UNIVERSIDAD DE ALCALÁ

ESCUELA POLITÉCNICA SUPERIOR

DEPARTAMENTO DE ELECTRÓNICA

Doctoral Thesis

Automatic body communication extraction through markerless motion capture

Álvaro Marcos-Ramiro

2014

UNIVERSIDAD DE ALCALÁ

ESCUELA POLITÉCNICA SUPERIOR

DEPARTAMENTO DE ELECTRÓNICA

Automatic body communication extraction through markerless motion capture

Author

Álvaro Marcos-Ramiro

Advisors

Daniel Pizarro-Pérez

Marta Marrón-Romera

Daniel Gatica-Pérez

2014

Doctoral Thesis

A mi familia

“Si quieres algo que nunca has tenido, haz algo que nunca hayas hecho.”
Anónimo

Acknowledgements

During the last almost four years I have been lucky to be surrounded by a number of people that made this journey worth experiencing. Without them, it would not have made any sense.

First I must thank my supervisors Daniel Pizarro Pérez, Marta Marrón Romera and Daniel Gatica Pérez for their help and orientation. They have helped me to learn so many things along this time. Thanks also to Javier Macías Guarasa and Manuel Mazo Quintas for their great help and patience.

I would like to thank my UAH lab friends David(es), José, José Ángel, Frank, Manuel and Raquel for their huge support and all the good and bad moments that we have shared, often in the distance. It would not have been as fun, even when the fridge overheats. Also in the Electronics Department: Ana, Ángel, Carlos, Guilherme, Eduardo, Enrique, Javier(es), Jesús, Manuel, Nacho, Noelia, Oscar, Pablo and Sergio. At Idiap I would like to thank specially Aleksandra, Darshan, Dayra, Dinesh, Gokul, Gulcan, Hari, Joan, Laurent, Minh-Tri, Oya, Paco and Rui, for their support and discussions.

I have met an enormous amount of incredibly great people in Lausanne. It is impossible to include all of them, and I am sorry in case I have forgotten someone. Thanks to those who I met during the first year: Daniel, Diego, Javier, Kajsa, Lianne, Nathalie, The Swedish Mafia, Oxana and Victoria; during the second year: Bruno, Céline, Chrissi, Francesc, Ilaria, Iñaki, Jan, Johanna, Kathrin, Lea, Michelle, Milena, Moritz, Outi and Tatjana; during the third year: Alberto, Almudena, Clara, Larissa, Patricia, Sarah and Spyros; and those who have been there along several: Aslam, Christelle, Damien, Edy, Hyunjin and the AG. Thanks to them, I have and will always carry with me very special memories.

Me gustaría además agradecer a mis amigos en Guadalajara su apoyo a lo largo de los años: Inés, Javier, José, Rodrigo, Samuel, Silvia y Teresa. Y especialmente, agradecer a mi familia toda la ayuda y paciencia: a mis padres, a mi hermano, a mi tía, y a mis abuelos.

Gracias, Thank You, Danke, Merçi, Grazie.

Resumen

En esta tesis se aborda el problema de la extracción automática de la comunicación no verbal en un contexto conversacional, gracias a distintos métodos de visión computacional. La comunicación no verbal juega un papel significativo en la percepción social de las personas, por lo que ha sido ampliamente analizado en psicología. Sin embargo, tradicionalmente ha sido necesaria la presencia de una persona que juzgue las características percibidas de los sujetos (es decir, un anotador), lo que supone una tediosa tarea e inconsistencias entre distintos evaluadores. Para tratar este problema, un elemento clave es el uso de métodos automáticos que permitan la abstracción sobre los anotadores, dotando de consistencia a los estudios de comportamiento.

En esta tesis se aborda esta tarea gracias a la captura de movimiento humana sin marcadores. La captura de movimiento sin marcadores consiste en la extracción de la posición de distintas partes del cuerpo a partir de imágenes y vídeos. Aunque existen sensores físicos aplicables directamente sobre los sujetos, han demostrado comprometer la naturalidad de los movimientos, algo fundamental a la hora de analizar el comportamiento conversacional.

Existen tres configuraciones en captura de movimiento sin marcadores: multi-cámara, cámara única, y cámara de profundidad. En esta tesis realizamos contribuciones en todas. Primero se ha propuesto un método multi-cámara basado en la reconstrucción 3D del entorno mediante Visual Hull. Utilizamos regresores no lineales para simplificar la búsqueda de la pose humana en el espacio altamente dimensional. De esta forma, conseguimos seguir múltiples personas simultáneamente con un único estimador. Gracias a un proceso de refinamiento, mejoramos el resultado del regresor y la capacidad de generalizar nuevas poses. Después, se ha desarrollado un método con cámara única, utilizando la idea de saliencia de manos: asumiendo que las manos son la parte de la imagen que más rápido se mueve a lo largo de una secuencia, hemos desarrollado nuevos seguidores basados en árboles de decisiones. Posteriormente se ha extendido este método con la información proporcionada por una cámara de profundidad. Finalmente, se ha desarrollado un método altamente invariante a la apariencia en el caso también de cámara única. Gracias al flujo óptico denso y un detector de torso, se ha obtenido la configuración de la pose a partir de la clasificación de las distintas partes corporales en la imagen. Hemos evaluado todas las contribuciones con bases de datos públicas y privadas, obteniendo o mejorando la precisión de estado del arte.

Adicionalmente, se han aplicado algunas de las ideas de los métodos mencionados para inferir una serie de variables sociales, a partir de una base de datos que contiene entrevistas de trabajo reales. Se han extraído y agregado una serie de características anotadas manualmente u obtenidas automáticamente, y se ha demostrado la correlación entre ellas y distintos rasgos de personalidad o rendimiento laboral. Finalmente, se ha conseguido predecir algunos de estos rasgos mediante un regresor.

Palabras clave: Visión computacional, computación social, captura de movimiento sin marcadores, comunicación no verbal, métodos automáticos.

Abstract

This thesis addresses the problem of automatic nonverbal communication extraction by means of different computer vision techniques. Nonverbal communication plays a significant role in how we perceive each other in a social context. It has therefore been intensively analyzed in social psychology and cognitive science. However, there has always been the need for an interpreter: a person that emits a judgment on the perceived traits of the analyzed subject, or that codes specific behaviors. This judgment always carries a degree of subjectivity, which can lead to inconsistencies across different evaluations. Also, depending on the amount of data available, it can be a cumbersome, time consuming task. In order to address this problem, the use of an automatic system that abstracts itself from human interpretation is a key element, providing consistency for studying the present behaviors.

We address this task by means of human markerless motion capture. Markerless motion capture extracts the position of the human body parts in images and videos. While there exist wearable sensors for the same purpose, the discomfort associated with them reduces the naturality of the movements.

There are three main sensor set-ups in markerless motion capture: multi-camera, single camera and depth camera. In this thesis we make contributions in all of them. We first designed a multi-camera approach based on 3D scene reconstruction through Visual Hulls. We took advantage of non-linear regression methods in order to simplify the search in the high-dimensionality human pose space. By doing this, we were able to track multiple subjects simultaneously with a single tracker. Helped by a refinement process, we were able to provide better generalization capabilities. Then we developed a single camera method, based on the idea of hand saliency: we hypothesized that the hands are the parts of the image that move quicker along a whole video. To this end, we designed a new hand tracker based on a Decision Trees algorithm, and performed simultaneously action recognition. We later extended this approach by fusing the information provided by a depth camera in the hand saliency map equations. Finally, we developed a highly appearance-invariant method for motion capture while using again a single color camera. Thanks to dense optical flow and a torso detector, we were able first to classify the body parts in the image and then obtain the body configuration. This latter contribution is a step in order to remove the appearance-related problems of markerless motion capture. We evaluated all the approaches with public and private datasets, showing or improving state-of-the-art performance.

Additionally, we applied some of the ideas behind of our methods in order to infer a series of social constructs from real job interviews. We extracted and aggregated a series of manually-annotated and automatic features from videos, and showed the correlation between them and personality traits or job performance. Finally, we were able to predict some of those traits with a regression scheme.

Keywords: Computer vision, social computing, markerless motion capture, nonverbal behavior, automatic methods.

Contents

Resumen ix

Abstract xi

Contents xiii

List of Figures xix

List of Tables xxiii

1 Introduction 1

1.1 Motivation 1

1.2 Problem statement 2

1.3 Approach 2

1.4 Contributions of the thesis 3

1.5 Organization of the rest of the document 5

2 Preliminary concepts 7

2.1 Introduction 7

2.2 Mathematical notation 7

2.2.1 Conventions 7

2.2.2 Glossary 7

2.2.2.1 Simultaneous motion capture and action recognition of several people: multi-camera approach (Chapter 5) 7

2.2.2.2 Upper body motion capture from hand tracking (Chapter 6) 9

2.2.2.3 Upper body motion capture from appearance- and scale-invariant features (Chapter 7) 10

2.2.2.4 Predicting social attributes (Chapter 8) 11

2.3 Classification techniques 11

2.3.1 Unsupervised techniques: extended K-means 11

2.3.2 Unsupervised techniques: Mean Shift 12

2.3.3 Supervised techniques: Random Forests 12

2.4 Regression techniques . . . 15
2.4.1 Gaussian Process Latent Variable Model (GPLVM) . . . 15
2.4.2 Random Forests . . . 17
2.5 Tracking techniques . . . 18
2.5.1 The Kanade-Lucas-Tomasi (KLT) tracker . . . 18
2.5.2 Bayesian tracking . . . 19

3 State of the art 21
3.1 Introduction . . . 21
3.2 Nonverbal communication and psychology . . . 21
3.3 Computational modeling of interaction . . . 23
3.4 Motion capture . . . 24
3.4.1 Monocular approach . . . 25
3.4.1.1 Non BPD-based methods . . . 26
3.4.1.2 BPD-based methods . . . 27
3.4.1.3 Conclusion . . . 29
3.4.2 Multi-camera approach . . . 30
3.4.2.1 Silhouette-based . . . 31
3.4.2.2 Other methods . . . 32
3.4.2.3 Conclusion . . . 33
3.4.3 Depth approach . . . 33
3.4.3.1 Purely depth . . . 33
3.4.3.2 Depth and color . . . 35
3.4.3.3 Conclusion . . . 35
3.5 Action recognition . . . 36
3.5.1 Action recognition from motion capture . . . 36
3.5.2 Conclusion . . . 36
3.6 Conclusion . . . 37

4 Datasets 39
4.1 Introduction . . . 39
4.2 Multiple and single viewpoints: HumanEva-I . . . 39
4.3 Single viewpoint: Job interview database . . . 40
4.3.1 Introduction . . . 40
4.3.2 Data retrieval . . . 40
4.3.2.1 Technical setup . . . 41
4.3.2.2 Social variables . . . 41
4.3.3 Data annotations and data splits . . . 46

4.3.3.1 Data subset A . . . 48
4.3.3.2 Data subset B . . . 49
4.3.3.3 Data subset C . . . 49
4.3.3.4 Data subset D . . . 50
4.4 Single viewpoint: ChaLearn 2011 . . . 50
4.5 Conclusion . . . 51

5 Simultaneous motion capture and action recognition of multiple people 53
5.1 Introduction . . . 53
5.2 Overview . . . 54
5.3 Observation system . . . 54
5.4 Priors . . . 55
5.4.1 Human body prior . . . 55
5.4.2 Movement prior . . . 56
5.5 Tracking process . . . 57
5.5.1 Initialization . . . 58
5.5.2 Prediction and re-initialization . . . 58
5.5.3 Probability Density Function formation and particle conversion . . . 59
5.5.4 Weighting . . . 60
5.5.5 Resampling with previous clustering and normalization . . . 61
5.5.6 Particle set clustering . . . 62
5.6 Pose refinement . . . 62
5.7 Results . . . 63
5.8 Conclusion . . . 66

6 Upper body motion capture from hand tracking 69
6.1 Introduction . . . 69
6.1.1 Contributions . . . 70
6.1.2 Chapter organization . . . 71
6.2 Overview . . . 71
6.3 Observation system . . . 71
6.3.1 Image motion retrieval . . . 72
6.3.2 Face detection . . . 72
6.3.3 Edge detection . . . 73
6.3.4 Skin segmentation . . . 73
6.3.5 Hand likelihood map formation . . . 74
6.4 Priors . . . 74
6.4.1 Body prior . . . 74

6.4.2 Movement prior 75

6.4.2.1 Dimensionality reduction 77

6.5 Tracking 78

6.5.1 Hand tracking 78

6.5.1.1 Per-frame hand likelihood map analysis 78

6.5.1.2 Full-sequence consistency with Decision Trees 78

6.5.2 Upper body pose tracking 79

6.5.2.1 Database lookup 79

6.6 Action recognition 80

6.7 Improvements of our initial method 81

6.7.1 Improvements in the observation system 81

6.7.1.1 Face detection 81

6.7.1.2 Addition of depth 82

6.7.1.3 Color and depth fusion in the hand likelihood map 82

6.7.2 Improvements in maxima extraction from the hand likelihood map 83

6.7.3 Addition of 3D hand tracking 83

6.7.4 3D upper body pose retrieval through model fitting 83

6.7.5 Action recognition 85

6.7.5.1 Hand height ($f_{h,t}$) 85

6.7.5.2 Hand movement ($f_{m,t}$) 85

6.7.5.3 3D face-hand distance ($f_{d,t}$) 85

6.7.5.4 Speaking status ($f_{s,t}$) 85

6.7.5.5 Ongoing activity ($f_{a,t}$) 85

6.8 Implementation and results 86

6.8.1 Data 86

6.8.2 RGB case 86

6.8.2.1 Hand likelihood map evaluation 86

6.8.2.2 Action recognition evaluation 86

6.8.3 RGBD case 88

6.8.3.1 Hand likelihood map evaluation 88

6.8.3.2 Action recognition evaluation 89

6.8.4 Comparison of the performance of RGB and RGBD cases 90

6.9 Conclusion 91

7 Upper body motion capture from appearance- and scale-invariant features 95

7.1 Introduction . . . 95

7.2 Overview . . . 96

7.3 Observation system . . . 97

7.3.1 Largely appearance-invariant image features . . . 97

7.3.2 Body part classification . . . 98

7.3.2.1 Training the forest for classification . . . 99

7.3.3 Body pose regression . . . 99

7.3.3.1 Obtaining images $\mathbf{I}_L$ and $\mathbf{I}_L^{'}$ in order to train the regressor . . . 100

7.3.3.2 Body part histograms from $\mathbf{I}_L$ and $\mathbf{I}_L^{'}$ . . . 100

7.4 Tracking . . . 101

7.4.1 Tracking the hands . . . 101

7.4.2 Stabilizing the body pose . . . 102

7.5 Results . . . 103

7.5.1 Classification . . . 103

7.5.2 Regression . . . 104

7.5.3 Tracking . . . 106

7.5.4 Computing time . . . 107

7.6 Conclusion . . . 108

8 Predicting social attributes 113

8.1 Introduction . . . 113

8.2 Extraction of features . . . 114

8.2.1 Manual annotations of body activity . . . 114

8.2.2 Automatic features . . . 115

8.2.2.1 Speaking status . . . 115

8.2.2.2 Hand speed . . . 115

8.2.2.3 Image activity histograms . . . 115

8.2.2.4 Beat gestures . . . 116

8.3 Inference of personality traits and hirability . . . 116

8.3.1 High level feature extraction . . . 117

8.3.1.1 Nonverbal cue encoding . . . 117

8.3.2 Prediction of tasks . . . 118

8.4 Results and discussion . . . 119

8.4.1 Annotation statistics . . . 119

8.4.2 Analysis of the effect of speaking status . . . 120

8.4.3 Prediction of hirability and personality . . . 120

8.4.4 Beat gesture detection performance 121
8.5 Future directions 122
8.6 Conclusion 123

9 Conclusions 125
9.1 Introduction 125
9.2 Contributions 125
9.3 Limitations and future work 126

Bibliography 127

List of Figures

1.1 Research framework . . . 3

1.2 General approach . . . 4

2.1 Extended K-means 2D example . . . 12

2.2 Mean Shift examples . . . 13

2.3 Failure examples . . . 14

2.4 Randomness levels in Random Forests . . . 14

2.5 Random Forests posterior composition . . . 15

2.6 Alignment to the principal components . . . 16

2.7 GPLVM example . . . 17

2.8 Example predictor models for regression . . . 18

2.9 Example of the KLT tracker . . . 19

2.10 Example of the XPFCP . . . 20

3.1 Automatic interaction modeling . . . 24

3.2 Motion capture techniques . . . 24

3.3 Evolution of part detectors . . . 26

3.4 Strong motion prior techniques . . . 26

3.5 Strong body model techniques . . . 27

3.6 BPD based techniques . . . 28

3.7 Strong motion priors . . . 28

3.8 Evolution of pictorial structures . . . 29

3.9 Closest works to ours . . . 29

3.10 Shape from silhouette methods . . . 31

3.11 Methods that use extra cues . . . 32

3.12 Closest techniques to ours . . . 32

3.13 Other approaches . . . 33

3.14 Range camera technology . . . 34

3.15 Body part classification from depth images . . . 35

3.16 Motion capture from RGBD images 35

4.1 HumanEva 39

4.2 Setting and data collection 41

4.3 Class examples 46

4.4 Annotated frame from subset A 48

4.5 Body part annotations 50

4.6 ChaLearn 2011 image examples 51

5.1 Overview 54

5.2 Volumetric reconstruction 55

5.3 Volumetric reconstruction with associated per-voxel voting 55

5.4 Body model 56

5.5 Actions compressed with GPVLM 57

5.6 Schematic overview of the tracking process 58

5.7 Activity changes handling 59

5.8 Cylinder-based body model weighting 61

5.9 Probability mode normalization before resampling. 61

5.10 Optimization effect in the final poses 62

5.11 Tracking error versus number of observed joints (n_{oj}) 63

5.12 Parameter tuning 64

5.13 Temporal tracking error in experiment B 65

5.14 Activity recognition results 66

5.15 Experiment 3 qualitative results 66

5.16 Output frames (1) 67

5.17 Output frames (2) 68

6.1 Proposed framework 70

6.2 Global data flow scheme 72

6.3 Optical flow retrieval 73

6.4 Skin color retrieval 73

6.5 Steps for building the hand likelihood maps in the RGB case 74

6.6 Torso Model 75

6.7 Training process 76

6.8 Key poses illustration 77

6.9 Hand tracking framework. 78

6.10 Decision tree example 80

6.11 Database lookup example for hands 80

6.12 Example of action recognition in the RGB case . . . 81
6.13 Face region removal . . . 82
6.14 Steps for building the hand likelihood maps in the RGBD case . . . 82
6.15 Hand likelihood map local maxima search with Mean Shift . . . 83
6.16 Segmented depth with its hand and head location . . . 83
6.17 Torso pose extraction overview . . . 84
6.18 Torso pose optimization . . . 84
6.19 3D upper body retrieval . . . 84
6.20 Graphical representation of the per-frame extracted features . . . 86
6.21 Failure examples . . . 87
6.22 Hand tracking performance in the RGB case . . . 87
6.23 Hand detection performance . . . 88
6.24 Results in ChaLearn 2011, as hand detection performance . . . 89
6.25 Action recognition performance . . . 90
6.26 Frame results in the RGB case . . . 92
6.27 Frame results in the RGBD case . . . 93

7.1 Data flow overview . . . 96
7.2 Pipeline of our proposal . . . 97
7.3 Torso context composition . . . 98
7.4 Example 4-C image . . . 98
7.5 Example of a 4-C image classification . . . 99
7.6 Example of body part histogram . . . 100
7.7 Body part histogram example . . . 101
7.8 Explanation of the KLT tracking method . . . 102
7.9 Integration with the KLT tracker example . . . 102
7.10 Explanation of the joints stabilization method . . . 103
7.11 Results of our method versus [Shotton et al., 2011] . . . 103
7.12 Sensitivity of the body part classifier to different parameters . . . 104
7.13 Accuracy of the body pose regressor . . . 105
7.14 Results in ChaLearn2011 . . . 106
7.15 Tracking results . . . 107
7.16 Qualitative results (1) . . . 109
7.17 Qualitative results (2) . . . 110
7.18 Qualitative results (3) and failure cases . . . 111

8.1 Proposed framework . . . 114
8.2 Illustration of the hand speed image computation . . . 115

8.3 Visualization of several nonverbal automatic features 116
8.4 Illustration for cues based on annotations of body activity 117
8.5 Example of correlation with social constructs 122

List of Tables

2.1 Chapter 5 notation . . . 7
2.2 Chapter 6 notation . . . 9
2.3 Chapter 7 notation . . . 10
2.4 Chapter 8 notation . . . 11

3.1 Big-Five traits and related adjectives [Gosling, 2003] . . . 22
3.2 Summary of the reviewed monocular motion capture works . . . 30
3.3 Summary of the reviewed multi-view motion capture works . . . 34
3.4 Summary of the reviewed depth motion capture works . . . 36
3.5 Summary of the reviewed action recognition works . . . 37

4.1 Subjects 1-15 information . . . 42
4.2 Subjects 16-30 information . . . 43
4.3 Subjects 31-45 information . . . 44
4.4 Subjects 46-60 information . . . 45
4.5 Subjects 1-45 interview length reduction . . . 47
4.6 Subjects 46-60 interview length reduction . . . 48
4.7 Per-class split of subset A . . . 49
4.8 Per-class split of subset B . . . 49
4.9 Per-class split of subset B . . . 50

5.1 Used parameters for experiments . . . 63

6.1 Labeled sub-actions in the training process . . . 75
6.2 Categories of actions . . . 76

7.1 Tracking results . . . 107

8.1 List of the manual and automatic nonverbal cues used in this study . . . 118
8.2 Descriptive statistics . . . 119
8.3 Speaking status analysis . . . 120

8.4 Prediction results 124
8.5 Performance of different feature sets 124

Chapter 1

Introduction

1.1 Motivation

This thesis addresses the problem of automatic nonverbal communication extraction by means of computer vision techniques. Nonverbal communication plays a significant role in how we perceive each other in a social context [Knapp and Hall, 2009], [Pentland, 2008]. Some key aspects of life take place in *vis-à-vis* conversations, in which the way we are socially perceived has a significant weight in our success [Curhan and Pentland, 2007b]. Nonverbal communication has therefore been intensively analyzed in social psychology and cognitive science [Knapp and Hall, 2009]. However, there has always been the need for an interpreter: a person that emits a judgment on the perceived traits of the analyzed subject, or that codes specific behaviors. This judgment always carries a degree of subjectivity, which can lead to inconsistencies across different evaluations. Also, depending on the amount of data available, it can be a cumbersome, time consuming task. In order to address this problem, the use of an automatic system that abstracts itself from human interpretation is a key element, providing consistency for extracting the behaviors displayed by people.

Different studies show that some of the most interesting cues in order to understand human behavior in conversational contexts are:

- *Adaptors*: unconsciously-used movements like nail biting and head scratching, that might provide information about the person's attitude, anxiety level, and self-confidence, which become a potentially rich source of information about the psychological state of the sender [Ekman et al., 1972].
- *Beat gestures*: movements that do not present a discernible meaning, i.e., small, low energy, rapid flicks of the hands and fingers that seem to beat along with the rhythm of the speech [McNeill, 1992]. They can be used to signal the temporal loci in speech of something the speaker considers important relative to the larger discourse [McNeill, 2005].
- *Posture*: intentionally or habitually acquired positions of the body, which can be an important clue about the emotional state of a person [Mehrabian, 1972].

In order to extract nonverbal information from recorded multi-modal data, plenty of automatic systems have been proposed. Audio-related nonverbal cues such as speaking status or voice pitch can be reliably extracted from microphone data, which has led to their extensive study. However, visual-related features are intrinsically difficult, since they involve image analysis in challenging scenarios. Some of those scenarios, such as facial tracking in frontal images, are on the way to be solved [Orozco et al., 2013]. However, other nonverbal cues such as body posture remain overly unsolved. Some works use marker-based motion capture in order to reliably explore body nonverbal communication. However, as it has been shown in [Fischer et al., 2003], placing markers on the body changes the locomotion patterns, which can lead to distorted results. Other works obtain reliable but coarse information about body activity, such as aggregated measures (mean, variance...) of the image activity along the video. This highlights the importance of markerless motion capture in body nonverbal communication analysis.

Markerless motion capture aims to retrieve body posture from images (thus without markers placed on the body), and has been a long-standing subject in computer vision and graphics. This is the consequence of a number of factors: high dimensionality nature of the problem; enormous variety of image observations that can be retrieved for a same body position depending on clothing, background, or lighting; presence of occlusions and self-occlusions; or dependence of camera viewpoint.

This problem has been addressed through numerous sensing techniques, which can be grouped into those using a single color camera (monocular case), multiple color cameras, and range cameras. Using a single color viewpoint provides the least amount of information: unlike the other cases, there is no 3D data easily available. The monocular case is therefore the most difficult and elusive case. However, as nonverbal analyses in communication and psychology studies have been typically recorded with a single color camera, it is the most relevant scenario for our goals. In the present thesis, the three approaches are investigated.

1.2 Problem statement

In order to address the several problems presented in the previous section, the aim of this thesis is two-fold:

First, we propose new methods to automatically analyze body nonverbal cues of people in a conversational context with markerless motion capture, using multi-camera and monocular configurations. This knowledge can be then applied to automatically extract the body pose from frontal videos of a person discussing around a table.

Second, we explore the new possibilities that extracting body pose has in nonverbal communication analysis. We develop a set of new computer vision algorithms in order to first extract, and then obtain a series of measurements that allow to analyze upper body movements and actions of a person with conversational meaning (see Figure 1.1).

1.3 Approach

Markerless motion capture represents a very challenging problem. We present here the road map of our approach in order to address it:

- We first approached the problem in the context of a multiple camera scenario. This made sense for several reasons: first, we have our own multi-camera system within the GEINTRA group[1]; secondly, even though we did not have a motion capture system in order to validate our approach, the highest quality motion capture database in the community (HumanEva), is provided in a virtually identical setting, allowing to seamlessly transfer to our setup the state-of-the-art based on HumanEva. In our approach, we simultaneously provide body pose and action recognition with high generalization

[1] http://www.geintra-uah.org

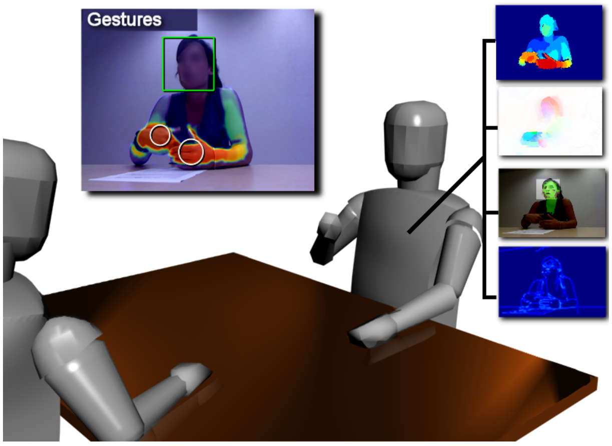


Figure 1.1: Research framework. Illustration of our approach in the upper-body setting: from different low-level computer vision cues (four plots in the right), we obtain higher level information such as the ongoing action and body part position (top plot).

capabilities, in a two-phase optimization scheme that relies on dimensionality reduction techniques. We map the high dimensional body pose into a low dimensional so-called latent space, simplifying the tracking process.

- The knowledge acquired in the multiple camera scenario was then applied to the most difficult monocular scenario. As in this case only one viewpoint per person is provided, we developed an alternative set of image cues, while retaining the dimensionality reduction concept coupled with a higher quality 3D body representation. We use new methods for head and hands tracking in order to use their position as a proxy for retrieving the full body pose, which has shown to be a very effective approach. We finally use range cameras for data fusion, showing a significant improvement in performance.
- After implementing the aforementioned multiple camera, single camera, and range camera systems, it became clear the need of new appearance- and scale-invariant image features in order to succeed in the monocular camera approach. Thanks to recent advances in dense optical flow and object detectors, and inspired by methods used with range cameras, we build a body part classifier that is largely invariant to appearance and scale. We show that its performance is comparable to that of the best method of range cameras, when there is movement present in the scene.
- Finally, we extract a number of nonverbal cues based on the obtained body pose information, and apply them in order to predict high-level information such as personality and performance. In order to validate the system, a real job interview dataset was collected, comprising 60 subjects.

As it can be seen in Figure 1.2, our approach follows a similar structure in the three scenarios considered. The followed pipeline is: (a) using several observation systems (single camera, multi-camera, range camera), we obtain a number of low level features; (b) we transform these features in order to get mid-level features (such as hand position or body pose); (c) finally, we obtain with them a higher level representation of the scene, in order to infer a series of social traits.

1.4 Contributions of the thesis

We have developed several algorithms in order to accomplish the goals mentioned in Section 1.1, producing the following contributions:

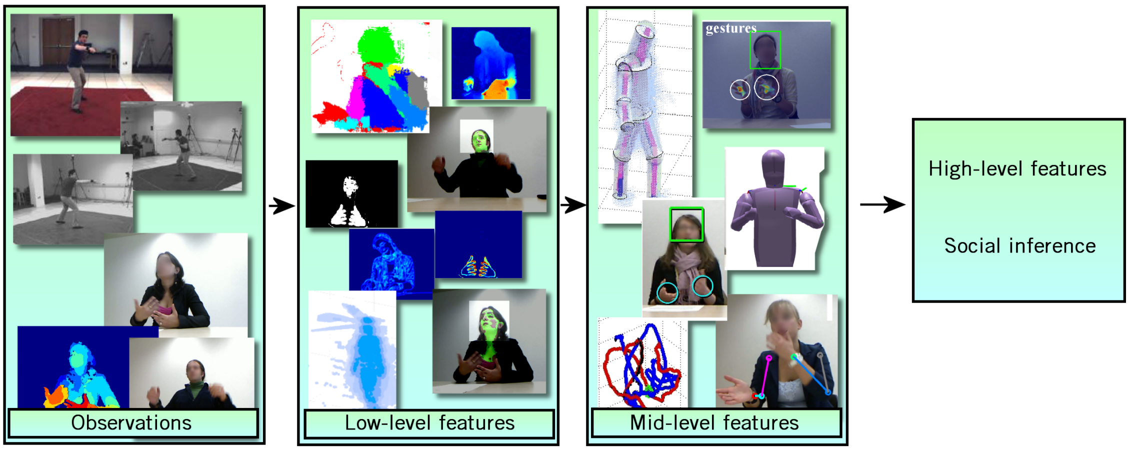


Figure 1.2: General approach. We use a wide selection of low- and mid-level features in order to automatically extract more descriptive high-level features and infer a series of social traits.

- A new method to perform simultaneous motion capture and action recognition from multiple cameras. The method involves using a single filter and movement priors for tracking several people. Thanks to a pose refinement process, generalization capabilities are maintained.
- A new method for extracting hand position from conversational video sequences, by exploiting the fact that optical flow is a strong indicator of where the hands are in conversation.
- A new method for object visual tracking, which assumes and exploits that the whole sequence is available beforehand (this is typically the case in psychology, management and cognitive science experiments).
- A new method for extracting 3D torso pose from 2D images in a seated person setting for action recognition.
- A new RGBD fusion method for hand tracking. We add the hypothesis of the hands being the closest part to the camera, when sensing with a range camera. We also improve the analysis of the hand likelihood map for better hand position extraction.
- An improvement over the pose and action retrieval performance, through a non-linear optimization scheme and a more robust action recognition method.
- An objective evaluation methodology of the above tasks using a real job interview dataset.

The aforementioned contributions resulted into the generation of a series of publications during the thesis, which are here listed:

- Works related to the development of the tracking algorithm
  - M. Marron-Romera, D. Pizarro-Perez, A. Marcos-Ramiro, R. Jalvo-Penin, J.C. Garcia-Garcia, M. Mazo-Quintas, **’Tracking multiple agents in an intelligent space with probabilistic algorithms and a camera ring’**, in IEEE International Symposium on Industrial Electronics, 2010. [Marron et al., 2010b]
  - M. Marron-Romera, J.C. Garcia-Garcia, M. Sotelo-Vazquez, D. Pizarro-Perez, M. Mazo-Quintas, J. Canas, C. Losada-Gutierrez, A. Marcos-Ramiro, **’Stereo Vision Tracking of Multiple Objects in Complex Indoor Environments’**, in Sensors, 2010. [Marron et al., 2010a]

- Works related to multi-camera motion capture
  - A. Marcos-Ramiro, M. Marron-Romera, D. Pizarro-Perez, **'Captura de movimiento de multiples personas mediante GPLVM y un PF mixto'**, in Seminario Anual de Automatica, Electronica Industrial e Instrumentacion (SAAEI), 2011. [Marcos et al., 2011]
  - A. Marcos-Ramiro, M. Marron-Romera, D. Pizarro-Perez, M. Mazo-Quintas, **'Captura de movimiento y reconocimiento de actividades para multiples personas mediante un enfoque bayesiano'**, in Revista Iberoamericana de Automatizacion Industrial (RIAI), 2013. [Marcos et al., 2013]
- Works related to upper-body motion capture
  - A. Marcos-Ramiro, D. Pizarro-Perez, M. Marron-Romera, L. Nguyen, D. Gatica-Perez, **'Body communicative cue extraction for conversational analysis'**, in IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2013. [Marcos-Ramiro et al., 2013]
  - A. Marcos-Ramiro, D. Pizarro-Perez, M. Marron-Romera, D. Gatica-Perez, **'Highly appearance- and scale-invariant monocular motion capture'**, submitted to European Conference on Computer Vision (ECCV) 2014.
  - A. Marcos-Ramiro, D. Pizarro-Perez, M. Marron-Romera, D. Gatica-Perez, **'Let your body speak: communicative cue extraction on natural interaction using RGBD data'**, submitted to IEEE Transactions on Human-Machine Systems.
- Works related to high-level nonverbal communication analysis
  - L. Nguyen, A. Marcos-Ramiro, M. Marron-Romera, D. Gatica-Perez, **'Multimodal analysis of body communication cues in employment interviews'**, in ACM International Conference on Multimodal Interaction (ICMI), 2013. [Nguyen et al., 2013b]

1.5 Organization of the rest of the document

The rest of the document is organized as follows:

- Chapter 2 introduces the most relevant preliminary knowledge required for the rest of the document. This includes mathematical notation and conventions along with an overview of some of the community algorithms in which we rely.
- Chapter 3 presents the state-of-the-art of the topics that are closest to our proposal.
- Chapter 4 introduces the datasets used for evaluating our approach and the way in which they were collected.
- Chapter 5 shows our first approach to motion capture and activity recognition from multiple, calibrated cameras.
- Chapter 6 explains our framework for extracting body pose and basic nonverbal cues from a new hand tracking scheme. It is applied in videos recorded with a single, uncalibrated point of view.
- Chapter 7 describes our motion capture system from monocular RGB data via robust appearance- and scale-invariant features.

- Chapter 8 presents real-world scenarios in which our cues are useful for behavioral analysis, namely in predicting job performance and personality traits in real job interviews.
- Finally, in Chapter 9 a conclusion of the work developed is presented, along with its limitations, which are discussed by proposing new research lines that could be started following our work.

Chapter 2

Preliminary concepts

2.1 Introduction

In this chapter we describe the preliminary knowledge that is necessary to read the rest of the document. Section 2.2 shows the mathematical conventions. This section can be used as a glossary while reading the rest of the work, as it is provided as a per-chapter, alphabetically ordered list. Section 2.3 briefly introduces the main classification algorithms that are used along the thesis. The aim of this section is to make the reader familiar with the concepts, not to give a comperhensive review, therefore references to works are provided in case that the reader needs more information. Section 2.4 presents a similar review to that of Section 2.3, describing different regression techniques also used in the thesis. Finally, Section 2.5 introduces several tracking techniques.

2.2 Mathematical notation

2.2.1 Conventions

$a, b, c,$ ... **Scalars** are typeset in regular italic lower-case.

$\vec{a}, \vec{b}, \vec{c},$ ... **Vectors** are typeset in italic lower-case and assumed column vectors: $\vec{a} = [a_1, a_2, ...]^T$.

$\mathbf{A}, \mathbf{B}, \mathbf{C}$ ... **Matrices** are typeset in non-italic boldface capitals $\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$.

$\mathcal{A}, \mathcal{B}, \mathcal{C}$ ... **Sets** are typeset in upper-case calligraphic font: $\mathcal{A} = \{a_1, a_2, ...\}$ or $\mathcal{A} = [\vec{a_1}, \vec{a_2}, ...]$.

2.2.2 Glossary

2.2.2.1 Simultaneous motion capture and action recognition of several people: multi-camera approach (Chapter 5)

Table 2.1: Chapter 5 notation.

Expression	Description
α_k	Percentage of particles to avoid kidnap
α_r	Percentage of particles for re-initialization
$\mathcal{A} = \{\vec{a_i}\}_{i=1}^{n_{joints}} = \{a_{\alpha i}, a_{\beta i}\}_{i=1}^{n_{joints}}$	Body pose in angle parameterization

Continued on next page

Table 2.1 – continued from previous page

Expression	Description
β	Regression parameter set
$\vec{c_t} = \{[\mathcal{L}_{i,t}, \vec{k_{i,t}}, \phi_{i,t}, d_t]\}_{i=1}^{n_p}$	Continuous part of the state space, in time t
d_t	Discrete variable in the state space, in time t
$\varepsilon \propto \sum_{i=1}^{n_m} \lVert 1 - \{\xi_i\}_{i=1}^{n_{joints}} \rVert^2 + \lVert \mathcal{P}_\rho - \mathcal{P}_o \rVert^2$	Cost function in optimization
$\mathcal{G} = \{\vec{g_i}\}_{i=1}^{n_g} = \{[x_i, y_i, \mathcal{V}_g]\}_{i=1}^{n_g}$	Visual Hull output group set
θ	Function that converts angles into 3D joint pose
$\vec{k_{i,t}} = [x_i, y_i]$	Position of the Visual Hull centroid
μ	Hyperparameter
n_a	Number of actions
n_c	Number of cameras
$n_{h,t}$	Number of Visual Hull points at time t
n_{joints}	Number of joints
$n_{f,tr}$	Number of training frames
n_{oj}	Number of observed joints
n_p	Number of particles per mode
n_{tp}	Total number of particles
n_m	Number of probability modes
$\mathcal{L}_{tr} = \{\vec{l_{tr,i}}\}_{i=1}^{n_{f,tr}} = \{l_{1i}, l_{2i}, ..., l_{qi}\}_{i=1}^{n_{f,tr}}$	Definition of latent space trained shape
$\mathcal{L}_{tr,o}$	Initialization of $\mathcal{L}$ with PCA
$\vec{l} = \{l_{1i}, l_{2i}, ..., l_{qi}\}$	Latent space point
$\mu_{\mathcal{X}_{tr}}$	Mean of the training pose set
o_{hi}	Number of cameras that observed the Visual Hull point i
$\Psi = \begin{cases} o_h & \text{if } \{x_j, y_j, z_j\} \subset \mathcal{Z}_i \\ 0 & \text{if } \{x_j, y_j, z_j\} \not\subset \mathcal{Z}_i \end{cases}$	Function that retrieves the weight associated with Visual Hull point j
$\mathcal{P} = \{\vec{p_i}\}_{i=1}^{n_{joints}} = \{x_i, y_i, z_i\}_{i=1}^{n_{joints}}$	Body pose definition, parameterized in 3D coordinates
$\mathcal{P}_o$	Output, final refined pose
$\mathcal{P}_\rho$	Reconstructed pose (PF output)
q	Number of latent dimensions
ρ_h	Visual Hull voxel size
$\vec{\sigma_{\mathcal{L}}} = [\sigma_{\mathcal{L}1}, \sigma_{\mathcal{L}2}, ..., \sigma_{\mathcal{L}q}]$	Standard deviation for each dimension of latent space
$\mathcal{S} = \bigcup_{i=1}^{n_m} \mathcal{S}_m$	Total particle set, as a composition of per-mode sets
$\mathcal{S}_m = \{\vec{c_{mi,t}}, d_t\}_{i=1}^{n_p}$	Set of mode particles, continuous plus discrete variable vector
$\mathcal{S}^{'}$	Resampled set of particles
$\mathbf{T} = \begin{bmatrix} t_{11} & ... & t_{1n_a} \\ ... & ... & ... \\ t_{n_a1} & ... & t_{n_an_a} \end{bmatrix}$	Probability transition matrix
t	Time instant
t_{re}	Re-initialization period
τ_i	Cylinder radius of body part i
$\mathcal{V}_t = \{x_{i,t}, y_{i,t}, z_{i,t}, \rho_{hi,t}, o_{hi,t}\}_{i=1}^{n_h}$	Set of Visual Hull observations at time t
$\mathcal{V}_g$	Visual Hull points associated with a Visual Hull group
$w_{i,t} = \prod_{i=1}^{n_{joints}} \xi_i$	Weight of particle i in time t

Continued on next page

Table 2.1 – continued from previous page

Expression	Description
$\xi_i = \frac{\mathcal{V}_t \subset \mathcal{Z}_i}{\xi_{i,max}} \frac{\sum_{j=1}^{n_h} \Psi(\vec{v_j}, \mathcal{Z}_i)}{\xi_{i,max} n_c}$	Weighted (by n_c) fill percentage of each body part
$\xi_{i,max} = \frac{\pi \tau_i^2 \omega_i}{\rho_h^3} + \frac{4 \pi \tau_i^3}{3 \rho_h^3}$	Maximum possible fill percentage of body part i
$\mathcal{X}_{tr} = \{\mathcal{P}_{tr}\}_{i=1}^{n_{f,tr}}$	Set of training points (in this case poses)
$\vec{x_t} = \{\vec{c_t}, d_t\}$	Extended state vector
$\mathcal{Z}_i = \{\omega_i, \tau_i\}$	Cylinder for body part i
ω_i	Length of body part i
Ω	Regression function

2.2.2.2 Upper body motion capture from hand tracking (Chapter 6)

Table 2.2: Chapter 6 notation.

Expression	Description
$\mathcal{A}$	Pose configuration parameterized with angles
$\delta_{ij} = \|dist(\vec{u_{f,i}}, \vec{u_{o,j}})\|$	Jump distance from tracklet i to j
e_n	Pose naturality energy
$\varepsilon \propto \|\vec{p_{hands_t}} - \vec{p_{hands,o,t}}\| + \|e_{n,t}\| + \|\mathcal{P}_{o,t} - \mathcal{P}_{o,t-1}\|$	Energy function
$f_{m,t}$	Hand motion
$f_{h,t}$	Hand height
$f_{d,t}$	3D distance from hands to head
fs, t	Speaking status
$\mathcal{F}_{ar} = \{f_{h,t}, f_{m,t}, f_{d,t}, f_{s,t}\}$	Feature set for action recognition, in this case nonverbal behavior set
$\mathcal{F}_{ar}$	Feature set for nonverbal processing
$\mathcal{H}_{L,R} = \{\vec{h_{L,R}}\}_{i=1}^{n_f} = \{[\vec{u_{L,R,i}}, t_i, \lambda_i]\}_{i=1}^{n_f}$	Left or Right hand trajectory along the sequence
$\mathbf{I}$	Image matrix (480×640×3)
$\mathbf{I}_D$	Range image (480×640×1)
$\mathbf{I}_D(\vec{u})$	Depth value at pixel $\vec{u}$
$\mathbf{I}_E$	Image edges (480×640×1)
$\mathbf{I}_F$	Face detection binarized image (480×640×1)
$\mathbf{I}_F^{'}$	Improved face detection image (480×640×1)
$\mathbf{I}_H$	Hand likelihood map (480×640×1)
$\mathbf{I}_H^{'}$	Regularized hand likelihood map (480×640×1)
$\mathbf{I}_{OF} = [\mathbf{I}_{OF,\rho}, \mathbf{I}_{OF,\phi}]$	Optical flow image (480×640×2)
$\mathbf{I}_S$	Skin segmentation image (480×640×1)
κ_1, κ_2	Constants for RGBD fusion
κ_d	Distance penalization factor in tracklet jumps
$\lambda_{T,ij} = \lambda_i + e^{-\kappa_d \delta_{ij}} \lambda_j$	Total accumulated likelihood for tracklets i, j with a jump
$\{\vec{l_{k,i}}\}_{i=1}^{n_a} \subset \mathcal{L}_{tr}$	Key points in latent space
$\mathcal{L}_{tr} = \Omega(\mathcal{X}_{tr}, \mathcal{B})$	Latent training points

Continued on next page

Table 2.2 – continued from previous page

Expression	Description
n_a	Number of actions
n_f	Number of total frames in the sequence
$n_{f,S}$	Number of random frames for skin modeling
$n_{f,tr}$	Number of training frames
n_{joints}	Number of joints
$n_{\mathcal{T}}$	Number of tracklets
$\mathcal{P}$	Pose configuration parameterized as 3D joints
$\mathcal{P}_o$	Final pose configuration
$\mathcal{T} = \{\vec{t_i}\}_{i=1}^{n_{\mathcal{T}}} = \{[t_{o,i}, t_{f,i}, \lambda_i, \vec{u_{o,i}}, \vec{u_{f,i}}]\}_{i=1}^{n_{\mathcal{T}}}$	Set of all tracklets
μ_S	Mean of the skin color distribution
σ_S	Standard deviation of the skin color distribution
t	Time index
t_w	Time window for training
$\vec{u} = (u, v)^\top$	Pixel at position u, v
$\mathcal{X}_{tr} = \{\mathcal{A}_i\}_{i=1}^{n_{f,tr}}$	Training set for the algorithm

2.2.2.3 Upper body motion capture from appearance- and scale-invariant features (Chapter 7)

Table 2.3: Chapter 7 notation.

Expression	Description
b_w, b_h	Torso bounding box width and height
$\vec{b} = [\vec{u_o}, b_w, b_h]$	Torso bounding box
$\mathcal{B}_{\mathcal{R}} = \{\vec{r}_i\}_{i=1}^{n_{\mathcal{B}}}$	Random Forests parameters
$\phi(\vec{u_\delta})$	Features of interest
$G_{j_{\mathcal{R}}}$	Objective function (information gain) of node $j_{\mathcal{R}}$
$\mathbf{I}_t$	Input image at time t (480×640×3)
$\mathbf{I}^1$	1-dimensional image (480×640×1)
$\mathbf{I}_{bw}$	Torso width context image (480×640×1)
$\mathbf{I}_{bh}$	Torso height context image (480×640×1)
$\mathbf{I}_C = [\mathbf{I}_{OF}, \mathbf{I}_{bw}, \mathbf{I}_{bh}]$	Appearance-invariant 4-channel image (480×640×4)
$\mathbf{I}_{OF} = [\mathbf{I}_{OF,\rho}, \mathbf{I}_{OF,\phi}]$	Optical flow image (480×640×2)
$j_{\mathcal{R}}$	Split node j of tree $\mathcal{R}$
l	Classification label for a given training point
$L(\vec{u}, \mathcal{U}, \mathbf{I}^1)$	Lookup function that returns feat vector $\mathcal{F}_c$
$\mathcal{M}_l = [\vec{m_{l,v}}, \vec{m_{l,h}}]$	Histograms of labels and scores
$n_{\mathcal{B}}$	Number of parameters for a given tree
n_δ	Number of offset features
$n_{\mathcal{R}}$	Number of trees
n_u	Number of pixels in an image
$\mathcal{P}_o$	Output body pose

Continued on next page

Table 2.3 – continued from previous page

Expression	Description
$\mathcal{R}$	Randomized random tree
$\upsilon_{j_{\mathcal{R}}}$	Threshold of node j, of tree $\mathcal{R}$
t	Time index
$\vec{u} = (u, v)^\top$	Pixel
$\mathcal{U} = \{\vec{u_{\delta i}}\}_{i=1}^{n_\delta} = \{(u_{\delta i}, v_{\delta i})\}_{i=1}^{n_\delta}$	Set of offset features
$\mathcal{X}_{tr} = \{\mathcal{F}_i\}_{i=1}^{n_u}$	Set of training data points
$\mathcal{X}_{tr,j_{\mathcal{R}}}, \mathcal{X}_{tr,j_{\mathcal{R}}}^L, \mathcal{X}_{tr,j_{\mathcal{R}}}^R$	Training data that arrive at node j, and (L,R) splits

2.2.2.4 Predicting social attributes (Chapter 8)

Table 2.4: Chapter 8 notation.

Expression	Description
$a_{h,t} = \lvert e_{h,t} - e_{h,t-1} \rvert$	Hand acceleration
$e_{h,t}$	Hand energy
κ	Inter-rater agreement
$\mathcal{M}_{A,t} = \lvert \mathcal{M}_{OF,t} - \mathcal{M}_{OF,t-1} \rvert$	Image acceleration vertical histogram
$\mathcal{M}_{OF,t}$	Optical flow vertical histogram
n_f	Number of training frames
s_t	Speaking status at t
t	Time instant
$\mathcal{X}_{tr}$	Feature vector for training

2.3 Classification techniques

2.3.1 Unsupervised techniques: extended K-means

K-means [Alsabti, 1998] is a linear, iterative clustering algorithm. Given a data set $\mathbf{X} = \{\vec{x_i}\}_{i=1}^{n_d}$ with n_d data points, and a number of groups or classes n_g, it aims to minimize the Euclidean distance from the data points to the centroids of each group, defined as their center of gravity. Algorithm 1 shows a simple version of K-means.

```
Data: Set of data points $\mathbf{X}$, number of groups $n_g$
Randomly initialize groups from data points;
while Maximum number of iterations not exceeded or group data unchanged do
    Calculate distance of each data point to every centroid;
    Assign to each data point the group with a closest centroid;
    Re-calculate centroids of every group;
end
Invalidate groups with no data associated
```

Algorithm 1: Simple K-means framework

The main limitation of this method however is that the number of classes n_g to find has to be predefined. In some applications such as the one addressed in this thesis, the number of output groups is a priori unknown. The appearance of the extended K-means algorithm [Pelleg and Moore, 2000] aims to solve this problem. The main difference with respect to the original algorithm is that a new group is created every time a data point falls outside a given distance threshold from every centroid, and is not related to any class. Figure 2.1 shows an illustration of this behavior.

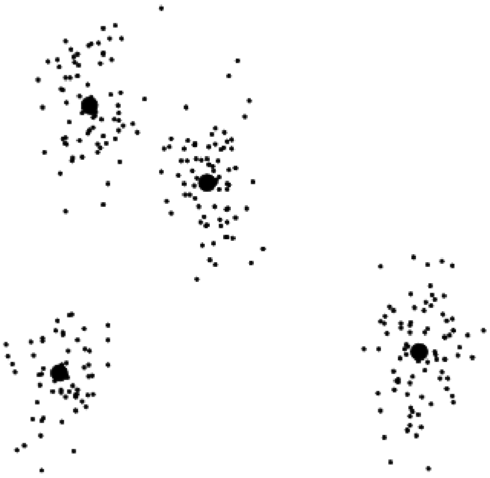


Figure 2.1: Extended K-means 2D example. With an unspecified number of input classes, but with a distance threshold parameter, the number of groups is estimated. Modified figure taken from [Pelleg and Moore, 2000].

2.3.2 Unsupervised techniques: Mean Shift

Mean Shift [Cheng, 1995] is a non-parametric clustering technique that does not require prior knowledge about the number of clusters, and does not constrain their shape. It is based on density estimation of the data points $\mathbf{X} = \{\vec{x_i}\}_{i=1}^{n_d}$ through a given kernel $\mathbf{K}(\mathbf{X})$.

For d-dimensional input points, the multivariate kernel density is obtained as:

$$f(\mathbf{x}) = \frac{1}{n_d h^d} \sum_{i=1}^{n_d} K(\frac{\mathbf{X} - \vec{x}_i}{h}), \tag{2.1}$$

where h is the kernel radius. The gradient of the density is estimated, and the mean shift vector always points towards the direction of the maximum increase in the density. The process of computing the gradient of the density and displacing the search window is iterative, and convergence is guaranteed. In Figure 2.2 a graphical illustration can be seen. In the end of the process, each data point is associated with an output class.

As the number of classes are automatically discovered and convergence is guaranteed, only a parameter h needs to be set, in contrast with the extended K-means approach in which a distance threshold, or the number of maximum iterations need to be defined. The main downside is its higher computational cost.

2.3.3 Supervised techniques: Random Forests

Decision trees have long been used [Breiman et al., 1984]. However, it was recently discovered that ensembles of slightly different trees have a higher generalization capability to previously unseen data.

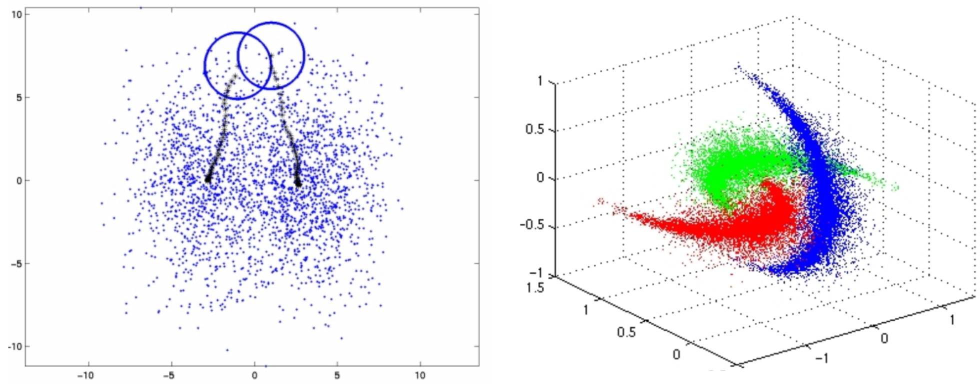


Figure 2.2: Mean Shift examples[1]. Left: displacement of the mean shift vectors towards the maximum density zones of the data. Right: clustering of arbitrarily-shaped groups. Best viewed in color.

A tree is a special type of graph, in which nodes are grouped into internal nodes (also known as splits) or terminal nodes (also know as leaves). Every node has a single input edge, therefore loops are not allowed within trees. In addition, commonly decision trees are binary, as each node outputs two edges.

A **decision tree** is a set of questions organized in a hierarchical manner and represented graphically as a tree [Criminisi et al., 2012]. It estimates a given unknown property of an entity by asking questions about its known properties. The decision on its unknown property is based on the terminal node of the chosen tree path. Therefore, the questions help moving towards the correct region of the decision space; the more questions, the higher the confidence in the response. The set of questions to make (i.e., the configuration of the nodes) is automatically learned during a training process.

A **data point** is denoted by vector $\vec{x}$, where its components represent attributes of the data point, named features. The dimensionality of $\vec{x}$ can be very large. However, not all features are equally good for making a split decision. Therefore, only a small portion of dimensions are extracted on an as-needed basis, also known as features of interest, denoted with $\phi(\vec{x})$.

Weak learners, also known as test functions or split functions, are a set of hierarchically organized tests. Each node has an associated test function that comprises a threshold $\upsilon_{j_{\mathcal{R}}}$, and whose output can be either 0 or 1 (”false” or ”true”), since we are considering binary trees. The node parameters can be grouped into the parameter set $\mathcal{B}$ for each tree $\mathcal{R}$.

$$\mathcal{B}_{\mathcal{R}} = \{\phi_j^*\}_{i=1}^{n_{\mathcal{B}}}, \tag{2.2}$$

where ϕ_j^* is the set of parameters for each weak learner at split node j.

A **training set** $\mathcal{X}_{tr} = \mathbf{X}_{tr} = \{\vec{x_{tr,i}}\}_{i=1}^{n_d}$ is a group of training points. That is, data points $\vec{x}$ which have an associated known label l, that denotes the attributes that we are looking for.

A **testing set** $\mathcal{X} = \mathbf{X}$ is a group of testing points. The difference with the training points is that the label l is no longer available. The property (or class) that we look for needs to be found through the series of questions that the decision tree comprises. Also, testing points are not intended to be included in the training set. See Figure 2.3 for an illustration of the process.

The purpose of training the decision tree is to maximize the information gain with the decision made in each node. That is, for a given input set of features, find the set of parameters ϕ_j^* that makes the output groups be as separated and ordered as possible. Therefore, in order to train a decision tree for classification, the parameters ϕ_j^* of the weak learner are optimized at each split node j with the function:

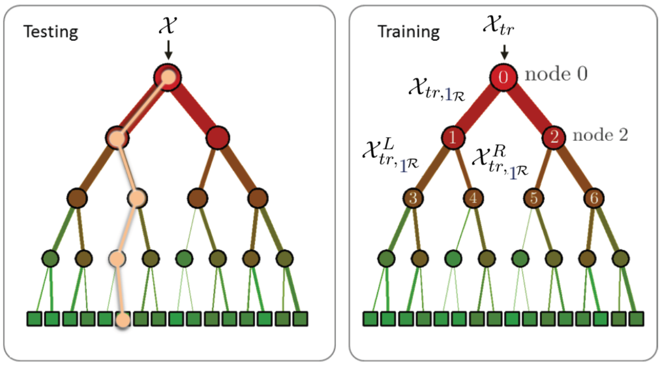

Testing
$\mathcal{X}$
Training
$\mathcal{X}_{tr}$
0
node 0
$\mathcal{X}_{tr,1_\mathcal{R}}$
1
2
node 2
$\mathcal{X}^L_{tr,1_\mathcal{R}}$
$\mathcal{X}^R_{tr,1_\mathcal{R}}$
3
4
5
6

Figure 2.3: Decision tree training and testing. Left: Testing. Right: training. Modified figure taken from [Criminisi et al., 2012].

$$\phi_j^* = \arg\max_{\phi_j \in (R)_j} G_{j_\mathcal{R}}, \tag{2.3}$$

where the objective function $G_{j_\mathcal{R}}$ used is the information gain, which is to be maximized in order to produce the highest confidence in the final distributions:

$$G_{j_\mathcal{R}} = H(\mathbf{X}_{tr,j} - \sum_{i \in \{L,R\}} \frac{|\mathbf{X}_{tr,i_j}|}{|\mathbf{X}_{tr,j}|} \mathbf{X}_{tr,i_j}), \tag{2.4}$$

where H is Shannon's entropy, $\mathbf{X}_{tr}$ the set of training points, and (L, R) are the left and right splits. This concept is at the basis of decision tree training. Once that the tree has been trained, each leaf node contains a posterior distribution $p_\mathcal{R}(y|\vec{x})$, which is a measure of the confidence in the predicted decision. We denote by y the predicted, a priori unknown, property of the input entity.

As mentioned, creating an **ensemble of decision trees** (or Random Forests), in which each tree is slightly different to others, is key to obtain the desired generalization capabilities. In order to obtain a set of trees, randomization is applied through two different approaches, namely random training set sampling and randomized node optimization. Given the entire parameter set $\mathcal{B}_\mathcal{R}$, for optimizing every node j only a small subset $\mathcal{B}_{j,\mathcal{R}}$ is used. Therefore, the ratio $\mathcal{B}_{j,\mathcal{R}}/\mathcal{B}_\mathcal{R}$ controls randomness and amount of correlation between different trees in the forest, as ilustrated in Figure 2.4.

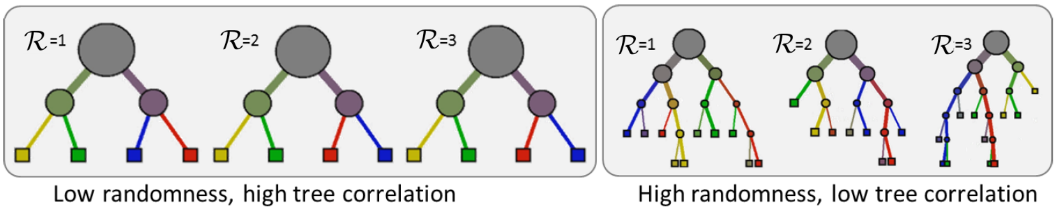


Figure 2.4: Randomness levels in Random Forests. Left: with little randomness the forest behaves as if it was made of a single tree. Right: with large randomness the trees are all very different from one another. Modified figure taken from [Criminisi et al., 2012].

The posterior probabilities of the different trees can be accumulated with different techniques (see Figure 2.5), in our case it is obtained as an average of all the outputs:

$$p(y|\vec{x}) = \frac{1}{n_\mathcal{R}} \sum_t^T p_\mathcal{R}(y|\vec{x}), \tag{2.5}$$

where $p_t(y|\vec{x})$ is the posterior of each leaf. After this step, the final predicted class y for input data $\vec{x}$ is obtained, constituting the output of the Random Forests.

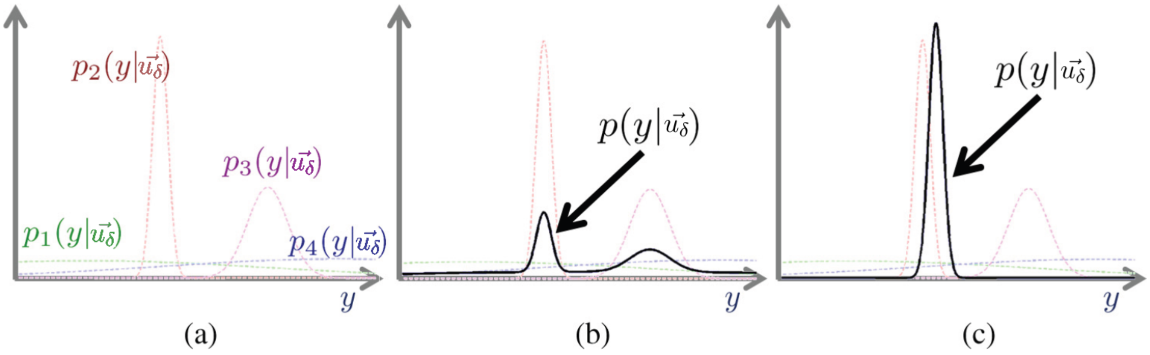


Figure 2.5: Random Forests posterior composition. (a) Posterior distributions of four different trees. (b) Combination as a sum of posteriors. (c) Combination as a product of posteriors. The latter shows higher peak probability, at the cost of being more influenced by noise. In this case, the data point is $\vec{x} = \vec{u_\delta}$. Modified figure taken from [Criminisi et al., 2012].

2.4 Regression techniques

2.4.1 Gaussian Process Latent Variable Model (GPLVM)

Visualization of high dimensional data can be achieved through projecting a dataset onto a lower dimensional manifold [Lawrence, 2005]. This projection is taken place through a mapping function Ω, whose aim is to explain as well as possible the variance of the data:

$$\mathbf{Y} = \Omega(\mathbf{X}, \beta), \tag{2.6}$$

where $\mathbf{X}$ is the original data in the observation space, $\mathbf{Y}$ is the data adapted to the new reference system, or so called ”latent variable”, and β are parameters. The function Ω can be linear or non-linear, depending on the used technique. For instance, in the technique known as Principal Component Analysis (PCA) [Jolliffe, 1986], a linear function is used, while in more advanced techniques, non-linear mappings are defined.

In this thesis we use PCA and a non-linear mapping technique, namely Gaussian Process Latent Variable Model (GPLVM) [Lawrence, 2005]. In PCA, the data is projected on the main directions of the data covariance with Ω. It therefore works well if the data has Gaussian characteristics, but might struggle otherwise. In Figure 2.6 a graphical representation of the principle can be seen.

GPVLM is a probabilistic reformulation of PCA that maximizes the likelihood of the data association between the low- and high-dimensionality manifolds. In order to understand GPLVM, first Probabilistic PCA (PPCA) needs to be introduced. PPCA is formulated as a latent variable model: given a set of centered d-dimensional data $\mathbf{Y} = \{\vec{y}_i\}_{i=1}^{n_d}$ and denoting the latent variable associated with each data point $\vec{x}_i$ (q-dimensional), the likelihood for an individual data point under the PPCA model becomes as follows, assuming Gaussian noise $\vec{y}_i = \Omega\vec{x}_i + \mathbf{N}$:

$$p(\vec{y}_i|\Omega, \beta_1) = \int p(\vec{y}_i|\vec{x}_i, \Omega, \beta_1)p(\vec{x}_i)d\vec{x}_i, \tag{2.7}$$

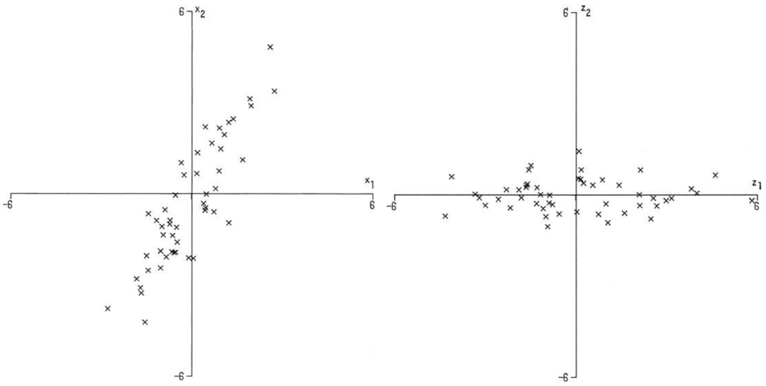


Figure 2.6: Alignment to the principal components. Left: 50 observations in their original space. Right: same observations aligned according to their principal components. Modified figure taken from [Jolliffe, 1986].

where $p(\vec{y}_i)$ is a Gaussian distribution with unit covariance, $p(\vec{x}_i) = N(\vec{x}_i|0, \mathbf{I})$, $\mathbf{I}$ is the identity matrix, and $p(\vec{y}_i|\vec{x}_i, \Omega, \beta_1) = N(\vec{y}_i|\Omega\vec{x}_i, \beta_1^{-1})$. The solution for Ω can be found by assuming that $\vec{y}_i$ is an independent and identically distributed by maximizing the dataset likelihood:

$$\Omega = \arg\max(p(\vec{y}_i|\Omega, \beta_1)). \tag{2.8}$$

In the case of GPLVM, the inner product kernel is modified to allow for non-linear functions. The product kernel is the kernelized version of the $\mathbf{U}$ matrix in the eigenvalue decomposition problem $\mathbf{X} = \mathbf{ULV^T}$, where the columns of $\mathbf{U}$ are the first eigenvectors of $\mathbf{YY^T}$, $\mathbf{L}$ is a diagonal matrix whose elements are also associated with the eigenvectors of $\mathbf{YY^T}$, and $\mathbf{V}$ is an arbitrary (as the solution is not unique) rotation matrix.

Later, ”sparsification” is made, which allows to represent the dataset as a subset $\mathbf{I}_s$ of d points, known as the active set, and denoted as $\mathbf{J}_s$. Then, a point j from the inactive set is used to project into the data space as a Gaussian distribution:

$$p(\mathbf{y}_j|\mathbf{x}_j, \beta) = N(\mathbf{y}_j|\mathbf{f}_j, \sigma_j^2\mathbf{I}), \tag{2.9}$$

whose mean is $\mathbf{f}_j = \mathbf{Y}^\top \mathbf{K}_{\mathbf{I}_s,\mathbf{I}_s}^{-1} \mathbf{k}_{\mathbf{I}_s,j}$, and the variance is denoted by σ. In the former expression, $\mathbf{K}_{\mathbf{I}_s,\mathbf{I}_s}^{-1}$ denotes the kernel matrix developed from the active set, and $\mathbf{k}_{\mathbf{I}_s,j}$ is a column vector consisting of the elements from the jth column of $\mathbf{K}$ that correspond to the active set. The matrix $\mathbf{K}$ is defined as:

$$\mathbf{K} = \beta\mathbf{XX}^T + \beta^{-1}\mathbf{I}. \tag{2.10}$$

In order to perform kernel optimization, the likelihood of the active set, given by:

$$p(\mathbf{Y}_{\mathbf{I}_s}) = \frac{1}{(2\pi)^{\frac{d}{2}}|\mathbf{K}_{\mathbf{I}_s,\mathbf{I}_s}|^{\frac{1}{2}}} exp(-\frac{1}{2}\mathbf{Y}_{\mathbf{I}_s}^T \mathbf{K}_{\mathbf{I_s},\mathbf{I_s}}^{\mathbf{-1}} \mathbf{Y}_{\mathbf{I}_s}) \tag{2.11}$$

is optimized with respect to parameters β with gradients evaluations. The core of the GPLVM work flow is represented in Algorithm 2. An example of GPVLM with real data can be seen in Figure 2.7.

```
Data: A size for the active set, $d$. A number of iterations, $T$
Initialize $\mathbf{X}$ through PCA;
for $T$ iterations do
    Select a new active set using the Informative Vector Machine algorithm [Lawrence et al., 2003];
    Optimize 2.11 with respect to the parameters $\mathbf{K}$ using scaled conjugate gradients;
    Select a new active set;
    for Each point not in active set, $j$ do
        Optimize 2.9 with respect to $\vec{x}_j$ using scaled conjugate gradients
    end
end
```

Algorithm 2: GPVLM framework

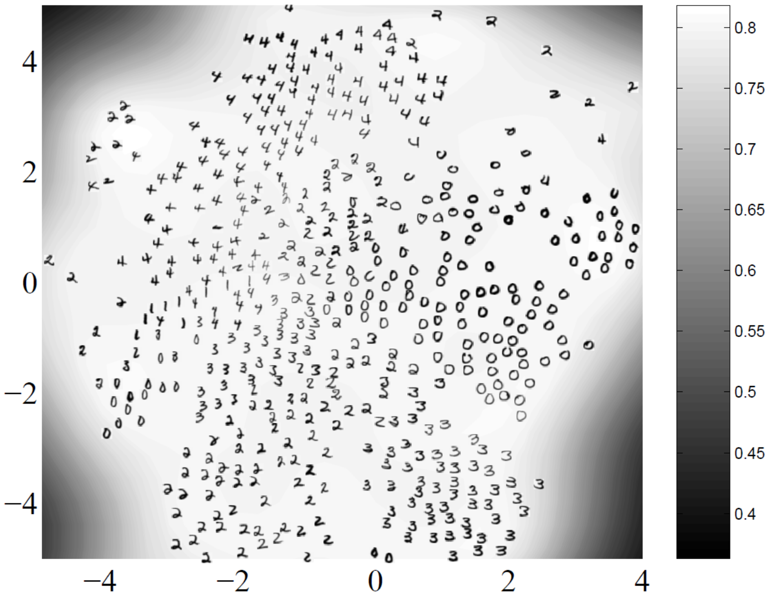


Figure 2.7: GPLVM example. Images containing handwritten digits are mapped into a low-dimensional 2D manifold through GPLVM. The digits are overlaid into their corresponding points, while gray level indicates likelihood. Modified figure taken from [Lawrence, 2005].

2.4.2 Random Forests

Regression Forests can also be used for non-linear regression. The main difference with respect to Random Forests used for classification is that in the regression, the labels l are continuous rather than discrete. Consequently, the objective function shown in eq. 2.4 has to be adapted appropriately as follows:

$$G_{j\mathcal{R}} = \sum_{v \in \mathcal{X}_{tr,j}} log(|\Lambda_y(\vec{x})|) - \sum_{i \in \{L,R\}} (\sum_{\vec{x} \in \mathcal{X}^i_{tr,j}} log(|\Lambda_y(\vec{x})|)) \tag{2.12}$$

where Λ_y is the conditional covariance matrix, $\mathcal{X}_{tr,j}$ is the set of training data that arrives at node j, and $\mathcal{X}^L_{tr,j}$, $\mathcal{X}^R_{tr,j}$ the left and right split sets. In Figure 2.8 it can be seen some examples of predictor models normally used in Regression Forests.

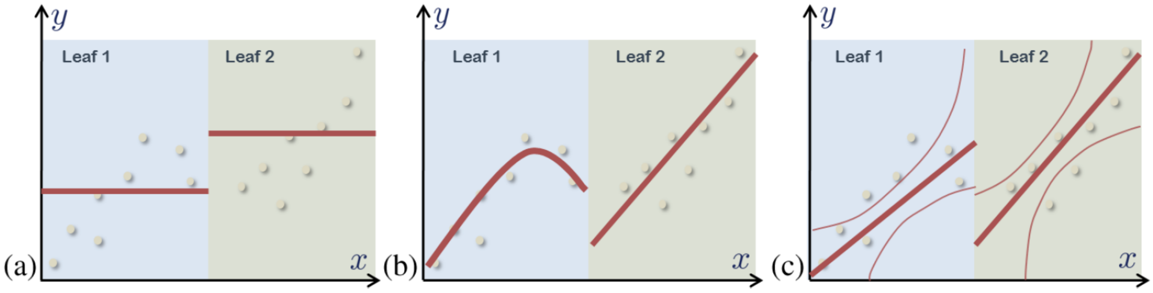


Figure 2.8: Example predictor models for regression. (a) Constant. (b) Polynomial and linear. (c) Probabilistic linear. Figure taken from [Criminisi et al., 2012].

2.5 Tracking techniques

2.5.1 The Kanade-Lucas-Tomasi (KLT) tracker

The KLT tracker is based on the early work from [Lucas and Kanade, 1981] and later fully developed in [Tomasi and Kanade, 1991]. It falls under the group of the feature trackers: it tracks the motion of features in an image stream. The method is based on the assumption that in a sequence of images, two consecutively-acquired frames $\mathbf{I}_t$ and $\mathbf{I}_{t+1}$ are very similar to each other. In this case, a fixed-size region of interest in image $\mathbf{I}_t$ can be defined, an used to look for its correspondent part in $\mathbf{I}_{t+1}$ by minimizing an error measure.

The KLT tracker uses the sum of squared intensity differences over the region of interest. The displacement is then defined as the one that minimizes this sum. For small motions, a linearization of the image intensities leads to a Newton-Raphson style minimization [Tomasi and Kanade, 1991]. Therefore, the displacement vector $\vec{d}$ obtained from times t and $t+1$ is chosen to minimize the residue error ε defined by the following double integral over the region of interest ROI:

$$\varepsilon = \int_{ROI} [\mathbf{I}_t(\vec{x} - \vec{d}) - J(\vec{x})]^2 \kappa d\vec{x} \tag{2.13}$$

where $\vec{x}$ is the camera position and κ a weighting function, that in the simplest case can be set to 1.

A crucial question is how to choose the feature windows that are best suited for tracking, as not all parts of the image contain motion information. Instead of defining a priori what a good region of interest is (for example, one that contains a high amount of texture), it is defined together with the tracking method: a good region of interest is one that can be tracked well.

Let a system 2x2 matrix $\mathbf{G}$ be defined as:

$$\mathbf{G} = \int_{ROI} \vec{g}\vec{g}^{\top} \kappa d\mathbf{A} \tag{2.14}$$

where $\vec{g}$ is the gradient of the intensity difference, and $\mathbf{A}$ is the area function of the region of interest. For good features, the matrix $\mathbf{G}$ must be above the image noise level and well-conditioned. The noise requirement implies that both eigenvalues of $\mathbf{G}$ must be large. In practice, when the smaller eigenvalue is sufficiently large to meet the noise criterion, the matrix $\mathbf{G}$ is usually also well conditioned, as the intensity variations in a window are bounded by the maximum allowable pixel value, so that the greater eigenvalue cannot be arbitrarily large. Therefore, a given region of interest is accepted if the eigenvalues fall above a predefined threshold. In Figure 2.9 it can be seen an example of the algorithm functionality.

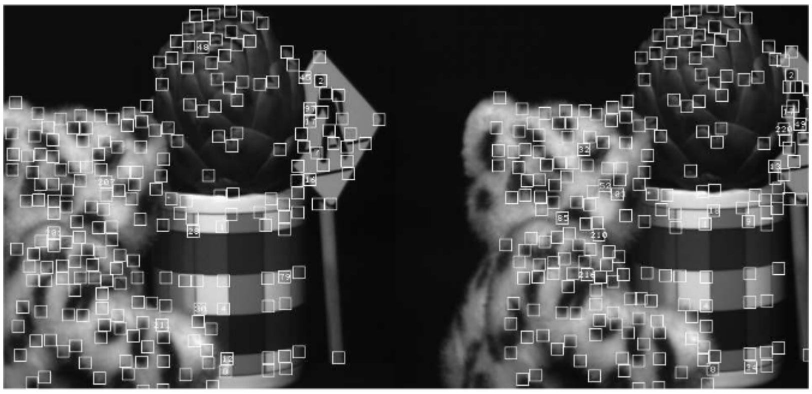


Figure 2.9: Example of the KLT tracker. Left: good features to track found at $t = 1$. Right: remaining features at $t = 100$. Of the initial 226 features, 9 disappear (6 of them off the right image boundary). Modified figure taken from [Tomasi and Kanade, 1991].

2.5.2 Bayesian tracking

The Particle Filter (PF) [Doucet et al., 2000] is a recursive, probabilistic estimator. It is based on a direct representation of the probability density function $p(\vec{x_t}|\vec{y_{1:t}})$ discretized through a set of n_p particles. The probability of each particle belonging to the probability density function is defined as a normalized associated weight $\vec{w_t}$. The probability density function is represented by pairs of particles and weights:

$$p(\vec{x_t}|\vec{y_{1:t}}) \sim \{\vec{x_t}, \vec{w_t}\}_{i=1}^{n_p} \tag{2.15}$$

The final value of the estimation is extracted from this function through different techniques, as well as its statistical analysis.

The PF employs a sequential Bayesian Filter. This formula avoids the discretization of the state space and the linearization of the system model, which is the case in a Kalman Filter for instance. This formation is made via Montecarlo sampling as follows:

$$p(\vec{x_t}|\vec{y_{1:t}}) \sim \sum_{i=1}^{n_p} \vec{w_{i,t}} \delta(\vec{x_{i,t}}) \tag{2.16}$$

Weights are then computed recursively with the following expression:

$$\vec{w_{i,t}} = \vec{w_{i,0:t}} = w_{i,\vec{0:t-1}} \frac{p(\vec{y_t}|\vec{x_t})p(\vec{x_t}|\vec{x_{t-1}})}{q(\vec{x_t}|\vec{x_{0:t-1}}, \vec{y_{1:t}})} \tag{2.17}$$

where $p(\vec{x_t}|\vec{x_{t-1}})$ is the probabilistic representation of the system model and $q(\vec{x_t}|\vec{x_{0:t-1}}, \vec{y_{1:t}})$ is the belief approximation function [Doucet, 1997], that allows for the real time execution of the Bayesian tracker. The weights are finally normalized:

$$\vec{w_t'} = \frac{\vec{w_{i,t}}}{\sum_{i=1}^{n} \vec{w_{i,t}}} \tag{2.18}$$

In order to avoid particle set degeneration (that is, the likelihood focusing on a single particle), the particle set is re-sampled: particles with a high associated weight are reproduced and dispersed, and particles with a low associated weight are deleted.

Another limitation of the PF is that it is unable to approximate well multimodal probability density functions. For example, it is not possible to simultaneously track multiple objects in a given scene for extended time periods. To solve that issue, several proposals have been made. In [Marron et al., 2010b] the eXtended Particle Filter with a Clustering Process (XPFCP) proposes to associate measurement groups with particles, and then regroup the output particles of the filter. In Figure 2.10 an image sequence of the XPFCP used as a multi-object tracker can be seen.

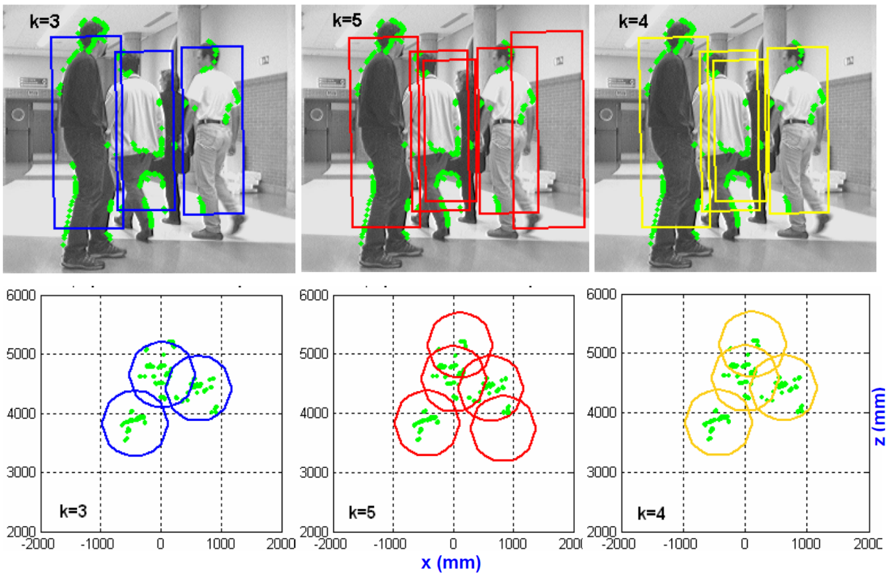


Figure 2.10: Example of the XPFCP. With a single tracker, several probability models are tracked. Figure taken from [Marron et al., 2010b].

Chapter 3

State of the art

3.1 Introduction

This thesis mixes several disciplines, such as motion capture, social computing and activity recognition. A complete and comprehensive review of the literature for each discipline would result in a too vast lecture. Instead, a weighted overview is provided for each subject. It is weighted in the sense that the amount of detail in which each discipline is reviewed is proportional to its importance to the current work. Therefore, we mainly focus on motion capture from cameras. Then, we describe previous works in social computing and action recognition. In any case, the surveys that appeared in the most prestigious publications are referenced, should the reader need a deeper knowledge.

Some parts of Sections 3.2 and 3.3 were originally published in [Nguyen et al., 2013b], co-authored with Laurent Nguyen (PhD student at Idiap Research Institute and Ecole Polytechnique Fédérale de Lausanne).

3.2 Nonverbal communication and psychology

Nonverbal communication plays an important role in face-to-face interactions as it conveys information in parallel with the spoken words. It is difficult to manipulate, as it involves unconscious processes [Knapp and Hall, 2009]. It has been shown to be a channel through which we reveal our internal state [Winters, 2005] or our personality traits [Naumann et al., 2009, DeGroot and Gooty, 2009]. It therefore has a strong influence on how we are socially perceived. In nonverbal communication, body communication plays an important role. It comprises what the face, head, eyes, limbs, and trunk transmit. Although the importance of head gestures, facial expressions, and gaze has been demonstrated in the literature, we focus here on the analysis of body posture and gestures.

Gestures are an essential component of body communication as they are used to enrich the vocal content and aid listener comprehension by augmenting the attention, activating images or representations in the listener's mind, and increasing the recall of what is being said [Knapp and Hall, 2009]. Moreover, restraining people from gesturing strongly affects the speakers' fluency [Knapp and Hall, 2009]. Body posture is another important component of body communication; various emotions such as fear, sadness, or happiness have been shown to be correctly inferred from a person's pose [Knapp and Hall, 2009]. In conversations, body posture can be used as a marker during a conversation: for instance, changes of body posture can precede a long utterance and may be kept for the duration of the speaking turn [Knapp and

Table 3.1: Big-Five traits and related adjectives [Gosling, 2003]

Trait	**Examples of Adjectives**
Extraversion	Active, Assertive, Enthusiastic
Agreeableness	Appreciative, Forgiving, Generous
Conscientiousness	Efficient, Organized, Reliable
Neuroticism	Anxious, Self-pitying, Tense
Openness to Experience	Artistic, Curious, Imaginative

Hall, 2009]. In this sense, both gestures and postures are inherently multimodal, in that they do not only occur in the visual modality, but are conditioned on the speaking status (*i.e.*, audio modality) of the person. For this reason, it could be necessary to consider the speaking status when analyzing posture and gestures.

Plenty of works have addressed the role of nonverbal body movements such as adaptors, which are movements like head scratching that provide information about attitude, anxiety level and self-confidence [McNeill, 1992]; and beat gestures, which are flicks of hands used to emphasize important parts of the speech with respect to the larger discourse [McNeill, 2005]. Body posture is also found to be an important indicator to the emotional state of a person [Mehrabian, 1972]

Specifically for the human communication wetting we study in the thesis (job interviews), the literature on nonverbal communication is rich. Used in nearly every organization, the employment interview is an interpersonal interaction between at least one interviewer and a job applicant, where the latter is evaluated for an open position. It aims to assess the candidate's suitability for the job at hand and is one of the most popular tools for this task [Wiesner and Cronshaw, 1988]. Interviews are inherently social as they require face-to-face interaction among the protagonists [Howard and Ferris, 1996]. Since applicants and recruiters typically meet for the first time, employment interviews are a form of *zero-acquaintance* interactions [Ambady et al., 1995]. Apart from the resumes, the applicant's verbal and nonverbal behavior is sometimes the only information that recruiters have to form an opinion.

In job interviews, applicant nonverbal behavior has a remarkable impact on the hiring decision. For instance, research shows that applicants who use more immediacy nonverbal behavior (*i.e.*, eye contact, smiling, body orientation toward interviewer, less personal distance) are perceived as being more hirable, more competent, more motivated, and more successful than applicants who do not [Imada and Hakel, 1977]. Organizational psychology literature suggests that the relation between nonverbal behavior and job interview outcomes can be based on the immediacy hypothesis, which claims that the applicant reveals through his or her immediacy behavior a greater perceptual availability between the applicant and the recruiter. This in turn leads to a positive affect in the recruiter and therefore to a more favorable evaluation [Imada and Hakel, 1977].

The five-factor structure, commonly known as the Big-Five, has received extensive support in psychology for describing personality [Gosling, 2003]. This framework is a hierarchical model of personality traits with five broad factors, which represent personality at its highest level of abstraction [Gosling, 2003]. The model suggests that most individual differences in human personality can be classified into five empirically-derived bipolar factors, namely extraversion, agreeableness, conscientiousness, neuroticism, and openness to experience (see Table 3.1). Previous studies have shown the importance of personality in employment interviews. In particular, conscientiousness and extraversion correlate significantly with employability ratings for conventional and enterprising jobs [Cole et al., 2004]. The validity of these inferences have been demonstrated, as conscientiousness is significantly correlated with job performance across all job occupations and extraversion is positively related to occupations requiring social skills [Barrick and Mount, 1991].

Body communication plays an important role in conjunction with spoken words to enhance the communication in face-to-face interactions [Knapp and Hall, 2009], which is also the case in employment interviews. Highly employable job candidates usually produce more hand gestures, both in frequency and amplitude [Anderson and Shackleton, 1990, DeGroot and Gooty, 2009], and have the tendency to lean forward more [Gifford et al., 1985]. These kinesic nonverbal cues were also found to be associated with some personality dimensions. For instance, rapid body movement is shown to be related to extraversion and agreeableness, while relaxed body posture has been associated with conscientiousness [Burnett and Motowidlo, 1998].

3.3 Computational modeling of interaction

Automatic nonverbal analysis made its appearance as an attempt to address the need for an interpreter, that can be a problem if large data is used, while also contributing with the consistency on judgments that a non-human annotator provides. As a result, a significant amount of literature has been published on the subject [Gatica-Perez, 2009]. Interactions among small groups and dyads are the main case studies. One of the approaches for automatic analysis came from the wearable computing field, where people wear sensors to be able to capture body motion and posture in conversations [Feese et al., 2011]. Despite being accurate, it involves placing and wearing intrusive devices, which compromises the naturality of behaviors. Basic image features (e.g. visual motion or basic hand gestures) have been studied to address this problem, as they can be robustly extracted from video. However, they correspond to rough representations of actual activity [Sanchez-Cortes et al., 2011]. Hand gesture recognition has also been the subject of a substantial amount of work [Mitra and Acharya, 2007], [Morency et al., 2008], although in general it has been applied to different problems (i.e. human-computer interaction or sign language recognition) than the one we address here.

Recent approaches in action recognition extract directly the ongoing activity from image cues. A more traditional method consists in first getting the body pose (i.e. perform motion capture) and then analyze it [Gall et al., 2012]. It has been recently shown [Angela Yao and Gool, 2011] that even though it is possible to perform activity recognition without knowing the body pose, it is still beneficial to know it.

The advent of cheap sensors, in combination with improved automated perceptual methods, have enabled the development of computational methods that predict social attributes. As a key construct to explain inter-individual differences, Big-Five personality has been studied in various settings. These include small group interactions [Pianesi et al., 2008], video blogs [Biel and Gatica-Perez, 2013], or human-computer interaction [Batrinca et al., 2011]. While most existing studies rely on audio nonverbal cues (prosody, speaking-turn-based features) and visual cues (head nods, visual activity, head pose) to predict personality, few studies have investigated the use of body communication. [Batrinca et al., 2011] used manually annotated hand movements and posture to predict personality in a human-computer interaction. Other human-computer interaction studies [Ball and Breese, 2000, Neff et al., 2010] have examined the link between specific personality dimensions (extraversion, friendliness, dominance) and body posture for enabling embodied conversational agents with the ability to produce postures given the personality tendency. To our knowledge, no computational study has specifically investigated the role of posture and gestures for the prediction of personality.

Despite the ubiquity and the importance of employment interviews, very few computational studies have examined the role of behavior in such settings. To our knowledge, two studies have approached this scenario. The work in [Curhan and Pentland, 2007a] studied the relation between nonverbal behavior and outcomes in a simulated dyadic negotiation configuration. The approach in [Batrinca et al., 2011] addressed short self-presentations in a human-computer interaction context, which resembles the job interview setting. Previous work have studied the employment interview setting in two ways. First in [Lu et al., 2012], a study on recognizing stress levels from acoustic features was done using 14 interviews, but no connection with hirability were explored. After, in [Nguyen et al., 2013a] (co-authored) we presented the first study on hirability prediction based on nonverbal cues. In this thesis, we extend this analysis by focusing on body communication cues, and by adding the task of predicting self-rated personality traits.

Relatively few automatic methods extract body pose from video with conversational constructs in mind. In [Quek et al., 2002, Xiong and Quek, 2006] several features are extracted thanks to among others, a hand tracker. Generic body pose retrieval has been the subject of enormous attention. As a proxy for body activity, some methods like [Biel and Gatica-Perez, 2013] measure basic image parameters, given a coarse estimation. The subject has also been approached by the wearable computing field [Feese et al., 2011]. However, the needed sensors can condition the naturality of the body movements.

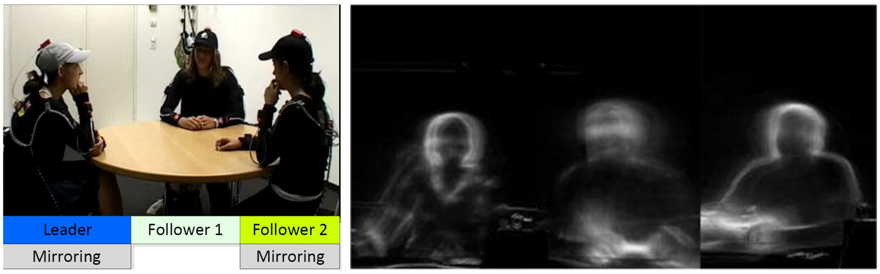


Figure 3.1: Automatic interaction modeling. Left: [Feese et al., 2011]. Right: [Sanchez-Cortes et al., 2011].

3.4 Motion capture

Motion capture can be performed using a number of different techniques (see Figure 3.2). As mentioned in the introduction, in the present work we are focused on retrieving the human body pose with computer vision techniques. Markerless motion capture has been a long-standing subject in vision and graphics. This is the consequence of a number of factors: the high dimensionality nature of the problem, the enormous variety of image observations that can be retrieved for a same body position depending on clothing, background, lighting..., the presence of occlusions and self-occlusions, the dependence of camera viewpoint, etc. Motion capture also allows for a number of interesting applications in many fields, such as of virtual-character animation and gaming, mood retrieval, elderly people assistance, or surveillance [Urtasun Sotil, 2006].

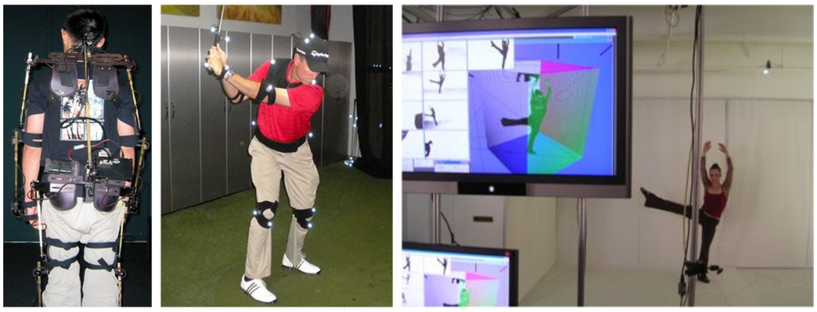


Figure 3.2: Motion capture techniques. Left: physical. Center: optical. Right: markerless

Traditionally, motion capture required wearing cumbersome sensors in the body, compromising the practical aspects of the concept. Recently, markerless motion capture solutions heavily removed the need for them, by being able to obtain the body pose with sensors in the environment (i.e. cameras of different kind), with relatively high precision. There exist several approaches, which can be grouped into single-camera (monocular) systems, multi-camera systems, and range camera systems.

Amongst all these approaches, finer classifications can be made, such as:

- Appearance dependent/independent works.
- Training reliant/free works.
- 3D/2D pose recovery works.
- Strong/weak/no body model used.
- Strong/weak/no motion model used.
- Camera calibration required/not required.
- Number of people recovered at the same time.
- Initialization required/not required.
- Whether it performs simultaneous action recognition.

In sections 3.4.1, 3.4.2 and 3.4.3, different criterion are used for their classification, in every case with the aim of having the most direct comparison with the work undertaken in this thesis.

Despite the size of the literature of markerless motion capture and multi-modal interaction, to our knowledge, a joint approach that explores the use and implications of having automatic body communication extraction in a social conversational context is still missing. Some related contributions have been proposed, like [Livne et al., 2012], which analyzes gender, age, or mood with gait cycles. The work here presented is designed to be a first solution to fill this void.

Surveys focusing on one or several of these techniques have been regularly released over the years: [Triggs, 2005] reviews the state of the art on the monocular approach from 2000 to 2004. In [Moeslund et al., 2006] an exhaustive overview of the main works in motion capture and activity recognition from 2000 to 2006 is presented. [Shaheen et al., 2009] the focuses on 3D model-based approaches in uncontrolled environments until 2009. In [Sigal and Black, 2010b] a complete review up to 2010 is published. [Ye et al., 2013] focuses on depth cameras. The aim of the present chapter is not to provide in-depth reviews, but rather to overview the most influential works in recent years. The reader is referred to the aforementioned surveys for more detailed information.

3.4.1 Monocular approach

Given all the challenges that recovering the human pose involves (as explained in 3.4), the single camera approach is considered the most difficult approach. The literature of monocular motion capture can be organized in many ways, but given the latest developments in the field, we classify it into methods that make use of *Body Part Detectors* (BPD) and the rest of approaches.

In this review a BPD is considered as a trained method that receives an image as input, and outputs the position (as a hard or soft bounding box) of a body part. [Dalal and Triggs, 2005] applied BPDs for human detection. One of the biggest contributions to BPDs appear in [Felzenszwalb et al., 2008]. The main idea is to compute the Histograms of Oriented Gradients (HOG) of the image, and then classify them to discriminate between different objects or body parts. Figure 3.3 shows two examples of part detectors.

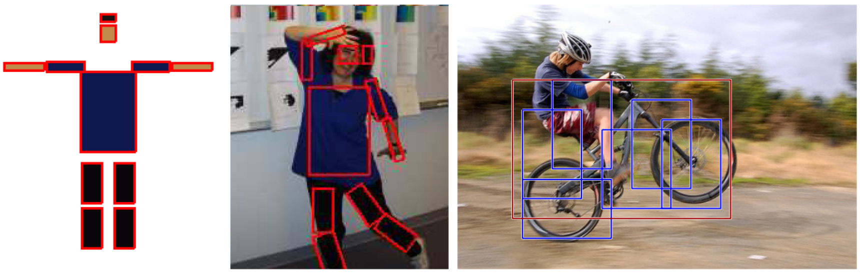


Figure 3.3: Evolution of part detectors. Left: initial approaches to BPDs from [Felzenszwalb and Huttenlocher, 2000]. Right: A sample image with a modern object detector (in this case, for a bycicle) from [Felzenszwalb et al., 2008].

3.4.1.1 Non BPD-based methods

Works based on strong motion models: These are methods that impose strong kinematic constrains (usually learned) in order to limit the search in the state space. For example, a walking motion can be trained and used to later disregard observations which do not explain a walking sequence. Motion models became popular since the first BPD-based methods, such as [Felzenszwalb and Huttenlocher, 2000], were largely appearance-dependent, leaving the need for more robust solutions. In [Sidenbladh et al., 2000] optical flow estimation and a strong motion and body model priors are used to track the human body, making it background-invariant. In [Elgammal and Lee, 2004], a direct non-linear mapping from silhouettes to body configuration is proposed, although it is viewpoint-dependent and contains many ambiguities in the observations. In [Grochow et al., 2004] a Scaled Gaussian Process Latent Variable Model (SGPLVM) is employed for the mapping between high and low dimensional state spaces. In [Urtasun Sotil, 2006] similar mappings are proposed with different image observations, while relying on the analysis of the motion models. In [Song et al., 2001], biological motion analysis is performed with graphs and motion capture data in order to infer and detect human movements. Using a similar idea, [Kulic et al., 2009] applies optical flow in constrained situations to recover the position of the body limbs. The work in [Agarwal and Triggs, 2006] uses strong motion priors to get 3D from 2D images by mapping silhouette descriptors to body pose configurations. Doing so, no explicit body model is required. In [Fossati et al., 2010] reliable sparse detections from silhouettes are obtained (what they call canonical poses) to get the 3D configuration, and then do reasoning between them, to give temporal consistency. Another approach uses feature selection from a large number of different visual features [Chen et al., 2011], in order to automatically obtain the most discriminative information.

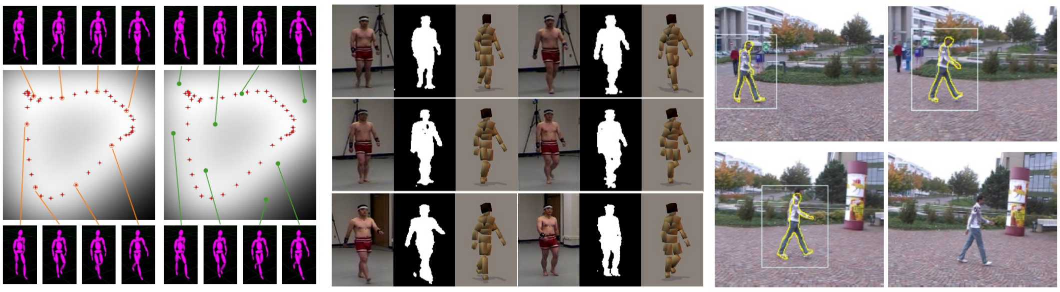


Figure 3.4: Strong motion prior techniques. Left: [Grochow et al., 2004]. Center: [Agarwal and Triggs, 2006]. Right: [Fossati et al., 2010].

Works based on strong body models: These are methods that impose a strong body definition prior. For example, by imposing a 3D mesh model of a person in order to explain the retrieved silhouettes. In [Howe, 2005], a strong body model prior and a tracking method are coupled with optical flow recognition for biological motion perception and 3D lifting. Instead of a strong motion prior, [Brox et al., 2007] uses a detailed upper body mesh model coupled with contours and optical flow to recover the pose. In [Noriega and Bernier, 2007] the face is located in the image, and the skin tones are then identified to compute skin segmentation. A Bayesian framework is coupled with some heuristics to track the arms. While promising, the hand detection is based only on skin segmentation, which limits the clothing possibilities. In [Sigal et al., 2007] a very detailed mesh-based body representation is proposed, and combined with generative and articulated methods for estimating 3D body pose. In [Zhang et al., 2008] a Dynamic Bayesian Network (DBN) is employed, which captures physical and anatomical constrains, and automatically detects self-occlusions. In [Kjellstrom et al., 2010] interaction with objects in the scene is exploited to impose constrains and retrieve 3D pose. In [Guan et al., 2010] the underlying intrinsic body configuration is learned to explicitly model clothing in the observed images.

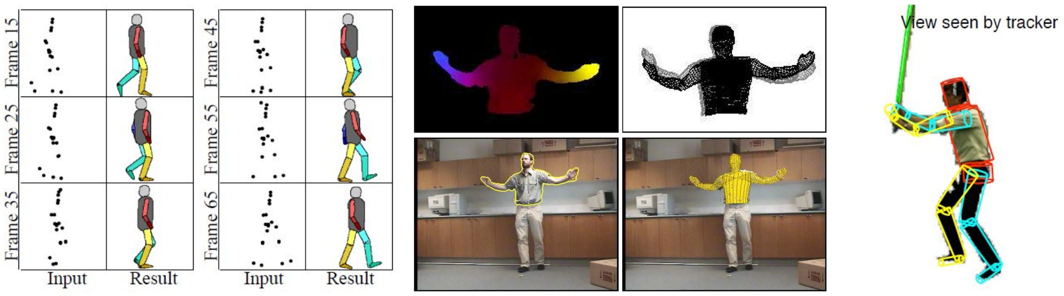


Figure 3.5: Strong body model techniques. Left: [Howe, 2005]. Center: [Brox et al., 2007]. Right: [Kjellstrom et al., 2010].

3.4.1.2 BPD-based methods

Origins: One of the earliest works with BPDs [Felzenszwalb and Huttenlocher, 2000] uses pictorial structures and spring-like connection between them to model appearance of body parts and their relationship. While it captures global coherence of the body pose, the part detector itself lacks enough detail to robustly track changes in orientation of the body parts and general clothing. In [Ramanan and Forsyth, 2003] a similar technique is used, although the emphasis is on parsing correct kinematics.

Methods focused on detection: In recent years, a number of HOG-based methods have arisen, which have helped to solve long-standing problems like automatic initialization. In [Tung and Matsuyama, 2008], the face and hands are tracked with a BPD aided with optical flow tracking. In [Kuo et al., 2011] the head is first localized by employing heuristic techniques, and then the torso pose is inferred from its position. In [Wang and Koller, 2011] the relationships between BPD and the image is modeled on multiple levels in order to improve the segmentation and detection. [Sun et al., 2012] reduces the computational cost by using a branch-and-bound algorithm while achieving performances comparable to those of the state of the art. In [Ladicky et al., 2013] extra image cues such as color models or texture potentials are used to improve the accuracy of the BPDs. In [Bourdev and Malik, 2009] 2D images and the estimated 3D pose of the underlying skeletons are annotated in order to infer the 3D pose. Their approach is the standard HOG extraction with SVM classification. In [Eichner and Ferrari, 2010] a person detector and modified pictorial structures are used to model interaction and occlusions between several people. It is one of the few methods that profit from image overlap and interactions to model multi-person body poses. In [Yang and Ramanan, 2011] the co-occurrence relations and spatial tree-structured relationships between part detectors are modeled to improve global coherence. Coherence can also be improved with HOG co-dependent body-part Random Forests regressors [Dantone et al., 2013], whereas [Ramakrishna et al., 2013] relies on symmetry analysis and [Hara and Chellappa, 2013] on dependency graphs and regressors. In [Gkioxari et al., 2013], HOG detectors are improved with skin color, contours, and contextual cues. In [Simo-Serra et al., 2013], 2D and 3D inference are simultaneously done with a generative Bayesian framework and discriminative HOG-based BPDs. In [Yu et al., 2013a], 3D poses are retrieved in unconstrained videos through BPDs, action classification, and pose regression with spatial-temporal features. In [Eichner et al., 2012], an upper body detector is used to obtain the body pose through several heuristics in weakly constrained images. In works like [Wang et al., 2013] simultaneous action recognition is also performed. Finally, in [Sapp and Taskar, 2013] global and local pose cues are included and a convex objective and joint training for mode selection and pose estimation is used to improve the ratio between performance and processing time.

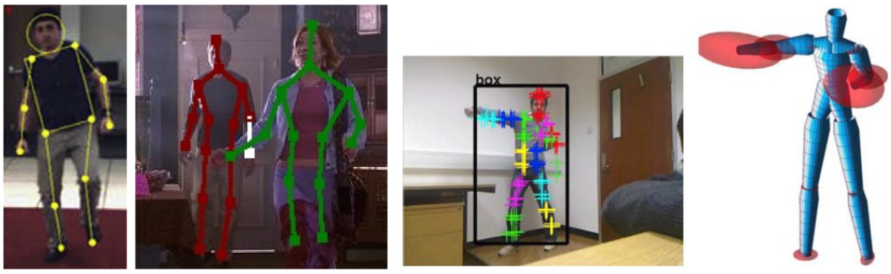


Figure 3.6: BPD based techniques. Left: [Kuo et al., 2011]. Center: [Ladicky et al., 2013]. Right: [Yu et al., 2013a].

Methods that profit from strong motion priors: Analyzing kinematics is another approach for solving the problem, since it adds extra constrains to reduce the search space and provides coherent detections. [Agarwal and Triggs, 2004] apply dimensionality reduction to the problem. Strong motion models are learned for trained actions, which in conjunction to a powerful tracking framework, allows for a more robust pose retrieval at the cost of generalization capabilities. However, in that work a BPD has to be manually initialized in the first frame. In [Andriluka et al., 2010] short paths of detections are extracted and then given spatial and temporal coherence. Using a movement prior (hGPLVM in their case) also allows to obtain the 3D pose.

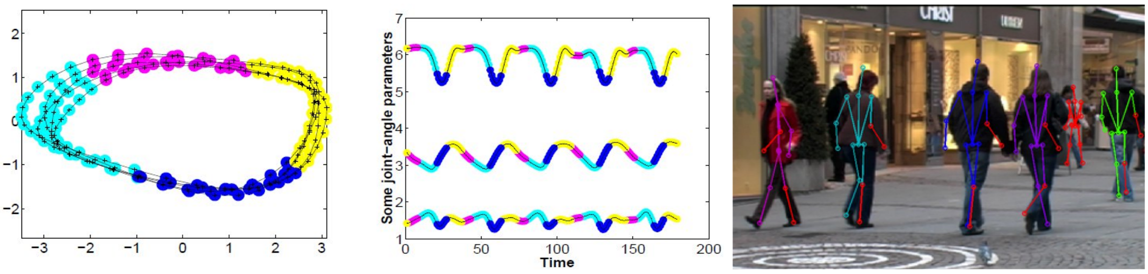


Figure 3.7: Strong motion priors. Left: Low dimensionality representation of walking cycles, from [Agarwal and Triggs, 2004]. Center: Their time correspondence with joint angles. Right: [Andriluka et al., 2010].

Advances on 2D human body representation: Given the large size of the state space of human poses, a compromise between prior information and generalization capabilities is usually made. Strong motion priors limit the generalization capabilities when observing untrained poses. It can be fixed used weak motion priors such as smoothness factors, but in return it introduces the need for a strong body model prior, in order to maintain the tractability of the proposal. To that end, several developments have recently been proposed. In Figure 3.8 (left) the traditional box-based pictorial structures model can be seen. While it provides the basic structure for the cinematic chain, it often leads to overlap problems because of the coarse representation. Because of its general usability across different subjects, viewpoints and scales it has however been extensively used. [Freifeld et al., 2010] proposed realistic contours, that provided a richer representation while keeping generalization capabilities. It enables therefore to better weight the pose likelihood resulting in a higher global accuracy. In [Guan et al., 2010] clothing is taken into account, by explicitly reasoning on the underlying body pose that is occluded by wearing different costumes. In [Zuffi et al., 2012] the problem is approached via deformable structures. See Figure 3.8 (center) for graphical representations. A different concept is to model the body with 3D meshes or primitives like in [Sigal et al., 2007], and then calculate its image projection. This however leads to computational and generalization problems.

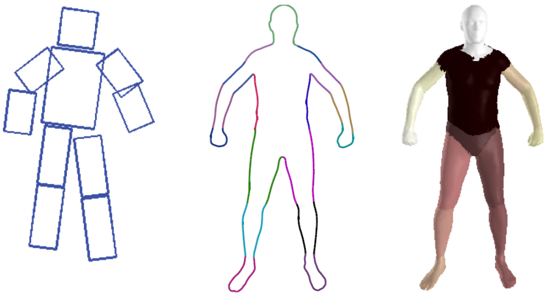


Figure 3.8: Evolution of pictorial structures, from [Freifeld et al., 2010]. Left: traditional box-based approach. Center: Eigenshapes. Right: Mesh-based approach.

Closest techniques to our proposal: In [Daubney et al., 2009] a motion-based part detector is used, by first tracking image features and then classifying them according to their motion patterns. However, the image features can get too sparse depending on the texture of the different surfaces, which does not guarantee good results in low texture sequences. In [Buehler et al., 2011], upper body motion capture is performed for language sign classification in long videos. Arm detections are used in order to disambiguate difficult hand detections. The torso shape is also inferred from a series of heuristics. However, this method assumes that the hands are generally visible, and the clothing is not problematic. In [Sapp et al., 2011], optical flow and color are used to detect hands, with body part detectors being one of the image features, and [Fragkiadaki et al., 2013] relies on optical flow and a precise 2D silhouette body model. In [Zuffi et al., 2013] a hand detector is trained with optical flow to then interpolate between correct guesses. However, they use image contours, which are prone to errors in certain situations, as we show in the results section. To conclude, in [Kazemi et al., 2013] multiple body part detectors are trained with Random Forests and per-pixel HOG features, which still inherits some of the appearance-dependence problems that edges entail.

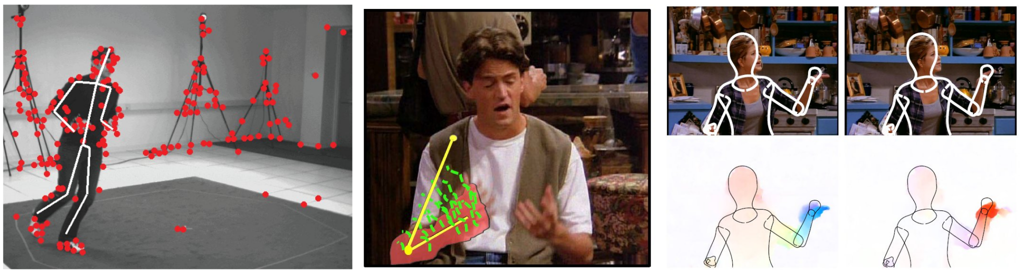


Figure 3.9: Closest works to ours. Left: [Daubney et al., 2009]. Center: [Fragkiadaki et al., 2013]. Right: [Zuffi et al., 2013].

3.4.1.3 Conclusion

The current tendencies for monocular RGB motion capture can be grouped into using multiple BPDs (when problems like overlapping, interference, or background-appearance dependency become important), having strong motion models, strong body models, or using appearance-relying methods with a torso detector. A full classification of the state-of-the-art can be seen in Table 3.2.

Two of our contributions follow the monocular motion capture approach. The RGB-only method that we present in Chapter 6 uses highly-robust appearance features, coupled with strong model and motion priors in order to efficiently extract the body pose from the position of the hands and face in the image. The method that we propose in Chapter 7 by contrast does not impose explicit motion or model priors, and benefits from robust appearance-invariant features. In addition, it improves the performance-to-computing time ratio of the state of the art.

It is not easy to predict the future of monocular markerless motion capture given the many approaches that generates. It would enormously benefit from advances in terms of feature description, similar to that from Haar to HOG. Techniques such as Deep Neural Networks (DNN) are taking off in many fields, which can be tailored for this problem, potentially increasing the performance.

Table 3.2: Summary of the reviewed monocular motion capture works.

Work	Appearance	Model prior	Mov prior	BPD
[Felzenszwalb and Huttenlocher, 2000]	Yes	Strong	No	Yes
[Sidenbladh et al., 2000]	Yes	Strong	Strong	No
[Song et al., 2001]	Yes	Weak	Strong	No
[Ramanan and Forsyth, 2003]	Yes	Strong	Weak	No
[Agarwal and Triggs, 2004]	Yes	Strong	Strong	Yes
[Elgammal and Lee, 2004]	Yes	Weak	Strong	No
[Grochow et al., 2004]	Yes	Weak	Strong	No
[Howe, 2005]	No	Strong	Strong	No
[Agarwal and Triggs, 2006]	Yes	Strong	Strong	No
[Urtasun Sotil, 2006]	Yes	Strong	Strong	No
[Brox et al., 2007]	Yes	Strong	Weak	No
[Noriega and Bernier, 2007]	Yes	Strong	Strong	No
[Sigal et al., 2007]	Yes	Strong	No	No
[Zhang et al., 2008]	Yes	Strong	Weak	No
[Bourdev and Malik, 2009]	No	Strong	No	Yes
[Daubney et al., 2009]	No	Strong	Strong	Yes
[Kulic et al., 2009]	No	Strong	Strong	Yes
[Andriluka et al., 2010]	Yes	Strong	Strong	Yes
[Eichner and Ferrari, 2010]	Yes	Weak	Strong	Yes
[Fossati et al., 2010]	Yes	Strong	Strong	No
[Freifeld et al., 2010]	Yes	Weak	No	Yes
[Kjellstrom et al., 2010]	Yes	Strong	Weak	No
[Chen et al., 2011]	Yes	Strong	Weak	No
[Kuo et al., 2011]	Yes	Strong	No	Yes
[Sapp et al., 2011]	Yes	Strong	Weak	Yes
[Wang and Koller, 2011]	Yes	Strong	No	Yes
[Yang and Ramanan, 2011]	Yes	Strong	No	Yes
[Eichner et al., 2012]	Yes	Weak	Weak	Yes
[Sun et al., 2012]	Yes	No	Strong	Yes
[Zuffi et al., 2012]	Yes	Strong	No	No
[Dantone et al., 2013]	No	Weak	No	Yes
[Fragkiadaki et al., 2013]	Yes	Weak	Strong	Yes
[Gkioxari et al., 2013]	No	Weak	No	Yes
[Hara and Chellappa, 2013]	No	Weak	No	Yes
[Ladicky et al., 2013]	Yes	No	No	No
[Ramakrishna et al., 2013]	Yes	Weak	Weak	No
[Sapp and Taskar, 2013]	No	No	No	Yes
[Simo-Serra et al., 2013]	No	Weak	No	Yes
[Wang et al., 2013]	Yes	Weak	Strong	Yes
[Yu et al., 2013a]	No	Weak	No	Yes
Ours (hand tracking)	Yes	Strong	Strong	No
Ours (Random Forests)	No	No	No	No

3.4.2 Multi-camera approach

Given that the human body movement is intrinsically non-linear and high-dimensional, having several viewpoints of the body helps to remove ambiguities. Although a lot of alternatives exist, the traditional approach when having multiple cameras, is to first obtain multi-view silhouettes of the body and then iteratively adapt a model to these silhouettes. We group the state of the art into silhouette-based approaches and other methods.

3.4.2.1 Silhouette-based

Traditional methods: The basis is to first extract body silhouettes from the images, and then try to explain those observations with body models of different kinds. Thus it consists of a high-dimensional optimization problem, that is usually tackled by minimizing an energy function. [Delamarre and Faugeras, 1999] proposed one of the first methods of this kind. Forces between the silhouette edge are modeled to adjust the observations to a 3D model. [Deutscher, 2000] proposed one of the biggest advances in the field: the Annealed Particle Filter (APF). It progressively searches along the high dimensionality space in order to avoid local minima. In addition to silhouettes, edges are used as cue. In [Cheung et al., 2003] a Visual Hull (VH) representation of the scene is retrieved in order to discover the joints of the body on the fly. In [Theobalt et al., 2004] 3D flow and appearance models are used to refine the estimation, which is initialized only with the silhouette. In [Corazza et al., 2006] VH and an APF are applied for a highly accurate biomechanical analysis. In [de Aguiar et al., 2007] SIFT features are added and optical flow is used to track 3D points over time. In [Michoud et al., 2007] a series of heuristics are proposed and coupled with a simple body model tracking to achieve real-time performance. In [Ballan and Cortelazzo, 2008] silhouettes are complemented with optical flow, by using a hierarchical bone model of the body. In [Gall et al., 2009] not only pose tracking is performed, but also surface estimation. To that end, they employ subject-specific high quality meshes obtained through laser scanning. In [den Bergh et al., 2009] the use of Haar-like features is extended to 3D with the use of VH volumes in order to recognize body positions. In [Wu et al., 2012b] the cameras in the room are sequentially triggered to simulate higher frame rates, in order to attenuate the motion blur effects in high speed movements.

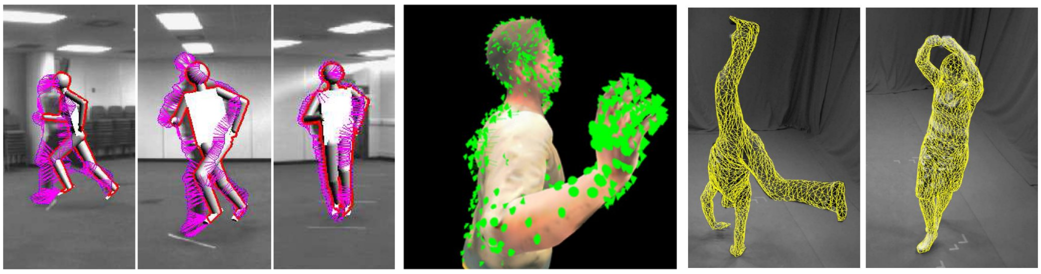


Figure 3.10: Shape from silhouette methods. Left: [Delamarre and Faugeras, 1999]. Center: [Theobalt et al., 2004]. Right: [Gall et al., 2009].

Extra cues: The difficulty of the problem has led to the search for extra cues. In some approaches the underlying articulated model is inferred by obtaining the spine of the volumetric reconstruction. In [Brostow, 2004] the spine of articulated entities is recovered. When applied to humans, it is capable of finding the underlying articulated structure. In [Matthias Straka and Bischof, 2011] a skeletal graph is extracted from the VH, which is later labeled and applied to a skeleton model in order to get the pose in real time.

Other approaches like [Balan et al., 2007] exploit the additional information that shadows provide as extra silhouettes. This effectively increases the number of viewpoints used, although shadow retrieval is not always possible, and their quality is often compromised by the lighting and floor material used.

Parsing physical feasibility of the motion is another approach that improves the search in the very large state space. In [Vondrak et al., 2008] and [Brubaker et al., 2009] a probabilistic physical simulation of the human body is built in order to correct wrong poses and estimate ongoing forces.

The clothing problem has also been addressed in the literature. In [Balan and Black, 2008] skin color information is added to the volumetric reconstructions, and then use a body shape database to estimate the body configuration under different types of clothing. In [Ukita et al., 2009] Visual Hulls are trained in order to obtain the intrinsic pose. This approach however is subject-specific.

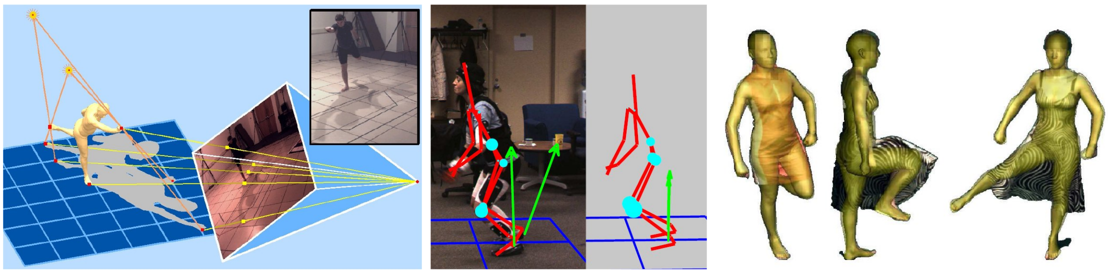


Figure 3.11: Methods that use extra cues. Left: [Balan et al., 2007]. Center: [Brubaker et al., 2009]. Right: [Balan and Black, 2008].

Closest techniques to our proposal: To the best of our knowledge, very few approaches perform simultaneous motion capture and action recognition, and none of them when having several people in the scene. In [Yao et al., 2012] the ongoing action is estimated from 2D views from features like optical flow, and the resultant prior information is used to constrain the search in the state space. It also allows to obtain an initial pose, that is later refined with the silhouette observations.

Regarding multi-person motion capture, in [Egashira et al., 2009] blob color information is analyzed in order to distinguish between weakly interacting people. In [Liu et al., 2011] the subjects interact heavily. Detailed subject-specific mesh body models are used with the help of color to apply a subject label to the silhouettes, and then perform the usual process with separated trackers.

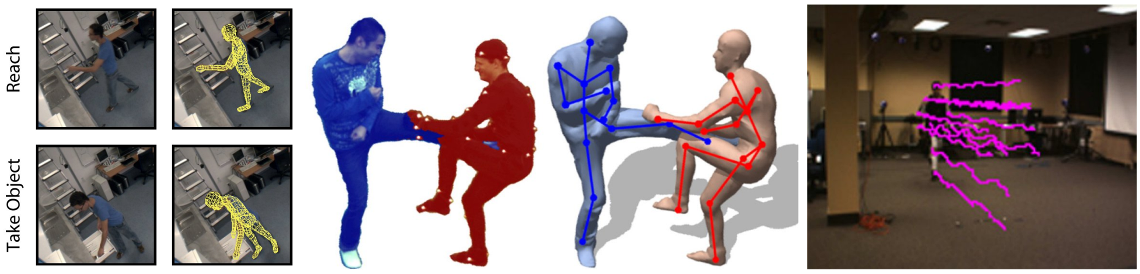


Figure 3.12: Closest techniques to ours. Left: [Yao et al., 2012]. Center: [Liu et al., 2011]. Right: [Livne et al., 2012].

3.4.2.2 Other methods

[Gavrila and Davis, 1996] is one of the first works using chamfer distances and edges to adjust a full body model to image observations, which contain two interacting persons. Manual initialization is required, and tracking is performed. In [Drummond and Cipolla, 2001] a real-time algorithm is developed, which propagates the probability distributions through a kinematic chain, to obtain maximum a posteriori estimates of the body configuration. They use edges as measurements. In [Sigal et al., 2004] part detectors are used with multi-scale edge and ridge filters. They combine the cues obtained in each viewpoint. In [Balan and Black, 2006] an adaptive appearance model is employed to detect and track the different body parts. In [Stoll et al., 2011] the tracker is manually initialized, modeling the body through a mixture of Gaussian that encode volume and color, and achieving good frame rates. In [Wu et al., 2012a] method invariant to uncontrolled and changing illumination is developed by using a color model and estimated lighting model of the person. This way they are also able to retrieve highly detailed geometry. In [Burenius et al., 2013], [Amin et al., 2013] and [Kazemi et al., 2013] part detectors in multiple views are used to extend the pictorial structures concept to 3D, showing that it is a tractable problem.

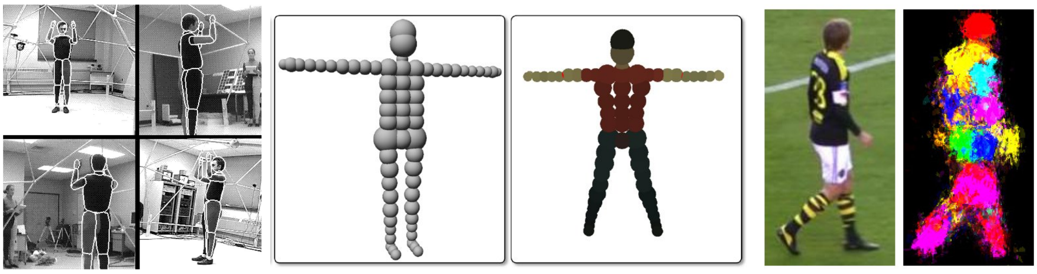


Figure 3.13: Other approaches. Left: [Gavrila and Davis, 1996]. Center: [Stoll et al., 2011]. Right: [Kazemi et al., 2013].

3.4.2.3 Conclusion

Along the years, there has been a focus on iteratively improving performance numbers in the silhouette-based case, as it can be seen in Table 3.3. Therefore, the working line generally shifted towards performance motion capture, in which controlled settings are used. Accuracies are improved at the cost of imposing a very strong body model, for example using subject-specific laser-scanned 3D body meshes. This results in smaller generalization capabilities, and heavily depend on the quality of silhouettes. A common approach to address the issue is to add strong motion priors, which can translate again into poor generalization.

Although our multi-camera proposal (described in Chapter 5) uses strong motion and body models, we provide an extra refinement layer that significantly improves the generalization capabilities, while retaining the advantages that the use of strong priors give. In addition, we track several persons simultaneously.

Recently, multi-camera approaches are integrating BPDs into the framework, just like in the monocular case. It becomes therefore the clear trend when looking for poses in color images, as it tackles the long-standing initialization problem while removing the need for strong body model priors. Issues with noise and background clutter are however still present in the BPDs.

3.4.3 Depth approach

Range cameras appeared recently, providing relative 3D information by using either structured lighting or time of flight technology (see Figure 3.14). Some sensors include also color information, generating images known as RGBD. We have grouped the literature into (a) methods that only use depth and (b) RGBD methods.

3.4.3.1 Purely depth

In one of the first works, [Zhu and Fujimura, 2007], every pixel of the image is classified into body parts, with a machine learning approach. In [Schwarz et al., 2011] anatomical key-points are detected with the help of geodesic distance. As geodesic distances are problematic in the case of self-occlusions, optical

Table 3.3: Summary of the reviewed multi-view motion capture works.

Work	Appearance	Model prior	Mov prior	BPD	#Persons	AR
[Gavrila and Davis, 1996]	Yes	Strong	Weak	No	>1	No
[Deutscher, 2000]	Yes	Strong	Weak	No	1	No
[Delamarre and Faugeras, 1999]	Yes	Strong	Weak	No	1	No
[Drummond and Cipolla, 2001]	Yes	Weak	Weak	No	1	No
[Cheung et al., 2003]	Yes	Weak	Weak	No	1	No
[Brostow, 2004]	Yes	No	Weak	No	1	No
[Sigal et al., 2004]	Yes	Strong	Strong	Yes	1	No
[Theobalt et al., 2004]	Yes	Strong	Weak	No	1	No
[Balan and Black, 2006]	Yes	Strong	Weak	No	1	No
[Corazza et al., 2006]	Yes	Strong	Weak	No	1	No
[de Aguiar et al., 2007]	Yes	Weak	Weak	No	1	No
[Balan et al., 2007]	Yes	Strong	Weak	No	1	No
[Michoud et al., 2007]	Yes	Weak	Weak	No	1	No
[Balan and Black, 2008]	Yes	Weak	Weak	No	1	No
[Ballan and Cortelazzo, 2008]	Yes	Strong	Weak	No	1	No
[Vondrak et al., 2008]	Yes	Strong	Weak	No	1	No
[Brubaker et al., 2009]	Yes	Strong	Strong	No	1	No
[den Bergh et al., 2009]	Yes	Weak	Weak	No	1	Yes
[Egashira et al., 2009]	Yes	Strong	Weak	Yes	>1	No
[Gall et al., 2009]	Yes	Strong	Weak	No	1	No
[Ukita et al., 2009]	Yes	Strong	Strong	No	1	No
[Stoll et al., 2011]	Yes	Strong	No	No	1	No
[Matthias Straka and Bischof, 2011]	Yes	Weak	Weak	Yes	1	No
[Liu et al., 2011]	Yes	Strong	Weak	No	>1	No
[Livne et al., 2012]	Yes	Strong	Weak	No	1	No
[Wu et al., 2012a]	Yes	Strong	Weak	No	1	No
[Yao et al., 2012]	Yes	Strong	Weak	No	1	Yes
[Burenius et al., 2013]	Yes	Strong	Weak	No	1	No
[Amin et al., 2013]	Yes	Strong	Weak	No	1	No
[Kazemi et al., 2013]	No	Weak	No	Yes	1	No
Ours	Yes	Strong	Strong	No	>1	Yes

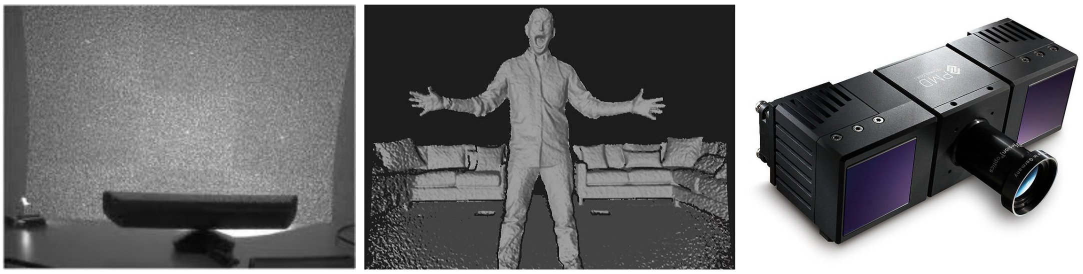


Figure 3.14: Range camera technology. Left: Kinect's structured light. Center: Kinect 2.0 increased resolution and precision. Right: Time of flight commercial camera.

flow is used to remove parts of the graph that are disconnected in practice. The work [den Bergh and Gool, 2011] is not strictly motion capture, but they explore techniques for segmenting and differencing hands from head in RGBD data. A hand detector is used together with depth distance from the face. [Ganapathi et al., 2010] proposes an almost real-time method, by combining a generative model with a body part detector that is first introduced in [Plagemann et al., 2010]. It is based in geodesic extrema detection and salient point analysis. In [Schwarz et al., 2010] action recognition is first performed by using a holistic feature based on depth silhouette analysis and a probabilistic framework. The pose is refined to provide a better generalization. In [Baak et al., 2011] the geodesic extrema are obtained to perform database lookup of the poses, that is later refined through a quick optimization process to get real-time performance. In [Lopez-Mendez et al., 2011] a heuristic approach is used to get the upper body pose in a seated setting. The most famous work on human motion capture is perhaps [Shotton et al., 2011,Shotton et al., 2012]. The problem is approached as per-pixel body part classification. A Random Forest is trained with an enormous amount of data. Thanks to motion capture and synthetically generated scenes they learn scale, clothing, or point of view invariance. Impressive performance is obtained at 200 frames per second, under the implementation of the popular Kinect device from Microsoft. In [Hernandez-Vela et al., 2012] the classification from [Shotton et al., 2011] is improved by adding a multi-label Graph Cuts as a post-processing step. In [Taylor et al., 2012] regression is used instead of classification to find corresponding points between a model and the observations, and then recover the pose. Finally, in [Holt et al., 2013] regression on joints position is performed to get the global pose.

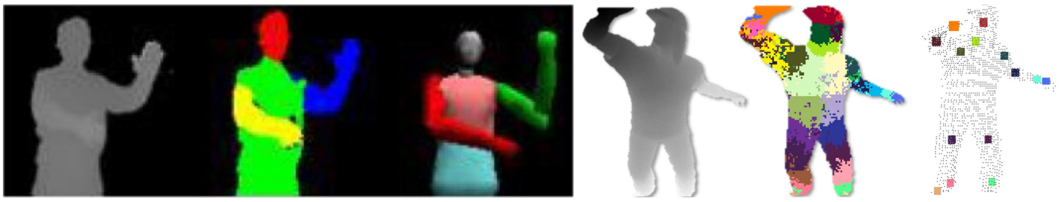


Figure 3.15: Body part classification from depth images. Left: [Zhu and Fujimura, 2007]. Right: [Shotton et al., 2011].

3.4.3.2 Depth and color

[Knoop et al., 2006] proposed one of the first works on human motion capture with range cameras. It also explores RGB and stereo configurations. It is based on Iterative Closest Point (ICP, an algorithm employed to minimize the difference between two clouds of points) to track the body configuration. To conclude, [Schwarz et al., 2011] uses optical flow to distinguish the limbs in moments where depth fails to give reliable geodesic extrema. This is therefore the **closest work to ours**, even if it does not take place in a torso-only setting. In [Gall et al., 2011] object identification is explored by looking at how a person interacts with them. Object detectors are used for obtaining sparse hand measurements, and then a mesh model is tracked in-between detections, requiring a calibrated camera.

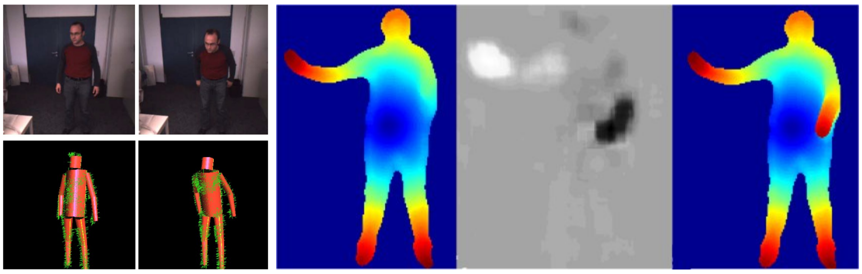


Figure 3.16: Motion capture from RGBD images. Left: [Knoop et al., 2006]. Right: [Schwarz et al., 2011].

3.4.3.3 Conclusion

Depth imagery is the new trend in markerless motion capture, and rightly so: it is highly appearance-invariant and provides reliable 3D dense information while only needing one point of view. Microsoft's development of a depth-based solution (formerly known as Natal project) gave an big momentum to the community in order to find new approaches.

In Table 3.4 a summary of the most relevant works can be found. In contrast with most proposals, in our approach (described in Chapter 6) we merge color and depth information in order to provide a robust hand tracker, which in turn is used to retrieve the body position in a database lookup approach.

However, there are some important issues with depth. There is a permanent problem regarding already recorded footage. For example, existing databases might only use traditional images, making it impossible to apply depth to obtain the body poses. Also, most range cameras do not work with sun light, restricting their use to indoor settings. In addition, there is a small distance range from the camera in which depth can be computed, limiting the scenarios in which it can be used.

Table 3.4: Summary of the reviewed depth motion capture works.

Work	Appearance	Model prior	Mov prior	RGBD
[Knoop et al., 2006]	Yes	Strong	Strong	Yes
[Zhu and Fujimura, 2007]	No	Strong	Weak	No
[Schwarz et al., 2011]	No	Weak	No	No
[den Bergh and Gool, 2011]	No	Strong	Strong	No
[Ganapathi et al., 2010]	No	Strong	Weak	No
[Plagemann et al., 2010]	No	Weak	No	No
[Schwarz et al., 2010]	No	Weak	Strong	No
[Baak et al., 2011]	No	Strong	Weak	No
[Gall et al., 2011]	No	Strong	Weak	Yes
[Lopez-Mendez et al., 2011]	No	Strong	Weak	No
[Shotton et al., 2011]	No	Weak	No	No
[Schwarz et al., 2011]	No	Weak	No	Yes
[Shotton et al., 2012]	No	Strong	Weak	No
[Hernandez-Vela et al., 2012]	No	Weak	Weak	No
[Taylor et al., 2012]	No	Weak	No	No
[Holt et al., 2013]	No	Weak	No	No
Ours	Yes	Strong	Weak	Yes

3.5 Action recognition

Activity recognition is an important objective in computer vision, motivated by the potential of many applications. An extensive review can be found in [Poppe, 2010]. Typically used sensors are single or multiple color cameras, and depth cameras. Low-level, mid-level and high-level features are extracted from sensor data, such as edges, color, HOG, geodesic extrema or time voxels. They are then usually used as input for a classifier, and their performance reported in a wide range of available activity recognition datasets. Most recent works focus on activity recognition in the wild [Liu et al., 2009], as it still is an open problem. Head and face information is commonly used to improve performance [Scherer et al., 2013]. In addition, to the best of our knowledge speaking status has never been used as a feature to improve the classification accuracy. This will be explored in Chapter 8.

3.5.1 Action recognition from motion capture

Action recognition with computer vision can be applied to automatically obtain body communication cues. The most traditional approach in these systems is to first get the body pose and then analyze it [Gall et al., 2012]. Most recent works are able to do this without getting the body pose (i.e. without performing motion capture) through diverse techniques such as [Sadanand and Corso, 2012, Laptev, 2005]. However, it has been shown in practice [Angela Yao and Gool, 2011] that, even though it is possible to perform activity recognition without knowing the body pose, it is still beneficial to know it. As [Muller et al., 2005] confirms, a very efficient action retrieval scheme from a very large database can be implemented by using simple relational features with motion capture data.

3.5.2 Conclusion

There are plenty of methods to perform action recognition: some in the wild, some in controlled settings (see Table 3.5). The general approach is to obtain spatio-temporal features from the image in order to later classify them. It is clear that they benefit from motion capture, as it makes the ongoing action a lot more explicit.

Table 3.5: Summary of the reviewed action recognition works.

Work	In the wild	Motion capture
[Laptev, 2005]	Yes	No
[Muller et al., 2005]	No	Yes
[Liu et al., 2009]	Yes	No
[Gall et al., 2012]	No	Yes
[Angela Yao and Gool, 2011]	No	Yes
[Sadanand and Corso, 2012]	No	Yes
[Scherer et al., 2013]	No	Yes
Ours	No	Yes

3.6 Conclusion

In this chapter we showed the need of automatic information extraction in the field of psychology. Motivated by that, we found that specifically body posture is key to extract. Given the difficulty of the task, it is a long-standing problem of the literature. Surprisingly, markerless motion capture has very rarely been used as part of an automatic annotation scheme in psychological studies. In the present thesis we aim to fill that void in several sensing scenarios.

Chapter 4

Datasets

4.1 Introduction

In this chapter we present the data used for evaluating and training the methods described in the thesis. Some of the databases (HumanEva and ChaLearn) are public, which allow for comparison with the state-of-the-art, and are useful to show generalization of performance. In addition to those, a private real job interview dataset plays an important role in the development of our algorithms, as explained in Section 4.3. Some of the material included has appeared originally in [Nguyen et al., 2013a], [Marcos-Ramiro et al., 2013] and [Nguyen et al., 2013b].

4.2 Multiple and single viewpoints: HumanEva-I

The HumanEva-I dataset contains 7 calibrated video sequences (4 grayscale and 3 color) that are synchronized with 3D body poses obtained from a motion capture system. The database videos contain 4 subjects (S1-S4) performing a 6 common actions (e.g. walking, jogging, gesturing, etc.). The error metrics for computing error in 2D and 3D pose are provided to users. The dataset contains training, validation and testing (with withheld ground truth) sets.

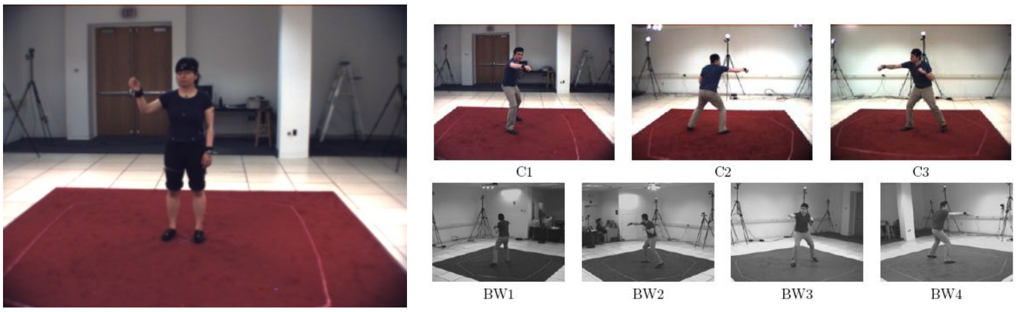


Figure 4.1: HumanEva. Left: S1 example. Right: Seven calibrated cameras.

We use two different subsets of the HumanEva-I database along this thesis:

- **Subset A**: The training and testing sequences of S1 and actions ”walking”, ”jogging” and ”box” are employed. All 7 cameras are used. It is used to evaluate the performance of our multi-camera approach (see Section 5.7).

- **Subset B**: The training and testing sequences of S1 and S3 and actions ”gestures” and ”box”. Only camera C1 is used, as it is the frontal, color camera recommended for use in monocular appications, like our appearance-invariant approach (see Section 7.5).

4.3 Single viewpoint: Job interview database

4.3.1 Introduction

The data that is presented in this section has been gathered within the context of the project ”SONVB: Sensing and Analyzing Organizational Nonverbal Behavior”, funded by the Swiss National Science Foundation. As a part of it, Laurent Nguyen (Idiap Research Institute), Denise Frauendorfer (University of Neuchatel) and Kenneth Funes (Idiap Research Institute) designed a intelligent room and a set of psychological experiments to record information of participants in a real job interview. Thanks to the existing collaboration between University of Alcala and Idiap Research Institute, this data has been used and enriched through different kinds of annotations, as described in the following sections.

In Section 4.3.2 an overview of the original data retrieval process is presented, while in Section 4.3.3 the additions that were made to that data for its use in the present work are discussed.

4.3.2 Data retrieval

The four-hour job at stake consisted in recruiting people on the street for psychology experiments, and was remunerated with 200 Swiss Francs. In order to gather subjects for the study, an open position was advertised in three different Swiss universities using multiple communication channels. Due to the large participation of students, the average participant age was 24 years, with a standard deviation of 5.7 years.

Before starting the interview, applicants were asked to fill in a consent form where they accepted that the interview would be audio- and video-recorded, and that the data could be used for research purposes. They then completed a questionnaire to assess their Big-Five personality scores (see Section 3.2).

For the interview itself a structured design was used, where the sequence of instructions and questions remained constant across interviews in order to ensure that comparisons could be made between job candidates. The job applicants were asked to answer four behavioral questions related to past experiences in specific situations requiring specific social skills, namely cases:

- Where communication skills were required.
- Where persuasion skills were required.
- Of conscientious/serious work.
- Where stress was properly managed.

The interviews were dyadic, and the recruiter was facing the job applicant. The protagonists were seated at both sides of a table (see Figure 4.2). For detailed information on some of the most important data that was collected, see Tables 4.1, 4.2, 4.3, 4.4, 4.5 and 4.6.

4.3.2.1 Technical setup

An intelligent room was set up at the University of Neuchatel in order to get multimodal information from job interviews. This set up consists on:

- A Kinect device: it provides RGB data (640x480, 30 fps) and range data (640x480, 30 fps).
- An HD color camera: it provides RGB data (720p, 26.7 fps).
- A Microcone [Mic,]: it provides directional audio recordings and automatic speaking turn segmentation (48 kHz).
- A qSensor[1]: it provides skin conductivity along time, skin temperature and acceleration.

The Kinect device and the HD color camera were not calibrated. However, all devices were time-synchronized via software. Recordings from the high quality color camera and Microcone are available for all the 60 subjects. Kinect recordings are available for 47 subjects, and qSensor data for only 20 subjects. An illustration of the interview setup is illustrated in Figure 4.2.

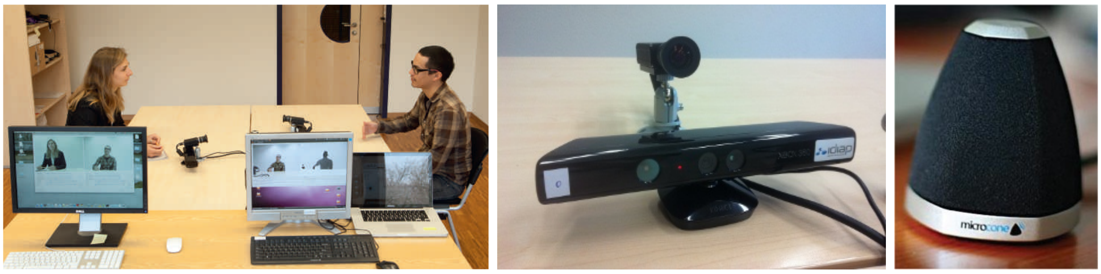


Figure 4.2: Setting and data collection. Audio and video for interviewer and participant were recorded and synchronized. Left: general setup. Center: kinect and HD cameras. Right: microcone.

4.3.2.2 Social variables

The Big-Five personality variables were assessed using the standard NEO-FFI-R questionnaire [Costa and McCrae, 1992], which is formed by 60 items (12 per personality dimension). Hirability scores were manually coded by a task-trained M.S. student in organizational psychology. The annotations were completed after watching the full audio-video recording of the interview, including both the recruiter and the job candidate.

Four hirability scores were coded based on the answers to the four behavioral questions: communication, persuasion, consciousness, and stress resistance. A score between 1 (very low) and 5 (very high) was assigned for each variable. In addition to these four scores, a general score (hiring decision) was given based on the general impression made by the applicant, ranging from 1 to 10. In order to validate the reliability of the annotator, a second expert coder rated 10 job interviews. The inter-rater agreement was satisfactory, with Pearson's correlation coefficient ranging from 0.69 for consciousness to 0.99 for persuasion.

The following Tables 4.1-4.4 show, among other data, the self-reported personality, the perceived personality (see Table 3.1), the action distribution (see Section 4.3.3), and the hirability scores from trained human resources personnel:

[1] http://www.qsensortech.com

Table 4.1: Subjects 1-15 information. Gray rows represent interviews without associated Kinect data.

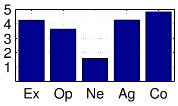
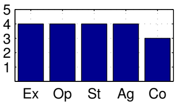
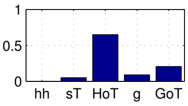
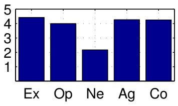
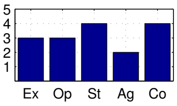
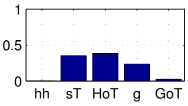
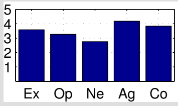
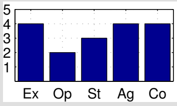
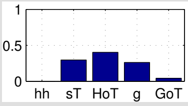
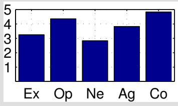
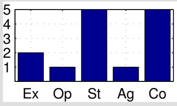
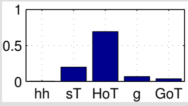
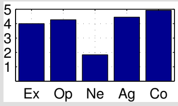
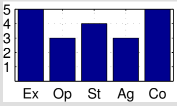
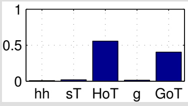
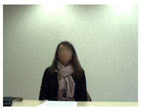
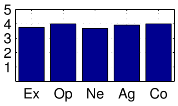
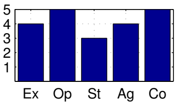
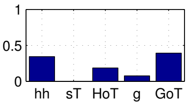
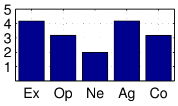
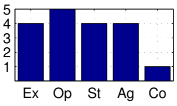
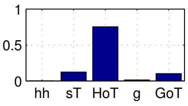
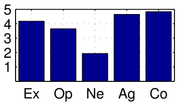
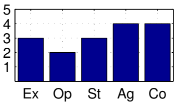
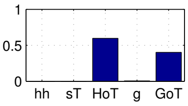
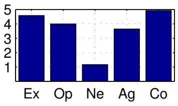
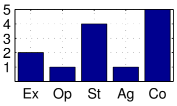
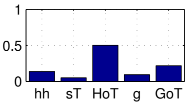
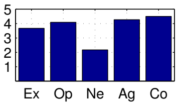
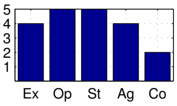
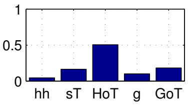
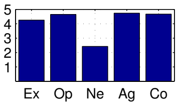
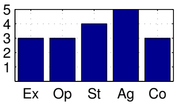
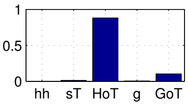
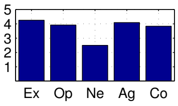
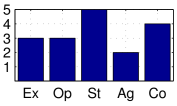
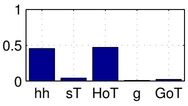
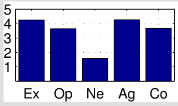
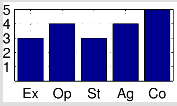
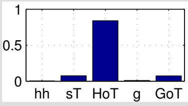
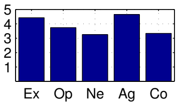
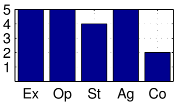
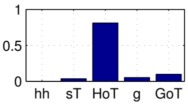
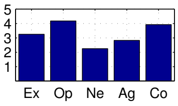
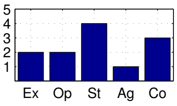
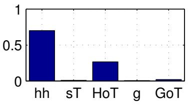
	Exp	Name	Pers. (self)	Pers. (perc)	Action distr.	Hir. score (1-5)
	01	ab17				4
	02	ac48				4
	03	ar05				2
	04	ar53				3
	05	as04				5
	06	as54				5
	07	av14				3
	08	az58				3
	09	ba13				4
	10	bg11				1
	11	bl15				3
	12	bl16				4
	13	bp08				4
	14	bs19				2
	15	bs60				1

Table 4.2: Subjects 16-30 information. Gray rows represent interviews without associated Kinect data.

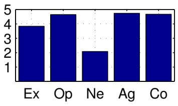
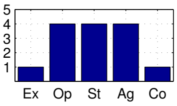
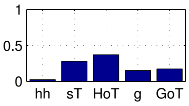
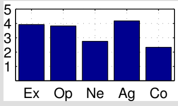
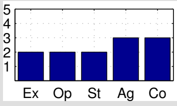
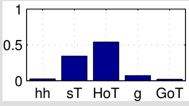
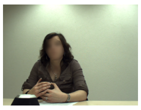
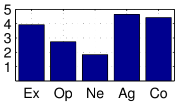
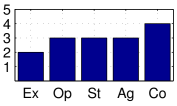
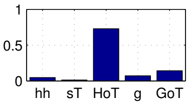
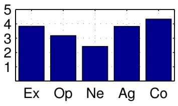
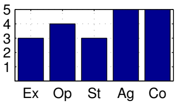
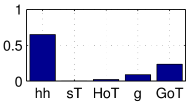
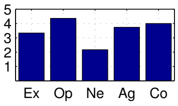
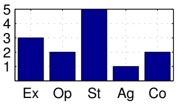
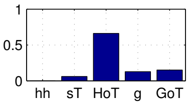
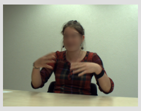
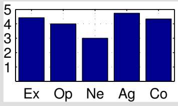
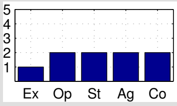
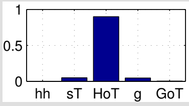
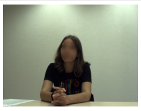
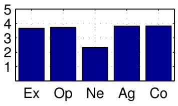
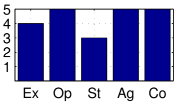
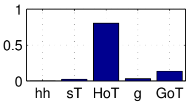
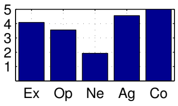
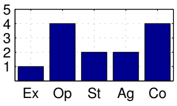
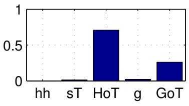
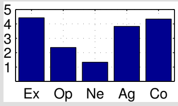
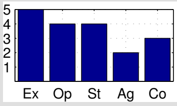
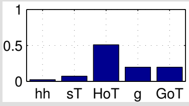
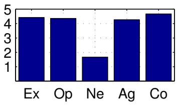
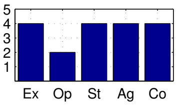
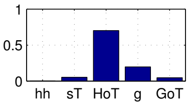
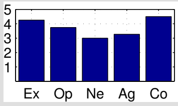
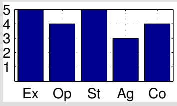
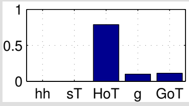
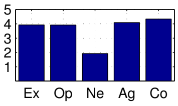
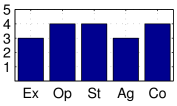
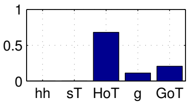
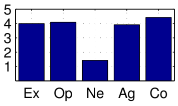
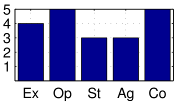
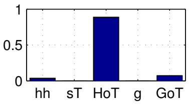
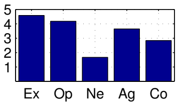
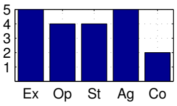
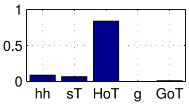
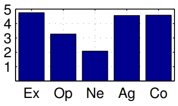
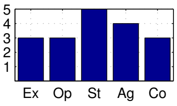
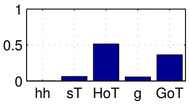
	Exp	Name	Pers. (self)	Pers. (perc)	Action distr.	Hir. score (1-5)
	16	ce34				1
	17	cg09				3
	18	cg35				4
	19	cm25				2
	20	cm40				3
	21	cv07				2
	22	dg18				3
	23	dj23				3
	24	eg10				4
	25	gs57				5
	26	hc02				4
	27	ia31				2
	28	jb27				3
	29	jr43				4
	30	kc51				3

Table 4.3: Subjects 31-45 information. Gray rows represent interviews without associated Kinect data.

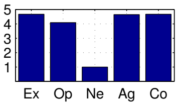
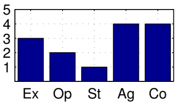
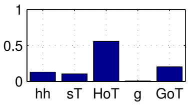
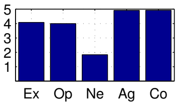
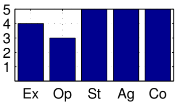
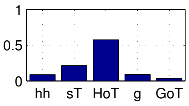
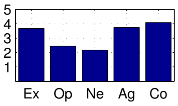
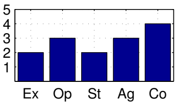
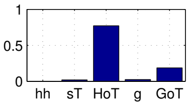
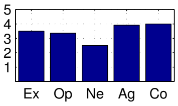
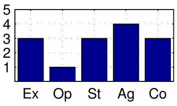
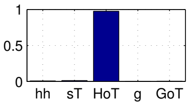
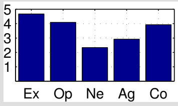
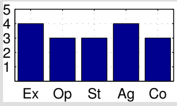
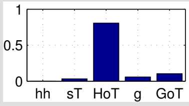
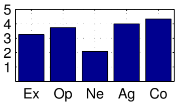
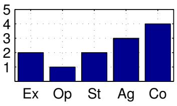
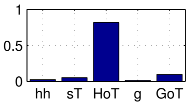
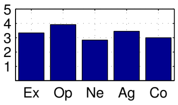
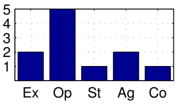
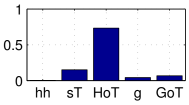
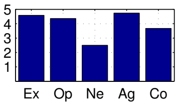
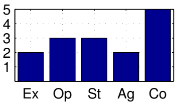
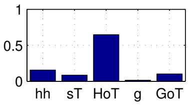
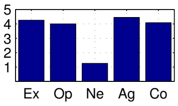
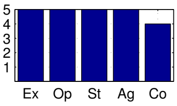
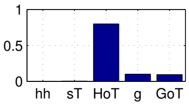
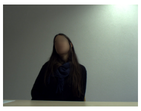
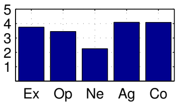
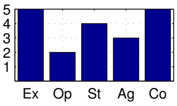
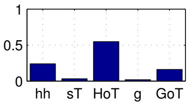
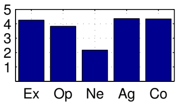
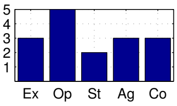
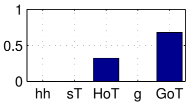
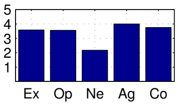
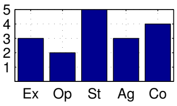
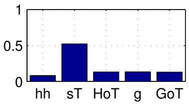
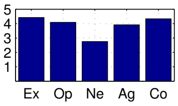
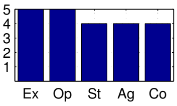
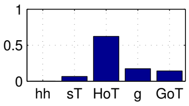
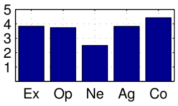
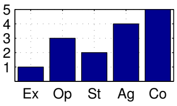
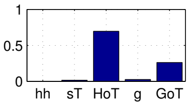
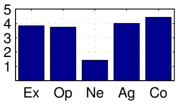
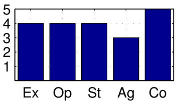
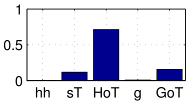
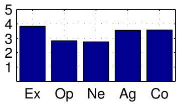
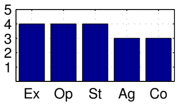
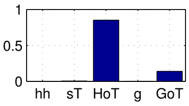
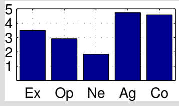
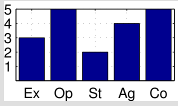
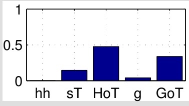
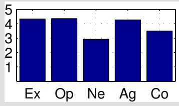
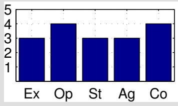
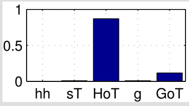
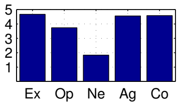
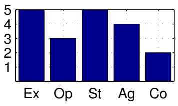
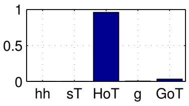
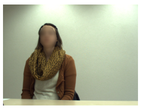
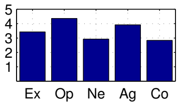
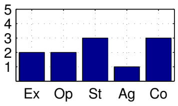
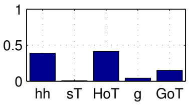
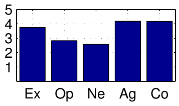
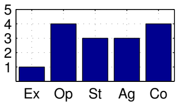
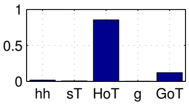
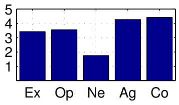
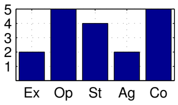
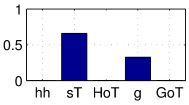
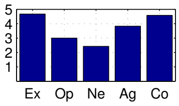
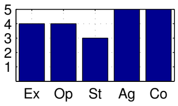
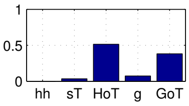
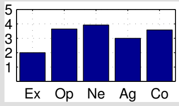
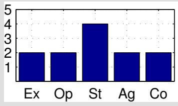
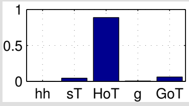
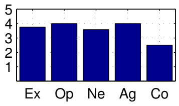
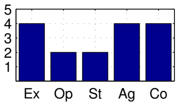
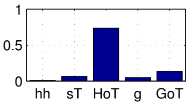
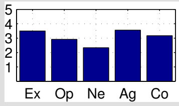
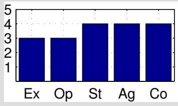
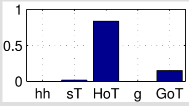
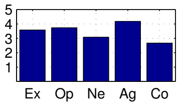
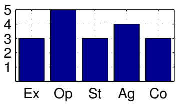
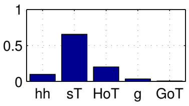
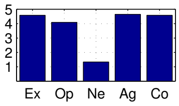
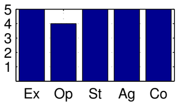
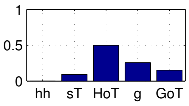
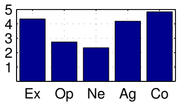
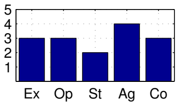
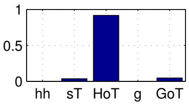
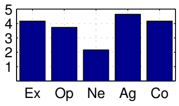
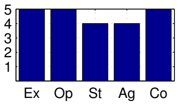
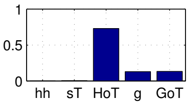
	Exp	Name	Pers. (self)	Pers. (perc)	Action distr.	Hir. score (1-5)
	31	kf29				3
	32	kr59				4
	33	la20				2
	34	lc21				3
	35	ld01				5
	36	ma37				4
	37	mf32				3
	38	mf39				3
	39	mg56				4
	40	mm33				3
	41	ms52				1
	42	ms55				3
	43	nd42				3
	44	nh41				3
	45	om47				3

Table 4.4: Subjects 46-60 information. Gray rows represent interviews without associated Kinect data.

	Exp	Name	Pers. (self)	Pers. (perc)	Action distr.	Hir. score (1-5)
	46	pc24				2
	47	pt01				2
	48	pt02				2
	49	pv26				2
	50	pv38				2
	51	rm12				3
	52	rm30				3
	53	rs50				3
	54	sc03				3
	55	sc36				3
	56	sn06				4
	57	vb45				3
	58	vn49				3
	59	wj46				3
	60	zp22				2

4.3.3 Data annotations and data splits

In order to test and validate the algorithms presented in this thesis, we annotate different subsets (A-D) of the collected information in order to have a groundtruth when evaluating our methods. Annotations consist either on specifying the ongoing activity of the participant, or the position in the image of the different body parts. This results in 4 different annotated sets of data, that we explain in Sections 4.3.3.1, 4.3.3.2, 4.3.3.3 and 4.3.3.4.

Data reduction with frame differences: As a pre-processing first step, we first reduce the amount of data to consider. By definition, in order to perform motion capture, there needs to be motion present in the image. In order to reduce the number of frames to process and annotate, we filter the segments of the video in which there is not enough image difference by using a manually set threshold on the frame difference signal. That is, we do not process the frames which do not show enough change, although we take them into account when computing the global signals.

Given the challenge that a very large database such as this one poses from a computational point of view, reducing the amount of data to process while retaining the relevant information is an important step for the pipeline of the work. In Tables 4.5 and 4.6 the video length reduction statistics can be seen: after this step, only a 46.6 % of the total data needs to be processed.

Ongoing action annotations: We annotated the ongoing actions for the whole dataset, in every half a second of the reduced length data. In total, approximately 27800 frames were manually annotated. After discussing with psychologists, all the body poses and movements were split into five different categories:

- Hidden hands (hh): the hands of the subject are not visible, for example when they are under the table or the arms are crossed.
- Gestures (g): the subject is gesticulating well above the table level.
- Hands on table (HoT): the subject is resting the hands on the table, and they are resting or showing low intensity fidgeting.
- Self touch (sT): generally, the subject is touching his/her face region.
- Gestures on table (GoT): the subject is gesticulating while the hands are slightly above or at the table level.

Initially, the last group ”gestures on table” was not included, while all the hand actions that took place on the table fell into the ”hands on table” class. However, upon development of the different algorithms and further discussion with psychologists, it became clear that the extra class ”gestures on table” was needed, in order to better group different conversational behaviors. In Figure 4.3 examples can be seen.

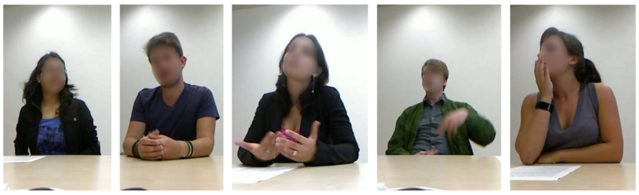


Figure 4.3: Class examples. From left to right: ”hidden hands”, ”hands on table”, ”gestures on table”, ”gestures”, ”self touch”.

Table 4.5: Subjects 1-45 interview length reduction. Gray rows represent interviews without associated Kinect data.

Exp	Original length (min)	Reduced length (min)	Percentage of original
01	10.39	4.12	39.6%
02	10.59	3.41	32.2%
03	10.40	6.53	62.7%
04	9.06	3.68	40.7%
05	11.60	6.18	53.3%
06	11.32	1.90	16.8%
07	8.20	4.33	52.8%
08	11.60	2.62	22.6%
09	14.98	8.24	55.0%
10	15.73	5.08	32.3%
11	10.04	5.56	55.3%
12	11.33	3.48	30.7%
13	14.43	8.32	57.6%
14	10.91	3.37	30.8%
15	7.59	1.42	18.7%
16	18.80	10.89	57.9%
17	9.65	6.50	67.4%
18	12.07	5.67	46.9%
19	10.48	3.48	33.2%
20	17.29	6.89	39.8%
21	10.38	6.38	61.4%
22	9.31	3.13	33.7%
23	10.73	5.09	47.5%
24	16.13	12.38	76.8%
25	8.71	5.55	63.7%
26	11.85	9.37	79.1%
27	15.57	3.70	23.8%
28	9.91	5.13	51.8%
29	10.26	5.79	56.4%
30	11.05	5.17	46.8%
31	11.92	4.28	35.9%
32	11.03	7.22	65.4%
33	7.66	4.32	56.4%
34	6.48	2.78	42.9%
35	13.02	12.40	95.2%
36	10.21	4.47	43.8%
37	16.30	7.93	48.6%
38	7.62	3.72	48.8%
39	13.85	6.99	50.4%
40	9.31	3.79	40.8%
41	10.20	1.43	14.0%
42	8.71	3.66	42.0%
43	15.41	10.90	70.7%
44	12.37	6.44	52.1%
45	10.82	4.15	38.3%

Table 4.6: Subjects 46-60 interview length reduction. Gray rows represent interviews without associated Kinect data.

Exp	Original length (min)	Reduced length (min)	Percentage of original
46	10.81	4.77	44.1%
47	9.80	8.74	89.1%
48	10.38	3.83	36.8%
49	9.57	1.56	16.3%
50	7.48	2.35	31.4%
51	18.83	10.08	53.5%
52	16.79	8.56	51.0%
53	9.49	3.61	38.1%
54	9.50	3.08	32.4%
55	10.77	6.18	57.3%
56	7.90	4.06	51.4%
57	11.50	4.84	42.1%
58	11.42	5.28	46.3%
59	8.72	3.16	36.2%
60	9.18	3.73	40.6%
Total	695.2	321.63	-
Average	11.58	5.36	46.65 %

4.3.3.1 Data subset A

The purpose of this subset is to evaluate the performance of the hand tracker and action recognition framework, when using only the HD RGB camera (see Section 6.8.2).

Hand likelihood map quality evaluation. We manually labeled the position of the hands as their pixel coordinates in the image, in a challenging 1450-frame sequence, where the interviewer wears a skin-colored scarf and has no sleeves (see Figure 4.4). It is therefore very useful to determine how well the proposed hand map behaves with the help of optical flow and edge information, in comparison to a color-based skin segmentation. The error will be measured in two ways:

- On the image plane, in pixels. It is defined as the euclidean distance between annotation and measurement.
- As a detection rate, that measures how often the number of detected hands (0, 1, or 2) is correct.

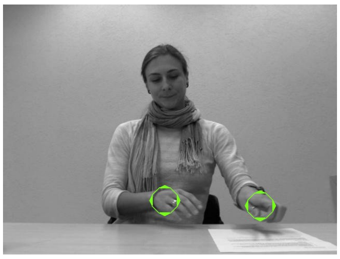


Figure 4.4: Annotated frame from subset A. Hands are manually marked along the sequence (green markers). Best viewed in color.

Action recognition performance. We manually labeled the actions performed by 8 different subjects, according to the categories ”hiddenHands”, ”gestures”, ”handsOnTable” and ”selfTouch”, previously introduced. To simplify the process, we labeled one every 15 frames (or approximately 6 tenths of a second) in the portions of the video which showed movement above the manually set threshold. This resulted into 2590 manually labeled frames, see Table 4.7 for the split per category. As performance measure, frame classification accuracy is used.

Table 4.7: Per-class split of subset A

hiddenHands	gestures	handsOnTable	selfTouch
2.23%	24.56%	67.53%	5.67%

4.3.3.2 Data subset B

A second, bigger split of data was obtained in order to have a better reliability method, when the RGB only case was extended with RGBD (see Section 6.8.3). It resulted in a total of 27 interviews (7 male and 20 female subjects) and almost 4.5 hours of audio-visual data, with an average interview length of 9.8 minutes. As in the previous case, it is designed to evaluate both the hand tracking precision and the action recognition performance.

Hand likelihood map quality evaluation. We hand-labeled a 2 minute sequence (3750 frames and 4 different subjects), in which visible 2D hand positions are annotated in every frame.

Action recognition performance. In order to evaluate action recognition algorithms, we manually labeled the actions performed by the 27 different subjects, according to the categories in Table 4.8. This resulted into 13900 manually labeled frames, see Table 4.8 for the split per category. In order to assess the reliability of annotations, a second person annotated 63 minutes of the dataset (around 5000 frames), resulting in a satisfactory inter-rater agreement with a Cohen's Kappa [Galton, 1892] of 0.81. As performance measure, we use frame classification accuracy.

Table 4.8: Per-class split of subset B

hiddenHands	gestures	handsOnTable	selfTouch	gesturesOnTable
4.97%	11.75%	51.89%	11.56%	19.83 %

4.3.3.3 Data subset C

This subset is used for training and evaluating our motion capture proposal with highly appearance-invariant features (see Section 7.5). In contrast to subsets A and B, the purpose of this subset is two fold:

- Train our methods: we need to obtain prior information about the human body in the image.
- Test our methods: in order to validate the performance after the training task.

We construct a database from a set of one-shot non-consecutive images, containing 34 different subjects (8 male, 26 female). We selected 1420 frames in which the subject is moving at least one part of the body, and distributed the resultant labeled frames as 1100 for training and 320 for testing.

The ground-truth of test images is obtained by manually labeling every pixel with 10 labels: right hand (RH) / right forearm (RF) / right arm (RA) / left hand (LH) / left forearm (LF) / left arm (LA) / head (H) / torso (T) / neck (N) / background (BG).

For the training images, sparse manual labeling has been followed. That is, we only label a few (around 100) pixels per image, in the parts that carry less uncertainty. For example, if there is a self-occlusion, only the pixels in which the different parts of the body are clearly distinguished are labeled.

Sparse manual labeling carries two main advantages: first, the labeling time is greatly reduced; secondly, it allows to choose the parts of the image that better represent each part. With the aid of a purpose-built script, labeling takes an average of only 25 seconds per image, with a speed close to 5 pixels per second. In contrast, dense labeling would take almost 1 minute per frame. Also, we effectively double the amount of training data by mirroring each image features and annotations, both in classification and regression. This approach can be considered semi-supervised training.

Finally, the ground-truth for the regression task is obtained by annotating the joint positions of the wrist, elbow and shoulder in every image. See 4.5 for a graphical representation.

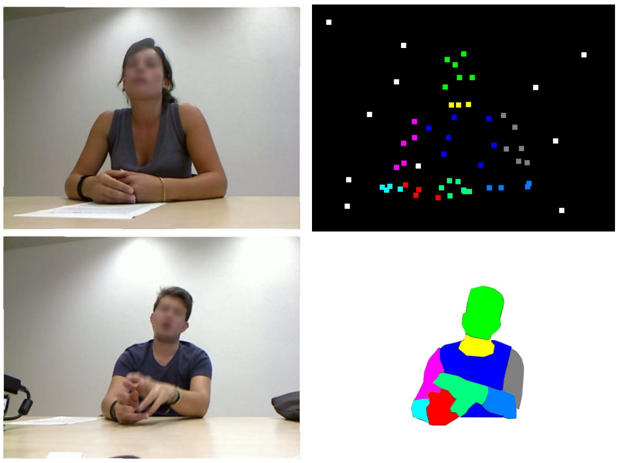


Figure 4.5: Body part annotations. Top: Sparse manual annotations. Only a few pixels per body part are annotated. This is used for training the body part classification algorithms. Bottom: dense manual annotations. Every pixel in the image is annotated through different regions. This is used in order to test the performance of the body part classificatoion algorithms.

4.3.3.4 Data subset D

As explained in Chapter 8 and originally described in [Nguyen et al., 2013a], raw action annotations were also employed in order to assess the relationship between nonverbal communication and social constructs. To that end, a subset of 43 subjects was used, containing over 23000 frame labels.

The class distribution of this corpus is shown in Table 4.9. We observe that "hands on table" accounted for more than half of the labels. The dataset was recorded in a real setting (shown in Figure 4.2), therefore it reflects the natural tendency of the participants while being seated. It should be noted that in 34.2% of the labeled data, the subject was silent while listening to the interviewer. The least represented class was "hidden hands" (4.1 %), while "self-touch" appeared almost as often as "gestures".

Table 4.9: Per-class split of subset B

hiddenHands	gestures	handsOnTable	selfTouch	gesturesOnTable
4.1%	13.7%	52.3%	11.0%	19.9 %

4.4 Single viewpoint: ChaLearn 2011

ChaLearn 2011 [cha, 2011] contains 437 non-time-consecutive, 320x240 color and depth frames, in which body joints have been manually annotated. The environment is uncontrolled as it can be seen in Figure

4.6, and different backgrounds, high variance of poses, clothing, positioning and lighting appear. In some of the frames, there is no movement at all. The database is extremely challenging due to other factors:

- Less training data: only 57% of the total labeled points are used in our approach, as we only use points that contain movement information in order to train the forest. This also gives an idea of the little movement information present in the dataset.
- Some subjects move out of frame (the head or one arm is not visible).
- In some cases the subject casts shadows in the background wall, producing spurious optical flow detections.

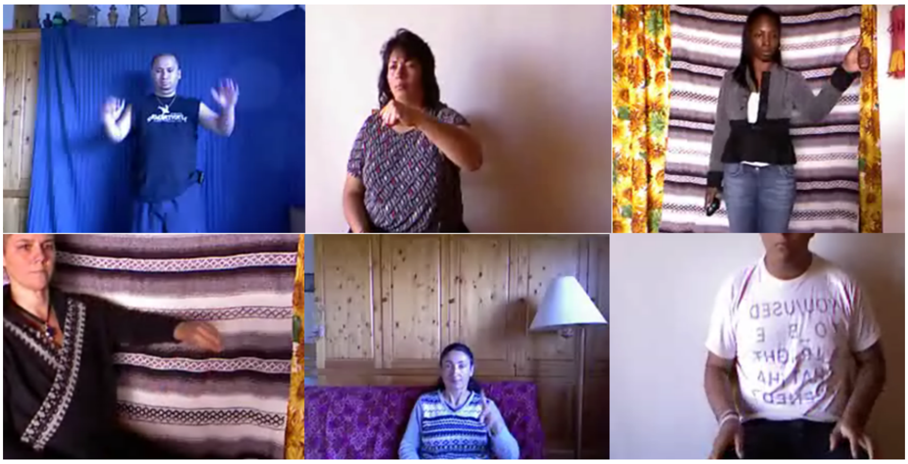


Figure 4.6: ChaLearn 2011 image examples. The image quality and variance of the data makes the database very challenging.

4.5 Conclusion

In the present chapter we have introduced the data that has been used for the development and evaluation of the proposed methods in this thesis. We combine public and private databases that contain single camera, multiple camera and depth camera information. The open datasets HumanEva-I and ChaLearn 2011 are used in order to provide a performance reference and allow for comparisons against the state of the art. A private job interview corpus, that had been collected in the context of a project in which Idiap Research Institute collaborates, is enriched through several manual annotations, which allow to evaluate and train our proposals. In addition, thanks to the metadata contained in the job interview database, we are able to look for correspondences between automatic computer vision methods and social constructs.

Chapter 5

Simultaneous motion capture and action recognition of multiple people

5.1 Introduction

Markerless motion tracking with a network of cameras has been extensively studied before. However, most of approaches focus on tracking a single person. To our knowledge, very little work has been done in multiple people frameworks. Being able to simultaneously obtain the body poses of several people widens the range of applications for such algorithms, as recently Microsoft's Kinect has shown [Shotton et al., 2011].

As reviewed in Chapter 3, there are three main sensing scenarios where markerless motion capture systems have been studied in the literature. There is the difficult monocular case, the successful range sensors, and the multiple camera scenarios, where multiple cameras allow to obtain 3D volumes. The proposal presented in this chapter is a general framework based on using prior human motion knowledge to perform tracking of multiple people from a variety of measurement systems. In this chapter we select the multiple camera scenario, sustained by the HumanEva dataset [Sigal et al., 2010], as it offers an affordable approach to motion capture and an easy way to compare our results with state-of-art methods.

The proposed method is based on a training-then-tracking philosophy. From motion capture datasets we train the human body pose under several labeled motions performed by different individuals. Due to the high number of degrees of freedom of the human articulated skeleton, a dimensionality reduction technique is used for training human poses. Studies [Urtasun Sotil, 2006] have shown that human movements are intrinsically non-linear, therefore a non-linear mapping outweighs linear alternatives such as PCA. From all the available alternatives, a GPLVM (Gaussian Process Latent Variable Model, [Lawrence, 2005]) framework has been chosen. Since its first appearance, several modifications have been made such as SGPLVM [Grochow et al., 2004], GPDM [Wang et al., 2006] or B-GPDM [Urtasun Sotil, 2006], but these modifications are not overly relevant to the purpose of the proposed work.

Training information is then used as the prior distribution in a Bayesian tracker, namely a mixed-state particle filter [Isard and Blake, 1998]. This filter uses discrete states that identify several motions and continuous states for parameterizing the pose (*i.e.* GPVLM state vector and global position and orientation of the body).

Finally, we refine the pose using an optimization algorithm with statistical priors (non-linear least squares, [Marquardt, 1963]). This allows for better generalization towards the real pose by using the tracker output as initialization. The data flow can be seen in Figure 5.1. The main contributions of our method are:

1. Multiple people pose estimation: The proposed framework allows for multiple people to be tracked at once.

2. Multiple different motions are identified, which can be used for online human activity recognition.

The rest of the chapter is organized as follows. In section 5.2 we present an overview of the complete framework. In section 5.3 we define the observation system that we use, that is, which features we extract from the image and how we extract them. In section 5.4 we describe the priors that we use to constrain the solution space. In section 5.5 the tracking process is detailed, while in section 5.6 it is shown how improve the initial estimation that the tracking provides. Finally, in section 5.7 we evaluate our system, and in section 5.8 we discuss our findings.

5.2 Overview

The main idea of the proposed framework is to use trained prior knowledge of the human motion as a building block for detecting several activities and track multiple people. This prior information is gathered by training several activities with a dimensionality reduction algorithm (*i.e.* GPLVM). The result of the training is a number of relatively simple activity trajectories in latent space, as shown in Figure 5.1. This data, together with the online observations, is used as input for a particle filter, which produces a coarse approximation to the observed poses (*i.e.* Visual Hull volumes) based on detecting highly probable hypotheses. These poses are then refined with an optimization technique, which makes the final output of the proposed work.

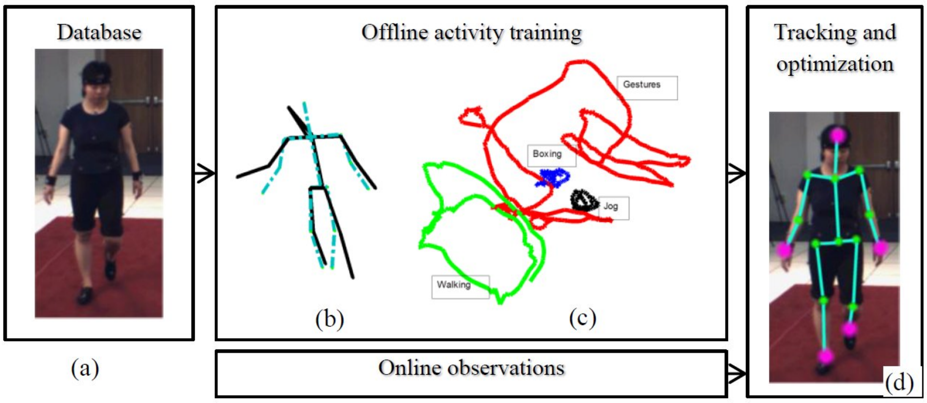


Figure 5.1: Overview: (a). Database information gathering. (b) Procrustes shape alignment. (c) Resultant offline-trained activities in the latent space. (d) Final poses estimation

5.3 Observation system

We use calibrated views from 7 different cameras, as described in section 4.2. We perform traditional background subtraction in order to get the body silhouettes. Then, thanks to the camera parameters, we obtain the volumetric 3D intersection of the silhouettes in several espatial heights, defining a volume called Visual Hull (see figure 5.2).

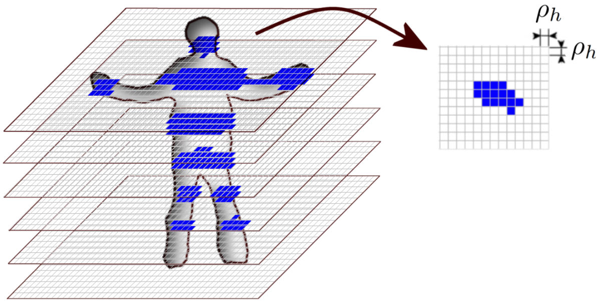


Figure 5.2: Volumetric reconstruction: Each layer corresponds to the intersection of the multiple view foreground figures, in different height levels. Combining several height levels a volumetric representation is obtained.

Each voxel is not only binary defined as ”full” or ”empty”, which is the usual practice in the literature. Instead, it also has an associated weight related to the number of cameras that have observed that voxel. The observation points set $\mathcal{V}_t$ at time t is therefore defined as:

$$\mathcal{V}_t = \{\vec{v}\}_{i=1}^{n_h} = \{x_{i,t}, y_{i,t}, z_{i,t}, \rho_{hi,t}, o_{hi,t}\}_{i=1}^{n_h} \tag{5.1}$$

Where $x_{i,t}, y_{i,t}, z_{i,t}$ is the 3D position of the voxel, $\rho_{hi,t}$ is the voxel size, which is constant among all voxels, $o_{hi,t}$ is the number of cameras that have observed each voxel, and n_h is the number of voxels. With this innovation in respect to typical Visual Hull observation models, it is possible to attach more confidence to parts of the scene that have been seen from more cameras. The proposed camera voting volume reconstruction is illustrated in Figure 5.3.

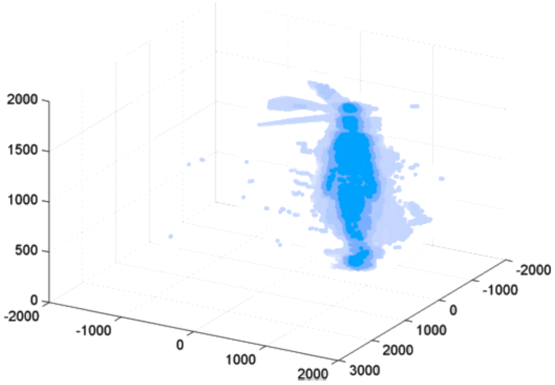


Figure 5.3: Volumetric reconstruction with associated per-voxel voting: more intense colors denote more information from different cameras, therefore a higher number of cameras that see the voxel. When using few cameras, the segmentation errors are unfiltered, and when using a high number of cameras certain body parts dissapear as a consequence of a bad segmentation from the black and white cameras.

5.4 Priors

5.4.1 Human body prior

There are certain constrains that the output pose has to satisfy, such as constant limb lengths or symmetry. We parameterize the human body by using a rigid articulated model consisting of 20 3D points. Each point is associated with one of the n_{joints} points of the body. Therefore, the human pose is defined with a 60D vector:

$$\mathcal{P} = \{\vec{p_i}\}_{i=1}^{n_{joints}} = \{x_i, y_i, z_i\}_{i=1}^{n_{joints}}, \tag{5.2}$$

where $[x_i, y_i, z_i]$ are the 3D coordinates of body joint i. As shown in Figure 5.4, in order to generate human volumes from a pose configuration $\mathcal{P}$, a cylindrical model is used, defined as:

$$\mathcal{Z}_i = \{\omega_i, \tau_i\}_{i=1}^{n_{joints}}, \tag{5.3}$$

where ω_i is the cylinder length and τ_i is the cylinder radius (see Figure 5.8). This cylindrical model is driven by the underlying articulated model, to reflect the body proportions. We choose 3D points as a parameterization for the model, as opposed to 3D angles. The cylindrical model we used does no change with some angle configurations, given the symmetry of cylinders. As an additional benefit, the numeric discontinuity in angles close to 0 and 360 degrees is removed, thus obtaining a better representation in the low dimensionality space.

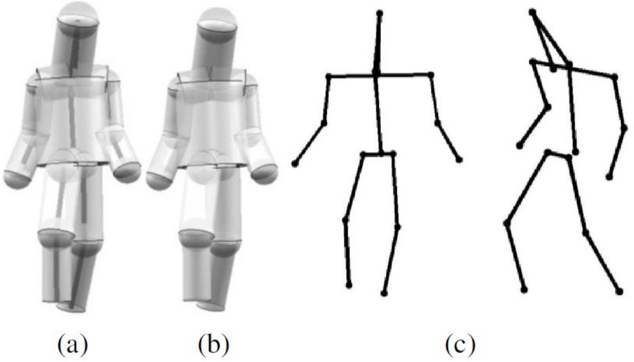


Figure 5.4: Body model. (a) Used cylindrical model with the underlying skeleton (b) Cylindrical model (c) Different views of the used 60-dimensional articulated model.

5.4.2 Movement prior

Equally to the body prior, the set of the output poses is constrained to those poses which are plausible, by using a strong movement prior through an off-line training phase. The goal of the training process is to find a way to generalize, with a small number of parameters, the position of the different joints under some labeled motions (e.g. walk, jogging...), where the accurate position of the 20 joints is known.

A training sequence consists of a set of poses $\mathcal{X}_{tr} = \{\mathcal{P}_{tr}\}_{i=1}^{n_{f,tr}} = \{\vec{p_{tr,i}}\}_{i=1}^{n_{f,tr}}$ of a person in a sequence of $n_{f,tr}$ frames. We propose to represent the pose vectors in a low-dimensional space also known as latent space $\mathcal{L}_{tr} = \{\vec{l_{tr,i}}\}_{i=1}^{n_{f,tr}}$ (see Figure 5.5). This mapping, also known as regression function Ω, is defined as:

$$\mathcal{L}_{tr} = \Omega(\mathcal{X}_{tr}, \beta) \tag{5.4}$$

where β are the mapping parameters, representing the statistics of human dynamics. Function Ω can be linear, as is the case with PCA, or non-linear, like in GPLVM. As it has been mentioned before, because of human movements being non-linear in nature, GPLVM is used to establish a mapping between $\{\vec{l_{tr,i}}, \vec{p_{tr,i}}\}$ pairs [Urtasun Sotil, 2006].

Previously to expressing the set of poses $\mathcal{X}_{tr}$ in a latent space, the whole set is aligned with Procrustes analysis [Bartoli et al., 2013] (see Figure 5.1). It removes translation and rotation components of each

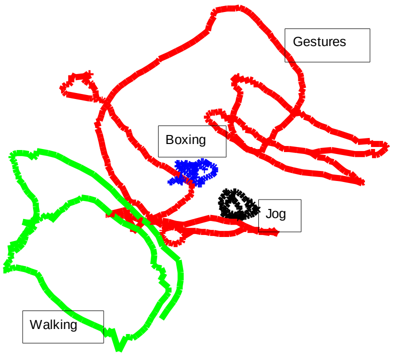


Figure 5.5: Actions compressed with GPVLM. View of the first 3 dimensions of the latent space for several trained actions.

motion capture data frame. Also, the mean sequence pose $\mu_{\mathcal{X}_{tr}}$ is subtracted from the data. It will later be used to regenerate the full pose from a latent space point.

The function Ω is iteratively trained with GPLVM, by refining a model $\mathcal{L}_{tr,o}$ that is first initialized with PCA. See Section 2.4.1 for more details.

5.5 Tracking process

A mixed-state Particle Filter (PF) is used to track the pose of multiple people (see Section 2.5.2). Unlike a regular PF, it allows for discrete states in the state vector. This is normally used to automatically switch between models [Isard and Blake, 1998]. We define the extended state space as follows:

$$\vec{x_t} = \{\vec{c_t}, d_t\}, \vec{c_t} \in \mathbb{R}^{n_m}, d_t \in 1, ..., n_m, \tag{5.5}$$

where $\vec{c_t}$ is continuous and d_t is discrete. d_t labels the model which is associated with the complete state $\vec{x_t}$. A transition probability matrix $\mathbf{T}$ describes the transition probability from the current state i to state j, when having n_a different actions:

$$\mathbf{T} = \begin{bmatrix} p(t_{11}) & ... & p(t_{1n_a}) \\ ... & ... & ... \\ p(t_{n_a1}) & ... & p(t_{n_an_a}) \end{bmatrix}. \tag{5.6}$$

The particle set associated with each probability mode m at time t can be defined as seen in Equation 5.7:

$$\mathcal{S}_{m,t} = \{\vec{c_{mi,t}}, d_t\}_{i=1}^{n_p}. \tag{5.7}$$

The proposed extended state space for each particle is defined as shown in Equation 5.8:

$$\vec{c_t} = \{\mathcal{L}_{i,t}, \vec{k_{i,t}}, \phi_{i,t}, d_{i,t}\}_{i=1}^{n_p}, \tag{5.8}$$

where $\{\vec{l_{tr,i}}\}_{i=1}^{n_{f,tr}} = \{l_{1i}, l_{2i}, ..., l_{qi}\}_{i=1}^{n_{f,tr}}$ is a q-dimensional latent state point, $\vec{k_{i,t}} = [x_i, y_i]$ are two-dimensional coordinates in the observation space in which the pose will be placed, $\phi_{i,t}$ is the pose orientation angle in its vertical anatomical axis (we assume that each person moves in the scene plane), and $d_{i,t}$

is the discrete state associated with the particle, which encodes the appropriate trained motion model. Finally, n_p is the number of particles per mode. Hence the total number of particles is $n_{tp} = n_m n_p$. The proposed algorithm can be seen in Figure 5.6, and is explained below:

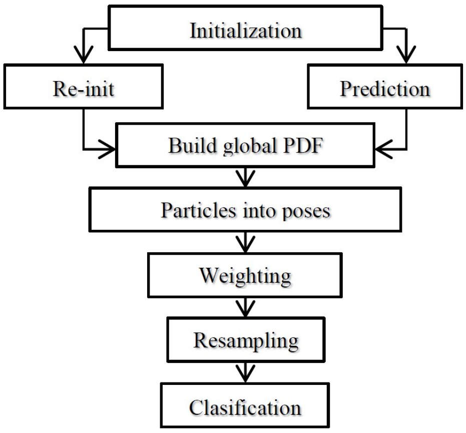


Figure 5.6: Schematic overview of the tracking process

5.5.1 Initialization

In the initialization step we obtain the initial particle set $\mathcal{S}_{m,t}$ for each mode m. Each component of the state vector of each particle is obtained as follows:

1. Discrete state variables $\{d_{i,t}\}_{i=1}^{n_p}$ are computed by sampling from the several training motions $(1, \cdots, n_a)$ according to a uniform distribution.

2. Latent space points $\{\mathcal{L}_{i,t}\}_{i=1}^{n_p}$ are computed by randomly picking samples from the trained motion $\mathcal{L}_{tr,d_{i,t}}$ associated with the discrete state variable, and then adding a dispersion:

$$\mathcal{L}_{i,t} \quad = \quad [l_{1i}, l_{2i}, ..., l_{qi}] + N(0, \sigma_{\vec{\mathcal{L}},i}), \tag{5.9}$$

   where $\vec{\sigma_{\mathcal{L}}} = [\sigma_{\mathcal{L}1}, \sigma_{\mathcal{L}2}, ..., \sigma_{\mathcal{L}q}]$ is the standard deviation of the chosen learned movement, in each dimension j of the latent space.

3. Global coordinates $\vec{k_{i,t}} = \{x_i, y_i\}_{i=1}^{n_p}$ are obtained adding a dispersion $\vec{\sigma}_{\vec{k}}$ to the centroids of the different Visual Hull volumes.

4. Finally, pose orientations $\{\phi_{i,t}\}_{i=1}^{n_p}$ are obtained by performing uniform sampling from the $\{0, 2\pi\}$ radians interval.

5.5.2 Prediction and re-initialization

The predicted particle set $\mathcal{S}_{m,t|t-1}$ is obtained for each mode m. The particle count for each mode is defined as:

$$n_{tp} = n_p^{'} + \alpha_k n_p + \alpha_R n_r, \tag{5.10}$$

where α_k is a proportion of the total particles available for the mode which is used to reduce kidnapping effect, and α_r is another portion which is used for the re-initialization step. The several components of the state vector of each particle are obtained as follows:

1. Discrete states $\{d_{i,t}\}_{i=1}^{n_p}$ are sampled according to the transition probability matrix $\mathbf{T}$, and then the latent space part of the state vector $\{\mathcal{L}_{i,t}\}_{i=1}^{n_p}$ is computed according to equation 5.11:

$$\mathcal{L}_{i,t} = \begin{cases} distmin(\Omega(\mathcal{P}_{o,t-1}, \mathcal{B}), \{\mathcal{L}_{i,t}\}_{i=1}^{n_p}) + N(0, \vec{\sigma_{\mathcal{L}}}) & \text{if } d_{t-1} \neq d_t \\ \mathcal{L}_{i,t-1} + N(0, \vec{\sigma_{\mathcal{L}}}) & \text{otherwise} \end{cases} \tag{5.11}$$

   Therefore, if the performed activity changes from $t-1$ to t, then $d_{t-1} \neq d_t$ is satisfied. It becomes necessary to find the latent point in the new activity latent shape that better describes the current pose. In order to compute it, the predicted output pose in the previous time instant $\mathcal{P}_{o,t-1}$ is mapped into the latent space with the Ω regression function associated with the new activity, thus obtaining a point in the latent space. The distances between this latent space point and every point in the trained latent shape associated with the current action are computed. The point of the trained shape associated with the minimum distance is chosen, and then added a dispersion $\vec{\sigma_{\mathcal{L}}}$. See Figure 5.7 for details.

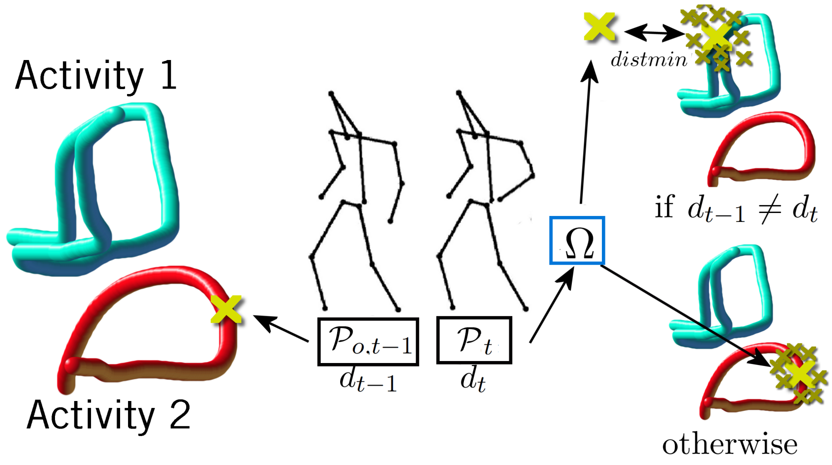


Figure 5.7: Activity changes handling: Two different actions in the latent space are represented with cyan and red shapes. Brighter crosses represent points in the latent state associated with current $\mathcal{P}$ poses. Darker crosses represent dispersion in the latent space. If there is a change of activity, the most representative point of the new action in its latent shape is found. Otherwise, only dispersion is added.

   If the performed activity does not change, only dispersion is added to the current state space point.

2. A given particle percentage α_k of the total n_p available for the mode, is used to distribute $\mathcal{L}_{i,t}$ uniformly among the whole trained latent space, as in the initialization step. The global coordinates $\{\vec{k_{i,t}}\}_{i=1}^{n_p}$ are obtained by dispersing the previous time instant state.

3. Pose orientation $\{\phi_{i,t}\}_{i=1}^{n_p}$ is processed in the same manner.

Finally, and known as the re-initialization step, a proportion α_r of the total particles available for the mode is used to search in the whole state space, by using the same method as in the initialization step every t_{re} iterations. This makes the algorithm more robust, and allows for for a faster recovery to tracking losses.

5.5.3 Probability Density Function formation and particle conversion

The several probability modes are joined before the weighting and resampling steps. The Probability Density Function is composed as follows:

$$\mathcal{S}_{t|t-1} = \{\mathcal{S}_{m,t|t-1}\}_{m=1}^{n_m}. \tag{5.12}$$

Once that $\mathcal{S}_{t|t-1}$ is formed, each particle is converted into an observation space reconstructed pose $\mathcal{P}_{o,t}$ by applying the mapping Ω^{-1} and adding the previously subtracted sequence mean $\mu_{\mathcal{X}_{tr}}$. Therefore, at this point a set of n_{tp} observation space poses is available, and ready for an importance weight to be assigned.

5.5.4 Weighting

A θ function is applied to every pose $\{\mathcal{P}_{i,t}\}_{i=1}^{n_{tp}}$ in order to obtain that very same pose in angle parameterization $\{\mathcal{A}_{i,t}\}_{i=1}^{n_{tp}}$, since it will be far more useful in the optimization process later on:

$$\mathcal{A} = \theta(\mathcal{P}) = \{\vec{a_i}\}_{i=1}^{n_{joints}} = \{a_{\alpha i}, a_{\beta i}\}_{i=1}^{n_{joints}}, \tag{5.13}$$

where $\{a_{\alpha i}, a_{\beta i}\}_{i=1}^{n_{joints}}$ is the angle parameterization in spherical coordinates of the n_{joints} body segments. A cylindrical model $\{\mathcal{Z}\}_{i=1}^{n_{joints}}$ is created in order to recreate the approximate human body volume proportions:

$$\{\mathcal{Z}\}_{i=1}^{n_{joints}} = \{\omega_i, \tau_i\}_{i=1}^{n_{joints}}. \tag{5.14}$$

Each member τ_i defines the cylinder radius or body part thickness, with length ω_i. Values of τ_i are set empirically, and lengths ω_i are obtained with motion capture data. Each particle weight $w_{i,t}$ is computed using the following expression:

$$w_{i,t} = \prod_{i=1}^{n_{joints}} \xi_i, \tag{5.15}$$

where ξ_i is the weighted fill percentage of each body segment, defined as follows:

$$\xi_i = \frac{\mathcal{V}_t \subset \mathcal{Z}_i}{\xi_{i,max}} \frac{\sum_{j=1}^{n_h} \Psi(\vec{v_j}, \mathcal{Z}_i)}{\xi_{i,max} n_c}, \tag{5.16}$$

where $\mathcal{V}_t =$ are the Visual Hull observations (see Section 5.3). The left multiplication term describes the filling percentage of the body part i, and ranges from 0 (empty) to 1 (full). The right multiplication term describes the likelihood percentage relative to the maximum possible, that is attached to the observations $\vec{v_j}$ related to the body part. It also ranges from 0 to 1. The function Ψ is defined as:

$$\Psi = \begin{cases} o_h & \text{if } \{x_j, y_j, z_j\} \subset \mathcal{Z}_i \\ 0 & \text{if } \{x_j, y_j, z_j\} \not\subset \mathcal{Z}_i \end{cases} \tag{5.17}$$

It therefore checks whether a voxel j lies within the body part $\mathcal{Z}_i$. If it does, a value o_h is output. It quantifies how many cameras have observed the voxel, as seen in section 5.3. For example, if voxel j has been observed by 6 cameras, o_h will have a value of 6. The second term of equation 5.16 is therefore normalized with the total number of cameras present in the scene n_c, in order to quantify the maximum possible likelihood of the observations.

The maximum possible fill percentage $\xi_{i,max}$ is defined as:

$$\xi_{i,max} = \frac{\pi \tau_i^2 \omega_i}{\rho_h^3} + \frac{4\pi\tau_i^3}{3\rho_h^3}, \tag{5.18}$$

which describes the volume of a cylinder (left term) and sphere (right term) in relation to the volume of a voxel if size ρ_h. The full sphere volume equation is used because each body part is modeled with a cylinder and two semi-spheres (see Figure 5.8). While the theoretical maximum is 1, the voxels will never fill completely the body part, since they are defined as cubes. Therefore, as can be seen in equation 5.18, $\xi_{i,max}$ is defined as the volume that the body part occupies, relative to the volume of the voxels that can be fitted inside it.

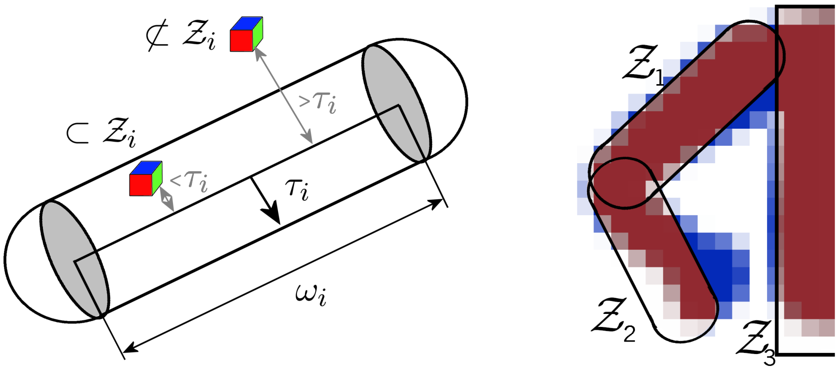


Figure 5.8: Cylinder-based body model weighting: Left: a body part cylinder $\mathcal{Z}_i$ is defined with its length ω_i and radius τ_i. The left voxel lies within the body part. Right: illustration of weighting. The Visual Hull observations are represented with squares, and probability of voxels is encoded with saturation. Color encodes voxels that lie within (red) or outside the body part (blue).

After computing the weight $w_{i,t}$ for each particle, the weight set is normalized so that the sum equals one.

5.5.5 Resampling with previous clustering and normalization

After weighting, it is possible for a probability mode to disappear after resampling if its weights are too low in relation with other modes. To avoid this, we cluster and then globally normalize each probability mode, as shown in Figure 5.9

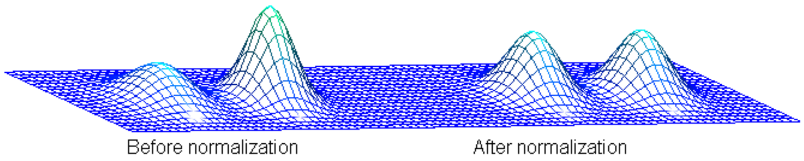


Figure 5.9: Probability mode normalization before resampling.

The resampling step is identical to that of the Bootstrap framework [Doucet, 1997]. However, the number of particles does not stay constant, as a number of them will be inserted in next-iteration re-initialization step. The obtained particle set is therefore:

$$\mathcal{S}'_{t|t-1} = \{c_{i,t|\vec{t}-1}, d_{t|t-1}\}_{i=1}^{n_m(n'_p+\alpha_k n_p)}, \tag{5.19}$$

where n'_p is the surviving number of particles after resampling.

5.5.6 Particle set clustering

The number of people present in the scene is set by using a linear clustering algorithm with support for a varying number of groups. The input data for this clustering algorithm is the global position part of the state space vector, $\{\vec{k}_{i,t|t-1}\}_{i=1}^{n_m(n'_p+\alpha_k n_p)}$. It is therefore assumed that the several people are far enough from each other to guarantee a successful classification, which produces a set of $\mathcal{G}$ groups:

$$\mathcal{G} = \{\vec{g_i}\}_{i=1}^{n_g} = \{x_i, y_i, \mathcal{V}_g\}_{i=1}^{n_g}, \tag{5.20}$$

where $[x_i, y_i]$ is the group centroid and $\mathcal{V}_g$ are the associated members in the group. Final poses in the high dimensional observation space, $\mathcal{P}_{\rho,t}$ are obtained by averaging the re-sampled particles that comprise each mode m. In this point there are n_g people and n_m total modes ready for the optimization process.

5.6 Pose refinement

In addition to the tracking process, an optimization process is run with the tracker output in order to fine tune the final poses. A non-linear least squares algorithm is used for each particle distribution mode, in order to minimize error ε:

$$\varepsilon = \sum_{i=1}^{n_m} \|1 - \{\xi_i\}_{i=1}^{n_{joints}}\|^2 + \mu \|\mathcal{P}_\rho - \mathcal{P}_o\|^2. \tag{5.21}$$

The first term is used to measure the fill percentage of the cylindrical model used, as the variable ξ_i quantifies said percentage: it has a value of 0 if there no point of the volumetric reconstruction lies within the cylinder, and 1 if the model is completely filled by the reconstruction. The bigger it is, the closer ξ_i will be to 1, therefore the term will move towards 0 when it gets optimized. This term therefore models the same criteria as used when evaluating particle weights. The second term is used to regularize the final pose $\mathcal{P}_o$ using the hyperparameter μ, in order for it not to be too different than the tracker output $\mathcal{P}_\rho$, maintaining some inertial properties. The number of nodes is represented with n_m, and the number of body segments by n_{joints}. Figure 5.10 shows the effects of this refinement on the poses:

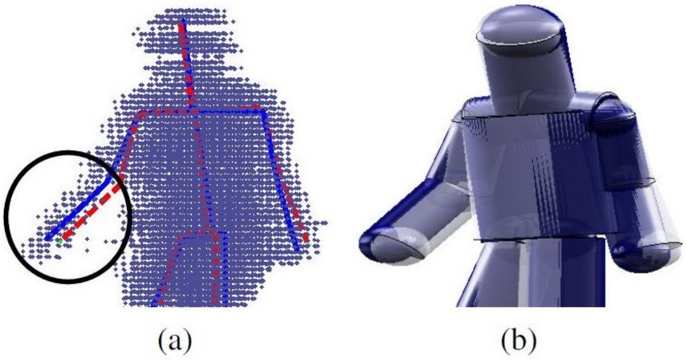


Figure 5.10: Optimization effect in the final poses. (a) In dotted red line: tracker output pose. In continuous blue line: optimized pose. Blue points: Visual Hull voxels. (b) In grey: volume model of the tracker output pose. In dark blue: volume model of the optimized pose. Best viewed in color.

5.7 Results

We designed three experiments in order to validate the concept and application of the proposed work. In Figure 5.15 several output frames can be seen. The experiments consist in:

1. Experiment A: The objective is to know the minimum amount of joints that is necessary to observe in order to track the human body pose with an acceptable level of precision.
2. Experiment B: Once that the minimum number of necessary joints is set through the experiment A, we apply our tracking method in order to analyze its behavior in the presence of more than one person in the scene.
3. Experiment C: Its purpose is to evaluate the performance regarding precision and activity recognition accuracy, with a real observation system such as Visual Hull. In order to accomplish the goal we follow a adjusting process for the most relevant parameters of our method.

To successfully perform reliable tests on our system, while also allowing for a better comparison with the state of the art, we use the publicly available HumanEva database. It provides images from three color cameras and four black and white cameras. In Table 5.1 it is shown the parameters of the algorithm that stayed fixed along the performance evaluation process.

For **experiment A** a sequence of 600 frames is used, during which there is a switch between ”walking” and ”running” actions. Performance is evaluated as the body pose reconstruction error in function of the number of observed joints. In each frame, the desired number of used joints is randomly sampled directly from the total of joints present in the groundtruth data (therefore no volume reconstruction has been used for this test, and no intrinsic body pose error). However, this experiment is valid in order to find a relationship between error and amount of available information. Results are shown in Figure 5.11.

Table 5.1: Used parameters for experiments

α_r	α_k	$\sigma_{\vec{k}_{xy}}$	$\sigma_{\vec{k}_z}$	σ_Φ	**T**
250%	15%	150mm	20mm	0.45rad	uniform

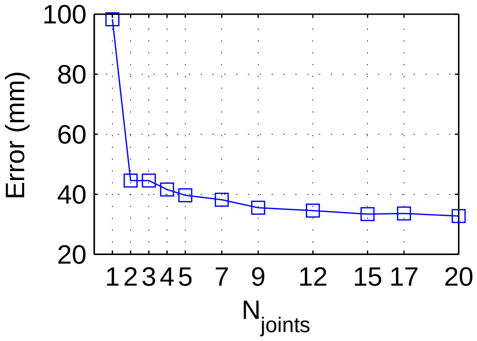


Figure 5.11: Tracking error versus number of observed joints (n_{oj})

As the results show, the error significantly decreases when observing more than one joint. If three or more joints are observed, the error decreases in a more progressive fashion. This is because the correct orientation $x_{\Phi mi}$ is already defined with two 3D points, because only the longitudinal anatomic axis needs to be defined, as explained in Section 5.5. Once that the angle is correctly defined, observing more joints allows for a more detailed characterization of the body pose. Given the difficulties of directly obtain

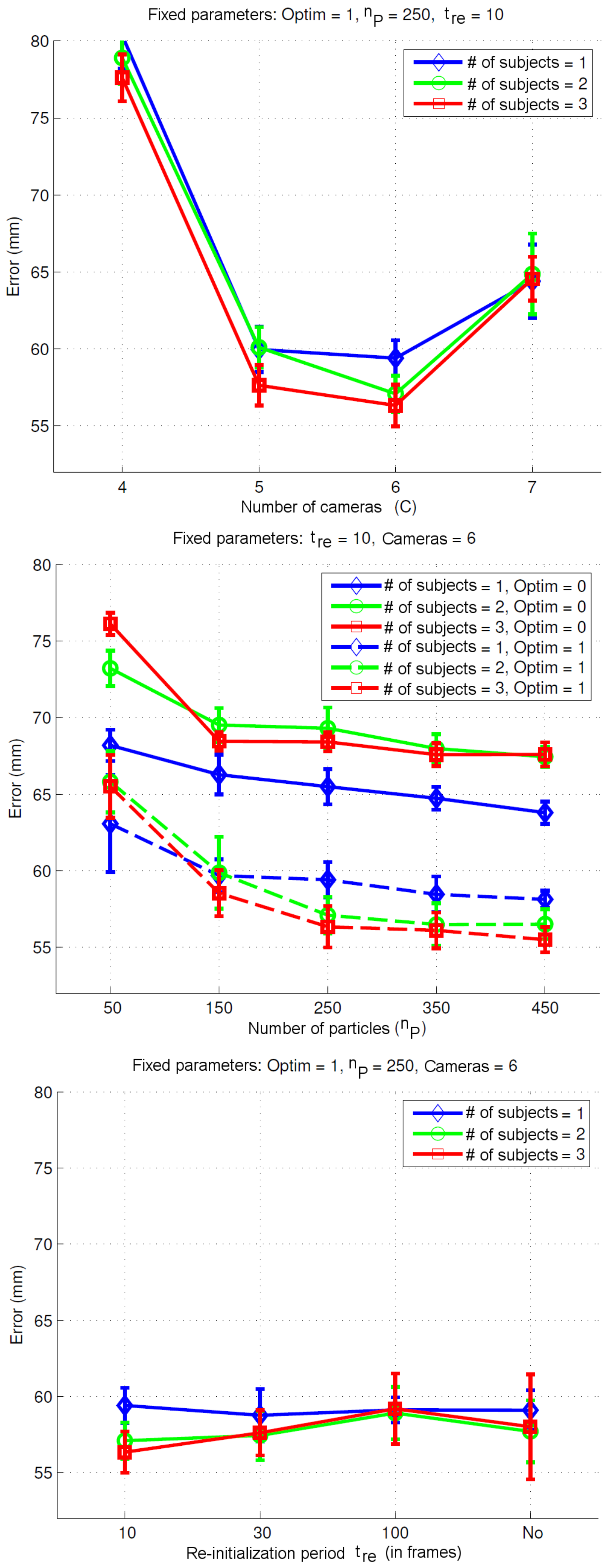


Figure 5.12: Parameter tuning. Top: variation of the number of cameras. Center: variation of the number of particles per mode n_p, and enabling or disabling the optimization (pose refinement) process. Bottom: variation of the reinitialization period t_{re}. The results are parameterized for one, two and three persons in the scene.

body joints observations, we consider that using 4 joints (such as hands and feet, obtainable through geodesic distances analysis of the reconstructed volumes), an acceptable balance between information and precision is achieved. The latter is close to 4 centimeters.

In the **experiment B** we evaluate the capacity of the proposed system to follow more than one person simultaneously. As a testing sequence, we employ 600 frames in which two persons are present in the scene. The first person starts walking and switches to running in frame 300. The second person initially runs and switches to walking in frame 340. The mean error is defined as the mean euclidean distance between the reconstructed body joints and the groundtruth, and is **41 mm** when using the 4 body joints that the experiment A justifies. It is a low value even considering that there are two persons present in the scene. The temporal evolution of the error can be seen in Figure 5.13.

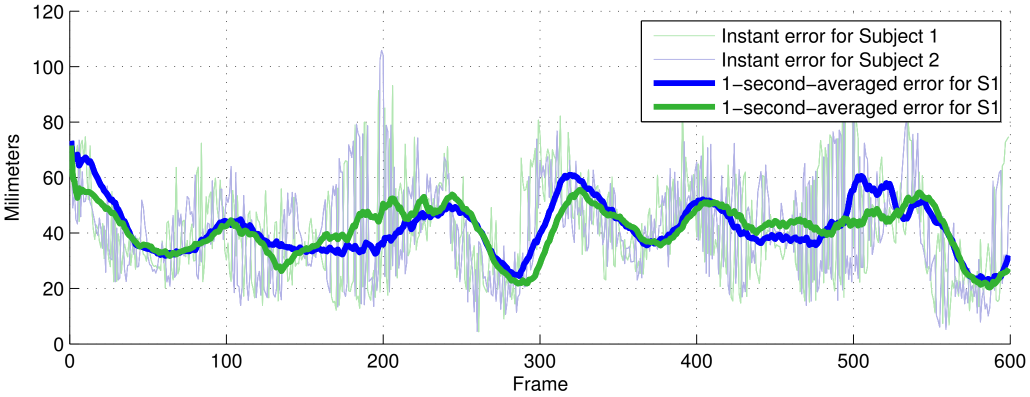


Figure 5.13: Temporal tracking error in experiment B. Instant and 1-sec averaged error are shown, for two different subjects (S1 and S2). Best viewed in color

Finally, the Visual-Hull-based observation system described in section 5.3 is used for **experiment C**. Three different sequences of 300 frames are used, each one with a different number of persons (one, two or three), performing actions ”walk” and ”run”. Given their similarities, these actions comprehensively test the capacity of the proposed action recognition framework. The results of the different tests that were used to adjust the system parameters are shown in Figure 5.12. Each experiment was repeated 10 times in order to obtain a measure of the confidence of the precision. In Figure 5.12 top it is shown the precision as a function of the number of cameras used for the volumetric reconstruction. As it can be seen, using 6 cameras provide the best precision, while in total there are 7 cameras available. Upon analysis, it became clear that this is a result of the bad background segmentation performance when black and white cameras (4 out of the 7) are used. With 6 cameras it is found therefore a sweet spot between amount of information and its quality.

Once set the number of cameras to 6, we explore the effect of changing the number of particles n_p, with and without the proposed pose refinement method. Results can be seen in Figure 5.12 center: when adding the optimization process, the error is systematically reduced, because of the improved generalization capabilities that it provides. Also, and typical for particle filtering trackers, the error decreases monotonously when increasing the particle number. Taking all this into account, we considered that $n_p = 250$ provides an acceptable balance between precision and execution time (at around 1.5 seconds per frame in a Matlab implementation).

Finally, we explore the effect of changing the number of iterations after which the tracker is re-initialized (see t_{re}, section 5.5.2). As it can be seen in Figure 5.12 bottom, there is not a clear trend regarding mean precision. However, short re-initialization periods reduce the variance of the error, as the particles are more often redirected towards the right portion of the state space, which in addition reduces tracking loses and kidnapping effects. More output frames can be seen in Figures 5.16 and 5.17.

Regarding the activity recognition system, we have observed a high precision with the already fixed parameters (6 cameras, $n_p = 250$, optimization enabled, $t_{re} = 10$). In 5.14 results are shown as a function of the number of cameras used and the use of the refinement methods, which are the parameters around which there appears the biggest variation of accuracy.

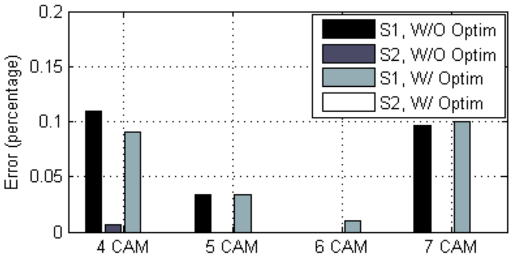


Figure 5.14: Activity recognition results. For two subjects performing different actions, with and without optimization and changing the number of cameras used.

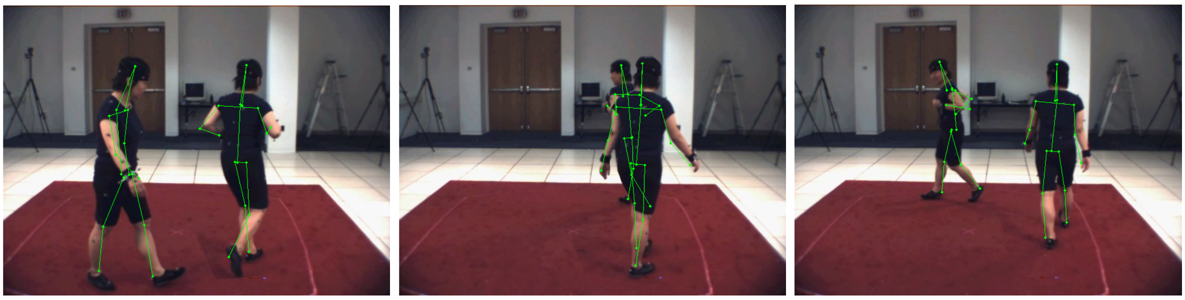


Figure 5.15: Experiment 3 qualitative results. HumanEva images have been altered in order to display two people simultaneously. The obtained poses have been overlaid in the images.

5.8 Conclusion

We presented a new method to simultaneously obtain the body pose and ongoing activity of multiple people, using data from multiple camera viewpoints. A Bayesian tracker is used to track the body posture quickly and efficiently, thanks to a non-linear regression such as GPVLM. This estimation is then improved with a non-linear optimization system, which in addition provides extra generalization capabilities. The obtained body pose error is comparable to that of the state-of-the-art when using similar body cylindrical models (as opposed to using a highly detailed body mesh, which is slower and requires manual laser scanning for each subject). The activity recognition method proved to be effective, being able to successfully distinguish between two very similar movements.

The main limitation of the proposed method is the lack of support for close interactions between different persons, which can be addressed with changes in the observation system, such as color-based person-specific segmentation. There is also room for improvement in the re-initialization step, in order to better take into account new observations, and in the amount of detail that the body model represents.

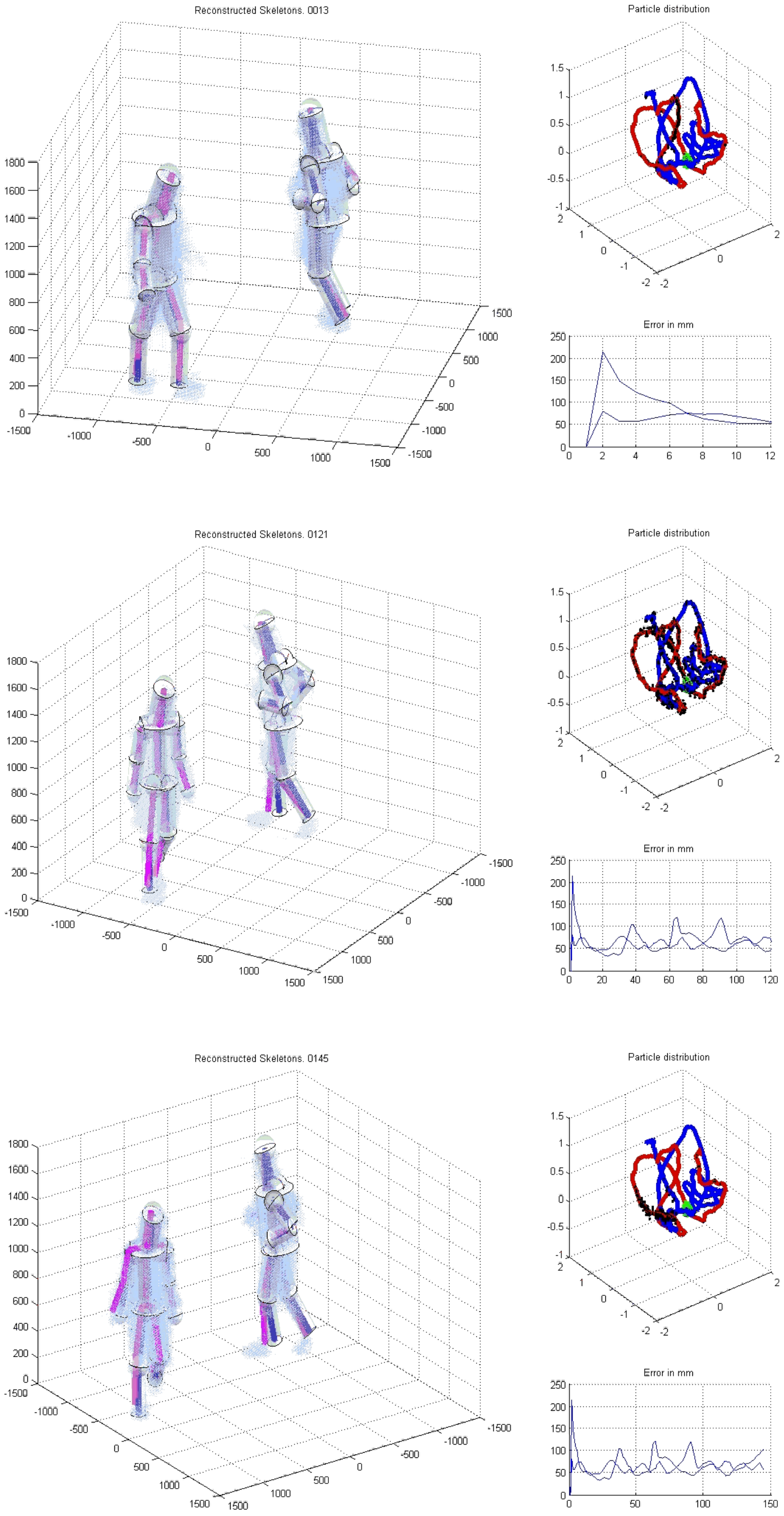


Figure 5.16: Output frames (1): Several output frames. Left: the groundtruth skeletons are represented in blue, and the reconstructed skeletons in magenta. Top right: latent space representation of the trained activities, with the particle distribution in black. Bottom right: temporal mean error.

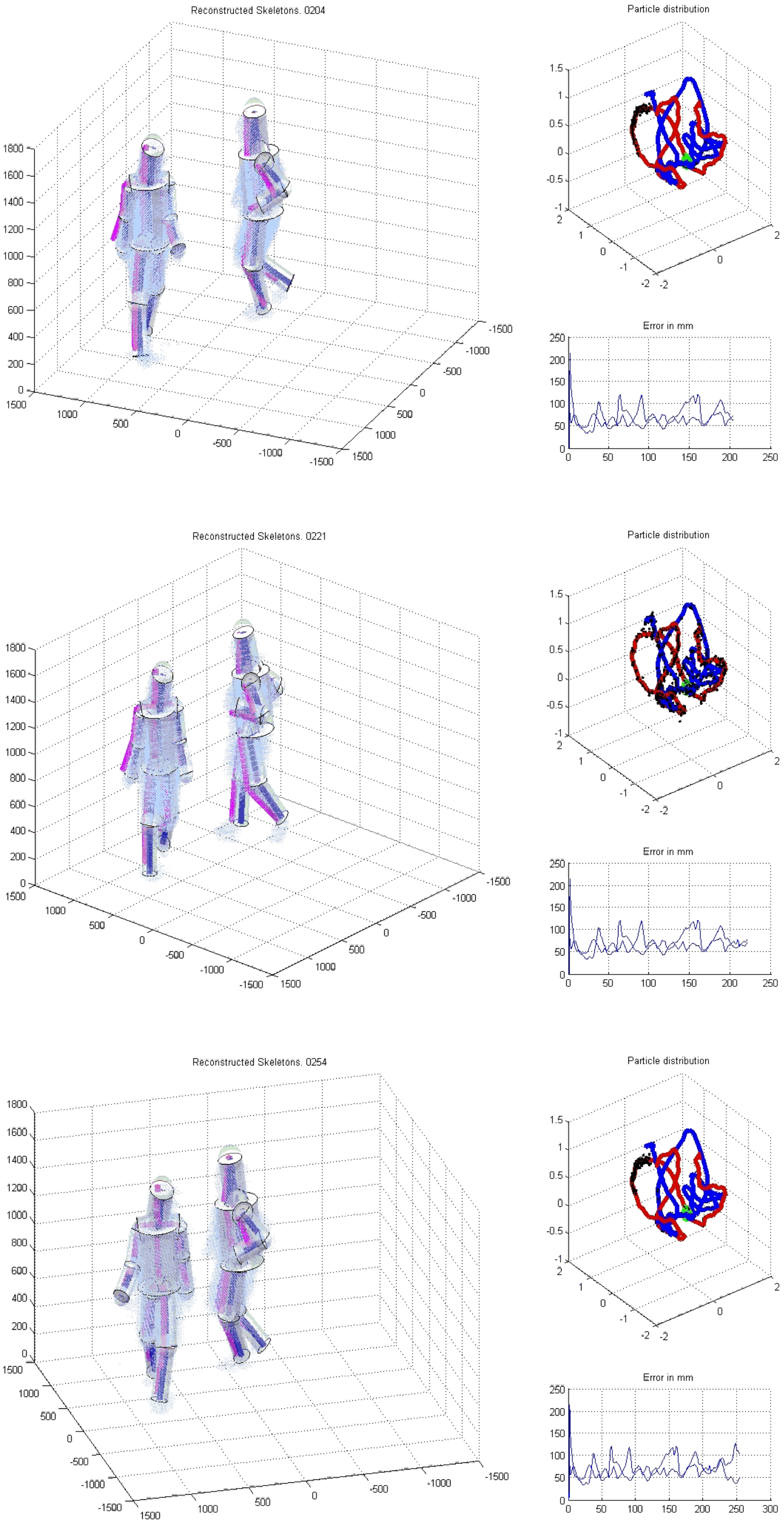


Figure 5.17: Output frames (2): Several output frames. Left: the groundtruth skeletons are represented in blue, and the reconstructed skeletons in magenta. Top right: latent space representation of the trained activities, with the particle distribution in black. Bottom right: temporal mean error.

Chapter 6

Upper body motion capture from hand tracking

6.1 Introduction

Social skills play an important role for success in our lives, and it is no exception in the workplace. In some jobs, having "people skills" can be as important as technical knowledge. Job selection processes are aimed to determine how valid a candidate is, but there is often limited interaction time between interviewer and interviewee before making a judgment. How the candidate portraits himself during this short period becomes crucial [Curhan and Pentland, 2007b], posing an interesting subject to study. In this work we are specifically focused on the ongoing nonverbal communication in job interviews and similar interactive settings, as it can influence how we are socially perceived [Knapp and Hall, 2009], [Pentland, 2008]. This matter has been intensively analyzed in social psychology and cognitive science [Knapp and Hall, 2009].

In these domains, however, there has traditionally been the need for an interpreter. That is, a person that emits a judgment on the perceived traits of the analyzed subject, or that codes specific behaviors. This judgment always carries a degree of subjectivity, which can lead to inconsistencies across different evaluations. Also, depending on the amount of data available, it can be a cumbersome, time consuming task. In order to address this problem, we propose a new method to analyze, in an automatic way, upper body nonverbal cues of people in a conversational context. Particularly, we are looking for (A) *adaptors*: unconsciously-used movements like nail biting and head scratching, that might provide information about the person's attitude, anxiety level, and self-confidence, therefore becoming a potentially rich source of information about the psychological state of the sender; (B) *beat gestures*: movements that do not present a discernible meaning, *i.e.*, small, low energy, rapid flicks of the hands and fingers that seem to beat along with the rhythm of the speech [McNeill, 1992] and can be used to signal the temporal loci in speech of something the speaker considers important relative to the larger discourse [McNeill, 2005]; and (C) *posture*, intentionally or habitually acquired positions of the body, which can be an important clue about the emotional state a person is in [Mehrabian, 1972].

In this chapter we developed a series of new computer vision algorithms in order to first extract and then analyze the upper body, human movements and actions with conversational meaning. We use frontal RGB and depth video sequences of discussions around a table as input (see Figure 6.1). A significant amount of vision research [Sigal and Black, 2010a] has been done in order to infer body movements, a topic generally known as markerless motion capture. While this problem has been identified to be hard to solve when using a single camera system, range sensors such as Kinect have proven to help by removing some of the ambiguities that monocular images cause.

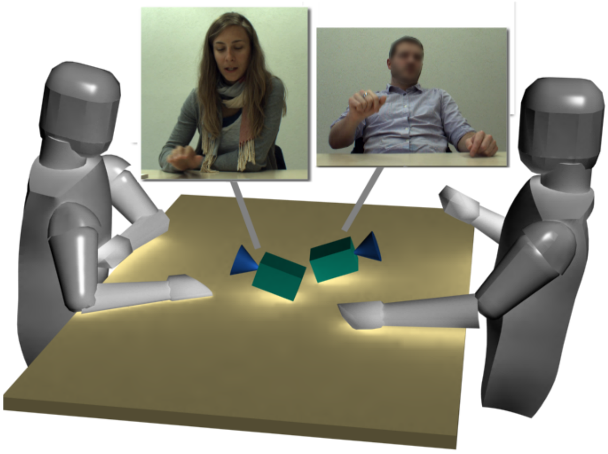


Figure 6.1: Proposed framework. Using several image and depth cues, our proposed framework outputs hand position, speed, approximate upper body 3D pose, and estimated ongoing activity, using conversational video sequences recorded with RGBD devices as input. Best viewed in color.

We tackle the problem by detecting the body extrema (head and hands) as a proxy to infer the complete body pose. This approach has been proven successful, as those body parts correspond to the end points of the cinematic chain, which accumulate most of the position error in the articulated body hierarchy. Therefore, if these extrema are correctly detected and used as a constrain, inferring the rest of the body pose is simpler.

Since our work has the ultimate goal of being applied in psychological studies, we must take into account a scenario where range cameras are not available, since a) pre-recorded footage exists and b) the use of these cameras is still not standard practice in psychological research. Therefore, we provide a method that is able to extract the hand and head position, and ultimately 3D body pose, when depth information is not available (referred as ”the RGB case”). However, we also extend our method for its use with depth for the cases in which it is available, thus increasing its performance (referred as ”the RGBD case”).

6.1.1 Contributions

We first developed the method for the RGB case and published it in [Marcos-Ramiro et al., 2013]. Then, we extended it for cases in which depth is also available, and we added a number of improvements along the processing pipeline (currently under review). Taking into account the related works previously reviewed, our contributions are as follows. In the **RGB case**:

- A new method for extracting hand position from conversational video sequences, by exploiting the fact that optical flow is a strong indicator of where the hands are in conversation.
- A new method for visual tracking, if the whole sequence is available from the start (typically the case in psychology, management and cognitive science experiments).
- A new method for extracting 3D torso pose from 2D images in a seated person setting for action recognition.
- An objective evaluation of the above tasks using a job interview dataset, containing 9 subjects.

In the **RGBD case:**

- A new RGBD fusion method for hand tracking, that improves and makes more robust the framework of the RGB case [Marcos-Ramiro et al., 2013]. We add the hypothesis of the hands being the closest part to the camera, while sensing with a range camera. We also improve the analysis of the hand likelihood map for better hand position extraction.
- An improvement over the pose and action retrieval performance in the RGB case [Marcos-Ramiro et al., 2013], through a non-linear optimization scheme and a more robust action recognition method.
- A significant increase of data for analysis. We recorded 27 real job interviews, which represent more than 4 hours and a half of RGBD video.

6.1.2 Chapter organization

In the rest of the Chapter we first present the RGB case, and later describe the improvements made for its extension with depth. In Section 6.2 we overview the RGB method. In Section 6.3 the observation system and feature extraction from the image are explained. In Section 6.4 we cover the prior information that we use as constrains in our system. In Section 6.5 the tracking scheme that gives consistency to the aforementioned features is introduced. In Section 6.6 we overview how the action recognition process is performed. In Section 6.7 we introduce the improvements made for the RGBD case. In Section 6.8 we test our system, and finally in Section 6.9 we present some conclusions.

6.2 Overview

As mentioned in Section 6.1, our approach to motion capture is to first localize the head and hands in the image, and then use their location to infer the complete upper body position. In order to measure the hands position in the image, we combine a series of computer vision cues to build what we call ”hand likelihood map”, which is an image in which higher values correspond to areas of the input image in which hands are more likely to be present. We later analyze the modes of this probability distribution to track the hands along a complete sequence, imposing spatial and temporal consistency. Finally, we employ a series of optimization methods coupled with pre-processed movement and body priors to obtain the final pose and to perform action recognition. A graphical representation of the pipeline can be seen in Figure 6.2.

Since the RGBD case is an extension of the RGB case, along the rest of the chapter we will present the RGB case first, and then explain the changes made for the RGBD case.

6.3 Observation system

Given an uncalibrated RGB video image $\mathbf{I}_t$, at time index t, we propose to obtain a per-frame measurement for the hands position in the image. We use a combination of several image cues, which should be as color/appearance invariant as possible to increase robustness. They should also take advantage of the specific constraints that the face-to-face upper body setting offers.

Our hypothesis is that (even if not necessarily true for every instant) while taking into account a whole sequence, hands are the parts of the image that show most motion and are closest to the camera. Two strong indicators are, respectively: a) hands are the furthest body part from the body's axis of

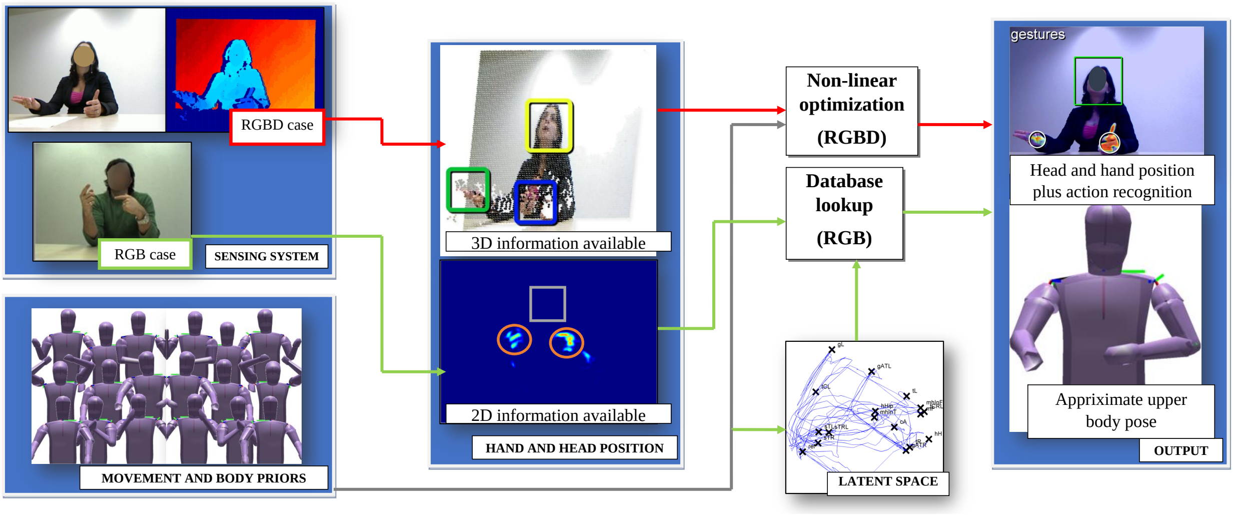


Figure 6.2: Global data flow scheme. We first extract and then analyze the body posture in conversational sequences. Faces are blurred only for displaying purposes, not for processing.

rotation, so they show the highest spatial speed for a given joint angular speed, and b) the nature of this specific setting with a frontal point of view shows a tendency of orienting the arms and hands closer to the camera than the rest of the body.

In order to formalize these hypotheses, we built a hand likelihood map, where numerical values are proportional to the expectancy of a hand being in that region. The hand likelihood map considers that the hands are the skin-colored parts that are not the face, show more amount of motion. In order to enforce this constraint, we need to compute the optical flow of the sequence to extract motion information, skin segmentation, and face detection. Also, given the natural appearance of the fingers, which have lots of edges, we explored image edge detection as a feature. We detail the steps in the rest of this section.

6.3.1 Image motion retrieval

We use a state of the art optical flow estimation framework [Chambolle and Pock, 2011] to measure image motion. It provides smooth optical flow by performing convex optimizations and it is resistant to outliers (see Figure 6.3). The output of the algorithm is a two-channel image as a function of two input RGB images:

$$\mathbf{I}_{OF,t} = f(\mathbf{I}_t, \mathbf{I}_{t-1}) = [\mathbf{I}_{OF,\rho,t}, \mathbf{I}_{OF,\phi,t}], \tag{6.1}$$

where the first channel $\mathbf{I}_{OF,\rho,t}$ encodes the modulus of the optical flow, and the second channel $\mathbf{I}_{OF,\phi,t}$ encodes its orientation. As we are only interested in detecting rapid moving zones of the image, regardless of the direction, we simply use $\mathbf{I}_{OF,\phi,t}$.

6.3.2 Face detection

In order to detect the face of the conversing person in the video, we use a probabilistic version of the Viola & Jones face detector [Viola and Jones, 2001]. This method uses likelihood information from the output of every AdaBoost cascade classifier, so that the output is probabilistic rather than binary. An initial mask $\mathbf{I}_{F,t}$ is set to 0 inside the face region of interest and 1 otherwise, so that later the face pixels are not taken into account when computing the hand likelihood map.

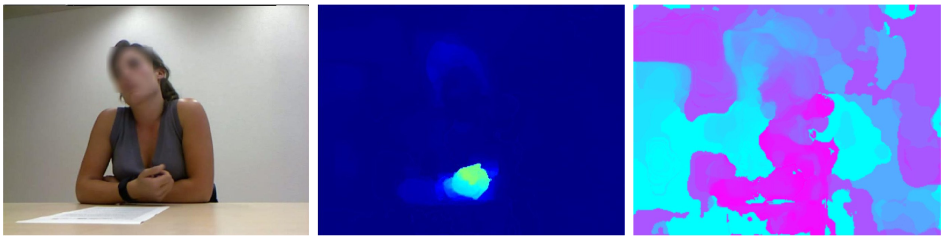


Figure 6.3: Optical flow retrieval. Left: one of the input RGB frames. Center: optical flow modulus. Right: optical flow orientation. Best viewed in color

6.3.3 Edge detection

We use a simple Canny edge detector [Canny, 1986] with a low threshold, to obtain an edge map. We then apply a smoothing filter in order to better search for maxima in the hand likelihood map, thus obtaining a Chamfer distance measure. We get the edge map $\mathbf{I}_{E,t}$. As it will be later explained, edges are only used in the RGB case.

6.3.4 Skin segmentation

Inspired by [Scheffler and Odobez, 2011], we use face detection to infer skin color, as the hue values of face and hands are usually similar. After having processed the face detections in the sequence, a number of $n_{f,S}$ frames frames are chosen randomly to obtain a skin color distribution the face ROI.

As established in [Gijsenij et al., 2011], skin color hue values usually fall within the $(0, 0.2)$ range of the hue channel in an HSV image. Therefore we set all hue values that satisfy that constraint to be skin candidate pixels. We get the hue mean and standard deviation (μ_S, σ_S) of candidate pixels, which constitutes our subject-specific skin model. After this has been computed, a per-pixel Mahalanobis distance is computed with every input image $\mathbf{I}_t$ to get a binary segmentation, which is later refined with simple morphological operations. The result is the binarized result image $\mathbf{I}_{S,t}$, see Figure 6.4 for details. This technique is used for both the RGB and RGBD cases.

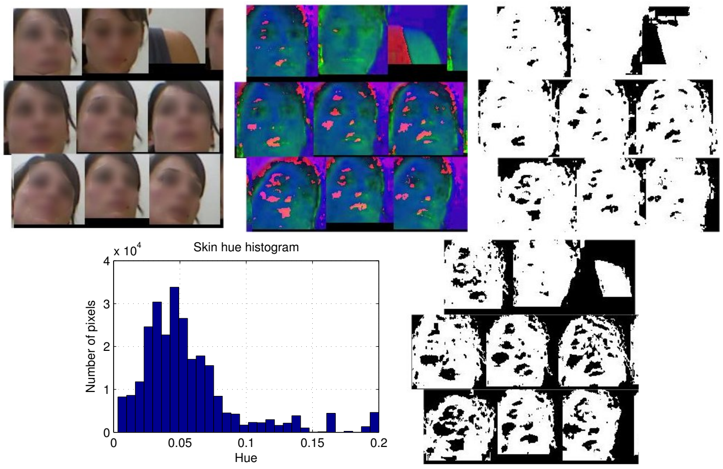


Figure 6.4: Skin color retrieval. Top left: composed image for skin statistics extraction. Top center: HSV converted image. Top right: hue zones that fall under the $(0, 0.2)$ range. Bottom left: hue histogram for skin candidate zones. Bottm right: skin segmentation with the skin statistics. Best viewed in color

6.3.5 Hand likelihood map formation

The hand likelihood map is obtained as the intersection of these cues, to account for the assumptions explained in Section 6.1 (see Figure 6.5 for a visual representation):

$$\mathbf{I}_{H,t} = \mathbf{I}_{OF,\rho,t} \cdot \mathbf{I}_{S,t} \cdot \mathbf{I}_{F,t} \cdot \mathbf{I}_{E,t}. \tag{6.2}$$

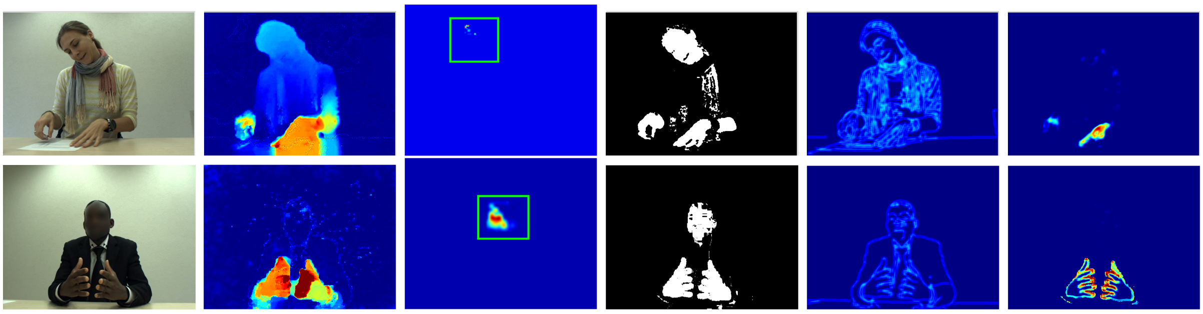


Figure 6.5: Steps for building the hand likelihood maps in the RGB case. All images in the same row corresponds to the same time instant. Columns, from left to right: input video frame, optical flow normalized modulus, probabilistic Viola & Jones output, skin segmentation and face ROI, edge map, and hand likelihood map (the intersection of the other cues). Best viewed in color.

6.4 Priors

As we learned from the state of the art and our motion capture approach in the previous chapter, using "a priori" information can greatly simplify the motion capture process, at the cost of a poorer generalization capability. The latter effect can be however diminished through diverse means.

As prior body information, we use an articulated structure that models the basic human body joints, on top of which we overlay a volume with a polygonal mesh. As prior movement information, we record a series of movements off-line, that are used both for body pose retrieval and action recognition. These processes are next explained in detail.

6.4.1 Body prior

We use a synthetic 3D polygonal torso mesh model, driven by an underlying skeleton with n_{joints} joints (see Figure 6.6). The skeleton pose $\mathcal{A}$ is parameterized by the 3D euclidean rotation angles:

$$\mathcal{A} = \{\vec{a_i}\}_{i=1}^{n_{joints}} = \{a_{\alpha i}, a_{\beta i}, a_{\gamma i}\}_{i=1}^{n_{joints}}. \tag{6.3}$$

The angles $[a_{\alpha i}, a_{\beta i}, a_{\gamma i}]$ correspond to pitch, yaw and roll, and are applied in a hierarchical manner. That is, to obtain the orientation of a given body part, angles must be composed in chain relative to the root node (the base of the neck joint). The root node is referenced to the world global coordinates by its 3D position and orientation.

We have not experienced any problems regarding Gimbal locks, therefore we did not deem necessary the usage of alternative angle representations such as quaternions [Liu and Prakash, 2003].

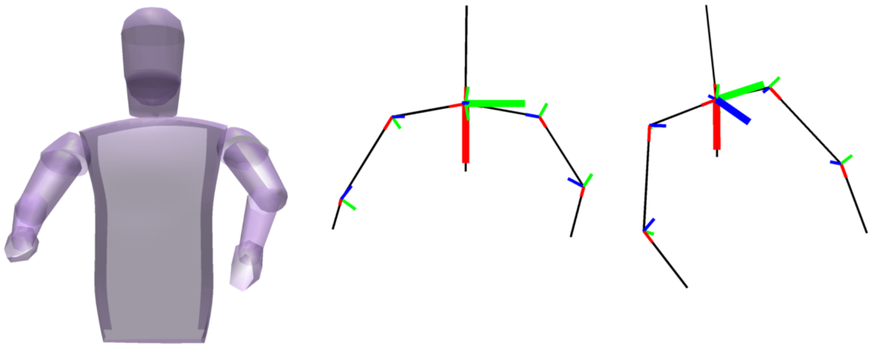


Figure 6.6: Torso Model. Left: 3D mesh. Center and right: underlying skeleton model. The base of the neck is the root node. Red, green and blue lines correspond to the X, Y and Z axes, respectively. Best viewed in color.

6.4.2 Movement prior

The idea of our motion capture approach is to first extract the body geodesic extrema, namely head and hands, and then adjust a model to explain the observations. However, freely adjusting the model to the hands and head position can lead to unrealistic, inaccurate poses. Therefore, we take advantage of prior information gathered by an off-line training process to constrain the state space. This is only done once, as a preprocessing step, and so it should not be mistaken with user-assisted methods. Four subjects (two male, two female) are recorded with a range camera in a similar setting to the target scenario (i.e. seated at the table, see Figure 6.7A left), while performing a number of sub-actions, as explained in Table 6.1. This set of sub-actions is later grouped into a set of n_a actions with similar conversational meaning, resulting in a total of $n_{f,tr}$ training frames, see Table 6.2.

Label	Description
refPose	Reference pose, hands separated, on table.
cA	Arms crossed.
gR, gL, gRL	Perform conversational gestures with one hand, then with the other hand, then with both.
gTR, gTL, gTRL	Same as previous, but resting the non-used elbow on the table.
hH	Hands touching the back of the head.
hHip	Hands touching the hips.
sTR, sTL, sTRL	Placing one hand, then the other, then both in different parts of the table.
thkr	One hand one the chin, another one supporting the elbow of the hand that touches the chin.
tCR, tCL, tCRL	Touching the chin with one hand, then the other, then both, with the non-used hand resting on the table.

Table 6.1: Labeled sub-actions in the training process. Later they will be grouped into bigger conversationally-relevant groups.

As seen in Figure 6.7C and mentioned in the introduction, the actions are all typical of conversational settings. The choice is aimed to capture the adaptors, beat gestures and body pose information from the sequences. The actions are recorded with a range camera, i.e. a set of $\mathbf{I}_{D,tr}$ training range images. These images are then manually annotated to label the joint position. For example, an annotation for joint i is expressed as:

Category	Actions in category
hiddenHands	No hands detected
gestures	cA, gR, gL, gRL, gTR, gTL, gTRL, hHip, hH
handsOnTable	sTR, sTL, sTRL, refPose
selfTouch	thkr, tCR, tCL, tCRL

Table 6.2: Categories of actions. With the help of psychologists we defined four basic actions with conversational meaning, in order to accomplish the objectives described in Section 6.1. See Figure 6.7C for a representation of each category.

$$\vec{p_{i,tr}} = [\vec{u}, \mathbf{I}_{D,tr}(\vec{u})] = [u, v, \mathbf{I}_{D,tr}(\vec{u})]. \tag{6.4}$$

It is therefore described as its pixel position in the image and its associated distance from the camera, forming a 3D vector. Manually annotating the position of every joint along all the training depth recordings allows to obtain the relative 3D location of the whole body, forming the set of training poses:

$$\mathcal{P}_{tr} = \{\vec{p_{i,tr}}\}_{i=1}^{n_{joints}} = \{x_{i,tr}, y_{i,tr}, z_{i,tr}\}_{i=1}^{n_{joints}} \tag{6.5}$$

If an occlusion occurs, we mark the position of the occluded joint either as the one used in the last frame, or as an estimated guess. Given that our torso model is parameterized by angles and at this point we have a set of 3D points $\mathcal{P}_{tr}$, we use an optimization fitting scheme, by using non-linear least-squares to get the angles parameterization $\mathcal{A}_{tr}$ from the 3D points $\mathcal{P}_{tr}$.

As more than one combination of angles could result in the same 3D joint positions, we establish an angle limit for every joint, and then build an energy function based on these constraints, so that the energy is minimum the furthest the joint is from the limit (see Figure 6.7B). The minimum energy is set for the resting "handsOnTable" position. The energy then increases linearly until reaching the empirically-defined maximum angles of rotation for each joint.

Even if rough, this setting produces good results in obtaining the desired parameterization (see Figure 6.7A, and 3D mesh in Figure 6.6, which shows natural limbs and head orientation, thanks to the joint angle energy function).

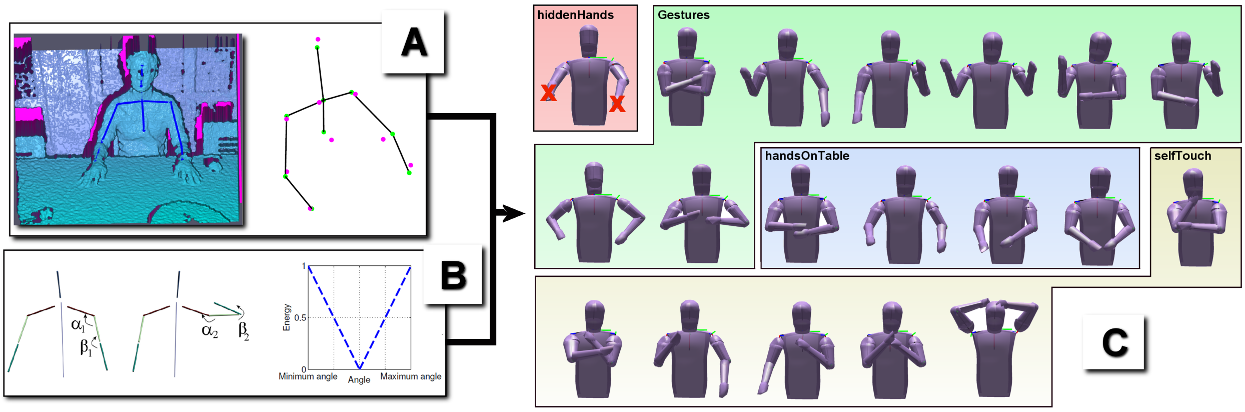


Figure 6.7: Training process. A: Labeling. Left: manually annotated 3D skeleton overlaid into the range camera 3D measurements. Right: optimized torso pose relative to the manually annotated points (in magenta). B: Natural pose attainment. Left skeleton: low energy arm pose. Right skeleton: high energy arm pose (given that α_2 and β_2 are closer to the maximum angle than α_1 and β_1). Right graph: energy function. C: Set of trained movements and their classification into four categories. Best viewed in color.

6.4.2.1 Dimensionality reduction

In the RGB case we aim to obtain the 3D body pose from 2D observations, while performing simultaneous action recognition. Leveraging on our experience in full body motion capture with multiple cameras we employ a similar strategy: we input the training poses parameterized with angles $\mathcal{A}_{tr}$ into the Principal Component Analysis (PCA) framework [Jolliffe, 1986], obtaining a low-dimensional latent space representation:

$$\mathcal{L}_{tr} = \{\vec{l_{tr,i}}\}_{i=1}^{n_{f,tr}} = \{l_{1i}, l_{2i}, ..., l_{qi}\}_{i=1}^{n_{f,tr}} \tag{6.6}$$

$$\mathcal{L}_{tr} = \Omega(\mathcal{X}_{tr}, \mathcal{B}), \; with \; \mathcal{X}_{tr} = \{\mathcal{A}_{tr}\}_{i=1}^{n_{f,tr}} \tag{6.7}$$

In Figure 6.8 it can be seen a graphical representation of our training method. While we are aware that there exist more modern alternatives to PCA such as [Yao et al., 2011], we consider PCA sufficient to reduce dimensionality.

After the movement is compressed by PCA, each action is described by a trajectory in the latent space. We manually mark the most characteristic instant of every action (totaling n_a points) in the PCA low-dimensional latent space, with what we call key points:

$$\{\vec{l}_{k,i}\}_{i=1}^{n_a} \subset \mathcal{L}_{tr}. \tag{6.8}$$

Marking key points is important. This is because along a movement sequence, there are intermediate instants in which the main characteristics of the final posture are not captured. For example, in a crossing arm movement from a hands on table position, the most representative instant of the action occurs when the arms are fully crossed. Intermediate instants such as when the hands start to move away from the table are thus not characteristic of the action. A graphical representation can be seen in Figure 6.8.

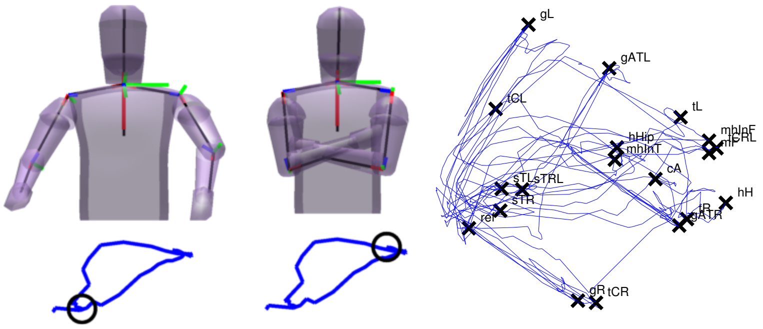


Figure 6.8: Key poses illustration. Left and center: two poses of the cross arms action, with their representation in the PCA space. Only the center one is characteristic of the action (the key point). Right: Whole set of trained motions in the latent space (see Figure 6.7C), with the labeled key points of every action (black crosses).

After the process just described, we construct, for every training frame, synthetic observations for the hand and face positions, projecting them from our torso model onto 2D images by using an estimation of the camera extrinsic parameters. Details can be seen in Section 6.5.2.

6.5 Tracking

We combine the described hands and head observations with the prior data in order to track the relevant information (i.e. hands position and body pose) along the full sequences. For the hands we profit from the fact that the whole information is available from the beginning by using a decision making algorithm. For the body pose we set different constrains that reduce the search in the state space.

6.5.1 Hand tracking

As usual with off-line settings like ours, the whole sequence is available since the beginning of the processing phase. This is a reasonable assumption for the practical application of our methodology (e.g. analyzing job interviews, see Section 6.1). We exploit that by analyzing the hand likelihood map sequence. As stated in Section 6.3, we search for the quickest and closest-to-the-camera body parts. The implementation is done through the steps that follow, as shown in Figure 6.9.

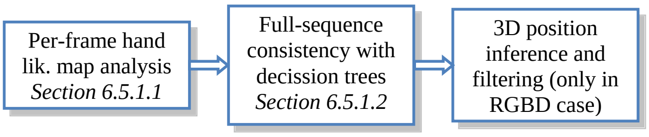

Per-frame hand lik. map analysis *Section 6.5.1.1*
Full-sequence consistency with decission trees *Section 6.5.1.2*
3D position inference and filtering (only in RGBD case)

Figure 6.9: Hand tracking framework.

6.5.1.1 Per-frame hand likelihood map analysis

For each hand likelihood map frame $\mathbf{I}_{H,t}$, we first perform a search for local maxima, first by using a smoothing filter (Gaussian kernel) in order to better show local tendencies, obtaining $\mathbf{I}'_{H,t}$. We then threshold $\mathbf{I}'_{H,t}$, and cluster local maxima using an adaptative k-means classifier with support for a variable, unspecified number of classes. K-means also provides identity consistency of local clusters along time. At this point we have a set of local maxima of the whole sequence of hand likelihood maps.

Aided by the identity consistency, we compute the paths of the modes in $\mathbf{I}'_{H,t}$ over time. This originates a set of $n_\mathcal{T}$ trajectories through the hand likelihood map. We call them tracklets, and are non continuous in the sense that detected local maxima will disappear and then re-appear in the image because of occlusions, being out of frame, and/or malfunction of the hand likelihood maps. This process outputs a group of tracklets, each of which is defined as follows:

$$\mathcal{T} = \{\vec{t_i}\}_{i=1}^{n_\mathcal{T}} = \{t_{o,i}, t_{f,i}, \lambda_i, u_{\vec{o},i}, u_{\vec{f},i}\}_{i=1}^{n_\mathcal{T}}, \tag{6.9}$$

where $n_\mathcal{T}$ is the number of tracklets in the sequence; $[t_{o,i}, t_{f,i}]$ are the time instants when the tracklet i starts and ends; λ_i is the accumulated likelihood along the tracklet i duration, $u_{\vec{o},i}$ is the pixel position where the tracklet i started, and $u_{\vec{f},i}$ is the pixel position where the tracklet i ended. Longer tracklets therefore usually have bigger λ_i. As tracklets do not have a maximum length value, λ_i is not upper-bounded.

6.5.1.2 Full-sequence consistency with Decision Trees

In order to obtain the best 2D paths for a hand in the image, we implement a decision tree algorithm, in which the tracklets are the branches. A decision of what tracklet to follow next is made in every node, based on several factors explained below. In Figure 6.10, a 1D example of how four tracklets look along time is shown. The goal is to find the path in which the accumulated likelihood is maximum. For this, we establish three basic rules:

- Once the hand is assigned to a tracklet, it is not possible to jump to another tracklet until the current one has reached its end. This is key to enforce the assumption that $\mathbf{I}_{H,t}$ encodes the most likely hand trajectories when taking the whole sequence into account, even though it does not have to hold true for every time instant.
- Once a tracklet has finished, it is possible for the hand to stay in that tracklet final pose until the end of the sequence, or jump to any other tracklet that has started afterwards.
- When jumping from one tracklet to another, jump distances (in pixel positions) are taken into account to penalize far jumps. The accumulated likelihood of a hand taking two tracklets, $\mathcal{T}_i$ then $\mathcal{T}_j$ (that is, following path from the initial point $\vec{u_{o,i}}$ of $\mathcal{T}_i$ to the final point $\vec{u_{f,j}}$ of $\mathcal{T}_j$ through points $\vec{u_{f,i}}$ and $\vec{u_{o,j}}$), separated by a distance $\delta_{ij} = \|dist(\vec{u_{f,i}}, \vec{u_{o,j}})\|$, is:

$$\lambda_{T,ij} = \lambda_i + \lambda_j e^{-\kappa_d \delta_{ij}}, \tag{6.10}$$

  where κ_d is a distance penalization factor (manually set in experiments). We then look for the path with the highest accumulated likelihood.

As the used job interview sequences are long (some up to more than 20 minutes), the number of tracklets can be large. Given that the number of paths increase exponentially with the number of nodes, we back-compute the accumulated likelihoods, retaining only the maximum path at each node.

Tracking two hands: As there are two hands to track, we look for the two trajectories (i.e. tracklet paths) with the highest likelihoods. We first define a priority hand, that is, the one that will evaluate the tracklet tree first, thus getting the best path. After it has been computed, we set to 0 the accumulated likelihood of the tracklets used by the optimal path, and then evaluate the modified tree for the other hand. This algorithm finally outputs the position of the visible moving hands (left or L and right or R) in the image at time t:

$$\mathcal{H}_{L,R} = \{\vec{h_{L,R}}\}_{i=1}^{n_f} = \{\vec{u_{L,R,i}}, t_i, \lambda_i\}_{i=1}^{n_f}, \tag{6.11}$$

where n_f is the total number of frames of the sequence.

6.5.2 Upper body pose tracking

At this point, the 2D position of the hands in the image is available, while for the latter the information is 3D. Therefore different approaches are used in order to obtain the torso pose. The distances between the observed head and hands, and the ones trained are computed. The final pose will be defined as the closest match in this database lookup process.

6.5.2.1 Database lookup

We compare the real inputs with our set of synthetic data, first by using discrepancies between hand and head positions, and then fine-tuning with the foreground/edges to choose the best match.

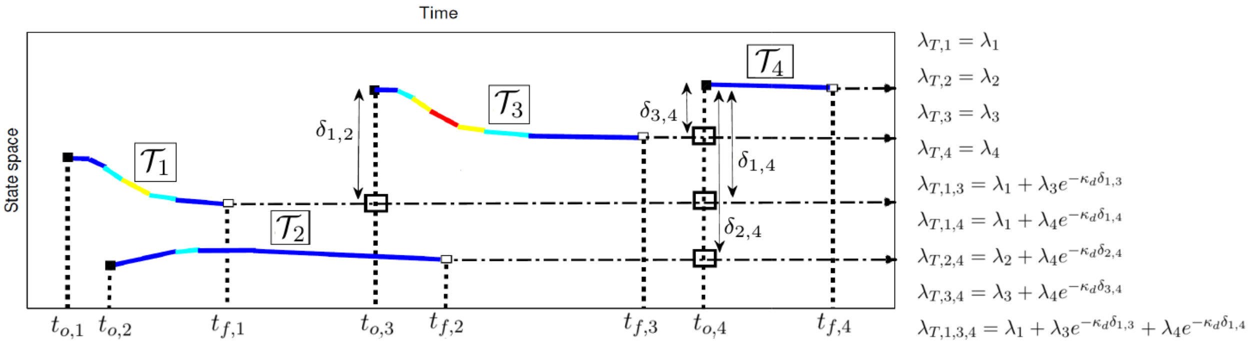

Time
State space
$\mathcal{T}_1$
$\mathcal{T}_2$
$\mathcal{T}_3$
$\mathcal{T}_4$
$\delta_{1,2}$
$\delta_{3,4}$
$\delta_{1,4}$
$\delta_{2,4}$
$t_{o,1}$
$t_{o,2}$
$t_{f,1}$
$t_{o,3}$
$t_{f,2}$
$t_{f,3}$
$t_{o,4}$
$t_{f,4}$
$\lambda_{T,1} = \lambda_1$
$\lambda_{T,2} = \lambda_2$
$\lambda_{T,3} = \lambda_3$
$\lambda_{T,4} = \lambda_4$
$\lambda_{T,1,3} = \lambda_1 + \lambda_3 e^{-\kappa_d \delta_{1,3}}$
$\lambda_{T,1,4} = \lambda_1 + \lambda_4 e^{-\kappa_d \delta_{1,4}}$
$\lambda_{T,2,4} = \lambda_2 + \lambda_4 e^{-\kappa_d \delta_{2,4}}$
$\lambda_{T,3,4} = \lambda_3 + \lambda_4 e^{-\kappa_d \delta_{3,4}}$
$\lambda_{T,1,3,4} = \lambda_1 + \lambda_3 e^{-\kappa_d \delta_{1,3}} + \lambda_4 e^{-\kappa_d \delta_{1,4}}$

Figure 6.10: Decision tree example. Left: Hand tracking tracklet decision tree example with 4 tracklets ($\mathcal{T}_1$ - $\mathcal{T}_4$) and 4 nodes, along a 1D state space. Color encodes tracklet likelihood in a given time instant (warmer means higher). Nodes are represented with squares. Right: Likelihood values λ for each possible path. Best viewed in color.

We compute the overlap between the body parts from which we compare observations and the database (namely head and hands) and the poses contained in the database. To that end, we first model the position of the body parts as 2D Gaussians. Later, we compute the difference between the observed and database-generated Gaussians, in order to obtain a overlap measure. This acts as a robust distance measure. An example with the hands can be seen in Figure 6.11.

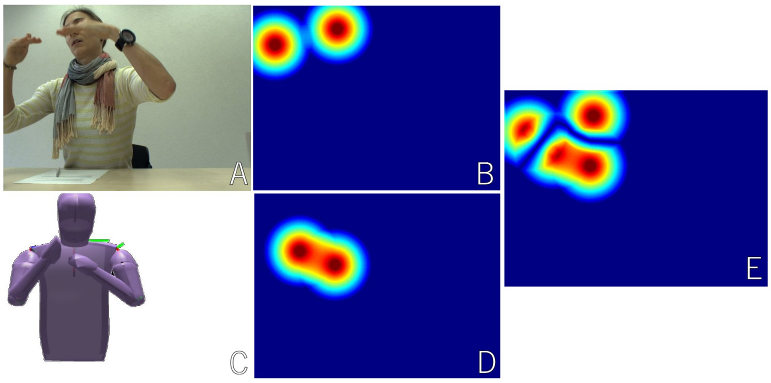


Figure 6.11: Database lookup example for hands. A: input RGB frame. B: likelihood for the hands build from their obtained position, not to be confused with $\mathbf{I}_{H,t}$. C: one of the frames in the database, corresponding with the action "gestures". D: synthetic likelihood for the hands in the database frame. E: difference image.

The value of the pixels of the overlap measure image (Figure 6.11E) are then aggregated. The resultant value is weighted by the difference between the observed and synthetic silhouettes. The database frame with a smaller numerical value is chosen as the current body configuration. It is then temporally smoothed with a basic filter, in order to reduce the flickering. This step means that the output pose is not restricted to identical poses to those found in the database, as intermediate positions are generated. After this process we obtain final body configuration $\mathcal{P}_o$.

6.6 Action recognition

We approach the problem in a similar way to that of the Chapter 5: by finding the best correspondence between the current pose and the training data, which is labeled. We profit from the dimensionality reduction scheme in order to better analyze the ongoing action, in a similar fashion as it was done in full body motion capture from multiple cameras, in the previous chapter. Instead of having a tracking framework with switching models, we obtain the performed action first by mapping the output pose $\mathcal{P}_o$ into the latent equivalent point $\vec{l_o}$ by using the inverse mapping function:

$$\vec{l_o} = \Omega^{-1}(\mathcal{P}_o) \tag{6.12}$$

As explained in Section 6.5.2, even if database lookup is performed to obtain the output pose $\mathcal{P}_o$, because of the use of a smoothing filtering, intermediate untrained poses are generated and $\vec{l_o}$ is not restricted to training points. Therefore, we define the ongoing action $f_{a,t}$ as:

$$f_{a,t} = argmin(dist(\vec{l_o}, \mathcal{L}_{tr})). \tag{6.13}$$

That is, we compute the distance of $\vec{l_o}$ to every point of the training latent shape $\mathcal{L}_{tr}$. The trained action associated with the minimum distance is set as the ongoing action $f_{a,t}$. See figure 6.12 for details.

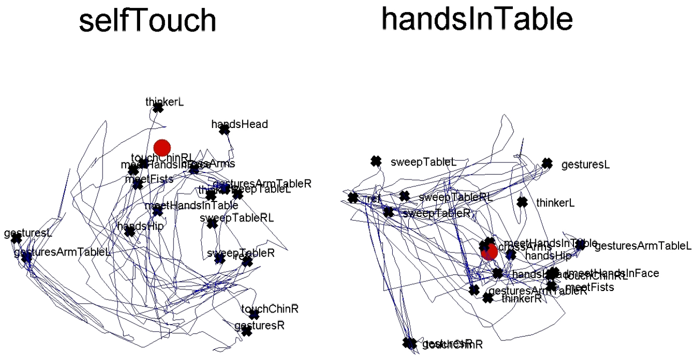


Figure 6.12: Example of action recognition in the RGB case. The blue line represents the whole training set $\mathcal{L}_{tr}$. The black crosses represent the keypoints of every sub-action (described in Table 6.1). The red sphere represents the current pose in the latent space, $\vec{l_o}$. Best viewed in color.

6.7 Improvements of our initial method

We have extended our RGB-only proposal from [Marcos-Ramiro et al., 2013] in order to profit from sequences in which depth data is also available (constituting what we call the RGBD case). In addition, several improvements are made in different steps, and described in this section.

6.7.1 Improvements in the observation system

6.7.1.1 Face detection

We improve the way to discard face region pixels when building the hand likelihood map. This is because some subjects wear open neck clothes, resulting in skin-colored pixels falling outside the face detector bounding box, thus taken into account as hands. To address the problem, we take several skin-colored pixels within the face bounding box, and use them as seed points for a region-growing segmentation algorithm, which employs a color similarity criterion. The growth is stopped in depth discontinuity points in order to account for the possible inclusion of hands in the growing region, as they are on a different depth level than the face. The output is a binary image $\mathbf{I}'_{F,t}$. The difference between both approaches can be seen in Figure 6.13.

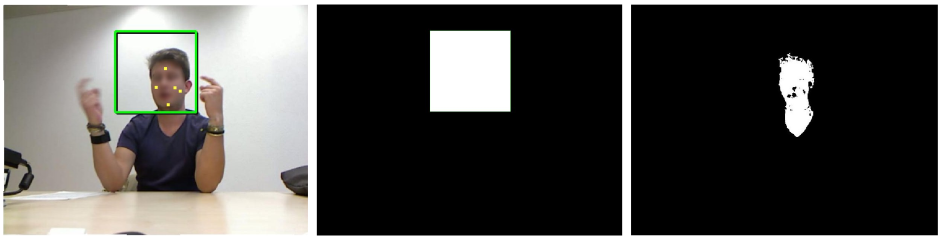


Figure 6.13: Face region removal. Left: input RGB frame (face blurred only for display purposes), with VJ detection overlaid as a region of interest. Seed points for the region growing are marked in yellow. Center: output face mask in the RGB case. Right: improved output mask for the RGBD case. Best viewed in color.

6.7.1.2 Addition of depth

We retrieve range images $\mathbf{I}_{D,t}$ by using a commercial Microsoft Kinect sensor. The background is segmented with a distance threshold, while the table is originally undetected because of the high angle of attack relative to the infrared beam of the range sensor. The resulting segmented torso can be seen in Figure 6.16. It should be noted that we inverted the range values so that closer parts relative to the camera have higher numerical values. Depth and RGB images are registered, so that color and depth information for the same pixel position refers to the same spatial point.

6.7.1.3 Color and depth fusion in the hand likelihood map

We added depth, weighted optical flow and depth fusion, and new face region segmentation, with removed edges:

$$\mathbf{I}_{H,t} = \mathbf{I}_{S,t} \cdot \mathbf{I}^{'}_{F,t} \cdot (\kappa_1 \mathbf{I}_{OF,\rho,t} + \kappa_2 \mathbf{I}_{D,t}), \tag{6.14}$$

where the constants κ_1 and κ_2 are set by hand and used to weight the importance that the optical flow and depth carries. See Figure 6.14 for an illustration of equation of the composition of the new hand likelihood map, and their advantages compared with the previous approach (Figure 6.5).

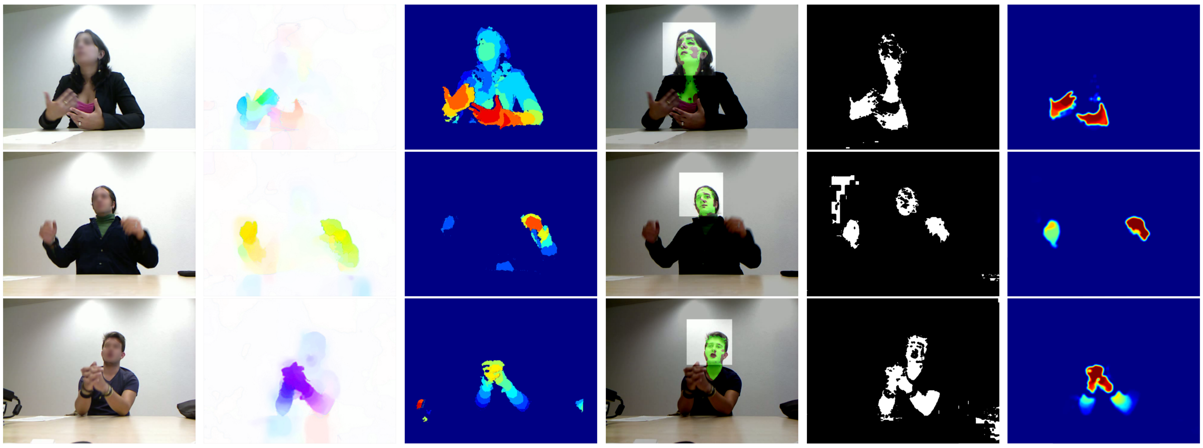


Figure 6.14: Steps for building the hand likelihood maps in the RGBD case. All images in the same row corresponds to the same time instant. Columns, from left to right: input video frame $\mathbf{I}_t$, optical flow $\mathbf{I}_{OF,\rho,t}$, depth map $\mathbf{I}_{D,t}$, skin segmentation and face ROI ($\mathbf{I}_{S,t}$, $\mathbf{I}^{'}_{F,t}$), and hand likelihood map $\mathbf{I}_{H,t}$ (the intersection of the other cues). Best viewed in color.

6.7.2 Improvements in maxima extraction from the hand likelihood map

We changed the way in which we look for local maxima in the hand likelihood map. For each hand likelihood map frame $\mathbf{I}_{H,t}$, we first regularize the hand map with a smoothing filter, obtaining $\mathbf{I}'_{H,t}$. We then use the Mean Shift framework (see Section 2.3.2) to find the different modes of $\mathbf{I}'_{H,t}$ (Figure 6.15). The number of modes is not restricted in this step in any way. We also provide identity consistency for the several local clusters along time. At this point we have a set of local maxima of the whole sequence of hand likelihood maps.

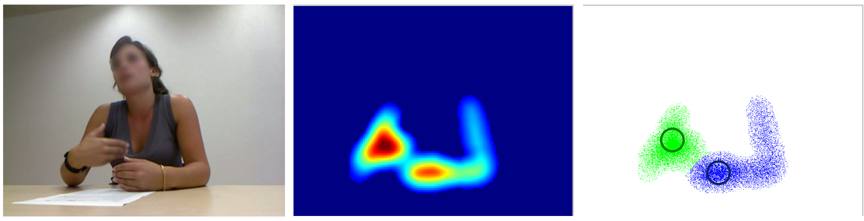


Figure 6.15: Hand likelihood map local maxima search with Mean Shift. Left: Original image. Middle: Regularized hand likelihood map. Right: Modes found with Mean Shift (in different colors) and their centroids, represented with circles. Best viewed in color.

6.7.3 Addition of 3D hand tracking

In the RGB-only case, the 2D hand paths are stored in the tracklets, while the 2D head position is obtained with a Viola & Jones detector, as explained in Sections 6.3 and 6.5.1. We use the depth value of the 2D tracks, obtained with the range sensor, in order to infer the hands and head 3D positions. As Figure 6.16 shows, regions of interest around the 2D locations are analyzed. Hands are deemed to be the closest point to the camera within their respective regions, and the head the furthest point. This is justified as after segmenting $\mathbf{I}_{D,t}$ the wall and table are excluded, so there is nothing behind the head or in front of a hand.

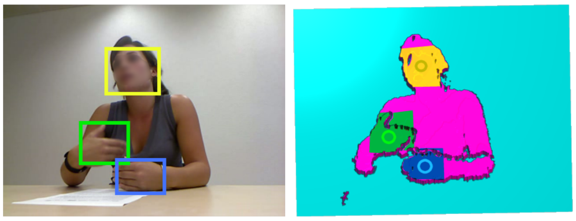


Figure 6.16: Segmented depth with its hand and head location. Left: Bounding boxes of the 2D hands and head locations. Right: Search for the 3D points in the depth image. Best viewed in color.

The information stored in the tracklets (both hand position and number of hands detected) can be noisy due to the tracking-by-detection nature of the system. In order to address this, we encode the number of detected hands as states in a Hidden Markov Model. By using the Viterbi framework, and establishing a tendency to stay in the current state (90% in the transition probability matrix) we decode the hand count, improving the estimation of visible hands. Finally, we implemented a Kalman Filter in order to obtain a smooth 3D hand position.

6.7.4 3D upper body pose retrieval through model fitting

In order to infer the torso 3D pose, we propose to adjust an articulated upper body model to the head and hands 3D measurements, helped by prior information gathered with custom-made training data. As explained, to build this data we first collect and label several typical, conversationally-relevant upper body poses, with the help of a range camera. Then we use a non-linear optimization technique to minimize a joint energy function. The process is presented as follows, and summarized in Figure 6.17.

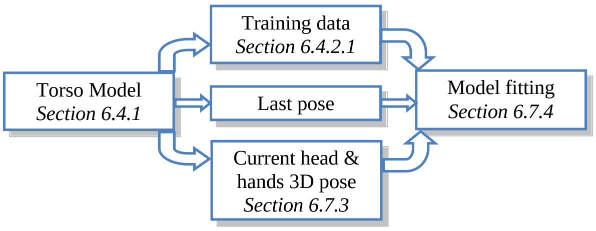


Figure 6.17: Torso pose extraction overview.

The 3D torso pose is extracted in a two-step process. First, an approximate pose is quickly estimated via database lookup, getting the best match by using 2D hands and head position and silhouette affinity as cues. The silhouette has been segmented by identifying the depth blob corresponding to the detected face of the subject. Then, this first guess is used to initialize a nonlinear least-squares optimization method that further refines the pose (see Figure 6.19). The cost function used is:

$$\varepsilon = \varepsilon_{hands} + \mu_1 \varepsilon_{prior} + \mu_2 \varepsilon_{tSmooth} \tag{6.15}$$

The term $\varepsilon_{hands} = \|p_{\vec{hands}_t} - p_{\vec{hands},o,t}\|$ comprises the 3D hands and head position difference between those detected in the image and the current position in the articulated model. The term $\varepsilon_{prior} = \|e_{n,t}\|$ penalizes the less natural positions by using the same function described in Figure 6.7B. The term $\varepsilon_{tSmooth} = \|\mathcal{P}_{o,t} - \mathcal{P}_{o,t-1}\|$ is the difference between the current estimate and that of the previous frame, used for temporal consistency. Hyperparameters μ_1 and μ_2 are set empirically. After this process, the approximate 3D upper body position of the participant $\mathcal{P}_o$ is retrieved, as seen in Figure 6.18.

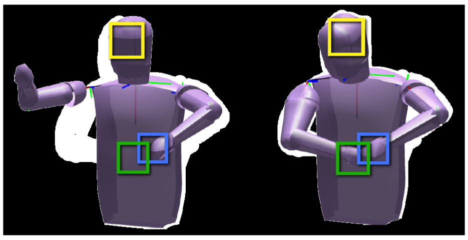


Figure 6.18: Torso pose optimization. Before iterating (left) and after converging (right). The face and hands 3D position and silhouette consistency are taken into account.

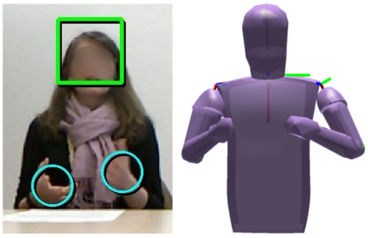


Figure 6.19: 3D upper body retrieval. Left: detected face and hands. Right: approximate 3D limbs configuration.

6.7.5 Action recognition

At this point we have obtained the hands' position in the image $\{\mathcal{H}_{L,R}\}_{i=1}^{n_f}$ and the approximate 3D torso pose $\{\mathcal{P}_o\}_{i=1}^{n_f}$. As the next step, we extract a series of features from the participant, that can schematically be seen in Figure 6.20.

6.7.5.1 Hand height ($f_{h,t}$)

By using a straight lines Hough detector, we obtain the table edge position in the image. We then compute the distance in pixels between the edge and the face of the participant. The height of each hand is expressed as a proportion relative to the face-table distance. A value of 0 means the hand is located in the edge of the table, while a value of 1 means that the hand is at the same height of the face. If the hand has not been detected, a -1 value is assigned. This feature is therefore two-dimensional, with one dimension per hand.

6.7.5.2 Hand movement ($f_{m,t}$)

The detected hand position time differences are not a reliable indicator for hand speed, since the tracker can focus on different parts of the hand, leading to inaccuracies. Also, it does not capture the nuances of small hand and finger movements, which are still hand activity. In order to circumvent that problem, we used the average optical flow modulus present in a region of interest around the detected hand coordinates as hand speed measure. This feature is also two-dimensional.

6.7.5.3 3D face-hand distance ($f_{d,t}$)

We compute the euclidean 3D distance between face and hands, and normalize it with respect to the face-table distance. Lower values indicate closeness to the face. This feature is two-dimensional.

6.7.5.4 Speaking status ($f_{s,t}$)

The commercial microphone array Microcone[1] provides automatic binary speaking status segmentation (talking or silent), that we use as a feature.

6.7.5.5 Ongoing activity ($f_{a,t}$)

We defined five activity classes (one extra respect to the RGB case), as explained in Chapter 4 and Table 4.7 shows. The choice is based on the relevant nonverbal cues that we want to extract (see Section 6.1) and in the observed frequencies in the recorded videos (see Section 4.3).

Up to this point we have therefore a 7-dimensional feature vector for instant t:

$$\mathcal{F}_{ar} = \{f_{h,t}, f_{m,t}, f_{d,t}, f_{s,t}\}. \tag{6.16}$$

For inferring the ongoing activity ($f_{a,t}$), we train a Random Forest classifier by concatenating $t + t_w$ and $t - t_w$ feature vectors $\mathcal{F}_{ar}$, where t_w denotes the size of a temporal window. We then obtain the final $(7 \cdot 2 \cdot t_w + 1)$-dimensional feature vector for time instant t, that we use for training:

[1] http://www.dev-audio.com/products/microcone/

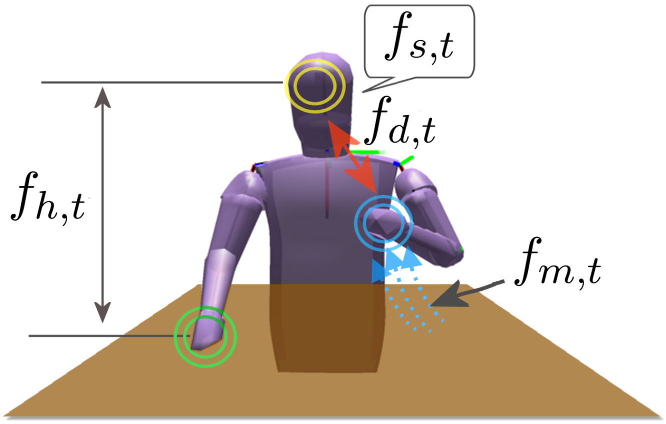


Figure 6.20: Graphical representation of the per-frame extracted features: hands height ($f_{h,t}$), hands movement ($f_{m,t}$), 3D face-hands distance ($f_{d,t}$) and speaking status ($f_{s,t}$). Best viewed in color.

$$\mathcal{F}'_{ar,t} = \{\mathcal{F}_{ar,t-t_w:t+t_w}\} \tag{6.17}$$

The labels for training are a set of manual annotations of the perceived ongoing activity along the whole video corpus (see Section 6.8 for details) The output is therefore a $f_{a,t_0:t_f}$ vector that encodes the action performed in every time instant t. Its performance is evaluated in Section 6.8.

6.8 Implementation and results

6.8.1 Data

We make use of the data described in Section 4.3. Specifically, we employ the subsets A and B to evaluate the performance of the proposed framework. They contain the hand-labeled actions and hand positions in the image. In order to validate the performance across different data, we also use the ChaLearn database (see Section 4.4).

6.8.2 RGB case

6.8.2.1 Hand likelihood map evaluation

The results are shown in Figure 6.22 and illustrated in Figure 7.11. Regarding hand tracking, Figure 6.22 (top) shows the error for both hands. As can be seen, the error remains below 20 pixels in many frames, except when error spikes appear. The mean error is **17.35 pixels**. Furthermore, the detection error is **8.75%**. Note that the chosen data for testing is specially challenging, so we would expect the method to perform better in many other situations.

6.8.2.2 Action recognition evaluation

Regarding action recognition, the overall classification accuracy is **72.5%**. The performance is significantly better than random (25%), but also than a majority-class method that would label every frame as 'handsOnTable' (67.5%, $p = 0.0238$). It is important to mention that correctly classifying the 'handsOnTable' action is not trivial, due to factors like slow movements, skin colored clothes, or sleeve-less

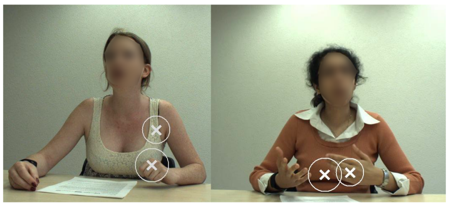


Figure 6.21: Failure examples. While the subject has the hands on the table, a combination of several image cues such as skin-colored clothing and movement in body parts that are not the hands makes our system fail in the shown frames.

shirts. As an illustration, we show two failure examples of the hand tracking in Figure 6.21. Examples of correctly recognized actions are shown in Figure 7.11.

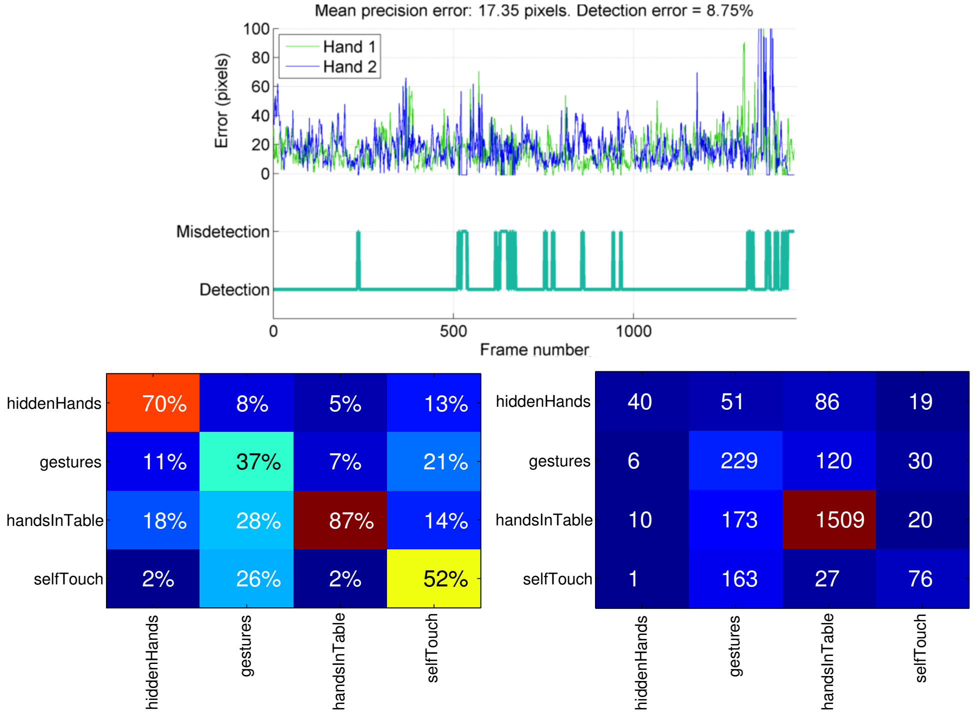


Figure 6.22: Hand tracking performance in the RGB case. Top: hand tracking error results. Middle and bottom: Confusion matrix, normalized by columns (middle) and not normalized (bottom). Warmer colors mean higher values. Best viewed in color.

Our algorithm finds two main challenging points:

- Given that we make a comparison with training data in order to obtain the torso 3D pose, the system has difficulties coping with body poses outside the training ones. This can be addressed in two different ways: by creating a larger training set, where using synthetically generated poses is an option [Shotton et al., 2011], or by using the current output to initialize an optimization scheme to better adjust the pose. The latter option could be viable only if the processing time is low enough, to keep the problem tractable given the large amount of data to process.

- As we perform the analysis on monocular video, the observed hand position if the subject makes hand gestures in front of his face is very similar to that of self touch. Similarly, judging exclusively the wrist joint position, it is challenging to differentiate between actions 'gestures' and 'handsOnTable', if the action is taking place near the table.

6.8.3 RGBD case

6.8.3.1 Hand likelihood map evaluation

Job interview database: the results of the hand detection are shown in Figure 6.23. We set 40 pixels as a threshold value, which visually is the accepted highest drift for a perceived correct detection. All the following numbers therefore will be provided for this pixel threshold value. Our method (RGBD+RG) is tested both with and without Decision Trees (DT) tracking. It has been compared to that of our previous work [Marcos-Ramiro et al., 2013], and also to the result obtained when leaving out the face skin region growing (RG) procedure (see Section 6.3 for details).

It is clear that the addition of depth helps the hand detection in a high degree. Our method gets an average **28.7%** higher detection rate than [Marcos-Ramiro et al., 2013]. The inclusion of RG increases the detection rate in an average of **5.6%**. The overall RGBD+RG detection rate is **78.2%** when using DT.

Upon visual inspection most of the non-detections are a cause of the person's hands being close together, being therefore detected as a single one. In practice this is however not a serious issue, as for action recognition usually the ongoing action is a function of where the highest hand is (for example, when gesturing or self-touching, it is irrelevant to leave a hand in the table for the action to be identified). Also, for upper body pose, the regularization term in the optimization method (see Section 6.7.4) usually dampens mis-detections. False positives on the other hand are quite costly. If a hand is mis-detected in the face region for several consecutive frames, there is high probability for the action to be incorrectly classified as 'self touch'. This is where the RG algorithm comes into effect. As seen in Figure 6.23B, it improves the false positive rate by **16.9%**. The same effect appears when applying the DT tracking scheme: while the outcome of the precision is largely unaffected, it reduces false positives by an average of **12.8%**. It therefore becomes a trade-off between missed hand rate and false positive rate. As the next steps of the pipeline are more affected by false positives, the RGBD+RG with DT becomes the best overall performer: while it misses hands more frequently than RGBD alone, it keeps the more relevant false positive rate much lower while offering better precision.

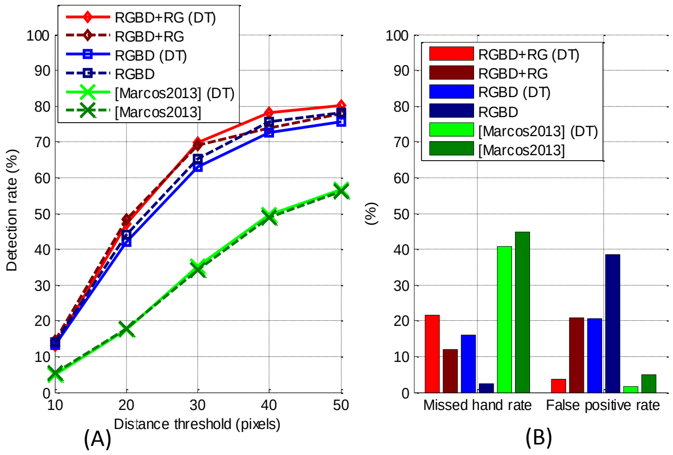


Figure 6.23: Hand detection performance. (A) Hand detection rate as a function of the pixel threshold. (B) Missed hand and false positive rates.

ChaLearn 2011: the metrics that we use to evaluate the performance are (a) detection rate as defined in Experiment 1, and (b) maxima average order. In (b), the local maxima of the hand likelihood map are obtained as described in Section 6.7.2, and ordered with respect to their likelihood value. Then, each hand annotation is associated with the closest (in its 2D position) maxima. For example, if the right hand is associated with a local maxima which contains the 3rd highest likelihood, and the left hand is associated with the best local maxima, the maxima average order is 2 for that frame. A maxima average order of 1.5 provides the best possible scenario, as the two hands would be associated with the two maxima that contain the highest likelihoods.

The results of the hand detection in ChaLearn are shown in Figure 6.24. We use [Ferrari et al., 2009], [Sapp and Taskar, 2013], as baselines, together with the body part classification of [Shotton et al., 2011] coupled to a 2D hand position regressor. As expected, methods that use entirely or partially depth outperform the 2D-only methods. In addition, our method is able to locate the hands very precisely, as the comparison with [Shotton et al., 2011] shows. The maxima average order is **2.4** for the right hand and **2.55** for the left hand, showing that in general, one of the best 3 maxima of the hand likelihood map are overlapped with the hand's positions. This enables high quality information to be passed to the tree-based tracker in order to reliably obtain the hand's position along time, as assessed using the job interview database.

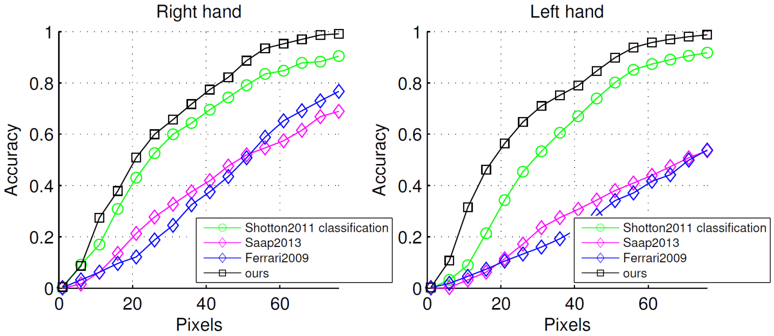


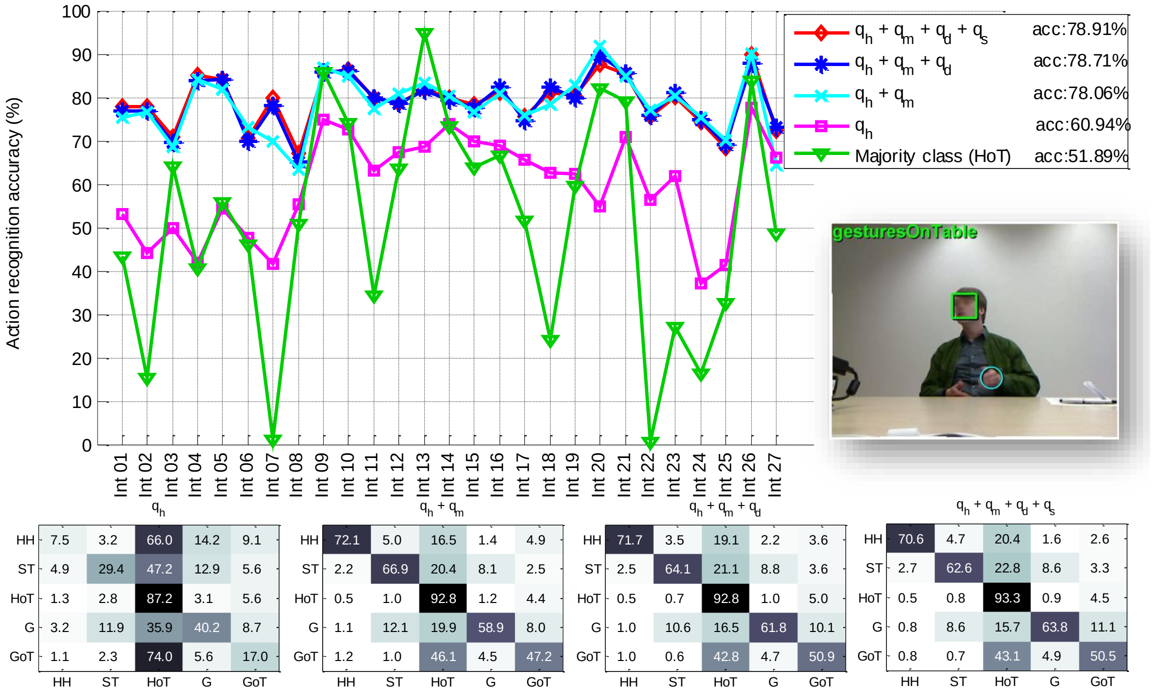
Figure 6.24: Results in ChaLearn 2011, as hand detection performance. We compare our method against baselines [Ferrari et al., 2009], [Sapp and Taskar, 2013] and a modification of [Shotton et al., 2011].

6.8.3.2 Action recognition evaluation

The results for action recognition can be seen in Figure 6.25. Several configurations for the feature vector have been tested. The value of t_w has been empirically set to 5, as higher numbers do not provide a significant precision improvement, while increasing the computational cost.

All variations of the used feature vector resulted in a statistically significant improvement over the majority class performance. The behavior of the algorithm is consistent along all interviews: under a high variation of clothing, skin color, gender, or class distribution, the mean and standard deviation of the accuracy for the best performing combination are **78.9%** and **5.9%**, respectively. The biggest performance jump is found when the motion of the hands $f_{m,t}$ is taken into account: on average, it improved the recognition by **18.9%**. Without this cue, the biggest source of error was the confusion between the two most populated classes, 'hands on table' and 'gestures on table' (see the confusion matrices in Figure 6.25). Understandably, it shows that hand speed is a big factor to distinguish between those classes. Nevertheless, even when the hand speed is used, the confusion matrices show that mixing both classes is still an issue. However, upon visual inspection of the sequences, the classification is consistent with the amount of movement of the hands. This suggests that the annotators used additional cues other than the amount of movement to distinguish between 'hands on table' and 'gestures on table', such as the orientation of the hand palms or finger position. This possibility offers grounds for future works.

Adding the 3D distance from hands to face ($f_{d,t}$) improved the recognition accuracy, but only by **0.65%**. This shows that although having the 3D position helps, the 2D hands position relative to the



q_h

	HH	ST	HoT	G	GoT
HH	7.5	3.2	66.0	14.2	9.1
ST	4.9	29.4	47.2	12.9	5.6
HoT	1.3	2.8	87.2	3.1	5.6
G	3.2	11.9	35.9	40.2	8.7
GoT	1.1	2.3	74.0	5.6	17.0

$q_h + q_m$

	HH	ST	HoT	G	GoT
HH	72.1	5.0	16.5	1.4	4.9
ST	2.2	66.9	20.4	8.1	2.5
HoT	0.5	1.0	92.8	1.2	4.4
G	1.1	12.1	19.9	58.9	8.0
GoT	1.2	1.0	46.1	4.5	47.2

$q_h + q_m + q_d$

	HH	ST	HoT	G	GoT
HH	71.7	3.5	19.1	2.2	3.6
ST	2.5	64.1	21.1	8.8	3.6
HoT	0.5	0.7	92.8	1.0	5.0
G	1.0	10.6	16.5	61.8	10.1
GoT	1.0	0.6	42.8	4.7	50.9

$q_h + q_m + q_d + q_s$

	HH	ST	HoT	G	GoT
HH	70.6	4.7	20.4	1.6	2.6
ST	2.7	62.6	22.8	8.6	3.3
HoT	0.5	0.8	93.3	0.9	4.5
G	0.8	8.6	15.7	63.8	11.1
GoT	0.8	0.7	43.1	4.9	50.5

Figure 6.25: Action recognition performance. Top: accuracy of the proposed algorithm in the different interviews (Int 1 ... Int 27), in function of different feature vectors configurations, and relative to the majority class performance ('hands on table'). In the right part, a frame of one of the video sequences with the hands, head, and ongoing activity overlaid. Bottom: confusion matrices for different feature vector configurations. Best viewed in color.

face is already enough to distinguish between actions. The same applies for the speaking status ($f_{s,t}$), although it crucially shows that multi-modality can be exploited in order to improve action recognition. This finding is also supported with other recent works like [Nguyen et al., 2012].

Overall, the present work significantly improves our previous work [Marcos-Ramiro et al., 2013], with a global accuracy of **78.9%** and baseline performance (majority class) of **51.9%**, in contrast to the **72.5%** and **67.5%** respectively of [Marcos-Ramiro et al., 2013], while adding an extra 'gestures on table' class. It is important to mention that correctly classifying the majority class 'hands on table' action is not trivial, as factors like slow motions, skin colored clothes, or sleeve-less shirts have to be dealt with. As an illustration, we show some failure examples of the hand tracking in Figure 7.11. Note that in some cases, even if the hands are wrongly detected, the temporal features used for action recognition recover the right action.

6.8.4 Comparison of the performance of RGB and RGBD cases

Comparing the performance of the RGB and RGBD cases is not straightforward. In the RGB case only a color camera is used, but it has a higher quality to the RGB camera of Kinect, which is used for the RGBD case. This is so because the used cameras are not calibrated, making it impossible to align the HD camera images to Kinect's depth information, and forcing the use of Kinect's integrated RGB camera.

In addition, the Kinect sensor and the HD RGB camera are placed in different points of the table. The consequence is that for some subjects the points of view are noticeably different, resulting in self-occlusions in one view that are not present in the other and making therefore the comparison of both methods by using the same sequences not completely rigorous. Another effect is that before the addition of depth information, the quality of the hand likelihood map for the RGBD case is lower than in the

RGB case. For instance, the skin segmentation method becomes less reliable.

However, it can be concluded that when depth is used, the reliability of the whole system increases. Even if it is not possible to comprehensively compare the RGB and RGBD cases, it has been demonstrated that in frontal views of the human upper body the assumption that in general the hands are the closest part of the body to the camera holds true.

6.9 Conclusion

We present a system that automatically analyzes the communicative cues of seated participants in conversational events recorded with regular, broadly available cameras, while accounting for the possible use of commercial RGBD sensors to increase the robustness. We studied our system in the context of real job interviews. We built original body part detectors which were used to get an approximate 3D upper body pose, which will be useful for developing more complex techniques in our future work.

We use different approaches depending on whether depth information is available or not. This is justified as generally, in psychological research, the addition of depth information is not yet considered. The sole use of RGB video also enables our framework to use existent prerecorded data in which range sensors were not available, while allowing for the possibility of using them in cases in which they have been used.

With our method we obtain the upper body pose and the ongoing activity during the job interview. With this information we intend to look for adaptors and beat gestures, which previous studies have shown to carry nonverbal communication information. Our system can recognize basic upper-body actions with an accuracy of 72.5%, in a dataset of 105 minutes of real job interviews in the RGB case. In the RGBD case, our system can recognize 5 upper-body actions with an accuracy of 78.9%, in a dataset of four and a half hours of real job interviews.

The results obtained with our proposal have shown to be useful as a first building block to automatically analyze psychological traits of the participants in the conversation, and psychologists that we have discussed with find this type of recognition and current performance quite promising. Our future work will deepen and explore the possibilities that this integration of disciplines give.

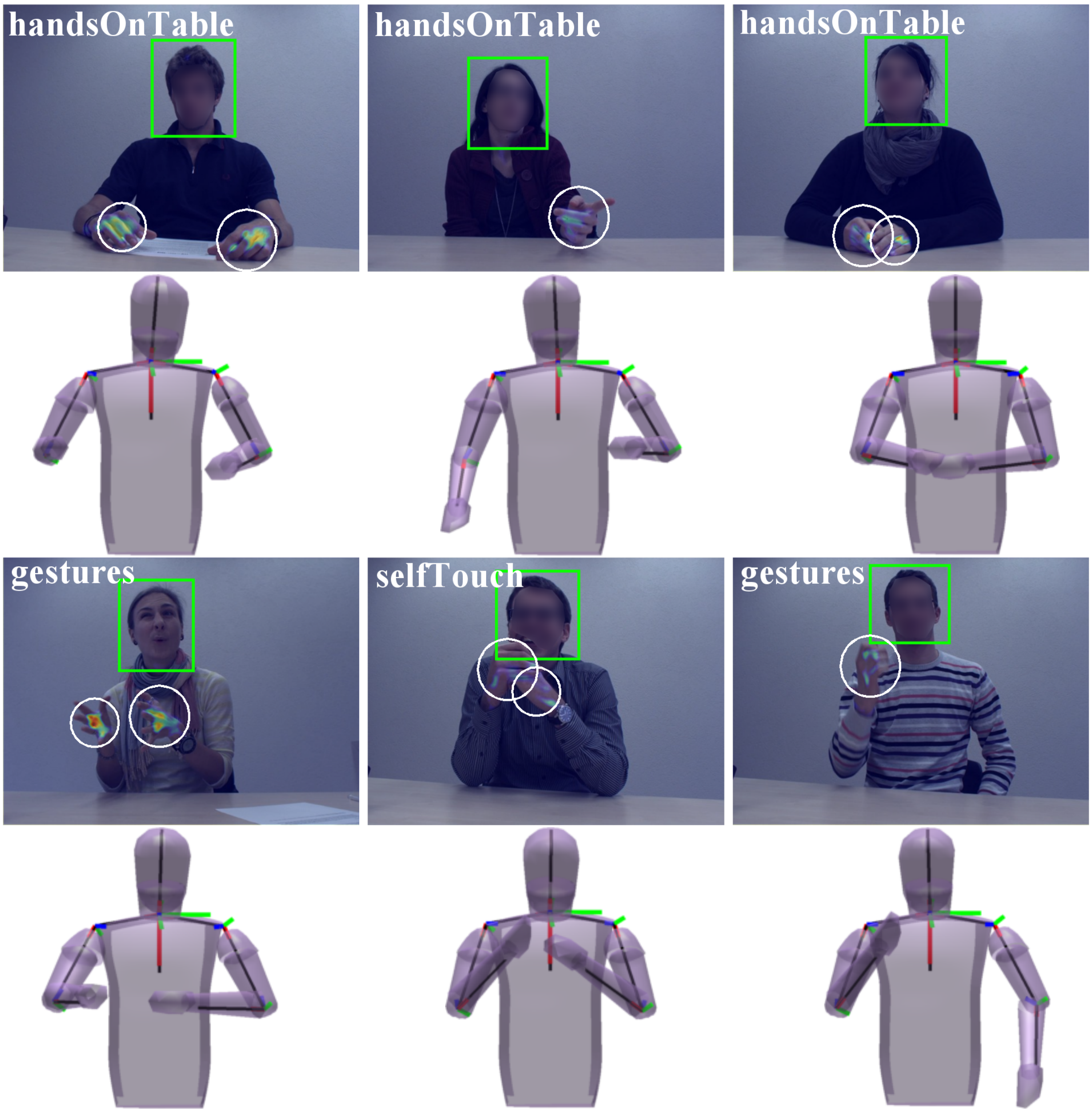


Figure 6.26: Frame results in the RGB case. Top row: input frame, with hand likelihood map, face detection, hand tracking and recognized action overlapped. Bottom row: output 3D torso pose.

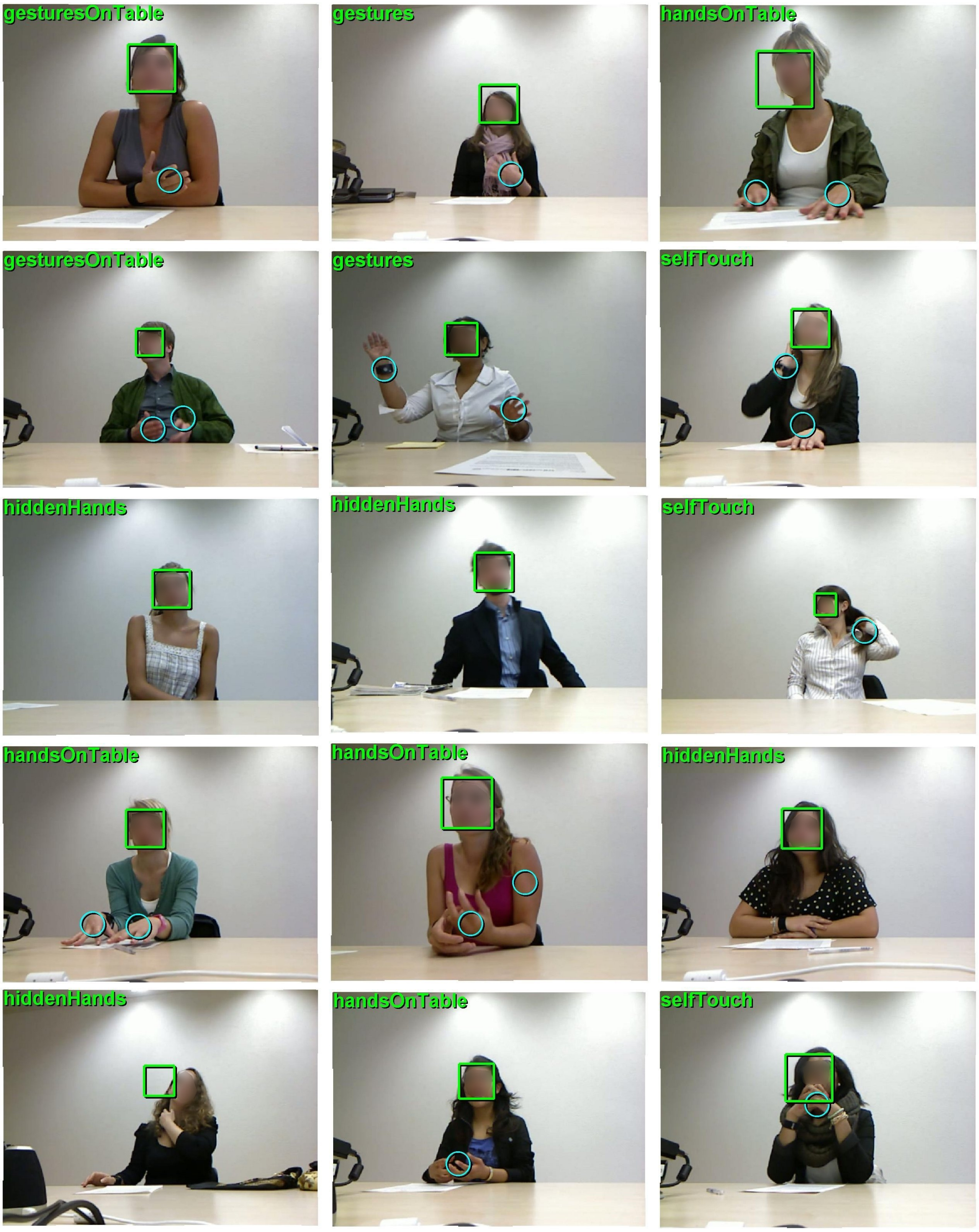


Figure 6.27: Frame results in the RGBD case. Overlaid to RGB images are the hand and face position, and the detected ongoing activity. The last five RGB images row shows failure examples (first: positive from the hand detector, second: missed hand detection; third: incorrect face detection; fourth and fifth: merged hands). Best viewed in color.

Chapter 7

Upper body motion capture from appearance- and scale-invariant features

7.1 Introduction

In Chapter 6 we have proposed a tracking system based on hand saliency, assuming that the whole video sequence is available from the beginning. In this chapter however, we propose an online, real-time capable system that is highly independent from the color appearance of the subjects and their relative scale in the image. The material in this chapter has been submitted for publication to an international conference.

As seen in Chapter 3, markerless motion capture from monocular images is a good solution to the nonverbal communication retrieval problem (as markers placed in the body can alter the behavior), but a challenge in computer vision. This is a consequence of many factors, such as the high-dimensionality of the data, camera projection distortions, appearance variability (*e.g.* clothing, skin, hair...) or external and self-occlusions. In the last few years range cameras have appeared in the mass market, obtaining depth without the need of motion or rich texture. Other difficulties remain, however, such as clothing, the many different contexts in which the same body part can appear, and the high dimensionality of the articulated motion. Recently, [Shotton et al., 2011] presented a solution to human pose estimation based on machine learning: a classifier was trained with a very large database from simple offset features capturing depth differences between near pixels. Invariance to clothing, body types, and the appearance of body parts was therefore learned. This resulted in a very robust solution with previously unseen performance levels.

However, in psychological studies there is still a need for addressing the problem in RGB images: most of the historical footage and even newer studies [Frank et al., 2012][Mihalcea and Burzo, 2012][Yu et al., 2013b] use traditional video, as it tends to be a discipline in which technological changes take time to be widely adopted. In RGB images, the approach of [Shotton et al., 2011] is not directly transferable: simple color differences as features would depend on the background and person appearance. In motion capture several techniques have been proposed in order to get invariant features from RGB images. HOG-based Body Part Detectors (BPDs) in particular are able to obtain a rough estimate of the pose, later refined with a number of different solutions such as global coherence [Yang and Ramanan, 2011], symmetry analysis [Ramakrishna et al., 2013] or iterative foreground registration [Wang and Koller, 2011].

However, the output of BPDs is often very noisy, with several parts interfering with each other. BPDs are also sensitive to changes in the background.

The main contributions presented in this chapter are therefore:

- Propose a motion capture system with a high degree of appearance and scale invariance while using only an uncalibrated RGB camera. To this end, we analyze image motion through dense optical flow. Our approach needs a single approximate human body detection in the form of a bounding box [Ferrari et al., 2009][Dollar et al., 2012], thus avoiding the clutter of using many part detectors. To overcome the lack of information when there is no motion, we integrate a Kanade-Lucas-Tomasi (KLT) tracker [Tomasi and Kanade, 1991].
- We evaluate our method with existing databases and a subset of the real job interview data corpus, which we make public. In our setting, the performance is comparable to that of [Shotton et al., 2011] without using a range camera, and state-of-the-art in HumanEva [Sigal et al., 2010] and ChaLearn 2011 [cha, 2011]. In Figure 7.1 an overview of the different processing stages can be seen.

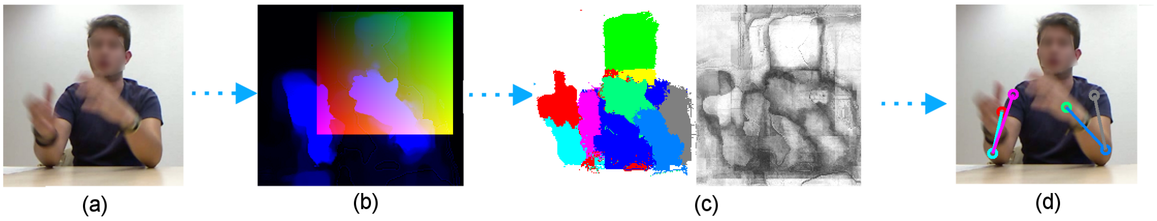


Figure 7.1: Data flow overview. (a) Original image. (b) Proposed 4-C image. (c) Body part classification and confidence scores. (d) Obtained pose.

The rest of the chapter is organized as follows: in Section 7.2 we overview the proposed method, in Section 7.3 we introduce the used observation system and its advantages in Section 7.4 we explain the tracking approach, in Section 7.5 the performance of our proposal is discussed, and finally in Section 7.6 we present some conclusions.

7.2 Overview

An overview of our method can be seen in Figure 7.2. Given two input consecutive frames $\mathbf{I}_{t-1}$ and $\mathbf{I}_t$, we compute the dense optical flow $\mathbf{I}_{OF}$ [Chambolle and Pock, 2011] and detect the subject torso bounding box $\vec{b}$:

$$\vec{b} = [\vec{u_o}, b_w, b_h] \tag{7.1}$$

where $\vec{u_o}$ is the top left pixel of the bounding box, b_w is its width and b_h is its height. We then define a 4-C image as:

$$\mathbf{I}_C = [\mathbf{I}_{OF}, \mathbf{I}_{bw}, \mathbf{I}_{bh}] \tag{7.2}$$

where $\mathbf{I}_{bw}$ and $\mathbf{I}_{bh}$ are images derived from the torso detection, that aim to give spatial context (see Section 7.3.1 for details). We then extract per-pixel offset features $\vec{u_\delta}$ in $\mathbf{I}_C$ from a training set (which constitutes the only prior used in this chapter) in a similar fashion to [Shotton et al., 2011], in order to predict the body part classification label image $\mathbf{I}_L$, a Random Forest classifier is used. The label image

$\mathbf{I}_L$ and its associated confidence scores $\mathbf{I}_L^{'}$ are used to train a Random Forest regressor that outputs the final body configuration. In Figure 7.2 the proposed work-flow can be seen.

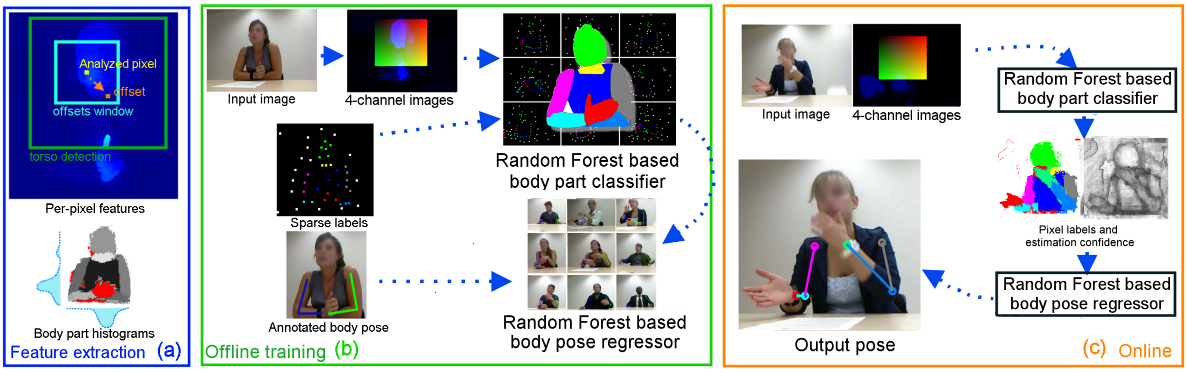


Figure 7.2: Pipeline of our proposal: (a) Image feature extraction, from the original RGB data and set of body part labels. (b) Offline training of the body part classifier and pose regressor, from sparsely labeled data. (c) Online usage: once the 4-C features have been computed, they are fed into the classifier. In turn, its output is used as input for the regressor, which estimates the body joints configuration. Best viewed in color.

7.3 Observation system

7.3.1 Largely appearance-invariant image features

As explained in Section 7.2, we aim to extract a series of features from the image that encode as much information as possible while maintaining a high appearance-invariance: they should be robust to clothing and skin color. Thanks to recent advances, dense optical flow and upper body detectors are good candidates. Therefore, we compose a 4-C image that merges the information that those low-level features provide. These channels are:

- Optical flow modulus $\mathbf{I}_{OF,\rho}$: as proven in Chapter 6, the assumption that in general hands are the quickest part of the image in our conversational settings is valid. Also, depth and motion in an image are closely related [Yoonessi and Baker, 2011]. In addition, we look at how other parts of the body move, also in relation to the hands. Therefore, the optical flow modulus is a strong cue for positioning the body pose.

- Optical flow orientation $\mathbf{I}_{OF,\phi}$: we hypothesize that in certain situations, such as when the hands move close to each other, the optical flow orientation is a useful cue in order to differentiate both.

- Torso vertical context $\mathbf{I}_{bh}$: We use the torso detector output to place and estimate the size of the person in the image, adding contextual information. Given the torso bounding box $\vec{b}$ (see equation 7.1), we encode the position relative to height, resulting in the image $\mathbf{I}_{bh}$ (see Figure 7.3). In $\mathbf{I}_{bh}$ pixels range from 0 to 1, from the bottom to the top within the bounding box, and are set to -1 outside the bounding box.

- Torso horizontal context $\mathbf{I}_{bw}$: Analogous to $\mathbf{I}_{bh}$, but providing horizontal information instead.

The largely appearance-invariant 4-channel image is then formed simply integrating the aforementioned channels, as equation 7.2 shows. Figure 7.4 shows an example.

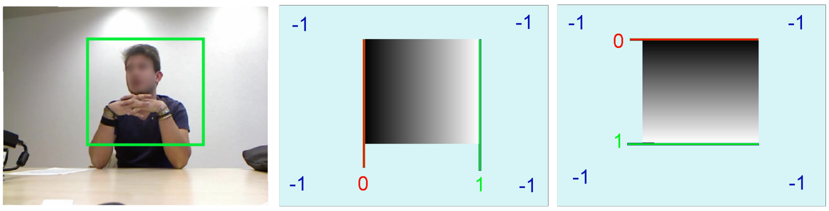


Figure 7.3: Torso context composition. Left: Input image and torso detection. Center: proposed horizontal context $\mathbf{I}_{bw}$. Right: proposed vertical context $\mathbf{I}_{bh}$.

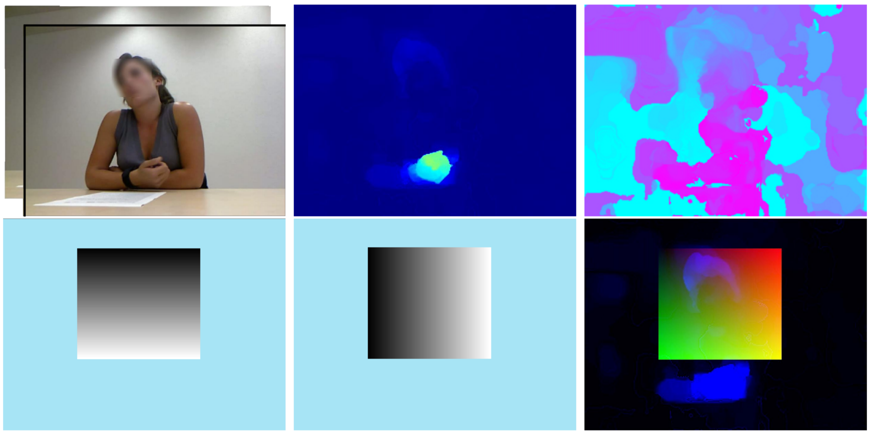


Figure 7.4: Example 4-C image. From the top to bottom and left to right: input image pair $\mathbf{I}_{t-1}$ and $\mathbf{I}_t$, optical flow modulus $\mathbf{I}_{OF,\rho}$, optical flow orientation $\mathbf{I}_{OF,\phi}$, torso vertical context $\mathbf{I}_{bh}$, torso horizontal context $\mathbf{I}_{bw}$ and final 4-channel image $\mathbf{I}_C$.

7.3.2 Body part classification

At this point, highly appearance-invariant images $\mathbf{I}_C$ are obtained. Similar to [Shotton et al., 2011], we use an offset sampling idea and a Random Forest classifier in order to associate every pixel with a body part label. A set $\mathcal{U}$ of pixel offset features is built:

$$\mathcal{U} = \{\vec{u_{\delta i}}\}_{i=1}^{n_\delta} = \{(u_{\delta i}, v_{\delta i})\}_{i=1}^{n_\delta} \tag{7.3}$$

For a given pixel $\vec{u}$, the feature response is computed with feature parameters $\vec{u_{\delta i}}$ that describe a number of n_δ 2D pixel offsets $(u_{\delta i}, v_{\delta i})$. In [Shotton et al., 2011], features are normalized with the distance to the camera in order to make them depth-invariant. In our case however it is not possible, as we use RGB only images. As a proxy for size of the person, we use the height of the torso bounding box b_h. We then extract the per-pixel features from each channel in a different way.

Optical flow modulus: both the the offset distance and optical flow value are normalized with b_h, since motions that take place far from the camera result in a lower optical flow modulus value. Let $L(\vec{u}, \vec{u_{\delta i}}, \mathbf{I}^1)$ be a lookup function that returns the feature associated with pixel $\vec{u}$, given a single-channel image $\mathbf{I}^1$ and an offset $\vec{u_{\delta i}}$ (see Figure 7.2a). An optical flow modulus feature becomes:

$$f_{OF,\rho}(\vec{u}|\vec{u_{\delta i}}) = L(\vec{u}, \frac{\vec{u_{\delta i}}}{b_h}, \mathbf{I}_{OF,\rho})\frac{1}{b_h} \tag{7.4}$$

It encodes the speed difference between pixel $\vec{u}$ and its associated offset $\vec{u_{\delta i}}$, after normalizing the offset distance and image speed with the torso size b_h.

Optical flow direction: in order to better differentiate parts of the body moving with similar speed modulus, but with different orientations, we also take the optical flow direction into account. For each pixel, we compute the direction similarity in relation to its offset features:

$$f_{OF,\phi}(\vec{u}|\vec{u_{\delta i}}) = L(\vec{u}, \frac{\vec{u_{\delta i}}}{b_h}, \mathbf{I}_{OF,\phi}) - \mathbf{I}_{OF,\phi}(\vec{u}) \tag{7.5}$$

Therefore, given the optical flow angle for pixel $\vec{u}$, the angle difference respective to the offset is computed. The discontinuity between 0 and 360 degrees is addressed so that the angle difference is less or equal than 180 degrees.

Position relative to the torso: the feature is obtained in an identical way to that of the optical flow direction:

$$f_{bh}(\vec{u}|\vec{u_{\delta i}}) = L(\vec{u}, \frac{\vec{u_{\delta i}}}{b_h}, \mathbf{I}_{bh}) - \mathbf{I}_{bh}(\vec{u}) \tag{7.6}$$

$$f_{bw}(\vec{u}|\vec{u_{\delta i}}) = L(\vec{u}, \frac{\vec{u_{\delta i}}}{b_h}, \mathbf{I}_{bw}) - \mathbf{I}_{bw}(\vec{u}) \tag{7.7}$$

where equations 7.6 and 7.7 correspond to the vertical and horizontal context, respectively. With this approximation, each feature gives an idea of where the pixel is positioned in relation to the main portion of the torso.

The feature vector for classification $\mathcal{X}_{tr,class}$ is formed by concatenating all the features, providing information about the speed, direction, and position relative to the torso of every pixel of the image, hence forming a rich representation of human motion. We demonstrate its capabilities in the results section.

7.3.2.1 Training the forest for classification

As described in Section 4.3.3.3, a subset of data from the real job interview dataset is annotated in order to serve as a training set (see Section 7.5). For classification, annotations consist in a number of manually-labeled pixels in the image, in which the nature of each label corresponds to a given body part.

We then compute the previously introduced per-pixel offset features for each labeled pixel in the training subset. As our method depends on the amount of movement in the image, we discard the pixels with low flow modulus. We then train a Random Forest with the extracted features and the associated body part labels. Given an unseen 4-C image, the classifier outputs the per-pixel predicted body part and an associated confidence score (see Figure 7.5).

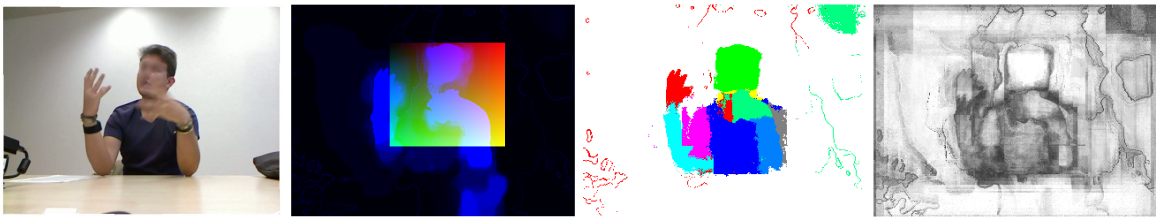


Figure 7.5: Example of a 4-C image classification. From left to right: input image $\mathbf{I}_t$, 4-channel image $\mathbf{I}_C$, label image $\mathbf{I}_L$ and its associated confidence scores $\mathbf{I}_L^{'}$ (darker zones mean less confidence). Best viewed in color.

7.3.3 Body pose regression

The next step is to obtain the final body configuration $\mathcal{P}_o$ (defined as the pixel position in the image for every joint) from the output of the body part classifier:

$$\mathcal{P}_o = \Omega(\mathbf{I}_L, \mathbf{I}_L^{'}, \beta) \tag{7.8}$$

where Ω is the regression model with β parameters. Therefore, we use a Random Forest regressor that takes information extracted from an image of densely classified body parts $\mathbf{I}_L$ and classification scores $\mathbf{I}_L^{'}$, resulting in the regression training set $\mathcal{X}_{tr,reg}$, through a process described below.

7.3.3.1 Obtaining images $\mathbf{I}_L$ and $\mathbf{I}_L^{'}$ in order to train the regressor

As shown in Figure 4.5 from Chapter 4, in order to train the classifier sparse labels were used. This responds to two reasons:

- Reduce annotation time: it can be reduced to less than half of the time required for dense labeling.
- Force the classifier to generalize: as during training only a few (∼100) pixels per image are labeled, the trained forest is shown later the same training images, obtaining the classification output for every pixel in the image. In this process, the provided classification scores $\mathbf{I}_L^{'}$ picture more realistically the confidence that the classifier will have in unseen images. This property is important when training the body pose regressor, as it is forced to learn from the mistakes that the body part classifier makes.

7.3.3.2 Body part histograms from $\mathbf{I}_L$ and $\mathbf{I}_L^{'}$

In order to capture more explicitly the characteristics of the predicted body parts placement in the image, we propose the use of vertical and horizontal per-class histograms of $\mathbf{I}_L$ and $\mathbf{I}_L^{'}$. The concept is shown in Figure 7.6.

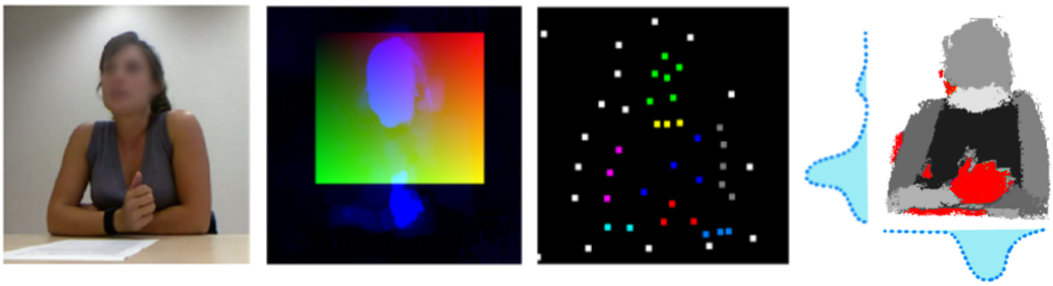


Figure 7.6: Example of body part histogram. From left to right: input image $\mathbf{I}_t$, 4-C image $\mathbf{I}_C$, sparse annotations, body part histograms for the right hand after the classifier was forced to generalize. Best viewed in color.

We build three sets of histograms. The first one, $\mathcal{M}_l$, measures the frequency in which each body part appears in the vertical and horizontal axis of the image, and it is defined as:

$$\mathcal{M}_l = [\vec{m_{l,v}}, \vec{m_{l,h}}] \tag{7.9}$$

The second one, $\mathcal{M}_l^{'}$ adds the scores $\mathbf{I}_L^{'}$ in the pixels of the image that belong to the relevant body part, along the vertical and horizontal axis of the image, and it is defined as:

$$\mathcal{M}_l^{'} = [\vec{m_{l,v}^{'}}, \vec{m_{l,h}^{'}}] \tag{7.10}$$

The third one, $\mathcal{M}_l^{''}$, is the product of the previous two:

$$\mathcal{M}_l^{''} = [\vec{m_{l,v}}\vec{m'_{l,v}}, \vec{m_{l,h}}\vec{m'_{l,h}}] \tag{7.11}$$

This results in the body part frequency histogram $\mathcal{M}_l$ to be weighted with its associated confidence scores $\mathcal{M}_l^{'}$. The effect can be seen in Figure 7.7: the resulting histogram $\mathcal{M}_l^{''}$ shows better where the most confident predictions are located in the image, for a given body part.

In order to reduce the dimensionality of the histograms, the images $\mathbf{I}_L$ and $\mathbf{I}_L^{'}$ are down-sampled to a resolution of 128x96 pixels. Therefore, a given vertical histogram becomes 96-dimensional, and a horizontal histogram becomes 128-dimensional. Since there are 10 different body parts classes, 10 different set of histograms $\mathcal{M}_l^{''}$ are obtained (one for each body part). This results in a 2240-dimensional feature vector $\mathcal{X}_{tr,reg}$ for regression:

$$\mathcal{X}_{tr,reg} = \{\mathcal{M}_{l,i}^{''}\}_{i=1}^{n_j} \tag{7.12}$$

where n_j is the number of body parts (classes). For each feature vector $\mathcal{X}_{tr,reg}$ there is an associated annotated body pose, consisting in the pixel position of 6 joints: shoulders, elbows and wrists. It is therefore 12-dimensional. Both the feature vector and the labels are used to train the regression model Ω.

As a summary, in order to obtain the body pose from a unseen pair of $\mathbf{I}_t$, $\mathbf{I}_{t-1}$ images, the associated 4-C image $\mathbf{I}_C$ is first composed, and then input into the body part classifier. The resulting densely-labeled image and confidence scores ($\mathbf{I}_L^{'}$ and $\mathbf{I}_L$) are fed into the body pose regressor, which outputs the predicted final pose configuration $\mathcal{P}_o$ as pixel positions of every joint.

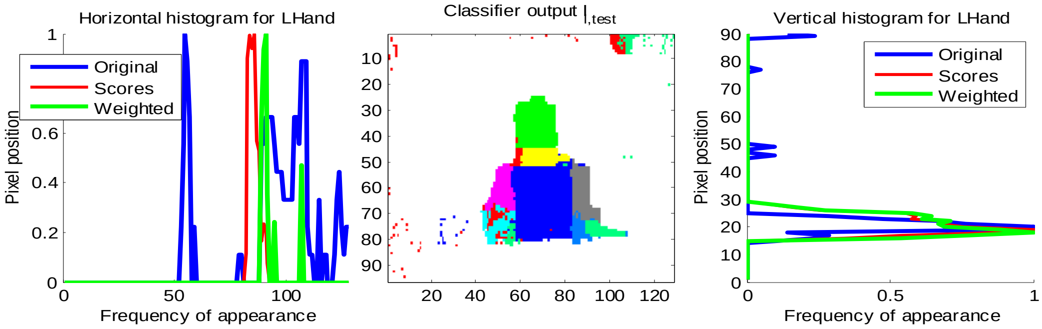


Figure 7.7: Body part histogram example. Left and right: $\vec{m_{l,v}}$ and $\vec{m_{l,h}}$ for the left hand. Center: classifier output. The confidence weighting helps to maintain the detection peak in the right position. Best viewed in color.

7.4 Tracking

In order to reliably obtain the body pose in every situation, we propose a tracking method where hands are tracked using the KLT framework when the body part classifier is not reliable. We then impose temporal smoothness to avoid sudden changes in the pose.

7.4.1 Tracking the hands

The main drawback of our proposal is the need for movement in the image, as the classifier needs optical flow measurements in order to classify different body parts. We extend the pose retrieval framework by adding a KLT tracker, based on image features (see Section 2.5.1). It therefore works better with slow motions, as motion blur and quick appearance changes become a problem when obtaining good features to track. On the other hand, it has shown to be very reliable when small motions are present. Since our pose detection method is most confident when movement is present, the KLT functions in a natural, complementary way with the pose regressor.

Taking this into account, a detection followed by tracking framework is proposed to obtain the body pose along a whole video, as shown in Figure 7.8. As finding the hands' position removes a high degree of uncertainty in the pose [Marinoiu et al., 2013], we employ this approach only in them.

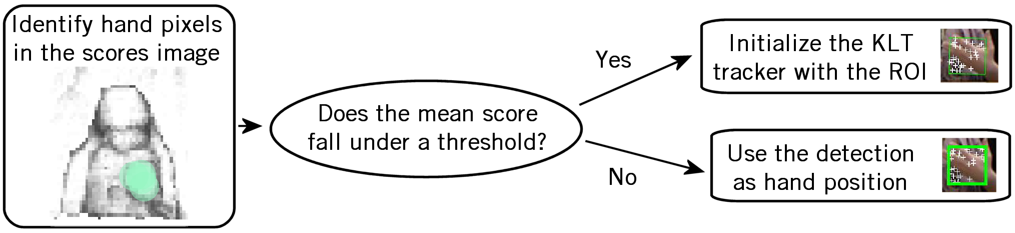

Identify hand pixels in the scores image
Does the mean score fall under a threshold?
Yes
Initialize the KLT tracker with the ROI
No
Use the detection as hand position

Figure 7.8: Explanation of the KLT tracking method. During the tracking process, the left and right hands are considered separately. Best viewed in color.

When the scores image $\mathbf{I}_L^{'}$ falls under a manually set threshold for a hand region, the detection is deemed unreliable, and the last reliable detection is used to initialize a KLT tracker, that takes control of the hand position until a new reliable detection is found, usually when the hand starts to move again. An example can be seen in Figure 7.9. Since both hands are considered separately, two KLT trackers are used, one for each hand.

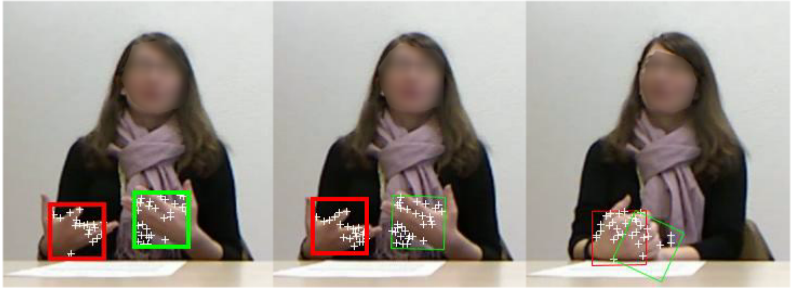


Figure 7.9: Integration with the KLT tracker example. Reliable detections are represented with a thick region of interest box. The good features to track are shown as white crosses. Is moments in which the detections are unreliable, such as when there is no movement in the image (right), the tracker keeps the correct location of each hand. Best viewed in color.

7.4.2 Stabilizing the body pose

At this point, a body pose configuration $\mathcal{P}_o$ is available. Given that the ultimate goal of our method is to serve as information to later analyze nonverbal communication, a further post-processing step is needed. When a subject stands still, ideally the output pose would remain perfectly still too. However, as the system involves a high degree of tracking by detection, $\mathcal{P}_o$ slightly fluctuates along time.

In order to compensate this effect, a simple yet effective approach is followed (see Figure 7.10). For each body joint, a small pixel radius κ_m is defined. If the predicted body joint falls under the radius longer than a small pre-defined time t_m, then the joint is defined as the center of the static circle that is configured with κ_m. The timing parameter t_m is necessary in order to avoid the discretization of continuous movements, as they would evolve in κ_m steps otherwise.

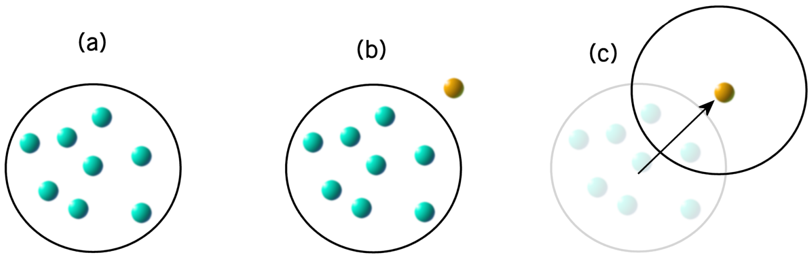


Figure 7.10: Explanation of the joints stabilization method. If the measurements (represented as spheres) fall within a given distance, the joint position is maintained (a). If a new measurement falls a given threshold (b), it is moved to its position (c). Best viewed in color.

7.5 Results

We evaluate the effects of different parameters in our system's performance, and compare it with the current state-of-the-art. For the one-shot-detection part of the approach, we define two experiments: one for classification and one for regression. For classification, we compare the output of our method to annotations of every body part. The result is given in per-pixel accuracy for each class. For regression, we compare the output pose that we get with manual annotations of the database we used, measuring joint detection rate. A joint is defined as detected if the predicted point falls within a given distance threshold.

7.5.1 Classification

The performance is evaluated in our job interview database. We train both our algorithm and [Shotton et al., 2011] with the same number of images. As our method depends on the amount of movement in the image, we discard the pixels that fall below a motion threshold. This results in fewer data points, albeit reliable ones. In total, more than 110k pixels were used for depth training, and 88k for optical flow training. The same classifier parameters were used in both cases: 75 trees, an offset window of 250 pixels and 700 features. We generate 320x240 classified images in order to have a reasonable resolution while maintaining a competitive processing time. The results can be seen in Figure 7.11 for the different cases.

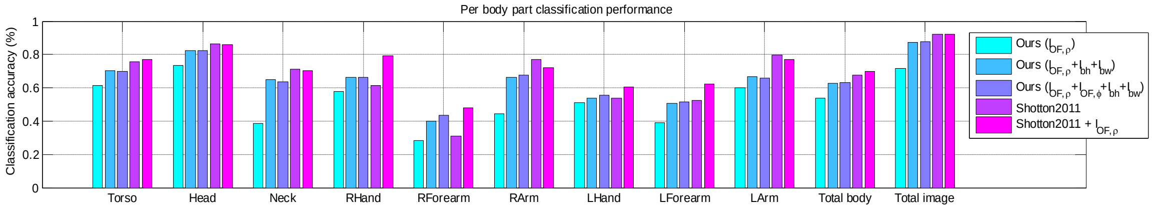


Figure 7.11: Results of our method versus [Shotton et al., 2011], when trained with the same number of images: Reported as classification rates for every body segment. Best viewed in color.

Using depth, a **67.7%** accuracy is achieved for the pixels associated with the person, which is consistent with the previous findings of [Shotton et al., 2011], when taking into account the much lower training information. Over the whole image, the accuracy is **92.3%**. Using optical flow and torso detections, the accuracy for the person is **63%**, and **87.7%** for the whole image. This shows that our method can achieve similar results to depth when there is body movement.

An interesting finding is that our method outperforms [Shotton et al., 2011] in hand detection. As the reviewed literature shows, hand position is a very good proxy in order to infer the rest of the body pose. The usefulness of optical flow in hand and arm detection is confirmed, as when combining depth and optical flow modulus, the body accuracy increases to **70%**, with the whole image at **92.2%**.

Sensitivity to parameters results can be seen in Figure 7.12. When considering offset window size before scale normalization, it is found that a size of 200x200 improves the body-related detections, but introduces more background noise than 250x250 sizes, making the regression system more prone to errors. Surprisingly, using a very low number of offset features did not cause big performance drops. Already with 150 features, very competitive results are obtained, as a result of the rich contextual information that the torso detector provides. As for the number of trees used, it is found that after 5 trees, there is only a slight performance increase to be found. Finally, when considering the amount of training data, competitive results can already be obtained with 300 training images (effectively 600, since they are mirrored). This highlights the generalization capabilities of the proposed features.

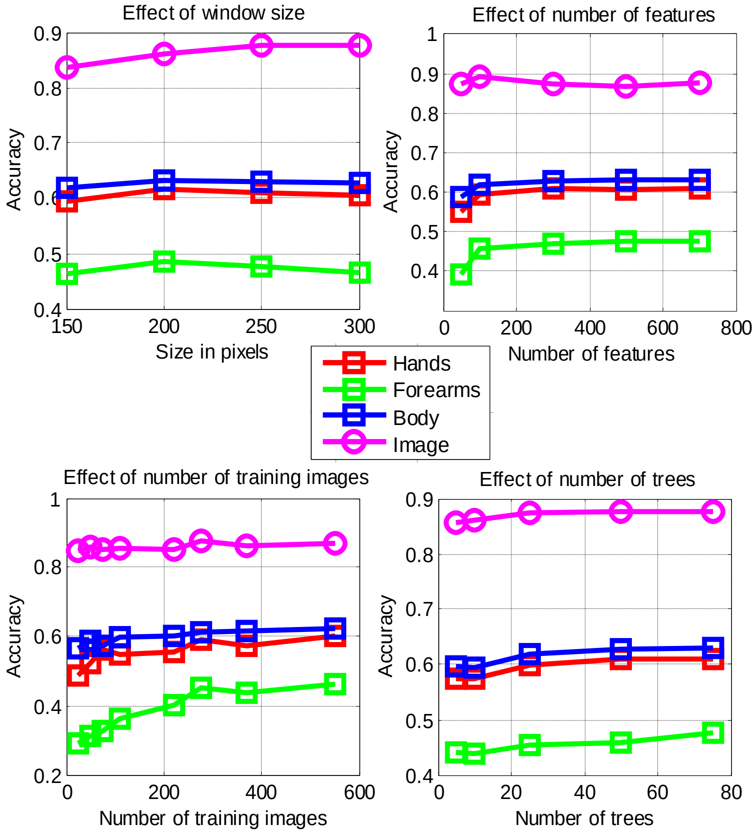


Figure 7.12: Sensitivity of the body part classifier to different parameters. Best viewed in color

7.5.2 Regression

Performance is measured with two different datasets: our job interview dataset, and ChaLearn 2011, which is non-conversational, but allows to show generalization of performance. Rates of correct detected parts can be found in Figures 7.12 and 7.14, respectively. We found 40 pixels to be the limit of a reliable guess while using 640x480 images.

Job interview dataset. Our method performs similarly to [Shotton et al., 2011], and outperforms [Ferrari et al., 2009] and [Sapp and Taskar, 2013]. The latter work had been trained with their FLIC database, and is considered as the best method in terms of performance to processing time ratio. Wrists are the hardest body part to detect, and our method achieves close to **20%** higher performance than

[Sapp and Taskar, 2013]. When compared against [Shotton et al., 2011], our method is less than **10%** behind. Using the classification weights scores, accuracy is improved by almost **5%** for the hands. In Figure 7.18 some qualitative results can be found.

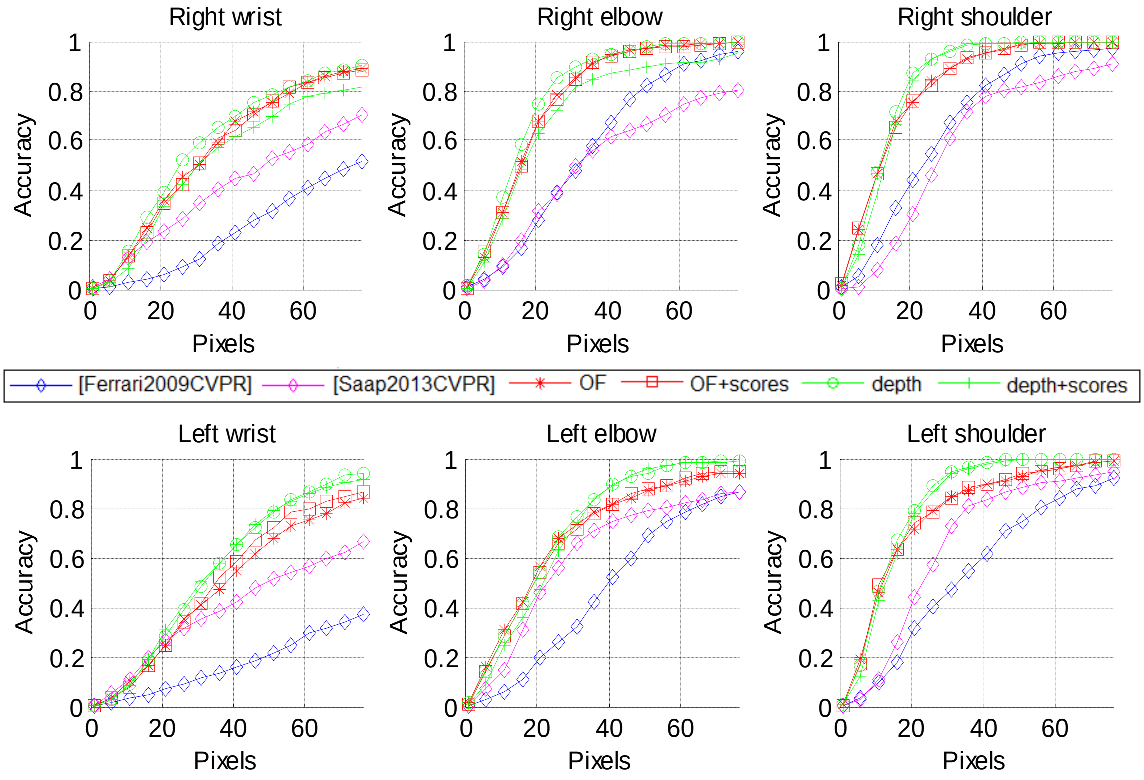


Figure 7.13: Accuracy of the body pose regressor, compared against [Ferrari et al., 2009] and [Sapp and Taskar, 2013] (methods are indicated in the central row). Results are reported as percentage of detections, parameterized with the pixel threshold within which a joint is considered as detected. Best viewed in color.

Finally, in Figure 7.18 some qualitative results can be found. While other methods struggle in situations motion blur, low image saliency zones, and the table edges, our method succeeds thanks to the use of largely appearance- and scale-invariant features.

ChaLearn 2011. We perform leave-one-out after dividing the data in 10 arbitrary groups. As it can be seen in Figure 7.14, the trends shown in our job interview dataset are reproduced when there is enough movement available. The gap to [Shotton et al., 2011] appears larger due to several factors:

- Less training data: only 57% of the total labeled points are used in our approach as we only use points that contain movement information in order to train the forest. This also gives an idea of the little movement information present in the dataset. As Figure 7.14 top right shows, in the left hand there is a higher amount of movement, greatly reducing the gap with [Shotton et al., 2011].
- Some subjects move out of frame (the head or one arm is not visible), making it a hurdle for the torso detection to be correctly placed, and reducing the accuracy of shoulders and elbows.
- In some cases the subject casts shadows in the background wall, producing spurious optical flow detections. Despite this, we found that in some instances the context provided by the torso detector is able to filter errors out. In any case, our method performs clearly better than the RGB-only baselines, and given the challenging conditions, remarkably close to [Shotton et al., 2011] when there is movement present. Also, it shows that our system performs well with different body scales.

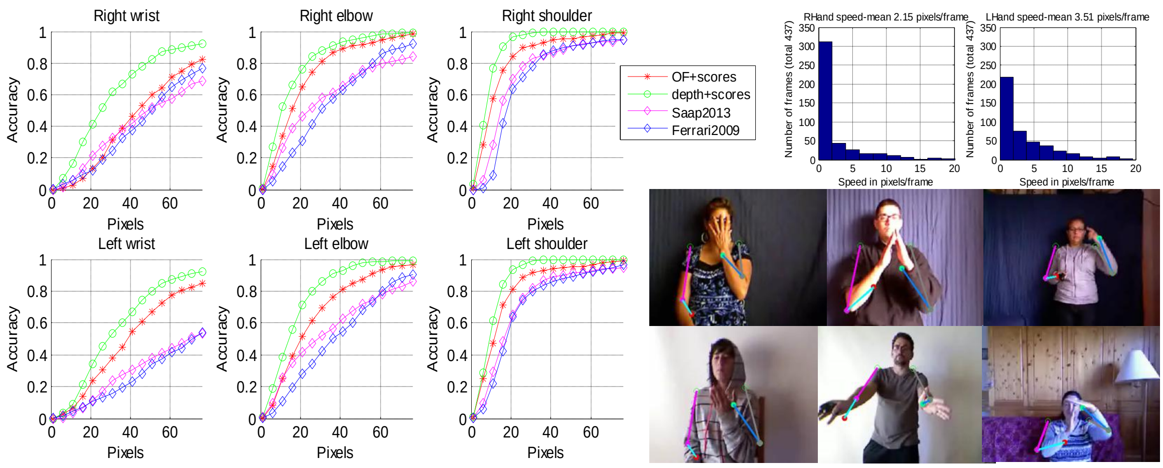


Figure 7.14: Left: Results in ChaLearn2011, compared against [Ferrari et al., 2009] (blue) and [Sapp and Taskar, 2013] (magenta). Top right: speed histograms for each hand. Bottom right: examples of obtained poses. Best viewed digitally in high zoom.

7.5.3 Tracking

In order to evaluate the tracking proposal, we define three experiments.

- **Experiment #1** evaluates hand positioning, as it is a very good proxy for the global pose [Marinoiu et al., 2013]. During two minutes of video from 4 subjects, the hand position was manually annotated (subset B of the job interview database, see Section 4.3.3.2).

- **Experiment #2** follows the same approach, but containing 30 seconds a very challenging subject (see column 3 of Figure 7.18 left), as she has no sleeves (therefore a lot of skin exposed, which can incur into appearance-dependence situations when using a skin segmentation scheme), moves her hair which is a similar color to that of the skin, and displays a series of unusual movements (such as the shoulders being closer to the camera than the hands at some points).

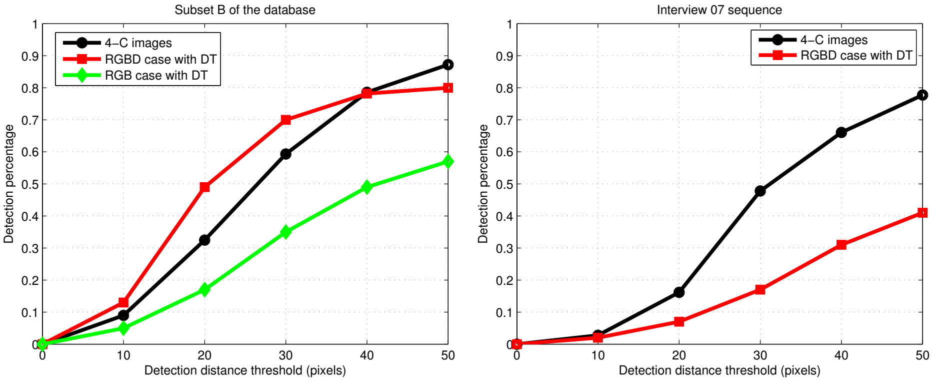
- **Experiment #3** uses the upper body joints of the 'gestures' and 'box' sequences of S1 and S3 from the public HumanEva I database, as they are the most relevant to our scenario. See Section 4.2.

The baselines for the experiments #1 and #2 are our hand tracking methods described in the previous chapter. Recently, in [Yin and Davis, 2013], a very similar hand saliency measure is defined, which shows that the baselines are state-of-the-art. We use our decision trees tracking method, denoted as DT.

The results of our tracking method applied for hand position can be seen in Figure 7.15. In experiment #1, the hand detection rate is **78.6%**, marginally higher than the baseline obtained with $\mathbf{I}_{H,t}$ (**78.2%**), even though no depth information is used in our approach. When compared to the performance of $\mathbf{I}'_{H,t}$, our framework shows a clear increase of performance.

In the challenging experiment #2, we found that our approach significantly increases the performance gap in respect to the $\mathbf{I}_{H,t}$ baseline, which uses depth, obtaining a detection rate of **66%**, compared to **31%**. This clearly shows that our method does not simply label body parts that move the fastest as hands, but rather takes the actual shape of the body part movements into account in the learning process. It also carries the extra advantage of being able to explicitly distinguish between right and left hands.

As Figure 7.15 shows, we obtain state-of-the-art performance in HumanEva I (in [Morariu et al., 2013] an average of 12.6 pixel error is reported). The largest errors occur when the performed pose substantially differs to those contained in the training set. Given the very few data samples that HumanEva I makes available for training (we use an average of 46 sparsely labeled frames per sequence), and the fact that our method requires large training sets [Shotton et al., 2011], we find the performance very encouraging overall. Qualitative results can be found in Figure 7.18.



Subject	Action	Error (pixels)
S1	Gestures	5.0
S1	Box	13.2
S3	Gestures	11.3
S3	Box	21.3

Figure 7.15 & Table 7.1: Tracking results. Left graph: first experiment. Right graph, using a specially challenging sequence. Table: Average error in HumanEva. Best viewed in color.

7.5.4 Computing time

Computing time is key in psychological studies, given the large amount of data. In [Sapp and Taskar, 2013], there appears an analysis of the state-of-the-art performance versus computing time, with their work being the best placed. Using the code they provide, our database is computed on a laptop with an Intel i7 processor in an average of 5.18 secs per frame (standard deviation 0.16 secs).

In our case, assuming that pre-computed optical flow is available (it can be comfortably processed in real time with modern GPUs), the average processing time from input features to body pose is **1.59 secs** per frame (standard deviation 0.11 secs). The same hardware has been used, with no GPU nor particular optimizations, with 700 features and 75 trees (it took 1.5 hours to train the forest). Our per-pixel feature retrieval implemented in Matlab takes most of the running time. Since regression needs the output of the classifier with a resolution of 128x96 pixels in order to build the histograms, we obtain features for every fifth pixel of the 4-C composed image. If only classification results are needed, we obtain comparable processing times to [Sapp and Taskar, 2013] by using a resolution of 320x240 for I_l, recording a mean of 5.05 secs and (standard deviation 0.14 secs). As shown in the results section, these times can still be reduced as the system continues to perform well with a very low number of offset features used. This offset feature-based approach is implemented on an XBox 360 in [Shotton et al., 2011] at 200 fps. Finally, an average of an extra **0.027 secs** per frame is required in order to track the hands with the KLT approach. It therefore does not cause a significant impact on our total reported processing time.

Limitations: At the moment our method requires a static camera and static background, but optical flow based methods in the literature have shown to overcome that problem by tracking background features. As the torso bounding box misplacements are one of the main sources of error, our approach can highly benefit from an elaborated torso bounding box tracking technique.

7.6 Conclusion

We addressed the problem of body communication detection of people engaged in conversation by means of a fast, largely appearance-invariant method for upper body monocular motion capture, that integrates detection and tracking. Detection was achieved through optical flow and body detectors, providing a proxy for depth information, visual context, and scale. This information was used to classify body parts in the image with a Random Forests classifier. The classification output and per-pixel confidence was later used to build per body part image histograms, and were fed to a regressor in order to infer the body pose. The integration of a KLT tracker allowed to follow the body pose when there are no reliable detections, thus resulting in a complementary framework.

We evaluated our method with different datasets, showing very close performance to that of the best depth-based method, while using only monocular information. We also clearly outperform the state-of-the art in the ratio accuracy to processing time. Our method is therefore attractive to process existing video data in typical psychology lab studies, where depth data is not available. Our job interview database will be made public, providing a reliable benchmark for real-world performance.

This chapter concludes our work in markerless motion capture. In the next chapter, we show how the presented methods can be used in order to process higher level information in conversational interactions.

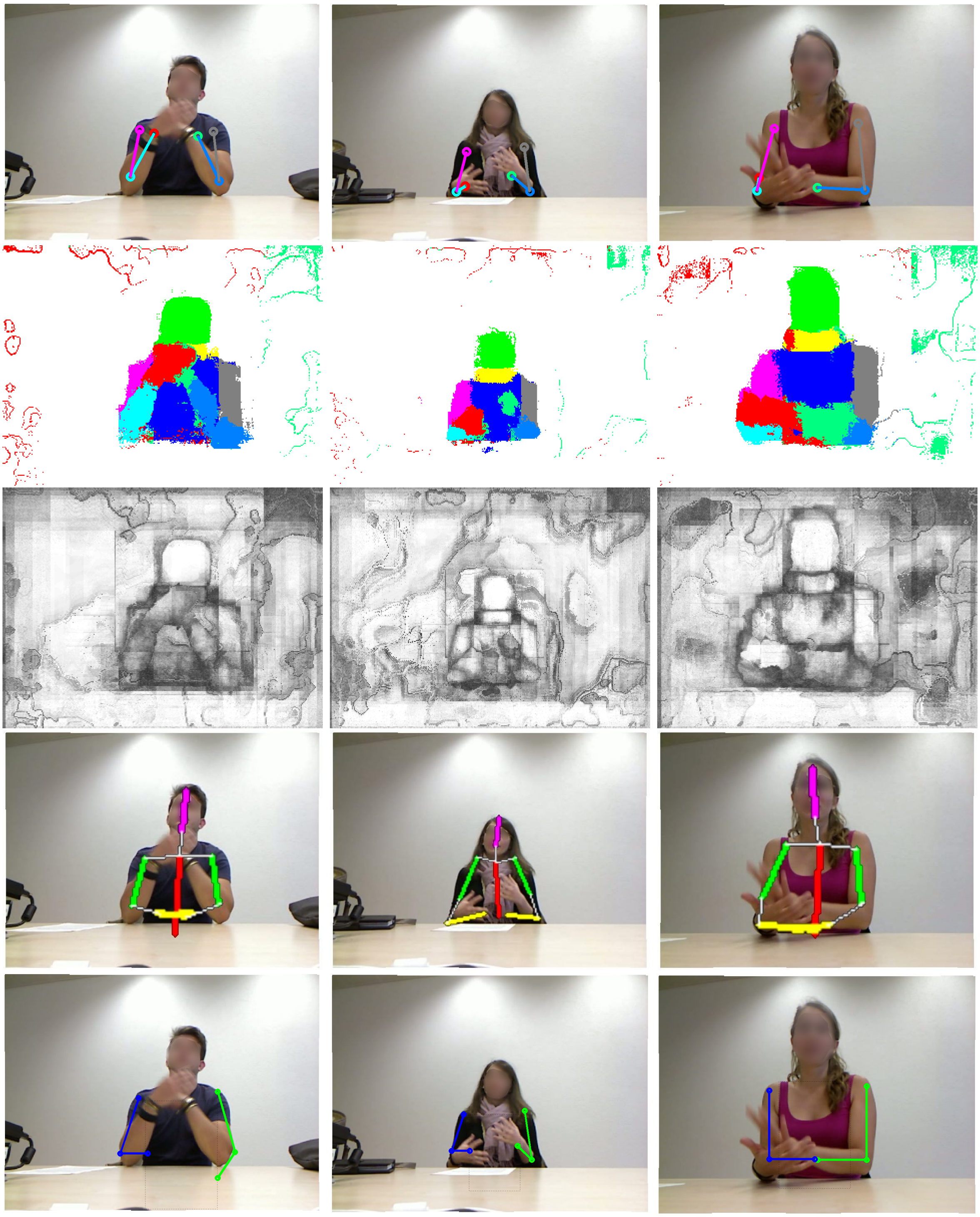


Figure 7.16: Qualitative results (1). First row: input RGB frames and the obtained pose with our method. Second and third rows: output labels and confidence scores (brighter means higher) of the body part classifier, respectively. Fourth row: output of [Ferrari et al., 2009]. Fifth row: output of [Sapp and Taskar, 2013].

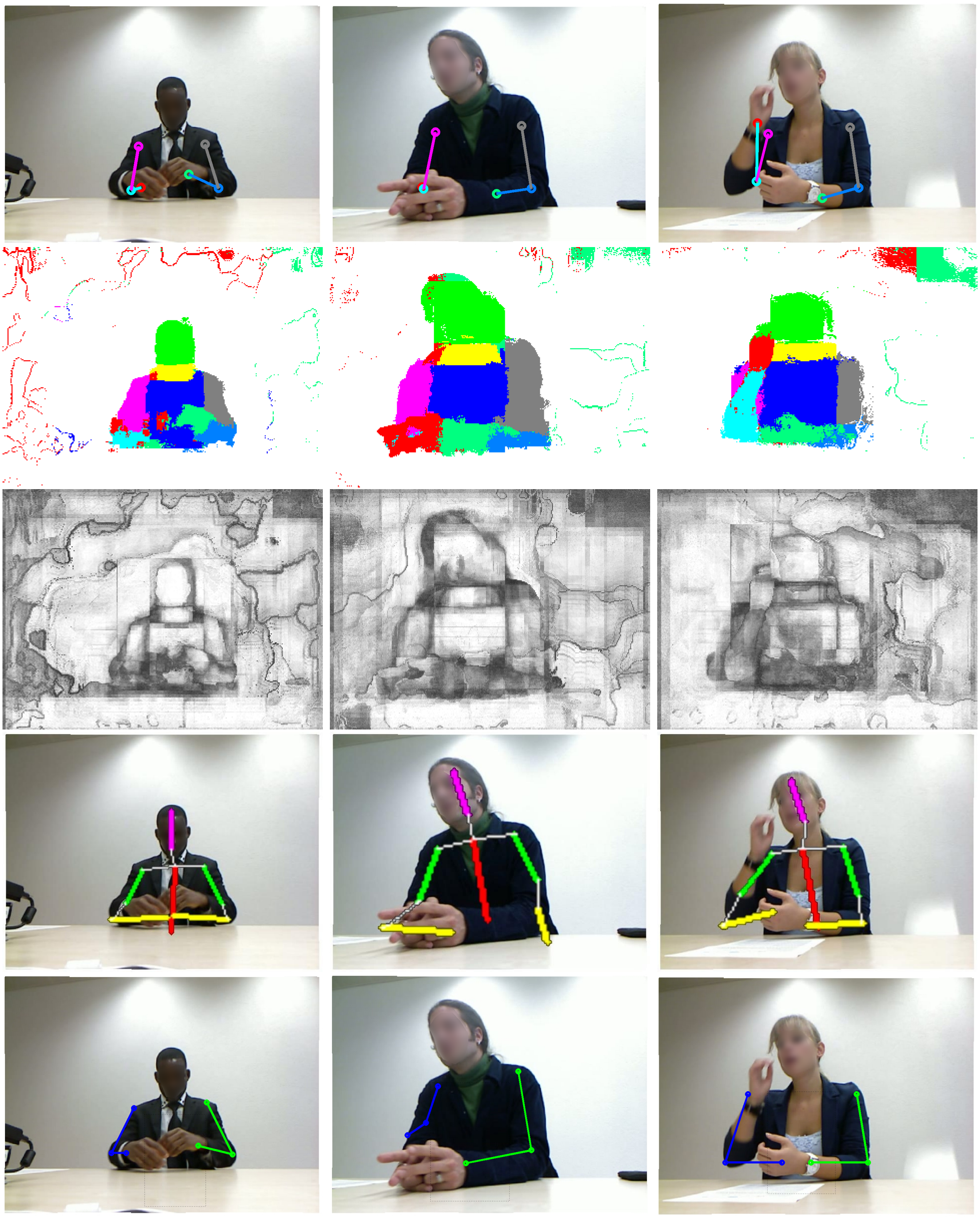


Figure 7.17: Qualitative results (2). First row: input RGB frames and the obtained pose with our method. Second and third rows: output labels and confidence scores (brighter means higher) of the body part classifier, respectively. Fourth row: output of [Ferrari et al., 2009]. Fifth row: output of [Sapp and Taskar, 2013].

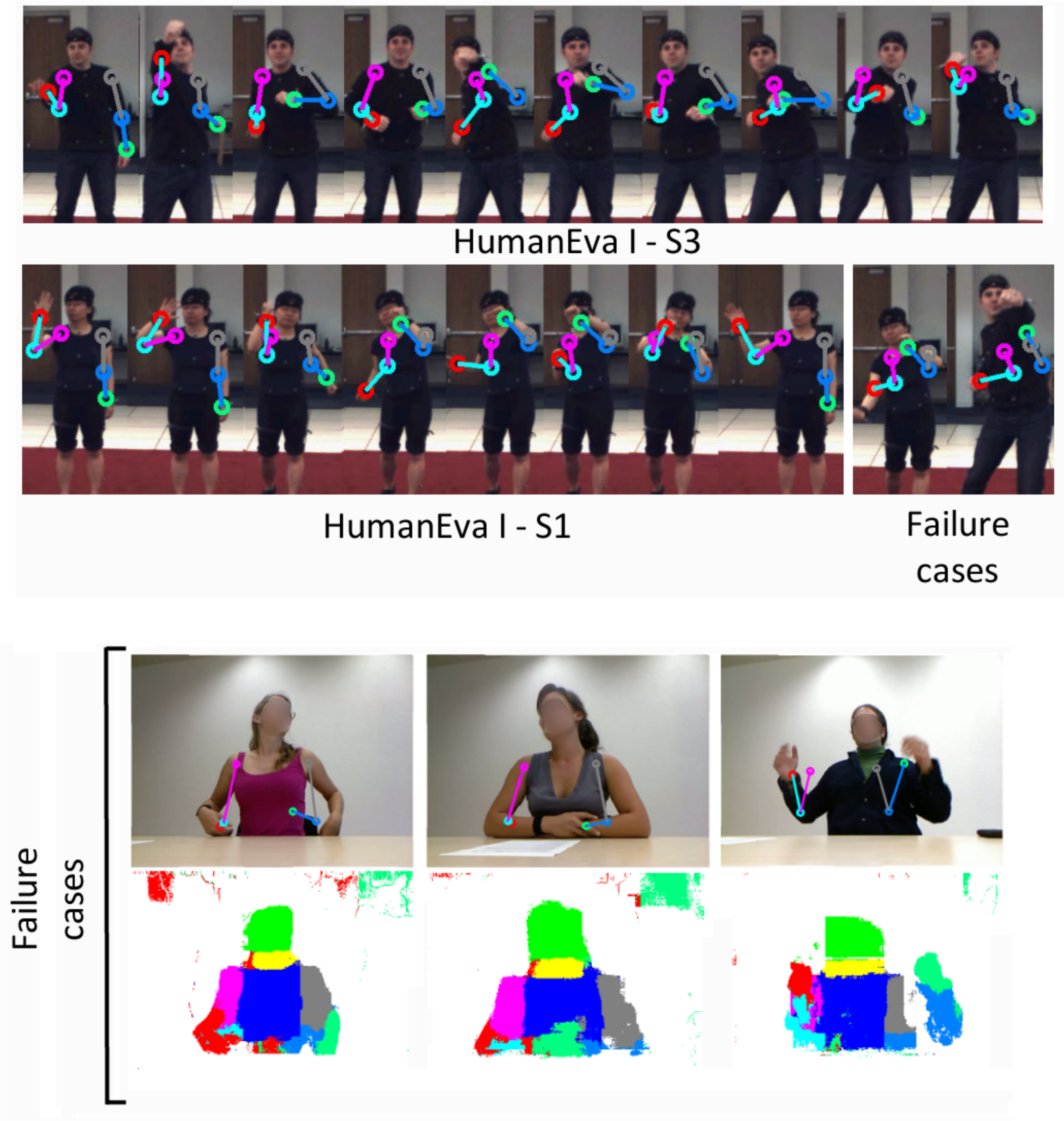


Figure 7.18: Qualitative results (3) and failure cases. Top: Tracking results with HumanEva. Bottom: Failure cases in the job interview dataset.

Chapter 8

Predicting social attributes

8.1 Introduction

In previous chapters we have presented the contributions made in this thesis in order to extract human body pose in different settings and sensing scenarios, and its usefulness for processing higher-level information in conversational set-ups. One of the goals of the thesis is to build features that are useful to later predict a series of social constructs and job performance measures. In this chapter we present our first findings in the matter. See Figure 8.1 for a quick overview of the process.

Two research questions are addressed: First, we investigate whether job hirability impressions and self-rated personality can be predicted using body communication cues; second, we examine whether the knowledge of the speaking status of an individual can be used to improve the prediction of personality and hirability.

To answer these research questions, several tasks were defined:

- First, we used the dataset of real job interviews, including the self-reported personality scores from questionnaire data and expert-rated hirability impressions (see Section 4.3.3.4).
- Then, we extracted a rich mixture of body cues from both manual annotations and automated extraction methods (see Section 8.2).
- Last, we evaluated the predictive validity of the extracted nonverbal cues with respect to hirability impressions and self-rated personality using a regression task (Section 8.3).

The contributions presented in this chapter are:

- The prediction of two organizational constructs in job interviews, namely personality and hirability, using body nonverbal cues. To our knowledge, [Nguyen et al., 2013a] is the first work in systematically analyzing audio-visual nonverbal behavior in employment interviews. In this present work, we extend that study by predicting personality traits in addition to hirability, and systematically focus on postures and gestures.
- The systematic analysis of body communication cues for the prediction of social constructs.
- The use of nonverbal cues such as speaking status to improve the performance of a computer vision algorithm, namely a beat gesture detector.

- The exploitation of the multi-modal nature of body communication to improve the prediction performance of personality and hirability.

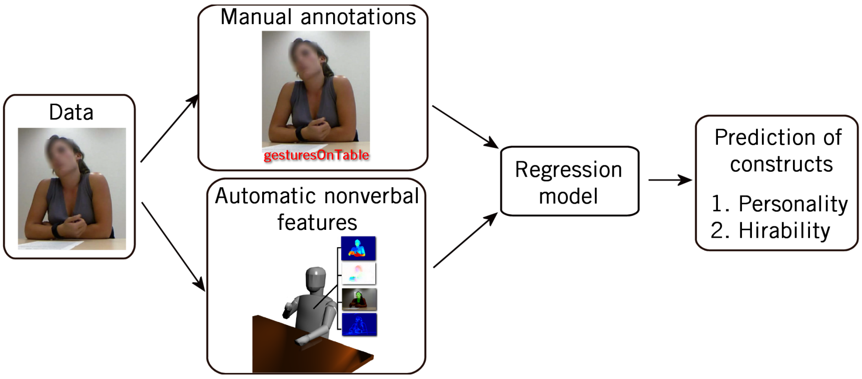


Figure 8.1: Proposed framework. Using nonverbal features of different kind (i.e. manually annotated and automatic), we propose to build a regression model to predict several social constructs.

Part of the work presented in this chapter was done in collaboration with Laurent Nguyen (PhD student at Idiap Research Institute and Ecole Polytechnique Fédérale de Lausanne), and has been published in [Nguyen et al., 2013b]. The rest of the material in this chapter has not been published elsewhere.

This chapter is organized as follows: in Section 8.2 we explain the extraction of low-level features, in Section 8.3 we show how we use those features for inference of social attributes, in Section sec:results we present our results, in Section 8.5 some research lines from our work are proposed, and finally in Section 8.6 we extract some conclusions.

8.2 Extraction of features

We use as features a combination of manual and automatic measurements. As a first evaluation for the validity of the proposed activity classes, we use their manual annotation instead of their automatic recovery. A similar reasoning was followed regarding automatic features: simpler body activity descriptors are used as a first approach. Both techniques are described during this section.

8.2.1 Manual annotations of body activity

As introduced in Chapter 4, five classes were defined based on the occurrences in the dataset and the relevance in the nonverbal communication literature [Knapp and Hall, 2009]: hidden hands, hands on table, gestures on table, gestures, and self-touch. They constitute an approximation for the applicant's body posture and gestures. Applicants were seated, therefore the posture was for a large part defined by the position of their arms. Other posture classes such as leaning forward or backward were also considered, but were discarded as the observed variability of such postures was low.

The data used in the present chapter is that of described in Section 4.3.3.4. As already described in Chapter 4, applicant body activity was annotated at the frame level by one person, with the help of a purpose-built script. To reduce the amount of frames to label, annotations were made every 15 frames (0.5 seconds); this temporal resolution was sufficient as no missing labels were observed while playing the full video at regular speed. To further reduce the amount of frames to label, we applied a motion threshold to the videos and annotated frames only when sufficient movement was present; unannotated frames in-between were assigned the same label as the latest annotated frame. This procedure allowed us to reduce the number of frames to annotate by 35%. In total, over 23000 frames were labeled. In order to assess the reliability of the annotations, a second person annotated 63 minutes of the dataset (approximately 5000 frames), and inter-rater agreement was satisfactory (using Cohen's Kappa [Galton, 1892]: $\kappa = 0.81$).

8.2.2 Automatic features

8.2.2.1 Speaking status

Speaking status was extracted using the Microcone, which automatically segments speaker turns by using a filter-sum beam former followed by a post-filtering stage in each of the spatial segments of the microphone array. The resulting speaker segmentations were stored in a file containing the relative time (start and end) and the speaker identifier. The objective performance of the speaker segmentation was not evaluated, but we manually inspected all segmentation files and observed only a small number of segmentation errors, even for short segments or overlapping speech. This has also been observed in previous research that used the Microcone device for speaker segmentation [Sanchez-Cortes et al., 2011].

8.2.2.2 Hand speed

To obtain estimates of hand speed, we used the method presented in Section 6.3 to compute the hand likelihood map for each frame of a video. This method assumes that the hands are the quickest parts of the video, that they are not the face, and that they have skin color. Based on these assumptions, the hand likelihood map can be computed as the product of the dense optical flow map, the binary face-mask image, and the skin-color segmentation binary image. Because the amount of data processing was substantially larger than in Chapter 6, we implemented a simple but effective method to obtain the hand speed image: we multiplied the hand likelihood map with the pixel frame difference, and normalized it by the distance between the head and the table to account for variations in the camera placement. An illustration of the procedure to compute the hand speed image is displayed in Figure 8.2. As a last step, we obtained the hand speed energy $e_{h,t}$ by aggregating the value of all pixels of the hand speed image, resulting in a single value for the hand speed estimate for each frame of a video.

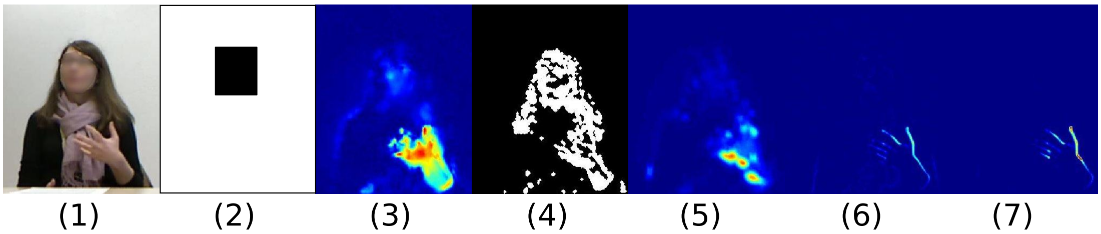


Figure 8.2: Illustration of the hand speed image computation (1) original image, (2) face mask, (3) optical flow map, (4) skin-color segmentation image, (5) hand likelihood map, (6) frame difference, and (7) resulting hand speed image.

8.2.2.3 Image activity histograms

In order to obtain information about the hand position of the participants, we created an image activity descriptor along the vertical axis of the image. We defined a 12-bin histogram $\mathcal{M}_{OF,t}$, which accumulates energy in different height bands of the dense optical flow image (normalized by the distance between the table and the head).

The histogram is able to capture two important factors which condition the applicant's visual activity, *i.e.* hand speed and hand height, which makes this feature suitable for the analysis of seated participants. Moreover, as the method is based on dense optical flow, it is appearance invariant, which makes it suitable for the analysis of subjects with different skin colors. An illustration of the image activity histogram can be seen in Figure 8.3.

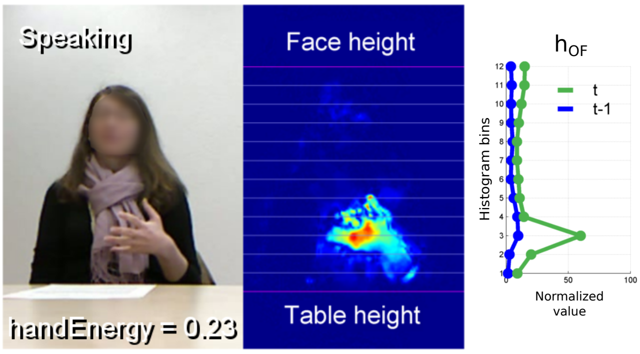


Figure 8.3: Visualization of several nonverbal automatic features. Left: input image with overlaid speaking status and hand energy value. Center: Dense optical flow and image height division. Right: Activity histograms $\mathcal{M}_{OF,t} and \mathcal{M}_{OF,t-1}$ that contain the optical flow aggregated along the vertical dimension of the image. Best viewed in color.

8.2.2.4 Beat gestures

In order to automatically detect beat gestures given an input video sequence, we combine the activity zone histograms $\mathcal{M}_{OF,t}$, hand energy $e_{h,t}$ and speaking status s_t to train a Random Forest classifier. The information that is automatically extracted forms therefore the feature vector $\mathcal{X}_{tr}$:

$$\mathcal{X}_{tr} = \{(\mathcal{M}_{OF,t}, \mathcal{M}_{OF,t-1}, e_{h,t}, s_t)\}_{i=1}^{n_f} \tag{8.1}$$

where n_f is the number of training frames. We therefore input the automatic extracted features described in 8.2.2.1, 8.2.2.2 and 8.2.2.3 as the feature vector, and manual beat gesture annotations in order to build the Random Forest model for classification.

In the modified annotationsm we merge classes 'gestures' and 'gestures on table' in a single 'beat gesture' class, since according to the literature they have the same conversational meaning [Knapp and Hall, 2009]. Classes 'hands on table', 'self touch', and 'hidden hands' are grouped as 'other'. The forest therefore becomes a binary beat gestures classifier. Its performance is evaluated in the results section.

8.3 Inference of personality traits and hirability

The following section studies the relationship between nonverbal behavior, personality, and hirability, seen from both the psychology and the social computing perspectives.

In Section 8.2 we have described the extraction process of basic nonverbal features. In this section, we describe the extraction of higher level information (such as statistics) from those basic features, which in turn are used to infer personality traits and hirability. Table 8.1 shows the list of all the high level features or body communication cues that we used in our study.

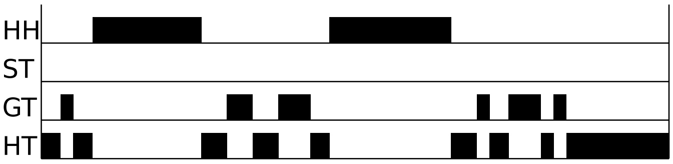


Figure 8.4: Illustration for cues based on annotations of body activity. In this example, there are two ”hidden hands” events, six ”gestures” events, nine ”hands on table” events, and no ”self touch” events. Statistics are computed from event duration. If no event occurred, the statistics are set to zero.

8.3.1 High level feature extraction

Our objective is to explore the use of multimodal body communication cues to predict hirability and personality. To this end, we leverage on automatic speaker segmentations provided by the Microcone device, manual annotations of postures and gestures, and automatic extraction of hand movement and hand activity zones. With this information, we present the method used to extract detailed gesture- and posture-based nonverbal cues.

8.3.1.1 Nonverbal cue encoding

Here, we describe the method used to encode the nonverbal cues from the manual annotations of body activity, the speaker segmentations, the hand speed estimates, and the image activity histograms. These cues will then be used as features for the prediction of personality and hirability.

Cues based on annotations of body activity: Nonverbal cues were extracted from the manual annotations of body activity. To capture a "big picture" of the body activity, they were based on statistics derived from event duration. Events were defined as a sequence of frames where the applicant showed the same type of body activity, and are characterized by their starting time and duration (see Figure 8.4). For all the activity classes, we computed the number of events, mean, median, standard deviation, lower and upper quartiles, minimum, maximum, range, position (in time) of shortest and longest events, and total relative time. It should be noted that it was possible for a given class to be missing in a given sequence. We addressed this by introducing a binary variable indicating whether at least an event occurred or not. The statistics on turn durations were set to zero if no event occurred. The list of body communication cues based on manual annotations is included in Table 8.1.

Cues based on hand speed and activity histograms: The hand speed approximation $e_{h,t}(t)$ and the image activity histograms $\mathcal{M}_{OF,t}$ (sections 8.2.2.2 and 8.2.2.3) not only provide information on *how much* hand movement occurred at a given instant, but also on *where* these hand movements occurred. We also extracted nonverbal cues based on the activity descriptors. To account for short bursts of hand movement characterized by quick changes of hand speed (which could be associated with beat gestures), we computed the hand acceleration. We defined the global hand acceleration at time t as:

$$a_{h,t} = |e_{h,t} - e_{h,t-1}|, \tag{8.2}$$

and the image-height-dependent acceleration as:

$$\mathcal{M}_{A,t} = |\mathcal{M}_{OF,t} - \mathcal{M}_{OF,t-1}|. \tag{8.3}$$

To extract nonverbal cues from the univariate time series $e_{h,t}$ and $a_{h,t}$, we computed their mean, median, standard deviation, minimum, maximum, range, quartiles, proportion of non-zero elements, and zero-crossing rate. We also computed the following statistics related to the histogram main mode (*i.e.*, the position of the maximum histogram bin) to account for hand position: mean, median, standard deviation, quartiles, and zero-crossing rate. Table 8.1 also shows the list of automatic body communication cues used in this study.

Exploiting the speaking status: To exploit the finding in psychology stating that body communication is conditioned on the speaking status [McNeill, 1985, Knapp and Hall, 2009], we computed the statistics on manual body activity event durations, hand speed and acceleration, and activity and acceleration histogram modes for four different cases:

- The unimodal case, *i.e.* without taking into account the speaking status.
- The speaking case, *i.e.* only using frames in which the applicant was speaking.
- The silent case, *i.e.* only using frames in which the applicant was silent.
- The aggregated case, *i.e.* aggregating the three previous cases.

Table 8.1: List of the manual and automatic nonverbal cues used in this study. Each statistical cue was computed for (a) the unimodal case (*i.e.* not taking the speaking status into account), (b) speaking case, (c) silent case, and (d) aggregated case (*i.e.* aggregating unimodal, speaking, and silent).

Manual features:		
Posture class	**Statistics**	**Speaking status**
Hidden hands (HH) Self-touch (ST) Hands on table (HT) Gestures on table (GT) Gestures (G)	mean, median, std, quartiles, # of events, min., max., range, rel. time, pos. of min./max., exists	Unimodal, Speaking, Silent, Aggregated

Automatic features:		
Time-series	**Statistics**	**Speaking status**
Hand velocity (HV) Hand acceleration (HA)	mean, median, std, quartiles, zero-crossing rate, min., max., range, non-zero proportion	Unimodal, Speaking, Silent, Aggregated
Histograms (type of aggregated information)	**Statistics**	**Speaking status**
Hand velocity histogram (HVH) Hand acceleration histogram (HAH)	mean, median, std, quartiles, zero-crossing rate,	Unimodal, Speaking, Silent, Aggregated

8.3.2 Prediction of tasks

In order to analyze the predictive validity of body posture with respect to self-rated personality traits and hirability impressions, we defined a regression problem which predicts hirability and personality scores, where each social variable is considered as an independent regression task. To this end, we used a leave-one-interview-out cross validation strategy. Two regression methods were used for predicting personality and hirability. We used Ridge regression [Hoerl and Kennard, 1970] as the first prediction model. It is a linear model where the parameters are estimated by minimizing the sum of squared errors plus an l_2 regularization term (referred as the Ridge parameter), which prevents the model from overfitting. The

Ridge parameter was estimated automatically using 10-fold inner cross-validation. As a second prediction method, we used random forest with 1000 trees, which is a robust non-linear regression model.

Given the large number of features (> 300) compared to the number of data points (43), we decided to analyze sub-groups of features independently. This allowed the regression model to be correctly learned, and enabled the analysis of the predictive validity of specific postures and speaking cases. For the nonverbal features based on the manual annotations, five feature groups were defined based on the annotated body activity classes. For the automatic cues, we used the hand movement e_h, the hand acceleration a_h, the activity histogram $\mathcal{M}_{OF}$, and the acceleration histogram $\mathcal{M}_A$ cues as four feature groups.

In order to test whether exploiting the speaking status improves the prediction accuracy, we further segmented the feature groups into four sub-groups:

- Unimodal features, *i.e.* obtained without taking into account the speaking status.
- Silent features only.
- Speaking features only.
- Aggregated features, *i.e.* the concatenation of unimodal, silent, and speaking cues.

Prediction results using specific posture cues, speaking status, and regression methods are reported and discussed in Section 8.4.3.

8.4 Results and discussion

8.4.1 Annotation statistics

In Table 8.2, we show descriptive statistics of the personality and hirability variables used in this study. We observe that except communication and conscientiousness, all hirability measures were significantly correlated with each other. Also, extraversion was found to be significantly and positively correlated with three hirability scores: hiring decision, conscientiousness, and stress resistance. This suggests that extraverts were seen as more employable by the coder. This finding is supported by the related psychology literature, which finds extraversion as a valid predictor of performance in jobs characterized by a high level of social interactions [Barrick and Mount, 1991], as it is the case here.

Table 8.2: Descriptive statistics (mean, std, and Pearson's correlation) of personality and hirability ($^*p < .05$, $^\dagger p < .005$)

	μ	σ	1	2	3	4	5	6	7	8	9	10
1. Hiring decision	6.209	1.753		$0.536^\dagger$	$0.770^\dagger$	$0.662^\dagger$	$0.707^\dagger$	$0.501^\dagger$	0.014	-0.199	0.136	0.119
2. Communication	2.953	0.872			$0.429^\dagger$	0.268	0.306^*	0.279	-0.113	-0.080	0.037	0.080
3. Persuasion	2.977	1.080				$0.493^\dagger$	$0.520^\dagger$	0.250	0.076	-0.210	0.059	0.113
4. Conscientiousness	3.070	1.033					$0.602^\dagger$	0.358^*	0.070	-0.235	-0.003	0.289
5. StressRes	3.047	0.722						0.341^*	0.124	-0.182	0.035	0.263
6. Extraversion	4.008	0.434							0.037	-0.292	$0.449^\dagger$	0.313^*
7. Openness	3.736	0.527								-0.049	0.112	-0.072
8. Neuroticism	2.210	0.573									-0.168	$-0.578^\dagger$
9. Agreeableness	4.144	0.418										0.346^*
10. Conscientiousness	4.106	0.640										

The class distribution of the data is shown in Figure 4.9. We observe that *hands on table* accounted for more than half of the labels. The dataset was recorded in a real setting, therefore it reflects the natural tendency of the participants while being seated. It should be noted that in 34.2% of the data the subject was silent while listening to the interviewer. Our proxy for beat gestures ("gestures" and "gestures on table") was present 33.6% of the time. The least represented class was "hidden hands", while "self-touch" appeared almost as often as "gestures".

8.4.2 Analysis of the effect of speaking status

In order to test whether our initial assumption stating that body communication was conditioned on the applicant's speaking status, we computed the Student's t-test [O'Connor and Robertson, 1908] to examine whether some significant differences in feature values between speaking and silent existed. In Table 8.3, we display the significantly different features ($p < 0.05$), and report whether the larger value was associated with moments when the job applicant was silent or speaking.

We observe that job applicants gestured more when they were speaking: more *gestures* and *gestures on table* time, longer events, larger range of durations; larger hand speeds; larger hand accelerations. Inversely, interviewees self-touched and kept their hands on the table longer when listening to the interviewer. This findings validate our main assumption of the multimodal nature of hand gestures and body posture, based on the nonverbal communication literature [Knapp and Hall, 2009, McNeill, 1985]. Furthermore, we observe that the automatic features based on hand speed and hand acceleration were also conditioned on the speaking status.

Table 8.3: Speaking status analysis. Features significantly different ($\mathbf{p} < \mathbf{.05}$) between speaking and silent, using Student's t-test.

Feature group	**Larger feature value for silent**	**Larger feature value for speaking**
Hidden hands		number of events
Self touch	relative time, median, maximum, minimum, quartiles, range	number of events
Hands on table	relative time, mean, standard deviation, upper quartile	number of events
Gestures		relative time, mean, median, standard deviation, maximum, upper quartile, range, exist, number of events
Gestures on table		relative time, mean, median, standard deviation, maximum, minimum, range, quartiles, number of events, exist
Hand speed	minimum, zero-crossing rate	mean, median, maximum, quartiles, range, non-zero proportion
Hand acceleration		mean, median, maximum, minimum, quartiles, non-zero proportion

8.4.3 Prediction of hirability and personality

One of the research questions of this study was to investigate whether hirability and personality could be inferred using body communication cues as predictors. In order to evaluate the prediction accuracy of our method, we used the standard coefficient of determination R^2, which can be seen as the amount of variance explained by the evaluated model. In Table 8.4, we report the results for which the R^2 values were higher than 0.1. From those findings, several observations can be made.

Except for communication ability, all hirability scores were inferred above the $R^2 = 0.1$ threshold using multimodal body communication cues. Importantly, we achieved $R^2 = 0.209$ for the hiring decision score using automatically extracted activity histogram features (aggregation of unimodal, speaking, and silent) and Ridge regression as a prediction method. This finding demonstrates the potential of predicting job interview outcomes using body communication cues.

For personality, we show that prediction can be achieved to some degree using body communication cues only, which was to our knowledge not analyzed systematically prior to this work. Using such nonverbal features showed prediction performance comparable to other existing work in social computing. More specifically, extroversion prediction ($R^2 = 0.165$) was found to be less accurate than in the state of the art (*e.g.* [Biel and Gatica-Perez, 2013]), but results obtained for openness to experience ($R^2 = 0.238$), agreeableness ($R^2 = 0.140$), and conscientiousness ($R^2 = 0.165$) can be considered as promising.

Only six prediction scores where $R^2 > 0.1$ were achieved using unimodal features (*i.e.* without taking into account the speaking status of the job applicant when extracting the cues). In comparison, 33 prediction scores were achieved by using body communication cues conditioned on the speaking status. This finding further shows the intimate link between speaking status and body communication in this job interview setting. Furthermore, we show that leveraging on this finding can improve the prediction of social constructs.

Overall, we observe that automatic hand activity cues were moderate yet promising predictors of hirability ratings. Indeed, the best prediction results for the hiring decision and stress resistance were achieved using automatic cues based on activity histograms. For the hirability variables of persuasion and conscience, the use of automatic features decreased the prediction accuracy compared to manual features (from 0.235 to 0.205 and 0.200 to 0.119, respectively). The use of automatic body communication cues was however found to show poor performance for self-rated personality traits, which calls for further work in this direction.

8.4.4 Beat gesture detection performance

We evaluated the performance of the beat gesture detector by using the same data (43 interviews). Our aim was not to perform a comparison against the state of the art of activity recognition, but to evaluate classification improvements when using previously unseen features for this task, such as speaking status.

A Random Forest with 100 trees was trained for each subject by using a leave-one-out strategy. We then compared the precision against random and majority class baselines, with different configurations for the elements of the feature vector $(\mathcal{M}_{OF,t}, \mathcal{M}_{OF,t-1}, e_{h,t}, s_t)$. As performance index, we chose global accuracy and the F_1 score, which is defined as the harmonic mean between precision and recall:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}. \tag{8.4}$$

It is therefore a strict measure that takes into account class balance. This is specially relevant in our case since beat gestures constitute only 33.6% of the test data. Performance results can be seen in Table 8.5.

The accuracy of the baseline methods, namely random and majority class, were 66.4% and 50%, respectively. We got a statistically significant improvement ($p < 0.001$) over both with every configuration of the feature vector $\mathcal{X}_{tr}$. We show that adding activity information from the previous time instant (i.e. taking into account image acceleration) improved the F_1 score by an average of 1.64% and the accuracy by an average of 0.9% ($p < 0.001$ using the one-sided test of a difference of proportions). Adding the aggregation of the hand energy $e_{h,t}$ increased the F_1 score an average of 0.12%, although it did not improve global accuracy.

Importantly, the results also show that by adding speaking status, there was an average increase of 1.66% in the F_1 score and 0.85% in accuracy ($p < 0.001$). We find this improvement revealing since in only 14% of the testing samples the subject was silent. This class imbalance is the consequence of applying the image movement threshold described in Section 4.3.3, as we ended up with the most challenging 35% of the original video (i.e. only the parts that contain movement are analyzed, which usually appear in conjunction with speech). It therefore imposes a strict performance measure, as by definition there are no beat gestures in the parts of the video that show no movement. By taking them into account, we obtained an accuracy of **92%**.

8.5 Future directions

As it has been demonstrated along this chapter, the ideas that we use in our automatic methods can be applied in order to infer important traits of candidates in job interviews. The next steps should include investigating in greater detail the characteristics and evolution of the body movement along time.

As an example, along this line of research, we have taken the output body poses obtained with our appearance-invariant method (described in Chapter 7) and normalized them to obtain consistency across subjects. Then, we have divided each interview length into 10 temporal intervals of equal duration, and computed the correlation between the mean position of one of the hands in each bin and the personality and hirability variables. Results are shown in Figure 8.5. As can be seen, there are significant effects for several slices and variables.

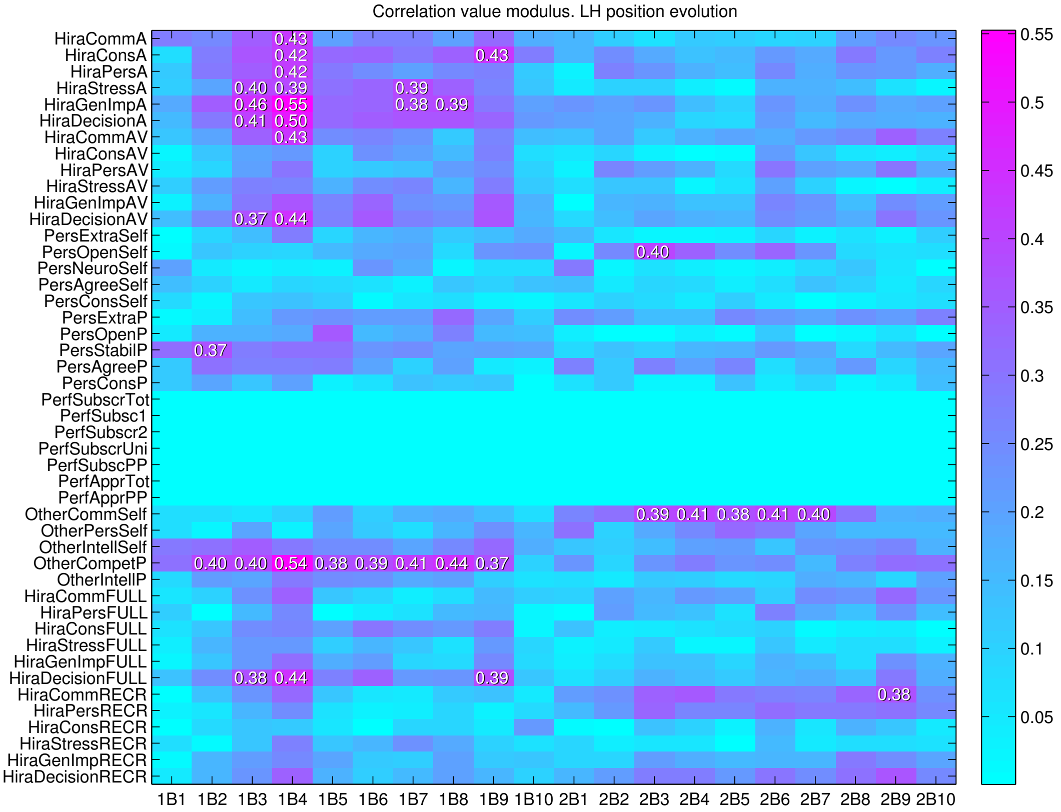


Figure 8.5: Example of correlation with social constructs. The horizontal and vertical position of the left hand is correlated with several constructs. Interviews have been divided into 10 time bins. Bins named 1B1...10B1 are associated with the horizontal position of the hand, and bins 2B1...2B10 are associated with the vertical position of the hand. Numerical values are displayed only for statistically significant correlations (p<0.05).

Aditionally, by using a Random Forests regressor and the hand position data, we obtain R^2 values (i.e. the amount of variance that the regressor can explain) of **0.27** for some dimensions of the perceived competence, or **0.1** for the general impression caused. This shows promising results that should be looked into and expanded.

Finally, whereas in this present work we limited our analysis to applicant behavior, we propose future work to analyze the other half of the dyad, *i.e.* the interviewer. Such an analysis could allow to determine whether the interviewer's behavior could be used to predict interview outcomes or provide information about the applicant's personality.

8.6 Conclusion

We conducted a systematic study on applicant body communication cues in job interviews with respect to hirability impressions and self-rated personality. Additionally, we leveraged on findings in psychology suggesting a link between body communication and speech to analyze body communication from a multimodal perspective.

We used a dataset of 43 real job interviews. We extracted a rich mixture of body communication features from manual annotations of body activity and automatically obtained hand speed descriptors. To account for the speaking status, these features were conditioned on whether the applicant was silent or speaking. By analyzing the corresponding differences in feature values, we validated our main assumption stating that speaking and silent differences existed.

We showed that the prediction of interview outcomes using body communication cues is promising. To our knowledge, the only work systematically analyzing employment interviews is that of [Nguyen et al., 2013a]; we have contributed new findings for the case when only body communication cues are used, as opposed to other nonverbal cues such as speaking activity, head gestures, or prosody. We also show that manual body communication cues can be used to predict applicant personality traits to some degree. The reported results also demonstrate that exploiting the intimate link between body communication and speaking status helps the inference of personality and hirability. Thanks to a beat gesture detector, we additionally show that audio information can be used in order to increase the performance of a computer vision detector.

The prediction of some of the constructs analyzed in this work rely on manual annotations of body activity. This is the case of personality traits, where no automatic feature could produce accurate prediction scores. However, results show that for hiring decision, using the automatic hand speed estimates yielded higher prediction results than manual features. This finding underlines the relevance of automatic hand speed estimates for the analysis of employment interviews, even if these estimates are coarse.

Table 8.4: Prediction results for hirability impressions (1-5) and self-rated personality (6-10) using manual (M) and automatic (A) cues. R^2 was used to evaluate the prediction performance. Only results with $\mathbf{R^2} > \mathbf{0.1}$ are reported.

Hirab. variable	**Feature group**	**Speaking status**	**Regression method**	R^2
1. Hiring Decision	Gesturing (M)	Silent	Ridge	0.196
	Activity histograms. (A)	Silent	Ridge	0.177
	Activity histograms. (A)	Silent	RF	0.141
	Activity histograms. (A)	Aggr.	Ridge	**0.209**
	Activity histograms. (A)	Aggr.	RF	0.139
	Acceleration histograms. (A)	Silent	Ridge	0.180
	Acceleration histograms. (A)	Silent	RF	0.114
2. Communication	-	-	-	-
3. Conscientiousness	Hidden hands (M)	Speak	RF	0.174
	Hidden hands (M)	Aggr.	RF	0.103
	Gestures on table (M)	Speak	RF	0.159
	Gestures on table (M)	Aggr.	Ridge	0.150
	Gestures on table (M)	Aggr.	RF	**0.200**
	Activity histograms (A)	Silent	RF	0.119
	Activity histograms (A)	Aggr.	RF	0.109
4. Persuasion	Hidden hands (M)	Aggr.	Ridge	**0.235**
	Activity histograms (A)	Silent	RF	0.204
	Activity histograms (A)	Aggr.	RF	0.109
	Activity histograms (A)	Aggr.	Ridge	0.127
	Acceleration histograms (A)	Silent	Ridge	0.144
	Acceleration histograms (A)	Silent	Ridge	0.109
5. Stress ressistance	Activity histograms. (A)	Aggr.	RF	**0.103**

Personality variable	**Feature group**	**Speaking status**	**Regression method**	R^2
6. Extroversion	Hidden hands (M)	Unimod.	RF	**0.165**
	Hidden hands (M)	Speak	RF	0.124
	Hidden hands (M)	Aggr.	Ridge	0.154
	Hidden hands (M)	Aggr.	RF	0.139
	Self touch (M)	Unimod.	RF	0.112
	Self touch (M)	Aggr.	RF	0.127
	Gestures on table (M)	Unimod.	RF	0.152
	Gestures on table (M)	Speak	RF	0.137
7. Openness	Self touch (M)	Unimod.	Ridge	0.109
	Self touch (M)	Silent	RF	0.103
	Hands on table (M)	Silent	RF	**0.238**
	Hands on table (M)	Silent	RF	0.177
8. Neuro	-	-	-	-
9. Agreeableness	Hidden hands (M)	Speak	RF	**0.140**
	Gestures on table (M)	Speak	RF	0.111
10. Conscientiousness	Hidden hands (M)	Unimod.	Ridge	0.136
	Hidden hands (M)	Silent	Ridge	**0.165**
	Gestures on table (M)	Unimod.	Ridge	0.136
	Gestures on table (M)	Silent	Ridge	**0.165**

Table 8.5: Performance of different feature sets

Feature combination	F_1 **score**	**Accuracy**	**Random accuracy**	**Majority class accuracy**
$\mathcal{M}_{OF,t}(t)$	56.73 %	75.1 %	50%	66.4%
$\mathcal{M}_{OF,t}(t), e_{h,t}$	56.80 %	75.2 %	50%	66.4%
$\mathcal{M}_{OF,t}(t), e_{h,t}, S$	58.43 %	76.0 %	50%	66.4%
$\mathcal{M}_{OF,t}(t, t-1)$	58.33 %	76.1 %	50%	66.4%
$\mathcal{M}_{OF,t}(t, t-1), e_{h,t}$	58.50 %	76.0 %	50%	66.4%
$\mathcal{M}_{OF,t}(t, t-1), e_{h,t}, S$	**60.05 %**	**76.9 %**	50%	66.4%

Chapter 9

Conclusions

9.1 Introduction

In this chapter, the conclusions obtained during the several studies conducted in the thesis are presented through a summary of our contributions. Finally, a discussion about limitations and future lines of work is presented.

9.2 Contributions

In this thesis we have justified the importance of computer vision methods for nonverbal communication analysis. We have approached the problem with markerless motion capture, focusing on the detection of body parts like face and hands, since in the literature and our own results they have proven to be an excellent proxy to obtain the body configuration. We summarize the conclusions obtained through our work as follows:

- There are three main sensor set-ups in markerless motion capture: multi-camera, single camera and depth camera. In this thesis we make contributions in all of them.

- We first designed a multi-camera approach based on 3D scene reconstruction through Visual Hulls. We took advantage of non-linear regression methods in order to simplify the search in the high-dimensionality human pose space. By doing this, we were able to track multiple subjects simultaneously with a single tracker. Helped by a refinement process, we were able to provide better generalization capabilities.

- We then developed a single camera method, based on the idea of hand saliency: we hypothesized that the hands are the parts of the image that move quicker along a whole video. To this end, we designed a new hand tracker based on a Decision Tree algorithm, and performed simultaneous action recognition.

- We later extended this approach by fusing the information provided by a depth camera in the hand saliency map equations.

- Finally, we developed a highly appearance-invariant method for motion capture while using again a single color camera. Thanks to dense optical flow and a torso detector, we were able first to classify the body parts in the image and then obtain the body configuration. This contribution is a step in order to remove the appearance-related problems of markerless motion capture, and improves the state-of-the-art in performance versus processing time ratio.

- We evaluated all the approaches with both public and private datasets, showing or improving state-of-the-art performance.
- Finally, we applied some of the ideas behind our methods to infer a series of social constructs from real job interviews. We extracted and aggregated a series of manually-annotated and automatically-extracted features from videos, and showed the connection between them and personality traits or job performance. We were able to predict some of those traits with a regression scheme.

9.3 Limitations and future work

It is interesting to emphasize some of the future research lines that arise from the work done in this thesis.

- Our multiple-camera approach would greatly benefit from different particle insertion strategies. For example, a rough approximation of the ongoing actions can be made from simple features from the image, which can be taken into account when injecting new particles.
- The processing time of our hand tracking method can be improved. For example, using greedy approaches instead of Decision Trees can speed up the process and might remove the need of having the whole sequence from the beginning. This in turn can open the doors for further applications.
- The research in color invariant features has shown potential to solve some of the problems in markerless motion capture. To this end, they would benefit from pre-processing the dense optical flow to reduce noise and offer more consistency. Also, more elaborated features in the body part classifier and alternative regression techniques can improve the accuracy of this method.
- In social construct prediction, inference results can be expanded by using the more detailed information that motion capture provides, as opposed to methods that aggregate information along the whole sequences. Also, looking at the interviewer could be useful to provide extra data about the interviewee.
- One of the most important limitations in computer vision in general is the definition of descriptive features, robust to variations. In markerless motion capture in particular, it became apparent in the big performance increase that Histograms of Oriented gradients provided when detecting body parts, compared to former state-of-the-art features like Haar. It is foreseeable that in the years to come, a similar leap will solve many of the current problems and open the door to completely new applications.
- In addition to low-level image features, machine learning techniques can greatly improve performance and perception understanding. Currently, significant research work is being undergone in Deep Neural Networks, or extensions of Random Forests like Random Jungles. Given its complexity, markerless motion capture is a perfect subject to benefit from such techniques.
- Advances in computer vision will most likely influence social computing. With the increase in accuracy and processing power, new sensing scenarios may arise, that could allow the prediction human behavior in unseen ways. While probably not conclusive, such prediction tasks might be used as reliable guidelines for the assessment of social attributes.

Bibliography

[Mic,] Microcone: intelligent microphone array for groups [online]. Available: http://www.dev-audio.com/products/microcone/.

[cha, 2011] (2011). Chalearn gesture dataset (cgd2011), chalearn, california, 2011. copyright (c) chalearn - 2011. http://gesture.chalearn.org/data.

[Agarwal and Triggs, 2004] Agarwal, A. and Triggs, B. (2004). Tracking articulated motion with piece-wise learned dynamical models. In *IEEE European Conference on Computer Vision (ECCV)*.

[Agarwal and Triggs, 2006] Agarwal, A. and Triggs, B. (2006). Recovering 3d human pose from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 28(1).

[Alsabti, 1998] Alsabti, K. (1998). An efficient k-means clustering algorithm. In *IPPS/SPDP Workshop on High Performance Data Mining*.

[Ambady et al., 1995] Ambady, N., Hallahan, M., and Rosenthal, R. (1995). On judging and being judged accurately in zero-acquaintance situations. *Journal of Personality and . . .* , 69(3):518–529.

[Amin et al., 2013] Amin, S., Andriluka, M., Rohrbach, M., and Schiele, B. (2013). Multi-view pictorial structures for 3d human pose estimation. In *British Machine Vision Conference (BMVC)*.

[Anderson and Shackleton, 1990] Anderson, N. and Shackleton, V. (1990). Decision making in the graduate selection interview: A field study. *Occupational Psychology*, 63(1):63–76.

[Andriluka et al., 2010] Andriluka, M., Roth, S., and Schiele, B. (2010). Monocular 3d pose estimation and tracking by detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Angela Yao and Gool, 2011] Angela Yao, Juergen Gall, G. F. and Gool, L. V. (2011). Does human action recognition benefit from pose estimation? In *British Machine Vision Conference (BMVC)*.

[Baak et al., 2011] Baak, A., Mueller, M., Bharaj, G., Seidel, H.-P., and Theobalt, C. (2011). A data-driven approach for real-time full body pose reconstruction from a depth camera. In *International Conference on Computer Vision (ICCV)*.

[Balan and Black, 2006] Balan, A. and Black, M. J. (2006). An adaptive appearance model approach for model-based articulated object tracking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Balan and Black, 2008] Balan, A. and Black, M. J. (2008). The naked truth: Estimating body shape under clothing,. In *European Conference on Computer Vision (ECCV)*.

[Balan et al., 2007] Balan, A. O., Black, M. J., Haussecker, H. W., and Sigal, L. (2007). Shining a light on human pose: On shadows, shading and the estimation of pose and shape. In *IEEE International Conference on Computer Vision (ICCV)*.

[Ball and Breese, 2000] Ball, G. and Breese, J. (2000). Relating Personality and Behavior : Posture and Gestures. *Lecture Notes in Computer Science*, 1814:196–203.

[Ballan and Cortelazzo, 2008] Ballan, L. and Cortelazzo, G. M. (2008). Marker-less motion capture of skinned models in a four camera set-up using optical flow and silhouettes. In *International Conference on 3D Vision (3DPVT).*

[Barrick and Mount, 1991] Barrick, M. and Mount, M. (1991). The big five personality dimensions and job performance: a meta-analysis. *Personnel Psychology*, 44(1):1–26.

[Bartoli et al., 2013] Bartoli, A., Pizarro, D., and Loog, M. (2013). Stratified Generalized Procrustes Analysis. *International Journal of Computer Vision (IJCV).*

[Batrinca et al., 2011] Batrinca, L., Mana, N., Lepri, B., Pianesi, F., and Sebe, N. (2011). Please, tell me about yourself: automatic personality assessment using short self-presentations. *Proc. Int. Conf. on Multimodal Interactions*, pages 255–262.

[Biel and Gatica-Perez, 2013] Biel, J.-I. and Gatica-Perez, D. (2013). The YouTube Lens: Crowd-sourced Personality Impressions and Audiovisual Analysis of Vlogs. *IEEE Transactions on Multimedia*, 15(1):41–55.

[Bourdev and Malik, 2009] Bourdev, L. and Malik, J. (2009). Poselets: Body part detectors trained using 3d human pose annotations. In *IEEE International Conference on Computer Vision (ICCV).*

[Breiman et al., 1984] Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees.* Wadsworth and Brooks, Monterey, CA.

[Brostow, 2004] Brostow, G. J. (2004). Novel skeletal representation for articulated creatures. In *European Conference on Computer Vision, (ECCV).*

[Brox et al., 2007] Brox, T., Rosenhahn, B., and Cremers, D. (2007). Contours, optic flow, and prior knowledge: cues for capturing 3D human motion in videos. In *Human Motion - Understanding, Modeling, Capture, and Animation.*

[Brubaker et al., 2009] Brubaker, M. A., Sigal, L., and Fleet, D. J. (2009). Estimating contact dynamics. In *International Conference on Computer Vision (ICCV).*

[Buehler et al., 2011] Buehler, P., Everingham, M., Huttenlocher, D. P., and Zisserman, A. (2011). Upper body detection and tracking in extended signing sequences. *International Journal of Computer Vision (IJCV)*, 95(2).

[Burenius et al., 2013] Burenius, M., Sullivan, J., and Carlsson, S. (2013). 3d pictorial structures for multiple view articulated pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[Burnett and Motowidlo, 1998] Burnett, J. and Motowidlo, S. (1998). Relations between different sources of information in the structured selection interview. *Personnel Psychology*, 51(4):963–983.

[Canny, 1986] Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI).*

[Chambolle and Pock, 2011] Chambolle, A. and Pock, T. (2011). A first-order primal-dual algorithm for convex problems withÂ applications to imaging. *Journal of Mathematical Imaging and Vision (JMIV).*

[Chen et al., 2011] Chen, C., Yang, Y., Nie, F., and Odobez, J.-M. (2011). 3d human pose recovery from image by efficient visual feature selection. *Computer Vision and Image Understanding (CVIU)*, 115(3).

[Cheng, 1995] Cheng, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17(8).

[Cheung et al., 2003] Cheung, G. K. M., Baker, S., and Kanade, T. (2003). Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[Cole et al., 2004] Cole, M., Feild, H., and Giles, W. (2004). Job type and recruiters' inferences of applicant personality drawn from resume biodata: Their relationships with hiring recommendations. *Journal of Selection and Assessment*, 12(4):363–367.

[Corazza et al., 2006] Corazza, S., Mundermann, L., Chaudhari, A. M., Demattio, T., Cobelli, C., and Andriacchi, T. P. (2006). A markerless motion capture system to study musculoskeletal biomechanics: visual hull and simulated annealing approach. *Annals of Biomedical Engineering*, 34(6).

[Costa and McCrae, 1992] Costa, P. T. and McCrae, R. R. (1992). *Neo PI-R Professional Manual.* Psychological Assessment Resources.

[Criminisi et al., 2012] Criminisi, A., Shotton, J., and Konukoglu, E. (2012). Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends in Computer Graphics and Vision*, 7(2-3).

[Curhan and Pentland, 2007a] Curhan, J. and Pentland, A. (2007a). Thin slices of negotiation: Predicting outcomes from conversational dynamics within the first 5 minutes. *Applied Psychology*, 92(3):802.

[Curhan and Pentland, 2007b] Curhan, J. and Pentland, A. (2007b). Thin slices of negotiation: Predicting outcomes from conversational dynamics within the first five minutes. *Journal of Applied Psychology (JAP).*

[Dalal and Triggs, 2005] Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR).*

[Dantone et al., 2013] Dantone, M., Gall, J., Leistner, C., , and Gool., L. V. (2013). Human pose estimation using body parts dependent joint regressors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[Daubney et al., 2009] Daubney, B., Gibson, D., and Campbell, N. (2009). Monocular 3d human pose estimation using sparse motion features. In *IEEE Workshop on Tracking Humans for Evaluation of their Motion in Image Sequences 2009 - Held in conjunction with ICCV.*

[de Aguiar et al., 2007] de Aguiar, E., Theobalt, C., Stoll, C., and Seidel, H.-P. (2007). Marker-less 3d feature tracking for mesh-based human motion capture. In *Workshop on Human Motion (HM).*

[DeGroot and Gooty, 2009] DeGroot, T. and Gooty, J. (2009). Can Nonverbal Cues be Used to Make Meaningful Personality Attributions in Employment Interviews? *Journal of Business and Psychology*, 24(2):179–192.

[Delamarre and Faugeras, 1999] Delamarre, Q. and Faugeras, O. D. (1999). 3d articulated models and multi-view tracking with silhouettes. In *IEEE International Conference on Computer Vision (ICCV).*

[den Bergh and Gool, 2011] den Bergh, M. V. and Gool, L. J. V. (2011). Combining rgb and tof cameras for real-time 3d hand gesture interaction. In *IEEE Winter conference on Applications of Computer Vision (WACV)*.

[den Bergh et al., 2009] den Bergh, M. V., Koller-Meier, E., and Gool, L. J. V. (2009). Real-time body pose recognition using 2d or 3d haarlets. *International Journal of Computer Vision (IJCV)*, 83(1).

[Deutscher, 2000] Deutscher, J. (2000). Articulated body motion capture by annealed particle filtering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Dollar et al., 2012] Dollar, P., Appel, R., and Kienzle, W. (2012). Crosstalk cascades for frame-rate pedestrian detection. In *European Conference on Computer Vision, (ECCV)*.

[Doucet, 1997] Doucet, A. (1997). *Monte-Carlo methods for bayesian estimation of Hidden-Markov models. Application to Radiation Signal*. PhD thesis, University of Paris-Sud, France.

[Doucet et al., 2000] Doucet, A., Godsill, S., and Andrieu, C. (2000). On sequential monte carlo sampling methods for bayesian filtering. *Statistics and Computing*, 10(3).

[Drummond and Cipolla, 2001] Drummond, T. and Cipolla, R. (2001). Real-time tracking of highly articulated structures in the presence of noisy measurements. In *IEEE International Conference on Computer Vision (ICCV)*.

[Egashira et al., 2009] Egashira, H., Shimada, A., Arita, D., and ichiro Taniguchi, R. (2009). Vision-based motion capture of interacting multiple people. In *International Conference on Image Analysis and Processing (ICIAP)*.

[Eichner and Ferrari, 2010] Eichner, M. and Ferrari, V. (2010). We are family: joint pose estimation of multiple persons. In *IEEE European Conference on Computer Vision (ECCV)*.

[Eichner et al., 2012] Eichner, M., Marin-Jimenez, M., Zisserman, A., and Ferrari, V. (2012). 2d articulated human pose estimation and retrieval in (almost) unconstrained still images. *International Journal of Computer Vision (IJCV)*, 99(2).

[Ekman et al., 1972] Ekman, P., Wallace, V., and Ellsworth, P. (1972). *Emotion in the human face: guide-lines for research and an integration of findings*. Pergamon Press.

[Elgammal and Lee, 2004] Elgammal, A. and Lee, C.-S. (2004). Inferring 3d body pose from silhouettes using activity manifold learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Feese et al., 2011] Feese, S., Arnrich, B., Troster, G., Meyer, B., and Jonas, K. (2011). Detecting posture mirroring in social interactions with wearable sensors. *Wearable Computers, IEEE International Symposium*.

[Felzenszwalb and Huttenlocher, 2000] Felzenszwalb, P. and Huttenlocher, D. (2000). Efficient matching of pictorial structures. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Felzenszwalb et al., 2008] Felzenszwalb, P., McAllester, D., and Ramanan, D. (2008). A discriminatively trained, multiscale, deformable part model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Ferrari et al., 2009] Ferrari, V., Marin-Jimenez, M., and Zisserman, A. (2009). Pose search: Retrieving people using their pose. In *CVPR*.

[Fischer et al., 2003] Fischer, D., Williams, M., and Andriacchi, T. (2003). The therapeutic potential for changing patters of locomotion. In *ASME Bioengineering Conference*.

[Fossati et al., 2010] Fossati, A., Dimitrijevic, M., Lepetit, V., and Fua, P. (2010). From canonical poses to 3d motion capture using a single camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 32(7).

[Fragkiadaki et al., 2013] Fragkiadaki, K., Hu, H., and Shi, J. (2013). Pose from flow and flow from pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Frank et al., 2012] Frank, M., Movellan, J., Bartlett, M., , and Littleworth, G. (2012). Ru-facs-1 database, machine perception laboratory, u.c. san diego, 2012. http://mplab.ucsd.edu/grants/project1/research/rufacs1-dataset.html.

[Freifeld et al., 2010] Freifeld, O., Weiss, A., Zuffi, S., and Black, M. J. (2010). Contour people: A parameterized model of 2D articulated human shape. In *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*.

[Gall et al., 2011] Gall, J., Fossati, A., and van Gool, L. (2011). Functional categorization of objects using real-time markerless motion capture. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Gall et al., 2009] Gall, J., Stoll, C., de Aguiar, E., Theobalt, C., Rosenhahn, B., and Seidel, H.-P. (2009). Motion capture using joint skeleton tracking and surface estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Gall et al., 2012] Gall, J., Yao, A., and Gool, L. J. V. (2012). 2d action recognition serves 3d human pose estimation. In *European Conference on Computer Vision (ECCV)*.

[Galton, 1892] Galton, F. (1892). *Finger Prints*. Macmillan.

[Ganapathi et al., 2010] Ganapathi, V., Plagemann, C., Koller, D., and Thrun, S. (2010). Real time motion capture using a single time-of-flight camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Gatica-Perez, 2009] Gatica-Perez, D. (2009). Automatic nonverbal analysis of social interaction in small groups: A review. *IVC, Special Issue on Human Behavior*.

[Gavrila and Davis, 1996] Gavrila, D. M. and Davis, L. S. (1996). 3-d model-based tracking of humans in action: a multi-view approach.

[Gifford et al., 1985] Gifford, R., Ng, C. F., and Wilkinson, M. (1985). Nonverbal cues in the employment interview: Links between applicant qualities and interviewer judgments. *Applied Psychology*, 70(4):729–736.

[Gijsenij et al., 2011] Gijsenij, A., Gevers, T., and van de Weijer, J. (2011). Computational color constancy: Survey and experiments. *IEEE Transactions on Image Processing (IP)*.

[Gkioxari et al., 2013] Gkioxari, G., Arbelaez, P., Bourdev, L., and Malik, J. (2013). Articulated pose estimation using discriminative armlet classifiers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Gosling, 2003] Gosling, S. (2003). A very brief measure of the Big-Five personality domains. *Research in Personality*, 37(6):504–528.

[Grochow et al., 2004] Grochow, K., Martin, S. L., Hertzmann, A., and Popović, Z. (2004). Style-based inverse kinematics. In *ACM Conference and Exhibition on Computer Graphics and Interactive Techniques (SIGGRAPH).*

[Guan et al., 2010] Guan, P., Freifeld, O., and Black, M. J. (2010). A 2D human body model dressed in eigen clothing. In *European Conference on Computer Vision, (ECCV).*

[Hara and Chellappa, 2013] Hara, K. and Chellappa, R. (2013). Computationally efficient regression on a dependency graph for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[Hernandez-Vela et al., 2012] Hernandez-Vela, A., Zlateva, N., Marinov, A., Reyes, M., Radeva, P., Dimov, D., and Escalera, S. (2012). Graph cuts optimization for multi-limb human segmentation in depth maps. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[Hoerl and Kennard, 1970] Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12.

[Holt et al., 2013] Holt, B., Ong, E.-J., and Bowden, R. (2013). Accurate static pose estimation combining direct regression and geodesic extrema. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG).*

[Howard and Ferris, 1996] Howard, J. and Ferris, G. (1996). The Employment Interview Context: Social and Situational Influences on Interviewer Decisions. *Journal of Applied Social Psychology*, 26(2):112–136.

[Howe, 2005] Howe, N. R. (2005). Flow lookup and biological motion perception. In *International Conference on Image Processing (ICIP).*

[Imada and Hakel, 1977] Imada, A. S. and Hakel, M. D. (1977). Influence of Nonverbal Communication and Rater Proximity on Impressions and Decisions in Simulated Employment Interviews responsibilities that produce different re-. *Journal of Applied Psychology*, 62(3):295–300.

[Isard and Blake, 1998] Isard, M. and Blake, A. (1998). A mixed-state condensation tracker with automatic model-switching. In *IEEE International Conference on Computer Vision (ICCV).*

[Jolliffe, 1986] Jolliffe, I. (1986). *Principal Component Analysis.*

[Kazemi et al., 2013] Kazemi, V., Burenius, M., Azizpour, H., and Sullivan, J. (2013). Multi-view body part recognition with random forests. In *British Machine Vision Conference (BMVC).*

[Kjellstrom et al., 2010] Kjellstrom, H., Kragic, D., and Black, M. J. (2010). Tracking people interacting with objects. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[Knapp and Hall, 2009] Knapp, M. and Hall, J. (2009). *Nonverbal Communication in Human Interaction.*

[Knoop et al., 2006] Knoop, S., Vacek, S., and Dillmann, R. (2006). Sensor fusion for 3d human body tracking with an articulated 3d body model. In *IEEE International Conference on Robotics and Automation (ICRA).*

[Kulic et al., 2009] Kulic, D., Lee, D., and Nakamura, Y. (2009). Whole body motion primitive segmentation from monocular video. In *IEEE International Conference on Robotics and Automation (ICRA).*

[Kuo et al., 2011] Kuo, P., Makris, D., and Nebel, J.-C. (2011). Integration of bottom-up/top-down approaches for 2d pose estimation using probabilistic gaussian modelling. *Computer Vision and Image Understanding (CVIU)*, 115(2).

[Ladicky et al., 2013] Ladicky, L., Torr, P. H. S., and Zisserman, A. (2013). Human pose estimation using a joint pixel-wise and part-wise formulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[Laptev, 2005] Laptev, I. (2005). On space-time interest points. *International Journal of Computer Vision (IJCV).*

[Lawrence, 2005] Lawrence, N. (2005). Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *The Journal of Machine Learning Research (JMLR)*, 6.

[Lawrence et al., 2003] Lawrence, N. D., Seeger, M., and Herbrich, R. (2003). Fast sparse gaussian process methods: The informative vector machine. In *Advances in Neural Information Processing Systems 15.*

[Liu et al., 2009] Liu, J., Luo, J., and Shah, M. (2009). Recognizing realistic actions from videos "in the wild". *IEEE International Conference on Computer Vision and Pattern Recognition.*

[Liu and Prakash, 2003] Liu, Q. and Prakash, E. (2003). The parameterization of joint rotation with the unit quaternion. In *International Conference on Digital Image Computing: Techniques and Applications (DICTA).*

[Liu et al., 2011] Liu, Y., Stoll, C., Gall, J., Seidel, H.-P., and Theobalt, C. (2011). Markerless motion capture of interacting characters using multi-view image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[Livne et al., 2012] Livne, M., Sigal, L., Troje, N. F., and Fleet, D. J. (2012). Human attributes from 3d pose tracking. *Computer Vision and Image Understanding (CVIU)*, 116(5).

[Lopez-Mendez et al., 2011] Lopez-Mendez, A., Alcoverro, M., Pardas, M., and Casas, J. R. (2011). Real-time upper body tracking with online initialization using a range sensor. In *Workshops of International Conference on Computer Vision (ICCV).*

[Lu et al., 2012] Lu, H., Rabbi, M., Chittaranjan, G. T., Frauendorfer, D., Schmid Mast, M., Campbell, A. T., Gatica-Perez, D., and Choudhury, T. (2012). StressSense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proc. Int. Conf. on Ubiquitous Computing.*

[Lucas and Kanade, 1981] Lucas, B. D. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *International Joint Conference on Artificial Intelligence (IJCAI).*

[Marcos et al., 2011] Marcos, A., Marron, M., and Pizarro, D. (2011). Captura de movimiento de multiples personas mediante gplvm y un pf mixto. In *Seminario Anual de Automatica, Electronica Industrial e Instrumentacion (SAAEI).*

[Marcos et al., 2013] Marcos, A., Marron, M., Pizarro, D., and Mazo, M. (2013). Captura de movimiento y reconocimiento de actividades para multiples personas mediante un enfoque bayesiano. *Revista Iberoamericana de Automatizacion Industrial (RIAI)*, 10(2).

[Marcos-Ramiro et al., 2013] Marcos-Ramiro, A., Pizarro-Perez, D., Marron-Romera, M., Nguyen, L., and Gatica-Perez, D. (2013). Body communicative cue extraction for conversational analysis. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG).*

[Marinoiu et al., 2013] Marinoiu, E., Papava, D., and Sminchisescu, C. (2013). Pictorial Human Spaces. How Well do Humans Perceive a 3D Articulated Pose? In *IEEE International Conference on Computer Vision (ICCV)*.

[Marquardt, 1963] Marquardt, D. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, 11(2).

[Marron et al., 2010a] Marron, M., Garcia, J. C., Sotelo, M. A., Pizarro, D., Mazo, M., Canas, J. M., Losada, C., and Marcos, A. (2010a). Stereo vision tracking of multiple objects in complex indoor environments. *Sensors*, 10(10).

[Marron et al., 2010b] Marron, M., Pizarro, D., Marcos, A., Jalvo, R., Garcia, J., and Mazo, M. (2010b). Tracking multiple agents in an intelligent space with probabilistic algorithms and a camera ring. In *IEEE International Symposium on Industrial Electronics (ISIE)*.

[Matthias Straka and Bischof, 2011] Matthias Straka, Stefan Hauswiesner, M. R. and Bischof, H. (2011). Skeletal graph based human pose estimation in real-time. In *British Machine Vision Conference (BMVC)*.

[McNeill, 1985] McNeill, D. (1985). So You Think Gestures Are Nonverbal? *Psychological Review*, 92(3):350–371.

[McNeill, 1992] McNeill, D. (1992). *Hand and Mind*.

[McNeill, 2005] McNeill, D. (2005). *Gesture and Thought*.

[Mehrabian, 1972] Mehrabian, A. (1972). *Nonverbal Communication*.

[Michoud et al., 2007] Michoud, B., Guillou, E., and Bouakaz, S. (2007). Real-time and Markerless 3D Human Motion Capture using Multiple Views. In *Wokrshop on Human Motion (HM)*.

[Mihalcea and Burzo, 2012] Mihalcea, R. and Burzo, M. (2012). Towards multimodal deception detection – step 1: Building a collection of deceptive videos. In *ACM International Conference on Multimodal Interaction (ICMI)*.

[Mitra and Acharya, 2007] Mitra, S. and Acharya, T. (2007). Gesture recognition: A survey. *IEEE System, Man and Cybernetics (SMC)*.

[Moeslund et al., 2006] Moeslund, T. B., Hilton, A., and Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding (CVIU)*, 104(2).

[Morariu et al., 2013] Morariu, V., Harwood, D., and Davis, L. (2013). Tracking people's hands and feet using mixed network and/or search. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35.

[Morency et al., 2008] Morency, L.-P., de, I. K., and Gratch, J. (2008). Context-based recognition during human interactions: Automatic feature selection and encoding dictionary. In *ACM International Conference on Multimodal Interaction (ICMI)*.

[Muller et al., 2005] Muller, M., Roder, T., and Clausen, M. (2005). Efficient content-based retrieval of motion capture data. In *ACM Conference and Exhibition on Computer Graphics and Interactive Techniques (SIGGRAPH)*.

[Naumann et al., 2009] Naumann, L. P., Vazire, S., Rentfrow, P. J., and Gosling, S. D. (2009). Personality judgments based on physical appearance. *Personality & social psychology bulletin*, 35(12):1661–71.

[Neff et al., 2010] Neff, M., Wang, Y., Abbott, R., and Walker, M. (2010). Evaluating the effect of gesture and language on personality perception in conversational agents. *Intelligent Virtual Agents*.

[Nguyen et al., 2013a] Nguyen, L., Frauendorfer, D., Schmid Mast, M., and Gatica-Perez, D. (2013a). Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior. Technical report, Idiap Research Institute.

[Nguyen et al., 2013b] Nguyen, L., Marcos, A., Marron, M., and Gatica, D. (2013b). Multimodal analysis of body communication cues in employment interviews. In *ACM International Conference on Multimodal Interaction (ICMI)*.

[Nguyen et al., 2012] Nguyen, L. S., Odobez, J.-M., and Gatica-Perez, D. (2012). Using self-context for multimodal detection of head nods in face-to-face interactions. In *ACM International Conference on Multimodal Interaction (ICMI)*.

[Noriega and Bernier, 2007] Noriega, P. and Bernier, O. (2007). Multicues 3d monocular upper body tracking using constrained belief propagation. In *British Machine Vision Conference (BMVC)*.

[O'Connor and Robertson, 1908] O'Connor, J. and Robertson, E. (1908). *Student's t-test*. MacTutor History of Mathematics archive, University of St Andrews.

[Orozco et al., 2013] Orozco, J., Rudovic, O., Gonzalez, J., and Pantic, M. (2013). Hierarchical on-line appearance-based tracking for 3d head pose, eyebrows, lips, eyelids and irises. *Image and Vision Computing*, 31(4).

[Pelleg and Moore, 2000] Pelleg, D. and Moore, A. W. (2000). X-means: Extending k-means with efficient estimation of the number of clusters. In *International Conference on Machine Learning (ICML)*.

[Pentland, 2008] Pentland, A. (2008). *Honest Signals: How They Shape Our World*.

[Pianesi et al., 2008] Pianesi, F., Mana, N., Cappelletti, A., Lepri, B., and Zancanaro, M. (2008). Multimodal recognition of personality traits in social interactions. In *Proc. Int. Conf. on Multimodal Interactions*, pages 53–60, New York, New York, USA. ACM Press.

[Plagemann et al., 2010] Plagemann, C., Ganapathi, V., Koller, D., and Thrun, S. (2010). Real-time identification and localization of body parts from depth images. In *IEEE International Conference on Robotics and Automation (ICRA)*.

[Poppe, 2010] Poppe, R. (2010). A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6).

[Quek et al., 2002] Quek, F., Bryll, R., Kirbas, C., Arslan, H., and McNeill, D. (2002). A multimedia system for temporally situated perceptual psycholinguistic analysis. *Multimedia Tools and Applications*, 18(2).

[Ramakrishna et al., 2013] Ramakrishna, V., Kanade, T., and Sheikh, Y. (2013). Tracking human pose by tracking symmetric parts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Ramanan and Forsyth, 2003] Ramanan, D. and Forsyth, D. A. (2003). Finding and tracking people from the bottom up. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Sadanand and Corso, 2012] Sadanand, S. and Corso, J. J. (2012). Action bank: A high-level representation of activity in video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[Sanchez-Cortes et al., 2011] Sanchez-Cortes, D., Aran, O., Schmid Mast, M., and Gatica-Perez, D. (2011). A nonverbal behavior approach to identify emergent leaders in small groups. *IEEE Transactions on Multimedia (TM).*

[Sapp and Taskar, 2013] Sapp, B. and Taskar, B. (2013). Modec: Multimodal decomposable models for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[Sapp et al., 2011] Sapp, B., Weiss, D., and Taskar, B. (2011). Parsing human motion with stretchable models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[Scheffler and Odobez, 2011] Scheffler, C. and Odobez, J.-M. (2011). Joint adaptive colour modelling and skin, hair and clothes segmentation using coherent probabilistic index maps. In *British Machine Vision Conference (BMVC).*

[Scherer et al., 2013] Scherer, S., Stratou, G., Mahmound, M., Boberg, J., Gratch, J., Rizzo, A., and Morency, L.-P. (2013). Automatic behavior descriptors for psychological disorder analysis. In *IEEE Conference on Automatic Face and Gesture Recognition*, Shanghai, China.

[Schwarz et al., 2010] Schwarz, L. A., Mateus, D., Castaneda, V., and Navab, N. (2010). Manifold learning for tof-based human body tracking and activity recognition. In *British Machine Vision Conference (BMVC).*

[Schwarz et al., 2011] Schwarz, L. A., Mkhitaryan, A., Mateus, D., and Navab, N. (2011). Estimating human 3d pose from time-of-flight images based on geodesic distances and optical flow. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG).*

[Shaheen et al., 2009] Shaheen, M., Gall, J., Strzodka, R., Gool, L. J. V., and Seidel, H.-P. (2009). A comparison of 3d model-based tracking approaches for human motion capture in uncontrolled environments. In *IEEE Workshop on Applications of Computer Vision (WACV).*

[Shotton et al., 2012] Shotton, J., Girshick, R., Fitzgibbon, A., Sharp, T., Cook, M., Finocchio, M., Moore, R., Kohli, P., Criminisi, A., Kipman, A., and Blake, A. (2012). Efficient human pose estimation from single depth images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 99.

[Shotton et al., 2011] Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A. W., Finocchio, M., Blake, A., Cook, M., and Moore, R. (2011). Real-time human pose recognition in parts from single depth images. In *CVPR.*

[Sidenbladh et al., 2000] Sidenbladh, H., Black, M. J., and Fleet, D. J. (2000). Stochastic tracking of 3d human figures using 2d image motion. In *IEEE European Conference on Computer Vision (ECCV).*

[Sigal et al., 2010] Sigal, L., Balan, A., and Black, M. (2010). Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision (IJCV)*, 87(1).

[Sigal et al., 2007] Sigal, L., Balan, A. O., and Black, M. J. (2007). Combined discriminative and generative articulated pose and non-rigid shape estimation. In *Conference on Neural Information Processing Systems (NIPS).*

[Sigal et al., 2004] Sigal, L., Bhatia, S., Roth, S., Black, M. J., and Isard, M. (2004). Tracking loose-limbed people. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Sigal and Black, 2010a] Sigal, L. and Black, M. (2010a). Guest editorial: State of the art in image- and video-based human pose and motion estimation. *International Journal on Computer Vision (IJCV)*.

[Sigal and Black, 2010b] Sigal, L. and Black, M. J. (2010b). Guest editorial: State of the art in image- and video-based human pose and motion estimation. *International Journal of Computer Vision (IJCV)*, 87(1-2).

[Simo-Serra et al., 2013] Simo-Serra, E., Quattoni, A., Torras, C., and Moreno-Noguer, F. (2013). A joint model for 2d and 3d pose estimation from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Song et al., 2001] Song, Y., Goncalves, L., and Perona, P. (2001). Learning probabilistic structure for human motion detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Stoll et al., 2011] Stoll, C., Hasler, N., Gall, J., Seidel, H.-P., and Theobalt, C. (2011). Fast articulated motion tracking using a sums of gaussians body model. In *International Conference on Computer Vision (ICCV)*.

[Sun et al., 2012] Sun, M., Telaprolu, M., Lee, H., and Savarese, S. (2012). An efficient branch-and-bound algorithm for optimal human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Taylor et al., 2012] Taylor, J., Shotton, J., Sharp, T., and Fitzgibbon, A. W. (2012). The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Theobalt et al., 2004] Theobalt, C., Carranza, J., Magnor, M. A., and Seidel, H.-P. (2004). Combining 3d flow fields with silhouette-based human motion capture for immersive video. *Graphical Models (GM)*, 66(6).

[Tomasi and Kanade, 1991] Tomasi, C. and Kanade, T. (1991). Detection and tracking of point features. *International Journal of Computer Vision (IJCV)*.

[Triggs, 2005] Triggs, W. (2005). Monocular human motion capture and other works. Technical report, Institut National Polytechnique de Grenoble.

[Tung and Matsuyama, 2008] Tung, T. and Matsuyama, T. (2008). Human Motion Tracking using a Color-Based Particle Filter Driven by Optical Flow. In *International Workshop on Machine Learning for Vision-based Motion Analysis (MLVMA)*.

[Ukita et al., 2009] Ukita, N., Hirai, M., and Kidode, M. (2009). Complex volume and pose tracking with probabilistic dynamical models and visual hull constraints. In *International Conference on Computer Vision (ICCV)*.

[Urtasun Sotil, 2006] Urtasun Sotil, R. (2006). *Motion models for robust 3D human body tracking*. PhD thesis, Lausanne.

[Viola and Jones, 2001] Viola, P. A. and Jones, M. J. (2001). Robust real-time face detection. In *International Journal of Computer Vision (ICCV)*.

[Vondrak et al., 2008] Vondrak, M., Sigal, L., and Jenkins, O. C. (2008). Physical simulation for probabilistic motion tracking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Wang et al., 2013] Wang, C., Wang, Y., and Yuille, A. L. (2013). An approach to pose-based action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Wang and Koller, 2011] Wang, H. and Koller, D. (2011). Multi-level inference by relaxed dual decomposition for human pose segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Wang et al., 2006] Wang, J., Fleet, D., and Hertzmann, A. (2006). Gaussian process dynamical models. *Advances in neural information processing systems (NIPS)*, 18.

[Wiesner and Cronshaw, 1988] Wiesner, W. H. and Cronshaw, S. F. (1988). A meta-analytic investigation of the impact of interview format and degree of structure on the validity of the lemployment interview. *Journal of Occupational Psychology*, 61(4):275–290.

[Winters, 2005] Winters, A. (2005). Perceptions of Body Posture and Emotion: A Question of Methodology. *The New School Psychology Bulletin*, 3(2):35–45.

[Wu et al., 2012a] Wu, C., Varanasi, K., and Theobalt, C. (2012a). Full body performance capture under uncontrolled and varying illumination: A shading-based approach. In *European Conference on Computer Vision (ECCV)*.

[Wu et al., 2012b] Wu, D., Liu, Y., Ihrke, I., Dai, Q., and Theobalt, C. (2012b). Performance capture of high-speed motion using staggered multi-view recording. *Computer Graphics Forum (CGF)*, 31(7-1).

[Xiong and Quek, 2006] Xiong, Y. and Quek, F. (2006). Hand motion gesture frequency properties and multimodal discourse analysis. *International Journal of Computer Vision (IJCV)*, 69(3).

[Yang and Ramanan, 2011] Yang, Y. and Ramanan, D. (2011). Articulated pose estimation with flexible mixtures-of-parts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Yao et al., 2012] Yao, A., Gall, J., and Gool, L. J. V. (2012). Coupled action recognition and pose estimation from multiple views. *International Journal of Computer Vision (IJCV)*, 100(1).

[Yao et al., 2011] Yao, A., Gall, J., Gool, L. V., and Urtasun, R. (2011). Learning probabilistic non-linear latent variable models for tracking complex activities. In *Advances in Neural Information Processing Systems*.

[Ye et al., 2013] Ye, M., Zhang, Q., Wang, L., Zhu, J., Yang, R., and Gall, J. (2013). A survey on human motion analysis from depth data. In *Tutorial in Computer Vision and Pattern Recognition*.

[Yin and Davis, 2013] Yin, Y. and Davis, R. (2013). Gesture spotting and recognition using salience detection and concatenated hidden markov models. In *ACM International Conference on Multimodal Interaction (ICMI)*.

[Yoonessi and Baker, 2011] Yoonessi, A. and Baker, C. L. (2011). Contribution of motion parallax to segmentation and depth perception. *Journal of Vision*, 11.

[Yu et al., 2013a] Yu, T.-H., Kim, T.-K., and Cipolla, R. (2013a). Unconstrained monocular 3d human pose estimation by action detection and cross-modality regression forest. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Yu et al., 2013b] Yu, X., Zhang, S., Yu, Y., Dunbar, N., Jensen, M., Burgoon, J., and Metaxas, D. (2013b). Automated analysis of interactional synchrony using robust facial tracking and expression recognition. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*.

[Zhang et al., 2008] Zhang, L., Chen, J., Zeng, Z., and Ji, Q. (2008). 2d and 3d upper body tracking with one framework. In *International Conference on Pattern Recognition (ICPR).*

[Zhu and Fujimura, 2007] Zhu, Y. and Fujimura, K. (2007). Constrained optimization for human pose estimation from depth sequences. In *Asian Conference on Computer Vision (ACCV).*

[Zuffi et al., 2012] Zuffi, S., Freifeld, O., and Black, M. J. (2012). From pictorial structures to deformable structures. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[Zuffi et al., 2013] Zuffi, S., Romero, J., Schmid, C., and Black, M. J. (2013). Estimating human pose with flowing puppets. In *IEEE International Conference on Computer Vision (ICCV).*