training Initialisation of Hidden Markov Models for Human Action Recognition", *Proc. of the IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2010, pp. 256-261.
- **Z. Moghaddam** and M. Piccardi, "Human Action Recognition with MPEG-7 Descriptors and Architectures", *Proc. of the 1st ACM International Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Streams (ARTEMIS)*, 2010, pp. 63-68.
- **Z. Moghaddam** and M. Piccardi, "Robust Density Modelling using the Student's t-distribution for Human Action Recognition", *Proc. of the 18th IEEE International Conference on Image Processing (ICIP)*, 2011, pp. 3261-3264.
- O.P. Concha, R. Xu, **Z. Moghaddam** and M. Piccardi, "HMM-MIO: An Enhanced Hidden Markov Model for Action Recognition," *Proc. of the 3rd International Workshop on Machine Learning for Vision-based Motion Analysis (MLvMA) in conjunction with CVPR 2011*, pp. 62-69.

Table of Contents

LIST OF FIGURES	XI
LIST OF TABLES	XIV
ABSTRACT	1
CHAPTER 1: INTRODUCTION	5
1.1. OVERVIEW	5
1.2. RESEARCH OBJECTIVES	8
1.3. THESIS STRUCTURE	8
CHAPTER 2: AUTOMATED HUMAN ACTION RECOGNITION	10
2.1. INTRODUCTION	10
2.2. CHALLENGES OF HUMAN ACTION RECOGNITION	11
2.3. HUMAN ACTION VIDEO DATASETS	12
2.3.1. KTH dataset.....	12
2.3.2. Weizmann dataset	13
2.3.3. INRIA XMAS dataset	14
2.3.4. UCF datasets	15
2.3.5. Hollywood dataset	16
2.3.6. MuHAVi dataset.....	17
2.3.7. Other datasets	18
2.3.8. Summary of datasets.....	18
2.4. FEATURE SET EXTRACTION	19
2.4.1. Global features	20
2.4.1.1. Silhouette.....	20
2.4.1.2. Contour	20

2.4.1.3. Projection histograms	22
2.4.1.4. Optical flow	22
2.4.1.5. Space-time volumes (STV)	23
2.4.1.6. Grid-based global features	23
2.4.2. Local features.....	23
2.4.3. Summary of feature sets	24
2.5. ACTION CLASSIFICATION	25
2.5.1. Direct classification	25
2.5.2. Recognising actions directly in the time domain	26
2.5.3. Temporal graphical models	26
2.5.3.1. Generative models.....	27
2.5.3.2. Discriminative models.....	28
2.5.3.3. Parameter initialisation issues in model training	29
2.5.4. Summary of action classification approaches	29
2.6. OTHER POSSIBLE STEPS FOR HUMAN ACTION RECOGNITION	30
2.6.1. Foreground extraction.....	30
2.6.2. Blob tracking.....	31
2.6.3. Time segmentation	31
2.6.4. Human action classification in real-time scenario	31
2.7. SUMMARY	32
CHAPTER 3: DENSITY ESTIMATION IN HIDDEN MARKOV MODELS.....	33
3.1. INTRODUCTION	33
3.2. PROBABILITY DENSITY FUNCTION.....	34
3.2.1. Mixture distribution	35
3.3. GAUSSIAN DISTRIBUTION.....	35
3.3.1. Univariate single component Gaussian distribution	35
3.3.2. Multivariate single component Gaussian distribution	36
3.3.3. Gaussian Mixture Model (GMM)	37
3.4. DENSITY ESTIMATION	38

3.4.1. Maximum likelihood density estimation (MLE)	39
3.4.2. MLE of a single Gaussian parameters.....	40
3.4.3. Expectation Maximisation (EM) algorithm for MLE	40
3.4.4. ML estimation of GMM parameters using the EM algorithm.....	42
3.5. HIDDEN MARKOV MODEL (HMM).....	44
3.5.1. HMM parameters.....	45
3.5.2. HMM basic problems	46
3.5.3. HMM evaluation	46
3.5.3.1. Forward procedure	47
3.5.3.2. Backward procedure	48
3.5.3.3. Forward –backward algorithm.....	49
3.5.4. Action classification using HMM evaluation	50
3.5.5. ML estimation of HMM parameters using EM algorithm	50
3.5.6. Correlation between parameter estimation and action classification	54
3.6. SUMMARY	55
CHAPTER 4: HMM TRAINING INITIALISATION FOR ACTION RECOGNITION.....	56
4.1. INTRODUCTION	56
4.2. LOCAL MAXIMA ISSUE IN THE LIKELIHOOD OF MIXTURE DISTRIBUTION	57
4.3. PARAMETERS INITIALISATION IN HMM TRAINING	59
4.3.1. Random initialisation method	60
4.3.1.1. The reference cluster initialisation method.....	60
4.3.1.2. The reference component dispatching method.....	61
4.4. ONE-OFF INITIALISATION METHODS	61
4.4.1. Time segmentation-based approach.....	62
4.4.1.1. The <i>average of training instances</i> cluster initialisation method.....	63
4.4.1.2. The <i>average of GMM parameters</i> cluster initialisation method	64
4.4.1.3. The <i>average of HMM parameters</i> cluster initialisation method	64
4.4.2. Histogram-based approach	65
4.5. COMPONENT DISPATCHING METHODS	68

4.5.1. The <i>nearest neighbours</i> method	69
4.5.2. The <i>feature sorting</i> method	70
4.6. UTILISED FEATURE SETS	70
4.6.1. The projection histograms feature set	71
4.6.2. The <i>sectorial extreme points</i> feature set	72
4.6.2.1. Feature processing by standardisation.....	74
4.6.3. Required processing time and real-time surveillance scenario	75
4.6.4. The STIP feature set	76
4.7. EXPERIMENTS ON HMM TRAINING INITIALISATION METHODS.....	77
4.7.1. Experiments over the Weizmann video dataset	77
4.7.1.1. Experiment on feature processing	78
4.7.1.2. Experiment on time segmentation-based cluster initialisation methods	79
4.7.1.3. Experiment on component dispatching methods	80
4.7.2. Experiments over the MuHAVi video dataset	82
4.7.2.1. MuHAVi dataset with automated foreground segmentation	82
4.7.2.2. Experiment on cluster initialisation methods	84
4.7.2.3. Divergence analysis.....	86
4.7.2.4. Experiment on component dispatching methods	89
4.7.3. Experiments over the Hollywood video dataset	90
4.8. SUMMARY	93
CHAPTER 5: STUDENT'S T-DISTRIBUTION FOR ROBUST DATA MODELLING	95
5.1. INTRODUCTION	95
5.2. THE SINGLE COMPONENT T-DISTRIBUTION DENSITY	96
5.2.1. Univariate single component <i>t</i> -distribution.....	97
5.2.2. Multivariate single component <i>t</i> -distribution.....	99
5.2.3. ML parameter estimation of a single component <i>t</i> -distribution.....	99
5.3. THE MIXTURE OF T-DISTRIBUTIONS (SMM)	101
5.3.1. ML parameter estimation of a SMM	101
5.4. THE HIDDEN MARKOV MODEL (HMM) WITH SMM DENSITY MODELLING	104

5.4.1. ML parameter estimation of an HMM-SMM	105
5.5. EXPERIMENTS OF ROBUSTNESS USING HMM-SMM	109
5.5.1. Experiment of using HMM-SMM with estimation of v	109
5.5.2. Experiments using HMM-SMM with fixed v	113
5.6. SUMMARY	115
CHAPTER 6: USING MPEG-7 TO STANDARDISE ACTION RECOGNITION	117
6.1. INTRODUCTION	117
6.2. MULTIMEDIA STANDARDISATION IN VIDEO SURVEILLANCE	118
6.3. MPEG-7 STANDARD FOR MULTIMEDIA CONTENT DESCRIPTION	119
6.3.1. MPEG-7 visual descriptors	121
6.3.1.1. Visual colour descriptors	121
6.3.1.2. Visual texture descriptors.....	122
6.3.1.3. Visual shape descriptors.....	122
6.3.1.4. Visual motion descriptors.....	123
6.3.2. MPEG-7 XM software	123
6.4. SOME EXAMPLE WORKS ON USING MPEG-7 FOR VIDEO SURVEILLANCE	125
6.4.1. "The Use of MPEG-7 for Intelligent Analysis and Retrieval in Video Surveillance"	125
6.4.2. "Evaluation of MPEG7 Colour Descriptors for Video Surveillance Retrieval"	125
6.4.3. "On the Use of MPEG7 for Visual Surveillance"	127
6.4.4. "Human Object Tracking Algorithm with Human Colour Structure Descriptor for Video Surveillance Systems"	128
6.4.5. "A framework for integrating MPEG-7 knowledge templates into video surveillance applications"	129
6.4.6. "Chaos-Synchronisation Based Representation of Objects and Events From MPEG-7 Low-Level Descriptors"	130
6.4.7. "Human Body Posture Recognition using MPEG-7"	131
6.5. USING MPEG-7 DESCRIPTORS FOR HUMAN ACTION RECOGNITION	131
6.5.1. Inadequacy of current MPEG-7 descriptors.....	132

6.5.2. Action recognition steps using MPEG-7	133
6.6. PROPOSED MPEG-7 DESCRIPTORS AND ARCHITECTURES FOR ACTION RECOGNITION	135
6.6.1. <i>Server-Intensive</i> architecture	135
6.6.2. <i>Client-Intensive</i> architecture	136
6.6.3. Architecture evaluation	139
6.7. SUMMARY	140
CHAPTER 7: CONCLUSIONS AND FUTURE WORK.....	141
7.1. CONCLUSIONS.....	141
7.2. FUTURE WORK	143
BIBLIOGRAPHY	145

List of Figures

FIGURE 2.1: SAMPLE FRAMES OF VARIOUS ACTION CLASSES FROM THE KTH DATASET (REPRINTED FROM [17]).	13
FIGURE 2.2: SAMPLE FRAMES OF VARIOUS ACTIONS FROM THE WEIZMANN DATASET.	13
FIGURE 2.3: SAMPLE FRAMES OF VARIOUS ACTIONS FROM THE INRIA XMAS DATASET.	14
FIGURE 2.4: SAMPLE FRAMES OF VARIOUS ACTIONS FROM THE “UCF SPORTS ACTION DATASET” (REPRINTED FROM [19]).	15
FIGURE 2.5: SAMPLE FRAMES OF VARIOUS ACTIONS FROM THE “UCF50 DATASET” (REPRINTED FROM [19]).	16
FIGURE 2.6: SAMPLE FRAMES OF VARIOUS ACTIONS FROM THE “HOLLYWOOD-2 DATASET” (REPRINTED FROM [22]).	17
FIGURE 2.7: SAMPLE FRAMES OF VARIOUS ACTIONS FROM THE “MUHAVI DATASET” (REPRINTED FROM [2]).	17
FIGURE 2.8: SAMPLES OF MANUALLY ANNOTATED SILHOUETTES FROM MUHAVI-MAS DATASET (REPRINTED FROM [2]).	18
FIGURE 2.9: THE ORIGINAL STAR SKELETON FEATURE SET WITH VARIOUS NUMBERS OF EXTRACTED EXTREMAL POINTS (REPRINTED FROM [26]). (A) THREE EXTREMAL POINTS. (B) FIVE EXTREMAL POINTS.	21
FIGURE 2.10: THE PROJECTION HISTOGRAMS FEATURE SET (REPRINTED FROM [28]).	22
FIGURE 2.11: A TIME WARPING BETWEEN TWO SEQUENCES IN DTW TECHNIQUE (REPRINTED FROM [50]).	26
FIGURE 2.12: A STATE TRANSITION AND OBSERVATION DIAGRAM.	27
FIGURE 2.13: HUMAN ACTION RECOGNITION STEPS.	30
FIGURE 2.14: TIME SEGMENTATION AND ACTION CLASSIFICATION STEPS.	31
FIGURE 3.1: (A) THE CUMULATIVE DENSITY FUNCTION (CDF) AND, (B) THE PROBABILITY DENSITY FUNCTION (PDF) OF A GAUSSIAN DISTRIBUTION.	34
FIGURE 3.2: THE PROBABILITY DENSITY FUNCTION (PDF) OF A 2-DIMENSIONAL GAUSSIAN DISTRIBUTION.	36
FIGURE 3.3: VARIOUS COVARIANCE MATRICES.	37
FIGURE 3.4: GAUSSIAN MIXTURE MODEL FOR MULTIMODAL GAUSSIAN DISTRIBUTIONS.	38
FIGURE 3.5: MULTIPLE MAXIMA FOR THE LIKELIHOOD FUNCTION OF A GAUSSIAN MIXTURE.	41
FIGURE 4.1: MULTIPLE MAXIMA FOR THE LIKELIHOOD FUNCTION OF A GAUSSIAN MIXTURE.	57
FIGURE 4.2: THE HISTOGRAM OF A SAMPLE SET FOR A MIXTURE OF GAUSSIANS DISTRIBUTION.	58
FIGURE 4.3: TWO RESULTING PDFS USING MLE FOR THE LIKELIHOOD FUNCTION OF A GMM.	58

FIGURE 4.4: HMM PARAMETER INITIALISATION IN MURPHY'S TOOLBOX USING RANDOM INITIAL CENTRES.	60
FIGURE 4.5: THE AVERAGE OF TRAINING INSTANCES TIME SEGMENTATION-BASED CLUSTER INITIALISATION METHOD.	63
FIGURE 4.6: THE AVERAGE OF GMM PARAMETERS TIME SEGMENTATION-BASED CLUSTER INITIALISATION METHOD.	64
FIGURE 4.7: THE AVERAGE OF HMM PARAMETERS HMM PARAMETERS INITIALISATION METHOD.	65
FIGURE 4.8: THE HISTOGRAM-BASED CLUSTER INITIALISATION.	65
FIGURE 4.9: HISTOGRAM-BASED INITIALISATION OF CLUSTERS' CENTRES.	66
FIGURE 4.10: AN EXAMPLE OF CLUSTER'S CENTRE CALCULATION USING MARGINAL HISTOGRAM OF OBSERVATIONS.	68
FIGURE 4.11: THE PROJECTION HISTOGRAMS FEATURE SET FOR ONE FRAME OF THE WEIZMANN ACTION VIDEO DATASET. (A) FRAME NO 11 OF VIDEO SEQUENCE 'DARIA-JACK'. (B) CALCULATED PROJECTION HISTOGRAMS.	71
FIGURE 4.12: THE EXTRACTED SECTORIAL EXTREME POINTS FEATURES FOR ONE FRAME OF THE WEIZMANN DATASET.	73
FIGURE 4.13: THE TIME-SEQUENTIAL VALUES OF ONE FEATURE FOR 10 ACTIONS PERFORMED BY ONE SUBJECT OF THE WEIZMANN ACTION DATASET.	74
FIGURE 4.14: THE SPACE-TIME INTEREST POINTS (STIPs) [3].	76
FIGURE 4.15: ACCURACY OF RUNNING THE EXPERIMENT WITH 100 RANDOM STARTS.	79
FIGURE 4.16: EXAMPLES OF FOUR FRAMES AND CORRESPONDING AUTOMATED MASKS FROM THE MUHAVI DATASET.	83
FIGURE 5.1: THE STUDENT'S T-DISTRIBUTION WITH VARIOUS VALUES OF N (REPRINTED FROM [82]).	96
FIGURE 5.2: (A) SIMILAR OBSERVATION FITTING FOR T -DISTRIBUTION (RED CURVE) AND GAUSSIAN (GREEN CURVE, HIDDEN BY THE RED CURVE) WHEN NO OUTLIER PRESENTS. (B) INFLUENCE OF OUTLIERS (REPRINTED FROM [82]).	97
FIGURE 6.1: SCOPE OF MPEG-7 (REPRINTED FROM [85]).	119
FIGURE 6.2: POSSIBLE SCENARIOS OF MPEG-7 APPLICATION (REPRINTED FROM [84]).	120
FIGURE 6.3: DIFFERENTIATING BETWEEN IMAGES USING COLOUR DESCRIPTORS (REPRINTED FROM [85]).	121
FIGURE 6.4: DIFFERENTIATING BETWEEN IMAGES USING TEXTURE DESCRIPTORS (REPRINTED FROM [85]).	122
FIGURE 6.5: IMAGE SIMILARITY MATCHING USING REGION-BASED SHAPE DESCRIPTOR (REPRINTED FROM [84]).	122
FIGURE 6.6: IMAGE SIMILARITY MATCHING USING CONTOUR-BASED SHAPE DESCRIPTOR (REPRINTED FROM [85]).	123
FIGURE 6.7: APPLICATION TYPES IN MPEG-7 XM SOFTWARE. (A) THE EXTRACTION FROM MEDIA SERVER APPLICATION. (B) THE SEARCH-AND-RETRIEVAL CLIENT APPLICATION (REPRINTED FROM [84]).	124
FIGURE 6.8: RETRIEVAL RATE TO MATCH PEOPLE ENTERING AND EXITING A ROOM (REPRINTED FROM [88]).	126
FIGURE 6.9: CORRECT AND INCORRECT MATCHES FOR TWO MPEG-7 COLOUR DESCRIPTORS (REPRINTED FROM [89]).	127
FIGURE 6.10: STEP-BY-STEP RUNNING OF THE HUMAN DETECTION ALGORITHM IN [90] (REPRINTED FROM [90]).	128

FIGURE 6.11: SCENE REGION DESCRIPTION (LEFT) AND OCCLUSION MAP (RIGHT) FOR A CITY VIDEO SURVEILLANCE SCENARIO (REPRINTED FROM [91]).....	130
FIGURE 6.12: TIME SEGMENTATION AND ACTION CLASSIFICATION STEPS	134
FIGURE 6.13: <i>SERVER-INTENSIVE</i> ARCHITECTURE.	135
FIGURE 6.14: AN EXAMPLE OF THE <i>ACTIONCLASS</i> MOTION DESCRIPTOR.....	136
FIGURE 6.15: <i>CLIENT-INTENSIVE</i> ARCHITECTURE.	136
FIGURE 6.16: AN EXAMPLE OF THE <i>OBJECTFEATURES</i> VISUAL DESCRIPTOR USING THE EXTRACTED “ <i>SECTORIAL EXTREME POINTS</i> ” FEATURE SET FOR ONE FRAME OF A VIDEO SEQUENCE IN WEIZMANN DATASET.	137
FIGURE 6.17: AN EXAMPLE OF THE <i>OBJECTFEATURES</i> VISUAL DESCRIPTOR USING THE PROJECTION HISTOGRAMS FEATURE SET FOR ONE FRAME OF A VIDEO SEQUENCE IN WEIZMANN DATASET.	138

List of Tables

TABLE 2.1: SUMMARY OF HUMAN ACTION VIDEO DATASETS.	19
TABLE 2.2: SUMMARY OF FEATURE SETS FOR HUMAN ACTION RECOGNITION.	24
TABLE 2.3: SUMMARY OF MAIN ACTION CLASSIFICATION APPROACHES.	29
TABLE 3.1: OVERALL LOG-LIKELIHOOD AND CORRESPONDING CLASSIFICATION ACCURACY FOR A CROSS-VALIDATION TEST USING HMM-GMM.	55
TABLE 4.1: DIFFERENT VERSIONS OF PROJECTION HISTOGRAM FEATURE SETS.....	72
TABLE 4.2: CLASSIFICATION ACCURACY (%) WITH THE ORIGINAL AND THE STANDARDISED FEATURES.	78
TABLE 4.3: CLASSIFICATION ACCURACY (%) USING RANDOM CENTRES AND VARIOUS TIME SEGMENTATION-BASED CLUSTER INITIALISATION METHODS.....	80
TABLE 4.4: CLASSIFICATION ACCURACY (%) WITH VARIOUS COMPONENT DISPATCHING METHODS.	81
TABLE 4.5: CONFUSION MATRIX (FOR THE AVERAGE OF TRAINING INSTANCES CLUSTER INITIALISATION AND THE FEATURE SORTING COMPONENT DISPATCHING METHODS).	81
TABLE 4.6: NUMBER OF ACTION SAMPLES WITH AUTOMATED MASKS FROM THE MUHAVI DATASET (CAMERA 4).....	84
TABLE 4.7: CLASSIFICATION ACCURACY (%) WITH VARIOUS CLUSTER INITIALISATION METHODS USING THE SECTORIAL EXTREME POINTS FEATURE SET.....	85
TABLE 4.8: CLASSIFICATION ACCURACY (%) WITH VARIOUS CLUSTER INITIALISATION METHODS USING THE PROJECTION HISTOGRAMS FEATURE SET.....	85
TABLE 4.9: DIVERGENCE EXPLOITATION FOR ACCURACY ASSESSMENT WITH VARIOUS CLUSTER INITIALISATION METHODS.	88
TABLE 4.10: ACCURACY COMPARISON (%) BETWEEN COMPONENT DISPATCHING METHODS: THE APPEARANCE ORDER (LEFT COLUMN) AND THE FEATURE SORTING (RIGHT COLUMN), FOR VARIOUS CLUSTER INITIALISATION METHODS.	90
TABLE 4.11: CLASSIFICATION ACCURACY (%) WITH VARIOUS CLUSTER INITIALISATION METHODS USING STIP FEATURE SET WITH THRESHOLD=1E-14.....	92
TABLE 4.12: CLASSIFICATION ACCURACY (%) WITH VARIOUS CLUSTER INITIALISATION METHODS USING STIP FEATURE SET WITH THRESHOLD=1E-9.....	92

TABLE 5.1: CLASSIFICATION ACCURACY (%) FOR HMM-GMM AND HMM-SMM WITH ML-ESTIMATED N OVER THE WEIZMANN DATASET USING THE PROJECTION HISTOGRAMS FEATURE SET.	110
TABLE 5.2: CLASSIFICATION ACCURACY (%) FOR HMM-GMM AND HMM-SMM WITH ML-ESTIMATED N OVER THE MUHAVI DATASET USING THE PROJECTION HISTOGRAMS FEATURE SET.	111
TABLE 5.3: CLASSIFICATION ACCURACY (%) FOR HMM-GMM AND HMM-SMM WITH ML-ESTIMATED N OVER THE WEIZMANN DATASET USING THE <i>SECTORIAL EXTREME POINTS</i> FEATURE SET.	111
TABLE 5.4: CLASSIFICATION ACCURACY (%) FOR HMM-GMM AND HMM-SMM WITH ML-ESTIMATED N OVER THE MUHAVI DATASET USING THE <i>SECTORIAL EXTREME POINTS</i> FEATURE SET.	112

Abstract

Modern video surveillance requires addressing high-level concepts such as humans' actions and activities. Automated human action recognition is an interesting research area, as well as one of the main trends in the automated video surveillance industry. The typical goal of action recognition is that of labelling an image sequence (video) using one out of a set of action labels. In general, it requires the extraction of a feature set from the relevant video, followed by the classification of the extracted features. Despite the many approaches for feature set extraction and classification proposed to date, some barriers for *practical action recognition* still exist. We argue that recognition accuracy, speed, robustness and the required hardware are the main factors to build a practical human action recognition system to be run on a typical PC for a real-time video surveillance application. For example, a computationally-heavy set of measurements may prevent practical implementation on common platforms.

The main focus of this thesis is challenging the main difficulties and proposing solutions towards a practical action recognition system. The main outstanding difficulties that we have challenged in this thesis include 1) initialisation issues with model training; 2) feature sets of limited computational weight suitable for real-time application; 3) model robustness to outliers; and 4) pending issues with the standardisation of software interfaces. In the following, we provide a description of our contributions to the resolution of these issues.

Amongst the different classification approaches for classifying actions, graphical models such as the hidden Markov model (HMM) have been widely exploited by many researchers. Such models include observation probabilities which are generally modelled by mixtures of Gaussian components. When learning an HMM by way of Expectation-Maximisation (EM) algorithms, arbitrary choices must be made for their initial parameters. The initial choices have a major impact on the parameters at convergence and, in turn, on the recognition accuracy. This dependence forces us to repeat training

with different initial parameters until satisfactory cross-validation accuracy is attained. Such a process is overall empirical and time consuming.

We argue that one-off initialisation can offer a better trade-off between training time and accuracy, and as one of the main contributions of this thesis, we propose two methods for deterministic initialisation of the Gaussian components' centres. The first method is a time segmentation-based approach which divides each training sequence into the requested number of clusters (product of the number of HMM states and the number of Gaussian components in each state) in the time domain. Then, clusters' centres are averaged among all the training sequences to compute the initial centre for each Gaussian component. The second approach is a histogram-based approach which tries to initialise the components' centres with the more popular values among the training data in terms of density (similar to mode seeking approaches). The histogram-based approach is performed incrementally, considering each feature at a time. Either centre initialisation approach is followed by dispatching the resulting Gaussian components onto HMM states. The reference component dispatching method exploits the arbitrary order for dispatching. In contrast, we again propose two more intelligent methods based on the effort to put components with closer centres in the same state which can improve the correct recognition rate.

Experiments over three human action video datasets (Weizmann [1], MuHAVi [2] and Hollywood [3]) prove that our proposed deterministic initialisation methods are capable of achieving accuracy above the average of repeated random initialisations (about 1 per cent to 3 per cent in 6 random runs experiment) and comparable to the best. At the same time, one-off deterministic initialisation can save the required training time substantially compared to repeated random initialisations, e.g. up to 83% in the case of 6 runs of random initialisation. The proposed methods are general as they naturally extend to other models where observation densities are conditioned on discrete latent variables, such as dynamic Bayesian networks (DBNs) and switching models.

As another contribution, we propose a simple and computationally lightweight feature set, named *sectorial extreme points*, which requires only 1.6 ms per frame for extraction on a reference PC. We believe a lightweight feature set is more appropriate for the task of action recognition in real-time surveillance applications with the usual requirement of processing 25 frames per second (PAL video

rate). The proposed feature set represents the coordinates of the extreme points in the contour of a subject's foreground mask. The various experiments prove the strength of the proposed feature set in terms of classification accuracy, compared to similar feature sets, such as the star skeleton [4] (by more than 3%) and the well-known projection histograms (up to 7%).

Another main issue in density modelling of the extracted features is the outlier problem. The extraction of human features from videos is often inaccurate and prone to outliers. Such outliers can severely affect density modelling when the Gaussian distribution is used as the model since it is short-tailed and highly sensitive to outliers. Hence, outliers can affect the classification accuracy of the HMM-based action recognition approaches that exploit Gaussian distribution as the base component. In contrast, the Student's t -distribution is more robust to outliers thanks to its longer tail and can be exploited for density modelling to improve the recognition rate in the presence of abnormal data. As another main contribution, we present an HMM which uses mixtures of t -distributions as observation probabilities and apply it for the recognition task. The conducted experiments over the Weizmann and MuHAVi datasets with various feature sets report a remarkable improvement of up to 9% in classification accuracy by using HMM with mixtures of t -distributions instead of mixture of Gaussians. Using our own proposed *sectorial extreme points* feature set, we have achieved the maximum possible classification accuracy (100%) over the Weizmann dataset. This achievement should be considered jointly with the fact that we have used a lightweight feature set.

On a different ground, and from the implementation viewpoint, surveillance software for automated human action recognition requires portability over a variety of platforms, from servers to mobile devices. The current products mainly target low level video analysis tasks, e.g. video annotation, instead of higher level ones, such as action recognition. Therefore, we explore the potential of the MPEG-7 standard to provide a standard interface platform (through descriptors and architectures) for human action recognition from surveillance cameras. As the last contribution of this work, we present two novel MPEG-7 descriptors, one symbolic and the other feature-based, alongside two different architectures: the *server-intensive* which is more suitable for "thin" client devices, such as PDAs and the *client-intensive* that is more appropriate for "thick" clients, such as desktops. We evaluate the proposed descriptors and architectures by way of a scenario analysis.

We believe that through the four contributions of this thesis, human action recognition systems have become more practical. While some contributions are specific to generative models such as the HMM, other contributions are more general and can be exploited with other classification approaches. We acknowledge that the entire area of human action recognition is progressing at an enormous pace, and that other outstanding issues are being resolved by research groups world-wide. We hope that the reader will enjoy the content of this work.