Universidad de Alcalá

PhD. Program in Information and Communications Technologies

Pedestrian Path, Intention and Pose Prediction Through Gaussian Process Dynamical Models and Pedestrian Activity Recognition

PhD. Thesis presented by

Raúl Quintero Mínguez

2017

Universidad de Alcalá

PhD. Program in Information and Communications Technologies

Pedestrian Path, Intention and Pose Prediction Through Gaussian Process Dynamical Models and Pedestrian Activity Recognition

PhD. Thesis presented by

Raúl Quintero Mínguez

Advisors

Dr. Miguel Ángel Sotelo Vázquez

Dr. Ignacio Parra Alonso

Alcalá de Henares, DAYth of MONTH, 2017

Universidad de Alcalá

Campus Universitario
Dpto. de Automática
Ctra. Madrid-Barcelona, Km. 36.6
28805 Alcalá de Henares (Madrid)
Telf: +34 91 885 65 94

D. MIGUEL ÁNGEL SOTELO VÁZQUEZ y D. IGNACIO PARRA ALONSO, Profesor Catedrático de Universidad y Profesor Ayudante Doctor respectivamente del Área de Conocimiento de Ingeniería de Sistemas y Automática de la Universidad de Alcalá,

CERTIFICAN

Que la tesis "**Pedestrian Path, Intention and Pose Prediction Through Gaussian Process Dynamical Models and Pedestrian Activity Recognition**", presentada por D. Raúl Quintero Mínguez, realizada en el Departamento de Automática bajo nuestra dirección, reúne méritos suficientes para optar al grado de Doctor, por lo que puede procederse a su depósito y lectura.

Alcalá de Henares, 25 de marzo de 2017.

Fdo.: Dr. D. Miguel Ángel Sotelo Vázquez Fdo.: Dr. D. Ignacio Parra Alonso

Universidad de Alcalá

Campus Universitario
Dpto. de Automática
Ctra. Madrid-Barcelona, Km. 36.6
28805 Alcalá de Henares (Madrid)
Telf: +34 91 885 65 94

D. Raúl Quintero Mínguez ha realizado en el Departamento de Automática y bajo la dirección del Dr. D. Miguel Ángel Sotelo Vázquez y del Dr. D. Ignacio Parra Alonso, la tesis doctoral titulada "**Pedestrian Path, Intention and Pose Prediction Through Gaussian Process Dynamical Models and Pedestrian Activity Recognition**", cumpliéndose todos los requisitos para la tramitación que conduce a su posterior lectura.

Alcalá de Henares, 25 de marzo de 2017.

EL COORDINADOR DEL PROGRAMA DE DOCTORADO

Fdo: Dr. D. Sancho Salcedo Sanz.

A todos aquellos que hicieron posible esta tesis.

"A computer would deserve to be called intelligent if it could deceive a human into believing that it was human."
Alan Mathison Turing (1912 - 1954)

Agradecimientos

A lo largo de los años, las personas van cumpliendo hitos que marcan el devenir de sus vidas. A veces, éstos son personales y otras veces profesionales. Sin embargo, la culminación de esta tesis doctoral supone para mi un hito en ambos aspectos. Profesionalmente significa un gran avance en mi carrera que abre nuevas etapas que a bien seguro serán exitosas. En lo personal me permite reafirmar que el esfuerzo, la constancia, el afán de conocimiento y la pasión por lo que uno hace son clave para conseguir resultados y satisfacciones. Durante el tiempo que he dedicado a la tesis doctoral no sólo he conocido a personas brillantes en sus campos, sino lugares únicos a los que tal vez nunca vuelva. Las experiencias que uno vive en los congresos internacionales o durante las estancias en centros de investigación extranjeros también marcan el éxito o fracaso personal que uno consigue al final de una tesis doctoral.

Sin duda, este hito profesional y personal no podría haberse cumplido sin la estimable ayuda de mis dos directores de tesis. Por ello, quiero dar mi más sincero agradecimiento al Dr. Miguel Ángel Sotelo Vázquez y al Dr. Ignacio Parra Alonso. Sus consejos, recomendaciones e ideas siempre fueron encaminados a lograr unos resultados de los que no sólo yo me pueda sentir orgulloso, sino ellos mismos también. Trabajar con ambos en esta tesis doctoral, además de en otros muchos proyectos, me ha permitido crecer profesionalmente y obtener mÃ¡s experiencia investigadora que seguro me servirá en el futuro.

Por otro lado, a pesar de no haber sido parte de esta tesis, me gustaría agradecer al Dr. David Fernández Llorca las oportunidades que me ha dado a lo largo de estos años. Trabajar conjuntamente en muchos y variados proyectos, desde el comienzo de mi Trabajo Fin de Carrera hasta hoy, siempre ha sido un placer. Los conocimientos que he adquirido gracias a él se ven reflejados indirectamente en esta tesis.

También me gustaría agradecer sus consejos y su tiempo al resto de miembros

presentes y pasados de los grupos INVETT y ROBESAFE de la Universidad de Alcalá. No daré sus nombres por no olvidarme de ninguno. Para mí, todos ellos han sido importantes a lo largo de estos años. Más destacable que la ayuda que puedan prestarte puntualmente con sus ideas, es el extraordinario ambiente de trabajo que hemos sido capaces de crear entre todos. Este ambiente no sólo se ve dentro del laboratorio, sino fuera de él. Espero que las comidas, las cenas y las rutas continúen sin importar donde nos encontremos cada uno de nosotros.

Tampoco quiero olvidarme en estas páginas del Dr. Eduardo Nebot. Gracias por darme la oportunidad de realizar una estancia de tres meses en el Australian Centre for Field Robotics. La experiencia de vivir en Sydney, así como trabajar allí, fue inolvidable. También quiero agradecer la oportunidad que me dio el Dr. Neil Lawrence de realizar otra estancia de tres meses en el grupo de Machine Learning del Sheffield Institute for Translational Neuroscience.

No podía continuar estas palabras sin acordarme de mi familia. Mis padres, Julián y Ángeles, y mis hermanos, Rubén y Sergio, siempre me han apoyado y preocupado por lo que hacía. El esfuerzo que mis padres siempre han realizado por sacarnos adelante se ve reflejado en esta tesis doctoral. Espero que se sientan tan orgullosos de este hito como yo. Tampoco quiero olvidarme de mis abuelos, ojalá vivieran para ver el final de esta etapa. Ellos también son partícipes de lo que soy y he conseguido.

Finalmente, me gustaría agradecer el apoyo que mis amigos siempre me han ofrecido. Nuevamente, no les nombraré por no olvidarme de ninguno. Sin embargo, seguro que cada uno de ellos se sentirá aludido cuando lea estás palabras. Espero seguir contando con todos en el futuro.

Ojalá que esta tesis doctoral sea de interés para cualquier lector que la tenga entre sus manos. En ella he dedicado mucho tiempo que robé a personas y aficiones que son muy importantes para mí. Espero que todas ellas sepan perdonarme.

Resumen

Debido al elevado número de muertes en carretera, a lo largo de los últimos años los vehículos han ido evolucionando hasta llegar a ser máquinas inteligentes con tecnologías avanzadas tales como Sistemas de Protección de Peatones, Sistemas de Frenado Automático de Emergencia u otro tipo de Sistemas Avanzados de Asistencia al Conductor. Mejorar estos avances tecnológicos es imprescindible ya que iniciar la frenada lo antes posible o evaluar de forma precisa las posiciones de los peatones antes de una colisión podrían ser tareas particularmente relevantes como aseguran varios trabajos.

Esta tesis describe un método basado en Balanced Gaussian Process Dynamical Models (B-GPDMs), los cuales aprenden información tridimensional y temporal procedente de diferentes puntos situados a lo largo de los cuerpos de los peatones con el objetivo de predecir sus trayectorias, posturas e intenciones futuras con una antelación de hasta un 1 segundo. Dado que los humanos no son objetos rígidos, es importante analizar el movimiento de cada parte del cuerpo. Por tanto, la información de los puntos sobre el peatón es significativamente valiosa a la hora de llevar a cabo dichas tareas. El B-GPDM permite reducir la dimensionalidad de un conjunto de vectores de características relacionados en el tiempo e inferir posiciones latentes futuras. Asimismo, el correspondiente vector de características puede ser reconstruido dada la posición en el espacio latente. Sin embargo, el aprendizaje de un único modelo genérico para todo tipo de actividades peatonales o la combinación de algunas de ellas en un único modelo normalmente produce estimaciones imprecisas de las observaciones futuras. Por esa razón, el método propuesto aprende múltiples modelos de cada tipo de actividad del peatón, éstas son: andando, parando, comenzando a andar y parado, y selecciona el modelo más apropiado en cada instante de tiempo con el objetivo de estimar estados de peatones futuros. El método funciona como sigue: dado un conjunto de entrenamiento compuesto de secuencias de movimientos de peatones, éste es dividido en 8 subconjuntos basándose en la orientación de cruce, ya sea, de izquierda a derecha o de derecha a izquierda, y tipo

de actividad. A continuación, se obtiene un B-GPDM por cada secuencia contenida en el conjunto de entrenamiento. Por otro lado, dada una nueva observación de un peatón, su actividad es determinada por medio de un algoritmo de reconocimiento de actividades basado en un Modelo Oculto de Markov. Así, la selección del modelo más adecuado entre todos los entrenados se realiza entre los pertenecientes a esa actividad. Finalmente, el modelo escogido se utiliza para predecir posiciones latentes futuras y, a partir de ahí, reconstruir las trayectorias y las posturas.

Los resultados verifican que la información de los hombros y las piernas es más valiosa que la información procedente de otras partes del cuerpo cuando se trata de reconocer la acción del peatón. Más específicamente, la mayor exactitud, 95.13%, se logra cuando las observaciones están compuestas de unos pocos puntos situados a lo largo de las piernas y los hombros. Sin embargo, esta exactitud cae hasta el 90.69% si se utilizan un mayor número de puntos localizados a lo largo de todo el cuerpo. El método propuesto en este documento detecta intenciones de comenzar a andar 125 milisegundos después de la iniciación del paso con una exactitud del 80% y reconoce intenciones de parado 58.33 milisegundos antes del evento con una exactitud del 70% cuando se utilizan únicamente puntos de los hombros y las piernas.

En cuanto a la predicción de las trayectorias, se han obtenido errores similares a otros trabajos. Sin embargo, algunas medidas de exactitud utilizas por otro métodos ofrecen a una idea confusa de cómo de bien funciona un sistema. Por ejemplo, la Mean Euclidean Distance (MED) da una interpretación física más precisa sobre las posiciones predichas de los peatones con respecto a la realidad que el Root Mean Squared Error (RMSE). Por tanto, en esta tesis, la medida de exactitud escogida para la evaluación de la trayectoria futura es la MED a diferentes Times To Event (TTEs) ya que ofrece información objetiva del rendimiento de la predicción de la trayectoria. Para actividades de andar, se han obtenido valores de MEDs a 0.25, 0.5, 0.75 y 1 segundos de 33.03±43.84, 70.87±89.69, 113.34±140.64 y 159.48±196.19 milímetros respectivamente. Para acciones de parando, el valor de MED es 238.01±206.93 milímetros para un TTE de 1 segundo y un horizonte temporal de 1 segundo. Finalmente, para acciones de comenzando a andar, se ha obtenido un valor de MED de 331.93±254.73 milímetros para un TTE de 0 segundos y un horizonte temporal de 1 segundo.

Palabras clave: Peatones, predicción, trayectorias, actividades, modelos.

Abstract

Because of the high number of fatalities on the road, during the last few years vehicles have been evolving to become intelligent machines with advanced technologies such as Pedestrian Protection Systems, Automatic Emergency Braking Systems (AEBSs) or other sort of Advanced Driver Assistance Systems (ADAS). Improving these technological advances is imperative since an early braking initiation or an accurate assessment about pedestrian positions before collisions could be particularly relevant as some works assert.

This thesis describes a method based on Balanced Gaussian Process Dynamical Models (B-GPDMs), which learns 3D time-related information from joints placed along the pedestrian bodies in order to predict their future paths, poses and intentions up to 1 second in advance. Given that humans are not rigid objects, the motion analysis of each body part should be taken into account. Hence, pedestrian joints are valuable information to perform these tasks. The B-GPDM can reduce the dimensionality of a set of feature vectors related in time and infer future latent positions. Likewise, given a latent position from the latent space, the corresponding feature vector can also be reconstructed. However, learning a generic model for all kind of pedestrian activities or combining some of them into a single model normally provides inaccurate estimations of future observations. For that reason, the proposed method learns multiple models of each type of pedestrian activity, i.e. walking, stopping, starting and standing, and selects the most appropriate among them to estimate future pedestrian states at each instant of time. The method works as follows: given a training dataset of pedestrian motion sequences, this is split into 8 subsets based on typical crossing orientations, that is, from left to right and from right to left, and type of activity. Then, a B-GPDM is obtained for each sequence contained in the dataset. On the other hand, given a new pedestrian observation, the current activity is determined by means of an activity recognition algorithm based on a Hidden Markov Model (HMM). Thus, the selection of the most appropriate model among the trained ones is centred solely on that activ-

ity. Finally, the selected model is used to predict the future latent positions and reconstruct the future pedestrian path and poses.

The results verify that shoulder and leg motions are more valuable sources of information than other body parts to recognise the current pedestrian action. More specifically, the maximum accuracy rate, 95.13%, is achieved when observations composed of a few joints placed along the legs and shoulders are taken into consideration. However, the accuracy rate falls to 90.69% whether a higher number of joints located along the whole body are used. The method proposed in this document detects starting intentions 125 milliseconds after the gait initiation with an accuracy rate of 80% and recognises stopping intentions 58.33 milliseconds before the event with an accuracy rate of 70% when joints from shoulders and legs are considered.

Concerning the path prediction results, similar errors are obtained with respect to other works. However, some measures of accuracy used by other methods provide a vague idea of how well a system works. For example, the MED gives a more precise physical interpretation of the predicted pedestrian positions with respect to a groundtruth than the RMSE. Hence, in this thesis, the measure of accuracy chosen for the path evaluation is the MED at different TTEs since it gives objective information of the path prediction performance. The MEDs achieved for walking activities at 0.25, 0.5, 0.75 and 1 second are 33.03±43.84, 70.87±89.69, 113.34±140.64 and 159.48±196.19 millimetres respectively. For stopping activities, a MED value of 238.01±206.93 millimetres for a TTE of 1 second and a time horizon of 1 second. Finally, for a starting action, the method described in this thesis achieves a MED value of 331.93±254.73 millimetres for a TTE of 0 seconds and a time horizon of 1 second.

Keywords: Pedestrians, prediction, paths, activities, models.

Contents

List of Figures

List of Tables

List of Acronyms

ADAS	Advanced Driver Assistance Systems.
AEBS	Automatic Emergency Braking System.
ANN	Artificial Neural Network.
B-GPDM	Balanced Gaussian Process Dynamical Model.
BF	Bayesian Filter.
BN	Bayesian Network.
CA	Constant Acceleration.
CGC	Constrained Gravitational Clustering.
CMU	Carnegie Mellon University.
CNN	Convolutional Neural Network.
CP	Constant Position.
CT	Constant Turn Rate.
CV	Constant Velocity.
DBN	Dynamic Bayesian Network.
DPPCA	Dual Probabilistic Principal Component Analysis.
EKF	Extended Kalman Filter.
EM	Expectation-Maximization.
EU	European Union.
FA	Factor Analysis.
FFA	Fuzzy Finite Automata.

GP	Gaussian Process.
GPDM	Gaussian Process Dynamical Model.
GPLVM	Gaussian Process Latent Variable Model.
GPS	Global Positioning System.
HMDP	Hidden Variable Markov Decision Process.
HMM	Hidden Markov Model.
HOG	Histogram of Oriented Gradient.
ICA	Independent Component Analysis.
IMM	Interacting Multiple Model.
IRL	Inverse Reinforcement Learning.
ITS	Intelligent Transportation Systems.
KF	Kalman Filter.
LDA	Linear Discriminant Analysis.
LDCRF	Latent-Dynamic Conditional Random Field.
LIDAR	Light Detection and Ranging.
LLE	Locally Linear Embedding.
LLS	Linear Least Squares.
MCM	Markov-chain Model.
MDP	Markov Decision Process.
MED	Mean Euclidean Distance.
MLP	Multi-layer Perceptron.
MMM	Mixed-Markov-chain Model.
PCA	Principal Component Analysis.
PDF	Probabilistic Density Function.
PF	Particle Filter.
PHTM	Probabilistic Hierarchichal Trajectory Matching.
PPCA	Probabilistic Principal Component Analysis.

QRLCS	Quaternion-based Rotationally Invariant Longest Commom Subsequence.
RADAR	Radio Detection And Ranging.
RBF	Radial Basis Function.
RMSE	Root Mean Squared Error.
SCG	Scaled Conjugate Gradient Algorithm.
SGM	Semi Global Matching.
SLDS	Switching Linear Dynamical System.
SLP	Single-layer Perceptron.
SSE	Sum of Squared Errors.
SVD	Singular Value Decomposition.
SVM	Support Vector Machine.
SVR	Support Vector Regression.
TPM	Transition Probability Matrix.
TTC	Time To Collision.
TTE	Time To Event.
UAH	University of Alcalá.
UKF	Unscented Kalman Filter.
VRU	Vulnerable Road User.
WHO	World Health Organisation.

Chapter 1

Introduction

According to the report about road safety that the Spanish General Division for Traffic publishes every year, Spain was the sixth country in the European Union (EU) with the lowest number of road fatalities per population in 2014 and had lower rates than other countries like the United States, Japan and Australia. Namely, there were 91.570 casualty accidents, which resulted in 1.688 fatalities at the time of the accident or within 30 days of its occurrence, 9.574 casualties were admitted to hospital and 117.058 people were slightly injured. Regarding pedestrians, 336 were fatalities (19.91%), 1.902 were hospitalised and 10.625 suffered minor injuries. It is noteworthy that 92.86% of the pedestrians involved in an accident were in urban roads. On the other hand, data are more dramatic when European statistics are analysed. According to the *Annual Accident Report 2015* published by the European Road Safety Observatory, almost 26.000 people died in road traffic accidents in the EU in 2013, including 5.712 pedestrians, which represent 22.02% of all fatalities. Concerning world statistics, data are more impressive. The *Global Status Report on Road Safety* published by the World Health Organisation (WHO) in 2015 indicates that more than 1.2 million people died in road traffic accidents worldwide in 2013. About 275.000 of these fatalities were pedestrians.

Because of the high number of fatalities, during the last few years vehicles have been evolving to become intelligent machines with advanced technologies such as traffic signs recognition, pedestrian protection systems, Automatic Emergency Braking Systems (AEBSs) or other sort of Advanced Driver Assistance Systems (ADAS). Likewise, more sophisticated mathematical algorithms in perception and machine learning, and their applications in the field of Intelligent Transportation Systems (ITS), have contributed to this evolution as well. In addition, the per-

formance gain on computers and sensors such as cameras, Radio Detection And Ranging (RADAR), Light Detection and Ranging (LIDAR) or Global Positioning System (GPS), have also improved on-board vehicle systems. Finally, all these new technological advances arise as a consequence of the promotion and funding launched by governments and worldwide organisations to increase the road safety. This evolution has not finished yet. Indeed, it started a few years ago and it will continue during the next decades. For example, an effective interaction with other traffic participants is an open challenge for intelligent vehicles. This is particularly true in urban environments that are not primarily dedicated to traffic and are populated with Vulnerable Road Users (VRUs) like pedestrians and cyclists. In order to cope with the wide variations in traffic situations and behaviours of traffic participants, scientific progress is required in perception, prediction and interaction techniques.

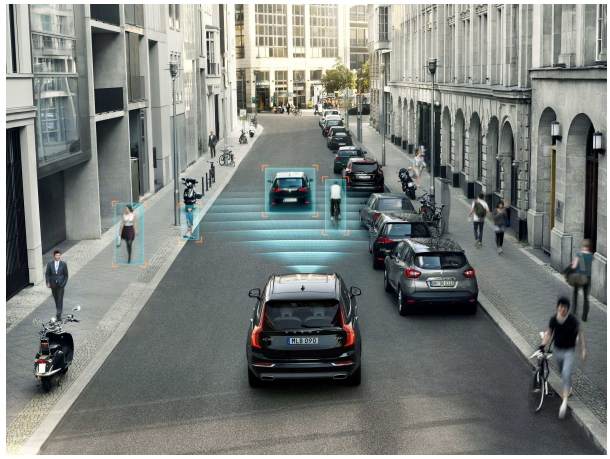


Figure 1.1: *Collision Warning with Full Auto Brake and Cyclist and Pedestrian Detection* developed by Volvo. The driver is warned when a road user or vehicle is in front and, if he does not take action to avoid the collision, an emergency braking is activated. www.automobilesreview.com.

Some of the sensors listed before, in particular those that enable to distinguish objects and determine their distances with high accuracy, are employed in the innovative solutions that have been developed by vehicle manufacturers over the last few years. The main motivation of these systems, called active safety systems, is to prevent accidents instead of mitigating them as passive safety systems do. For example, regarding pedestrian protection systems and AEBSs, Toyota recently presented the *Pre-Collision System with Pedestrian-avoidance Steer Assist* that warns the driver when a pedestrian or object is in front of the vehicle and, if he does not take action to avoid the collision, an AEBS in addition to automatic steering is

activated. Volvo has also equipped some of its vehicles with the *Collision Warning with Full Auto Brake and Cyclist and Pedestrian Detection*, described in [11], which assists the driver when there is a risk of collision with a VRU or vehicle in front, regardless of whether the object is stationary or moving in the same direction. The system combines a long-range RADAR and a forward-viewing wide-angle camera to continuously monitor the frontal area of the vehicle. The driver is first warned of a potentially imminent collision with a flashing red warning and an acoustic signal. But, in case he did not start an evasive action, then the automatic braking function would be deployed. The system can automatically avoid collisions if the driving speed is less than 35 km/h and mitigate injuries above that threshold. Beyond this, its effectiveness is assessed by means of the reconstruction of real-world accidents in [43]. This work asserts that the system may completely avoid 30% of the impacts involving pedestrians and could reduce up to 24% of the fatalities for crashes where pedestrians were struck by the front of a vehicle assuming that the system has been universally adopted.

Improving these technological advances is imperative since an early braking initiation or an accurate assessment about pedestrian positions before collisions could be particularly relevant as some works assert. For example, the studies developed in [24, 52] evaluate the potential effectiveness of AEBSs using real pedestrian-vehicle crashes. The first study analyses the functionality of these systems considering different attributes such as the sensor field of view, detection, reaction and braking initiation. The study determines that those systems based on a camera with a field of view of 35° need a reaction time between 0.5 and 1 second from the instant when a pedestrian is visible to the braking initiation in order to achieve a collision avoidance rate of 75% and 64% respectively. Additionally, the work concludes that 50% of these accidents would be avoided if the brakes were triggered 1 second before the impact. Finally, it asserts that a period between 1.5 and 0.5 seconds before a collision is critical regarding pedestrian positions since, for example, at 1 second before a crash, people are mainly located no more than 3 meters of the side of the vehicle and less than 20 meters far away. It is worth mentioning that the vehicle travel speeds in the dataset ranged from 20 to 60 km/h with an average value of 40 km/h. On the other hand, the second study evaluates the effectiveness as a function of the sensor field of view, maximum braking deceleration and braking initiation time assuming that the brake force has a linear ramp up time of 300 milliseconds. As expected, the longer the braking initiation time, the higher the impact speed and thus, the injury risk. These results are in accordance with those of [24]. It is noteworthy that, concerning the *Global Status Report on Road Safety* published by the WHO in 2015, an adult has less than a 20% chance of dying if struck by a car

at less than 50 km/h but almost a 60% risk of dying if hit at 80 km/h. Hence, a precise assessment about the current and future pedestrian positions and an early detection of people entering a road lane is a major challenge in order to increase the effectiveness of AEBSs. Similarly, an early recognition of pedestrian intentions can lead to much more accurate active interventions in the last second automatic manoeuvres. Therefore, with the aim of addressing these challenges, a lot of effort has been put into recognising pedestrian activities and predicting trajectories and intentions in the last few years so that strong gains are expected to be made in the performance and reliability of VRU protection systems.

This thesis describes a method to predict pedestrian positions, poses and intentions up to 1 second ahead in time applying a novel probabilistic modelling technique called Balanced Gaussian Process Dynamical Model (B-GPDM) and a Hidden Markov Model (HMM). The B-GPDM enables to estimate future observations from pedestrian motion sequences previously modelled. These sequences, in which different pedestrian dynamics were captured, are composed of 3D positions and displacements of several joints placed along the pedestrian body. On the other hand, an activity recognition based on a HMM makes possible to select the most accurate model to estimate future pedestrian states.

It is worth doing a distinction between the terms 'intention' and 'activity'. Hereafter, the former will be referred to a future pedestrian action and the latter will be referred to the current one. In this thesis, intentions and activities are classified into different categories, i.e. start crossing (or starting), stop before crossing (or stopping), crossing (or walking) and waiting (or standing). On the other hand, 'positions', 'paths' and 'trajectories' are considered as synonyms and make reference to pedestrian locations with respect to an origin in different future, current or past instants of time. In addition, the term 'pose' is referred to the pedestrian posture. Finally, the 'pedestrian state' comprises all the attributes mentioned above.

The present document is organised as follows: Chapter 2 presents a brief overview of previous works focused on pedestrian intention and path prediction. Chapter 3 introduces the theoretical basis of the Gaussian Process Dynamical Model (GPDM) and B-GPDM to reduce the dimensionality of a set of observations related in time in a non-linear way. Chapter 4 explains in detail how the activity recognition and pedestrian path and intention prediction are carried out applying the HMM and B-GPDM. Chapter 5 describes extensive results obtained by the proposed method. Finally, Chapter 6 lists the conclusions of this thesis and future research lines that may spring from it.

Chapter 2

Previous Works

The problem of vision-based pedestrian detection for ADAS has been extensively researched in the past. Indeed, outstanding surveys on this field can be found in the literature such as [12, 15, 18]. As a consequence, many manufacturers have equipped their vehicles with commercial systems that warn the driver when a pedestrian or object is in front. Nonetheless, as mentioned previously, improving these systems with the estimation of future VRU states could activate effective automatic manoeuvres earlier. Despite this, not many works have been published so far about intention, path and pose prediction once pedestrians are detected.

Throughout this chapter, a brief overview of previous works orientated towards estimating future pedestrian states is presented and glossed. Firstly, in Section 2.1, the most relevant features for that purpose are analysed. Beyond that, the creation of appropriate models from one or several of these features is a common task in order to obtain accurate descriptions of pedestrian motions. Therefore, different modelling techniques are explored in Section 2.2. After that, the path prediction accuracies and time horizons accomplished by significant works are examined in Section 2.3. Finally, some conclusions about the analysis of these works and the main objectives of this thesis are presented in Sections 2.4 and 2.5 respectively.

2.1. Features and Information

A wide range of features and information can be extracted from pedestrians to infer their future states. However, some of them are certainly more significant than others. Studies such as [4, 16, 21, 23, 45, 53, 56, 65, 66] give some useful clues.

Several of these works prove that pedestrians relied on the distance to vehicles rather than the Time To Collision (TTC) to cross the road or wait. Specifically, the data analysis of some experiments carried out in [53] shows that, independently of vehicle travel speeds, the road crossing action is unlikely when the TTC is below 3 seconds, however it is almost sure when the TTC is above 7 seconds. Between these limits, the distance to vehicles normally determines the decision. A similar conclusion is reached in [65], where the crossing probability is almost 100% when the TTC is longer than 6 seconds.

Other parameters, e.g. the direction and size of oncoming vehicles, pedestrian gender and age, step frequency, head-turning, gait and presence of other pedestrians, have important effects in road crossing decisions. Regarding the vehicle direction, the study developed in [53] asserts that, when vehicles go in the same direction as pedestrians, shorter distances and TTCs are chosen. It also appears that pedestrians accept longer TTCs when facing larger vehicles or are accompanied by others, as demonstrated in [65]. Besides, the gender-based analysis of this last work reflects that men usually take fewer risks than women. Likewise, concerning the age, as claimed in [45], elderly pedestrians select more dangerous decisions than younger people despite the fact that they normally take more time to make them. Another important variable is the step frequency. The results showed in [56] confirm that people tend to use higher step frequencies when they are crossing the road, especially when vehicles are moving towards them or when they are crossing without right-to-way. In [23], the head-turning is examined in crossing activities when vehicles are approaching. The work states that the head-turning frequency increases towards the entry of crosswalks and at collision points. Moreover, the analysis indicates that the head-turning frequency at nighttime, when vehicles are approaching from behind or by elderly people tends to be low. On the other hand, four studies focused on pedestrian gait are presented in [4, 16, 21, 66]. The first one establishes reference values for both comfortable and maximum human gait speeds. The second one evaluates pedestrian behaviours and gait responses at signal-controlled intersections by analysing the elapsed time between the illumination of a pedestrian walk sign and the gait initiation, the rate of acceleration to reach a steady state velocity and the number of steps required to reach that velocity. The third study shows that the mechanism of gait termination is a combination of a decrease in the step length and an increase in the step time. It is also remarkable that, during the last three steps of deceleration, the behaviours of children, adults and elderly people are very similar. Additionally, the analysis of pedestrian velocity profiles indicates three typical motion patterns in the way people slow down from steady state velocities: stopping with constant deceleration,

stopping with increasing deceleration and fast stopping. Finally, the last study evaluates pedestrian speeds in steady motions and accelerations from stationary positions taking into account the age and gender.

Although all these variables are examined from a pedestrian's point of view, significant clues can also be collected from a driver's perspective since they are capable of understanding complex traffic situations and forecasting paths and intentions of other road users. The study elaborated in [53] addresses this issue and concludes that the observation of only pedestrian trajectories is unreliable for drivers to estimate forthcoming positions. Therefore, future paths and intentions are mainly predicted with motion parameters and body language.

Taking into account these studies, it seems that the most relevant features to compute path and intention predictions can be mainly extracted from two sources. The first one is directly obtained from pedestrians whose body languages, positioning information, orientations, head poses and motions determine the variables that a driver commonly uses to infer intentions and to know whether pedestrians are aware of oncoming vehicles. The second source emerges from the situation criticality and the environment at each instant of time, where vehicle-pedestrian and curbside-pedestrian distances, existence of zebra crossing, road width or size of approaching vehicles are significant data. An analysis of recent works focused on the task of predicting future pedestrian states confirms these conclusions. Figure 2.1 shows a classification of the main features and information that these works use to estimate pedestrian paths and intentions. In the following sections, "positioning information" makes reference to one or several of the next pedestrian features: position, velocity and acceleration.

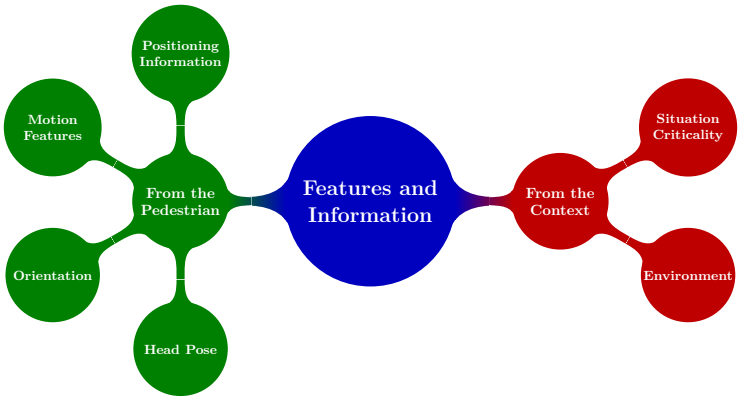


Figure 2.1: Features and information that relevant works focused on pedestrian path and intention prediction use to infer future states. These features can be mainly extracted from two sources: from the pedestrian and from the context.

2.1.1. Pedestrian Features

The path prediction of a pedestrian walking towards the road curbside, when viewed from an oncoming vehicle, and the assessment of whether he will cross or stop are fundamental tasks for innovative ADAS. To carry them out, pedestrian motion features are regularly extracted applying image processing instead of computing only pedestrian positions and velocities as less accurate approaches do. For example, in [1,54], positioning information is only considered to predict pedestrian-vehicle collisions and paths at short time horizons respectively. Nonetheless, in [31, 32], apart from using that information, augmented motion features derived from dense optical flow fields are also processed for path and intention predictions. These studies compare the proposed approaches with two simpler methods based only on positioning information. On the one side, the path prediction results indicated in [31] show a similar performance for walking trajectories in all algorithms, however, the approaches based on augmented motion features achieve more accurate results for stopping trajectories. On the other side, in [32], the results indicate that the addition of motion features does not improve the accuracy in the estimations. Despite this, the intention prediction results from both studies show that the proposed approaches outperform the other methods. Another example of the use of motion features can be found in [36] where a method to recognise starting, stopping and bending in intentions from a moving vehicle is implemented. The motion features are gathered using the overlapping of pedestrian silhouette images which are based on depth maps at consecutive time instants.

Beyond that, the orientations in which pedestrians are facing could be evaluated to predict where people may move in the future or determine the situational awareness of oncoming vehicles. When pedestrians are moving, motion directions and orientations can be easily approximated with their position histories. However, when pedestrians are static, only the orientations they are facing offer information about possible future paths. For instance, the applicability of pedestrian orientations and head poses to predict intentions is investigated in [17, 55]. In the first work, Histogram of Oriented Gradient (HOG) features are fed to an 8-class Support Vector Machine (SVM) classifier whose probabilities allow to model a HMM to infer future orientations. The second work presents an approach that combines intention recognition and path prediction for pedestrians that are walking along or towards the road curbside on their way to cross, stopping or just keeping on going in the same direction. The proposed model integrates positioning information from a stereo vision system and situational awareness computed by head pose estimations.

Moreover, many dangerous situations arise from the fact that the driver's view of the road scene may be obstructed by objects. In these cases, it is difficult or even impossible to detect pedestrians from the inside of a vehicle and avoid potential collisions. For that reason, infrastructural sensors in combination with roadside units can be mounted at urban hazard spots and send the appropriate signals to vehicles through wireless communication channels. This solution is proposed in [34, 35, 37] with the aim of predicting starting intentions. The algorithms extract pedestrian motion features by overlapping a sequence of edge images or depth-based foreground images. This spatial-temporal information implicitly comprises the body language of a pedestrian gait initiation. On the other hand, positioning information is extracted in [19, 20, 22] to create velocity-time-based and position-based models which are able to predict paths in the course of a gait initiation at crosswalks or for typical pedestrian motions. Furthermore, apart from using positioning information, heading angle is also considered in [7]. The work proposes a method to avoid vehicle-pedestrian collisions that learns and predicts pedestrian intentions while their motion instances are being observed. Positioning information and heading angles are taken into account to extract trajectories that are clustered into motion patterns later. Thereby, a future trajectory can be predicted by means of the matching between the current pedestrian path and a classified motion pattern.

Although this section has been centred on ITS, path and intention predictions are regularly carried out in other areas. Applications orientated towards surveillance, robotics, vehicle motion prediction or human motion analysis compute future trajectories as well. For example, a shopping mall is the scenario chosen in [8, 9] to test two improved approaches based on the work developed in [7]. Whereas the former uses the same features, positioning information at discrete time steps is only deployed in the latter. A similar strategy is also applied in [14] using static surveillance cameras. Additionally, the work developed in [27] assumes that different head pose patterns reflect different intentions. In this case, positioning information and head poses are taken into account to model a Dynamic Bayesian Network (DBN) to predict intentions in a shopping mall as well. It is noteworthy that all these approaches may also be implemented for stationary ITS infrastructures set up in intersections or streets since pedestrians are normally located along crosswalks or sidewalks. Regarding vehicle motion prediction, this task is addressed in [26] by means of a trajectory matching algorithm based on positioning information and orientations relative to the ego-vehicle. Finally, a people motion tracking and path prediction approach designed for robot applications is presented in [51]. The method computes 3D positions of different points located along the human body.

2.1.2. Contextual Features

Where and when a pedestrian will cross the road is highly related to its specific context location. For example, a pedestrian walking towards a crosswalk is more likely to cross. However, a pedestrian with that intention might not do so if there is a car approaching fast, but might cross if the car is still far away. Therefore, although urban environments are generally very complex, exploiting and analysing the situational criticality and the structure of streets, sidewalks, intersections or crosswalks, i.e. the spatial layout of the environment, can also provide some valuable information to innovative ADAS.

In this sense, the factors introduced before and the pedestrian situational awareness are computed in [38] using an on-board stereo camera with the aim of predicting pedestrian paths from an approaching vehicle. Concretely, the situational awareness is assessed by the pedestrian head orientation, the situation criticality by the vehicle-pedestrian distance at the expected collision point, and the spatial layout by the curbside-pedestrian distance. Furthermore, pedestrian features and contextual information are combined in [5, 39, 61] as well. The first work fuses two models to predict crossing intentions from a moving vehicle. One is a generic context-based model fitted for inner-city and the other is a specific model fitted for crosswalk environments. Contextual information such as lateral distances and times that pedestrians need to reach some goals (collision point, curbstone, ego-lane or crosswalk) and pedestrian features such as tridimensional positions, velocities and directions are processed by a stereo vision system. In the second work, curbside-pedestrian and vehicle-pedestrian distances, head orientations and their variations, and pedestrian speeds are computed to predict intentions using a stereo thermal camera mounted on the front-roof of a car. Finally, the last work is focused mainly on identifying those features from the environment that are necessary to learn the best model which is able to determine whether a pedestrian will cross the road at a crosswalk. The features are divided into two different basic types: those that describe pedestrian motions (velocities, distances to curbs, distances to crosswalks and distances travelled between consecutive time steps) and those that characterise the interaction between pedestrians and vehicles (closest vehicles to pedestrians, velocities and distances travelled by those vehicles, distances between vehicles and crosswalks and distances between vehicles and pedestrians). The work confirms that the features related to pedestrians provide better results in inferring intentions.

On the other side, as previously mentioned, many dangerous situations arise from the fact that the driver's view of the road scene may be obstructed by objects

Figure 2.2: Given a single pedestrian detection, the approach proposed in [33] forecasts plausible paths and destinations from vision-input. Physical attributes of the scene are able to encode agent preferences like using the sidewalks.

and, hence, it could be impossible to avoid a collision. However, only pedestrian features have been considered so far for these types of situations. Including prior knowledge about the scene such as objects, sidewalks, roads, entries and destinations might provide richer information to systems focused on predicting pedestrian trajectories. For example, in [33], the task of inferring paths and intentions from a static camera is addressed by incorporating physical scene features and noisy tracker observations. Thereby, the effect of physical environments on pedestrian intentions is modelled through the information that is gleaned from physical scene features and prior knowledge of possible destinations. The scene understanding is done by means of a semantic scene labelling algorithm combined with ideas from an Inverse Reinforcement Learning (IRL) framework.

2.2. Modelling Techniques

Unlike animals, whose behavioural patterns have their origins on primitive instincts and emotions, humans normally act according to their reasoning. This is related to the learning and experience that they acquire from many different situations along their lives. Indeed, the knowledge and observation of events allow humans to understand and predict future situations, intentions, motions or trajectories and react correctly in each case. For example, they can avoid collisions with moving objects when they are walking on a street predicting future paths a few seconds before and changing their trajectories accordingly.

Based on these considerations, providing the capability of prediction to computers has been a growing topic among researchers in the last few years. Getting to understand the underlying intent of an observed agent is of paramount interest in a large variety of domains that involve some sort of collaborative and competitive scenarios, e.g. robotics, surveillance, human-machine interaction and intelligent vehicles. This capability of prediction is normally carried out by machine learning algorithms. The machine learning is a branch of computer science that studies how to give computers the capability of learning in order to make predictions and identify patterns on data. The machine learning techniques are classified into three categories in [3]. These are:

- **Supervised learning**: In this case, input data and their corresponding target values are provided in order to infer a function or model that relates both. The algorithms associated with this category can be divided into classification and regression. Whereas the former assign each input vector to one of a finite number of discrete classes, the latter assign each input vector to one or more continuous variables. Techniques like the SVM, linear regression, boosting, Artificial Neural Networks (ANNs), naïve Bayes classifiers or decision trees are included within this group.

- **Unsupervised learning**: In this instance, input data are also provided but, unlike supervised learning, target values are unknown. Some subfields of this category are clustering, which consists in discovering groups of similar features within the data, density estimation, which determines a distribution of data within the input space, or visualisation, which projects the data from a high-dimensional space to another of lower dimensions. The Expectation-Maximization (EM) algorithm, k-means, Gaussian Process Latent Variable Model (GPLVM), GPDM and Principal Component Analysis (PCA) are some examples of techniques included in this group.

- **Reinforcement learning**: Concerning this category, agents interact with their environments through actions that change the environment states and, as a result, these agents receive some rewards. Learning how to maximise the future rewards is the final goal of the algorithms belonging to this group. Some representative techniques are the Markov Decision Process (MDP) and Monte Carlo algorithm.

As mentioned before, the creation of appropriate models from one or several of the features glossed in the previous section by means of machine learning techniques is a common task in systems focused on intention and path prediction. This task

allows to achieve accurate descriptions of pedestrian motions. Hence, it demands a prior step of gathering information about the process which will be modelled. Choosing the best modelling technique, regardless of the features selected, is an important decision because not all approaches yield as good results as others. This decision is especially critical concerning pedestrians since simple models may not deal correctly with changes in human dynamics. Therefore, throughout this section, different techniques for pedestrian motion modelling, which are applied to predict intentions and paths, are examined.

2.2.1. Linear Models

Simple linear models have been proposed in several works obtaining interesting results in path prediction. For example, in [20], a piecewise linear model is fitted to a velocity-time curve whose data are derived from pedestrians in the course of a gait initiation at crosswalks. It assumes motions with Constant Acceleration (CA) during the first stride and motions with Constant Velocity (CV) afterwards. Thus, pedestrian trajectories emerge integrating their velocities. Besides, a sigmoid model is also fitted to the curve and compared with the previous approach, achieving the linear model better results at the beginning of starting motions.

SVMs are also linear models that have been extensively applied in many vision-based applications. Specifically, some systems focused on starting intention recognition are found in [34,35,37]. In these works, linear 2-class SVM classifiers are used in order to determine whether a motion-based descriptor belongs to a pedestrian which is starting to walk or not. Additionally, some of them include a class probability estimation through the transformation of the SVM outputs into probability distributions over classes which are interpreted as pedestrian intention probabilities. On the other hand, the number of pedestrian motions contemplated is expanded in [36]. Therefore, learning a SVM classifier with class probability estimation for each type of motion is an essential condition to infer future pedestrian intentions.

2.2.2. Non-linear Models

Non-linear models have also been implemented to predict pedestrian trajectories. In fact, polynomial approximations are well suited to model temporal trends providing an extraction of the principal information of the underlying time series in the form of polynomial coefficients, high independence of input data and additional noise resistance. These models are formed as compositions of basis polynomials that can consist of various functions, such as polynomials, wavelets or sinusoidal

functions. Besides, the coefficients are optimised by minimising the least-squares error between the time series data and the polynomial. For example, a simplified model of pedestrian motions is learned by applying this technique in [19]. Beyond that, in [22], the polynomial coefficients are trained in an ANN and an ϵ-Support Vector Regression (SVR) model with a non-linear Radial Basis Function (RBF) kernel in order to predict pedestrian paths. The ϵ-SVR model obtains slightly higher prediction errors than the approach based on the ANN.

2.2.3. Dynamic Bayesian Networks

A Bayesian Network (BN) is a probabilistic graphical model represented with a directed acyclic graph composed of nodes and edges. Whereas the former correspond to random variables that can take both discrete and continuous values, the directed edges express probabilistic relationships between these variables. Furthermore, extending the scope of these models, a DBN is a sequential BN whose nodes can also have connections with nodes at adjacent time steps, thus making possible to model time-series data. Because of their flexibility, DBNs are also implemented to predict future pedestrian states. For example, a DBN, which capture contextual information as latent states on top of a Switching Linear Dynamical System (SLDS), is proposed in [38] to predict pedestrian paths from vehicles (see Figure 2.3). The SLDS uses the top-level DBN to select per time step the underlying system dynamics.

Additionally, special cases of DBNs are the Markov-chain Model (MCM), the HMM and the recursive Bayesian Filters (BFs) which are used in many ITS and robotics applications to model pedestrian motions. In a MCM, the future state of a process depends solely on its present state. Hence, it is assumed that pedestrians just choose their next positions on the basis of their current ones. In contrast, a HMM copes with transitions of unobservable states, i.e. pedestrian thoughts, and the observations, which correspond to measured positions, are dependent on these thoughts. Likewise, in a HMM, the future state of a process also depends solely on its current state. Unfortunately, the hypothesis that all pedestrians behave similarly is assumed when these models are considered. To solve this problem, modelling different motions and selecting the most appropriate one at each instant of time is a recurrent approach. For example, a method to predict future pedestrian positions is developed in [2] by means of a Mixed-Markov-chain Model (MMM). In this work, the pedestrian motions, composed of positioning information, are classified into patterns corresponding to groups of similar activities. Because of its simplicity, the prediction accuracy is rationally low at earlier steps since the

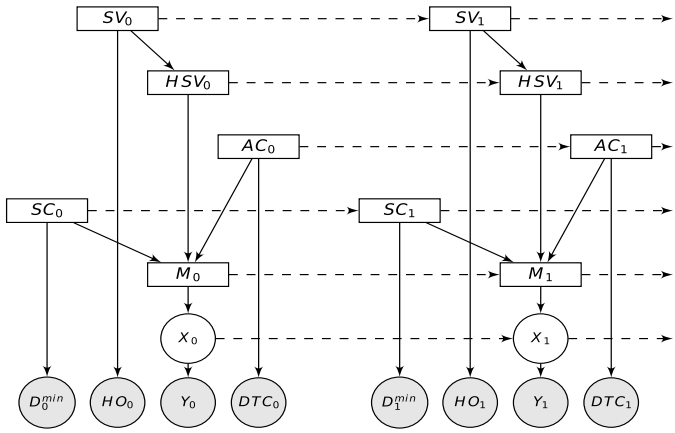


Figure 2.3: The DBN and SLDS proposed in [38] for two time slices. Two sets of variables are distinguished: those related to the SLDS (consisting of the discrete switching state M, the continuous hidden state X and the associated observation Y) and those related to the spatial layout, situation criticality and pedestrian awareness (consisting of the following discrete latent variables: SV (Sees-Vehicle), HSV (Has-Seen-Vehicle), SC (Situation-Critical) and AC (At Curb)) that influence the SLDS switching state. The observations HO (Head-Orientation), D^{min} (Minimum Vehicle-Pedestrian Distance) and DTC (Distance-To-Curb) provide evidence for the context and pedestrian awareness.

process can not gather enough information about the pedestrian to estimate the next position. However, the accuracy is rapidly improved in later steps. Finally, a recursive BF is a general probabilistic method to deal with the problem of extracting information about parameters, or states, in a dynamical system given noisy measurements. In order to make inference of future states at least two probabilistic models are normally required. One for describing the transitions between states, i.e. the process model, and the other for relating the current state to the noise measurement, i.e. the measurement model. A recursive BF lies essentially in estimating the posterior probability associated with the state by means of two stages: prediction and update. The former applies the process model to project the previous posterior probability forward in time, thus predicting the next state. In contrast, the latter uses the latest measurement to tighten the posterior probability obtained in the prediction stage by means of the measurement model. It is noteworthy that simple recursive BFs take into account the current position and velocity of a dynamic object to estimate the most probable next position in a discrete-time domain. However, since these approaches are based only on physical observations, they are accurate for objects with low dynamical behaviours. Thereby, these models reach their limits in the context of pedestrian path and intention prediction since unexpected or very fast changes could occur.

Likely, the most popular recursive BF to estimate future states are the Kalman Filter (KF). This is a discrete-time linear model in which the current state of a dynamic system can be propagated to the future by means of the underlying linear dynamical model without the incorporation of new measurements. Beyond that, whereas the KF normally assumes that an object moves at Constant Position (CP), CV, CA or Constant Turn Rate (CT), the Interacting Multiple Model (IMM)-KF takes into account the capability of some objects to suddenly change their dynamical behaviour. This model combines different KFs by means of a Transition Probability Matrix (TPM) that captures the probability of transition from one type of motion to another. Some examples of the use of the KF and IMM-KF for intention and path prediction can be found in [19,31,32]. All these works implement the KF with CV motion model to estimate paths of moving pedestrians. However, the IMM-KF is also considered in [31,32] in order to include an additional KF with CP model for non moving pedestrians. Besides, this last model enables to derive the pedestrian intentions from the transition probabilities at each instant of time.

The Extended Kalman Filter (EKF) and the Unscented Kalman Filter (UKF) are two non-linear versions of the KF that have been used in many applications focused on pedestrian tracking and path prediction. Specifically, these tasks are accomplished in [51] by means of an EKF. Likewise, an IMM-EKF or IMM-UKF can be implemented to combine different filters that model different pedestrian motions. For example, in [34], an IMM-EKF with CP and CV as motion models recognises starting intentions. In [54], pedestrian paths at short time horizons are predicted using an EKF and an IMM-EKF with CV, CA and CT models corresponding to crossing, stopping, bending in and starting activities. Besides, in [55], an IMM-EKF is implemented in combination with a Latent-Dynamic Conditional Random Field (LDCRF) model for the intention recognition and path prediction tasks in different scenarios. The LDCRF output has a direct impact on the transition probabilities which control the behaviour of the IMM-EKF. The method is able to integrate the features extracted from the pedestrian dynamics as well as the context-based interaction to learn inner connections within a specific type of scenario and external correlations between different types of environments.

On the other hand, the MDPs are an extension of the MCMs. The difference is the addition of actions and rewards. At each time step, the process is in a state and the decision maker chooses an action that is available for that state. The process responds at the next time step by randomly moving into a new state and giving the decision maker a corresponding reward. For example, in [33], the task of inferring the future actions of people from noisy visual input is addressed by means of a

Hidden Variable Markov Decision Process (HMDP). The method models the effect of static environments, instead of dynamic environments like moving people, on the future pedestrian intention. The learning of how much a physical scene feature affects a person action is done by training the parameters of a cost function.

2.2.4. Trajectory Matching Models

Pedestrians generally follow well-defined paths, either to cross the road, walk along sidewalks or turn at an intersection. For this reason, the learning of pedestrian motion patterns has been widely carried out in relevant path prediction systems over the last few years. These methods apply a matching algorithm to compare the current pedestrian trajectory (in terms of spatial position, velocity or heading angle) with trajectories previously learned. When the best matched trajectory is found, this is used to predict future positions and estimate the risk of collision. In other words, future positions arise looking ahead on matched trajectories that are contained in a dataset. Nonetheless, the main drawback of these algorithms is the need to define a temporal window prior to the prediction to achieve significant results.

Some examples of the use of these models can be found in [7–9, 14, 26, 31, 32]. In particular, path prediction methods based on the clustering and classification of observable pedestrian trajectories into motion patterns are developed in [7–9]. Firstly, in the learning stage, pedestrian states are extracted and associated with existing trajectories using a distance-based procedure. After that, the derived trajectories are clustered by applying a Constrained Gravitational Clustering (CGC) algorithm and classified into motion patterns. In the prediction step, the similarity between the current pedestrian trajectory and the motion patterns makes possible to select an appropriate model to predict future positions. On the other side, a probabilistic model based on Gaussian Process (GP) regression is proposed in [14] to describe typical motion patterns and predict pedestrian trajectories using solely positioning information. After the execution of the trajectory clustering, a model of each cluster is built. Furthermore, a long-term vehicle motion prediction approach based on a combination of a trajectory classification and a Particle Filter (PF) framework is proposed in [26]. The method learns and uses motion patterns to estimate future vehicle positions. As a measure of similarity, the Quaternion-based Rotationally Invariant Longest Commom Subsequence (QRLCS) metric is introduced. Similar strategy is addressed in [31, 32] where a trajectory matching and filtering framework called Probabilistic Hierarchichal Trajectory Matching (PHTM) is developed. In these works, each trajectory is composed of pedestrian

lateral and longitudinal positions and features extracted from optical flow fields at different instants of time. Besides, the matching of an observed test trajectory and a trajectory included in a dataset is computed by a measure of similarity and a probabilistic search framework.

2.2.5. Social Force Models

People are usually driven by an inner motivation towards some goal, are influenced by obstacles and other people along their paths, and follow social rules. In other words, human motions are influenced by physical and social constraints related to the environment. Based on these considerations, pedestrian motions are represented in [25] by simple social force models. These models describe the interactions between pedestrians using the concept of social forces or social fields. These forces model different aspects of pedestrian behaviours, such as the motivation of people to reach a goal or the repulsive effect of walls and other people.

2.2.6. Gaussian Process Dynamical Models

The GPDM is also a suitable non-linear option since it reduces the dimensionality of feature vectors related in time into a latent space, thus modelling the underlying dynamics. Besides, this model provides smooth predictions of future observations which can be effective to estimate future pedestrian states. Nonetheless, the absence of a direct mapping from the feature space to the latent space is an obstacle that should be overcome when new observations are captured. In [31], lateral pedestrian dynamics are trained into two GPDMs, one for walking motions and the other for stopping motions due to the fact that combining data belonging to different activities could result in degenerated models. To overcome the absence of mapping from the original space to the latent space, each model is combined with a PF that finds the latent position given an observation. Finally, an IMM-PF makes possible to combine both GPDMs to determine what model is used at each instant of time.

2.2.7. Fuzzy Finite Automatas

A Fuzzy Finite Automata (FFA) is implemented in [39] to predict pedestrian intentions using a stereo FIR camera mounted on the front-roof of a vehicle. The FFA connects four states corresponding to standing and crossing intentions in different contexts. Whereas the states are represented by nodes in a FFA, the transitions

between nodes are represented by arcs. The states have corresponding membership values and are interpreted as the probability of a pedestrian event at a particular instant. Moreover, changes between states are controlled by various transition Probabilistic Density Functions (PDFs) based on spatial-temporal feature variations.

2.2.8. Neural Networks

The use of ANNs, such as the Single-layer Perceptron (SLP) or Multi-layer Perceptron (MLP), are alternative approaches for activity recognition or time series forecasting. They provide the development of path prediction methods that are capable of dealing with all kind of intentions included in a training dataset without a prior state classification as demonstrated in [19, 22]. The results show that the ANNs outperform simpler models such as the KF. This is due to the fact that an ANN has the capability to handle intentions by learning a single implicit motion model independent of a specific motion type. In [19], a preprocessed selection of n trajectory points at defined time steps prior to an event is used as input pattern and estimated m points of the future path as output pattern. Furthermore, in [22], a polynomial least-squares approximation is combined with a MLP. The velocity profiles of past and future time windows are approximated with polynomials in order to learn the relation of in- and output coefficients. Hence, the future velocity profiles can be estimated by the reconstruction of the output polynomials based on the predicted coefficients. Additionally, in [5], a SLP is trained to classify context and pedestrian features with the intent of obtaining crossing or non crossing outputs.

2.3. Prediction Accuracies and Time Horizons

Previously, it was mentioned that two main sources of information can be used to make predictions of future pedestrian states. Nonetheless, each of these sources involves getting different prediction horizons. The approaches based on pedestrian features can normally cope with a higher variety of intentions but they have the drawback of achieving shorter time horizons. The opposite occurs in the case of context-based algorithms which normally obtain the longest time horizons but only for limited pedestrian intentions in controlled scenarios.

Additionally, none of the works reviewed in this thesis offers a discussion about the best method of event-labelling, i.e. when a pedestrian starts or finishes an event

such as crossing, starting or stopping, and the best evaluation method of path predictions. Addressing these issues is imperative in order to establish a standard criterion which enables to make comparisons among approaches in similar conditions. Regarding the event-labelling, in [31,32,54], the last placement of the foot on the ground at the curbside is labelled as non-crossing event when pedestrians are stopping. However, when they continue walking, the closest point to the curbside, before entering the roadway, is selected as crossing event. Finally, when pedestrians are bending in or starting to walk, the first moment of visually recognisable body turning or leg movements is chosen to label the event. On the other hand, the frame where a human observer recognises the initial foot movement is labelled as starting action in [34, 35]. Furthermore, the initiation of a crossing activity is defined when the foot of a pedestrian touches the ego-lane in [5]. Despite these examples, there are several events that are harder to label, e.g. transitions from walking to stopping or from starting to walking actions. Establishing a criterion to label these transitions would allow to model each pedestrian activity appropriately.

Concerning path evaluations, the RMSE and MED between estimated pedestrian positions and the groundtruth are often chosen as measure of accuracy. For example, the MED used in [19, 20, 55] gives a more precise physical interpretation of the predicted pedestrian positions with respect to a groundtruth than the RMSE used in [22]. Likewise, the mean and standard deviation of the per-sequence RMSE used in [31, 32] provide vague information of the system performance since the RMSE for each sequence does not offer information about the similarity between predicted positions and the groundtruth at discrete time steps. Besides, although most of the works consider that the evaluation must be done for each type of intention separately, it is not clear what methodology is the most appropriate in order to standardise the path evaluation. Hence, a reliable comparison of path prediction approaches has not been done for the moment.

2.3.1. Short-term Predictions

As mentioned previously, due to the fact that humans have highly dynamic behaviours, the approaches based on pedestrian features are only suitable for short prediction horizons, normally up to a few seconds ahead in time. These approaches are analysed in Table 2.1, where the path accuracies and time horizons are showed. The features, modelling algorithms and evaluation methods that have been used by these works are also included.

Ref.	Features	Algorithm	Error	Time	Starting	Stopping	Walking
[31]	Position	KF. Stationary veh.	Mean combined lat. and long. RMSE	0.77 s	-	0.93 ±0.15	0.28 ±0.12
[31]	Position	IMM-KF. Stationary veh.	Mean combined lat. and long. RMSE	0.77 s	-	0.87 ±0.12	0.25 ±0.12
[31]	Motion	PHTM. Stationary veh.	Mean combined lat. and long. RMSE	0.77 s	-	0.58 ±0.17	0.29 ±0.07
[31]	Motion	GPDM. Stationary veh.	Mean combined lat. and long. RMSE	0.77 s	-	0.51 ±0.07	0.34 ±0.18
[31]	Position	KF. Veh. moving.	Mean combined lat. and long. RMSE	0.77 s	-	1.25 ±0.33	0.62 ±0.29
[31]	Position	IMM-KF. Veh. moving.	Mean combined lat. and long. RMSE	0.77 s	-	1.19 ±0.17	0.77 ±0.26
[31]	Motion	PHTM. Veh. moving.	Mean combined lat. and long. RMSE	0.77 s	-	0.74 ±0.23	0.43 ±0.17
[31]	Motion	GPDM. Veh. moving.	Mean combined lat. and long. RMSE	0.77 s	-	0.66 ±0.32	0.62 ±0.25
[32]	Position	IMM-KF	Mean combined lat. and long. RMSE	0.77 s	-	1.54 ±1.23	1.33 ±0.87
[32]	Motion	PHTM	Mean combined lat. and long. RMSE	0.77 s	-	0.88 ±0.43	1.07 ±0.39
[20]	Position	Piecewise linear model	MED	0.6 s	0.19	-	-
[20]	Position	Piecewise linear model	MED	1.2 s	0.20	-	-
[20]	Position	Piecewise linear model	MED	2.4 s	0.28	-	-
[20]	Position	Sigmoid model	MED	0.6 s	0.08	-	-
[20]	Position	Sigmoid model	MED	1.2 s	0.19	-	-
[20]	Position	Sigmoid model	MED	2.4 s	0.47	-	-
[19]	Position	KF	MED	1.2 s	0.374	0.296	0.296
[19]	Position	KF	MED	2.5 s	1.294	0.881	0.820
[19]	Position	Polynomial approx.	MED	1.2 s	0.408	0.310	0.305
[19]	Position	Polynomial approx.	MED	2.5 s	1.300	0.871	0.814

Continue on next page

Continued from previous page

Ref.	Features	Algorithm	Error	Time	Starting	Stopping	Walking
[19]	Position	MLP	MED	1.2 s	0.315	0.224	0.230
[19]	Position	MLP	MED	2.5 s	1.131	0.647	0.755
[22]	Position	KF	RMSE	1.0 s	0.458	0.415	0.373
[22]	Position	KF	RMSE	2.5 s	1.617	1.429	1.226
[22]	Position	Polynomial approx.+MLP	RMSE	1.0 s	0.334	0.292	0.250
[22]	Position	Polynomial approx.+MLP	RMSE	2.5 s	1.227	0.937	0.984
[55]	Position and head	IMM-EKF	Lateral MED	1.0 s	-	0.31 ±0.20	0.25 ±0.22
[55]	Position and head	IMM-EKF +LDCRF	Lateral MED	1.0 s	-	0.14 ±0.18	0.23 ±0.21

Table 2.1: Short-term path prediction errors (means and standard deviations) in meters for different pedestrian activities.

Analysing the results obtained in [31], path predictions at 0.77 seconds ahead in time for stopping intentions have a mean combined lateral and longitudinal RMSE of 0.51 ± 0.07 and 0.66 ± 0.32 meters from stationary and moving vehicles respectively. On the other hand, regarding walking intentions, the path predictions at the same time horizon have a mean combined lateral and longitudinal RMSE of 0.25 ± 0.12 meters for stationary vehicles and 0.43 ± 0.17 meters for moving vehicles. All algorithms compared in this work show a similar performance in this last case due to the fact that the dynamical pedestrian behaviours do not change as abruptly as in stopping intentions. Moreover, inspecting the results obtained in other works, low errors in path predictions are found as well. For example, in [32], the approach based on the PHTM outperforms the IMM-KF. However, in contrast to the work developed in [31], the results correspond to pedestrian positions manually extracted and perturbed by artificial uniform noise from moving and stationary vehicles. In [20], mean prediction errors of absolute walking distance are computed in the course of a gait initiation. The work presents errors of 0.19 and 0.28 meters at 0.6 and 2.4 seconds respectively using a piecewise linear model. The sigmoid model, also proposed in the work, achieves errors of 0.08 and 0.47 meters at the same instants of time respectively. Furthermore, path predictions up to 2.5 seconds are estimated in [19, 22]. The results show that bigger deviations for starting than stopping or walking intentions are produced. Specifically, in [19], by means of an ANN, the MED for starting, stopping and walking intentions at

a time horizon of 1.2 seconds are 0.315, 0.224 and 0.23 meters respectively. As expected, longer errors are accomplished at 2.5 seconds. In [22], the RMSE for starting, stopping and walking intentions at 1 second are 0.334, 0.292 and 0.25 meters respectively. Once again, the RMSE for starting, stopping and walking intentions at 2.5 seconds are longer. Finally, in [55], the lateral prediction error is computed in meters when predicting 1 second ahead around event occurrences for crossing and stopping pedestrians. At the moment of the event, the lateral MED are 0.23±0.21 and 0.14±0.18 meters for crossing and stopping events respectively.

Figure 2.4: Prediction of pedestrian path during a gait initiation with an interval of 0.2 seconds computed by the algorithm described in [20].

Nonetheless, not all the works reviewed in this chapter have the capability of predicting pedestrian paths. In fact, most works only predict pedestrian intentions such as starting, stopping, walking and bending in. The studies developed in [31, 32] also test the different approaches in the task of recognising pedestrian walking and stopping intentions, providing the capacity of human experts as baseline, who reach an accuracy of 80% in predicting the correct intention about 570 milliseconds before the event. This precision is only reached about 200-230 milliseconds in advance by the algorithms based on augmented motion features and 0-90 milliseconds by the algorithms based on positioning information. In [36], stopping intentions are detected between 500 and 125 milliseconds before standing still within an accuracy range of 80% and 100% respectively. Bending in intentions are recognised from 320 to 570 milliseconds after the first visible lateral body motion in the same accuracy range. Finally, starting intentions are detected from 125 to 250 milliseconds after the event with an accuracy range of 75% and 100%. On the other side, the approach developed in [35] recognises starting intentions 120

and 340 milliseconds after the gait initiation with an accuracy of 80% and 99% respectively in a controlled scenario. Similar results are obtained in [34, 37] despite more realistic scenarios are tested. Finally, the method developed in [22], which is focused on the early recognition of the gait initiation, is also evaluated and compared with the approach developed in [34]. The first algorithm outperforms the second one achieving a precision of 80% at the moment of the event.

2.3.2. Long-term Predictions

Unlike the approaches based on pedestrian features, the context-based systems have the advantage of making long-term predictions, up to 3 or 4 seconds ahead in time. Nonetheless, they are unable to deal with changes in the pedestrian dynamics correctly and estimate future paths. For example, an event-based evaluation is done in [5] for the two models developed to predict crossing intentions. Prediction horizons of 0.77 seconds are accomplished by the inner-city model when pedestrians are close to a crosswalk and 0.67 seconds otherwise. On the other hand, longer prediction horizons are achieved by the specific crosswalk model. In this case, the system can predict all crossing intentions on average 3.23 seconds in advance. Analysing the performance of the combination of both models the method predicts crossing intentions 2.59 seconds ahead in time.

In addition, the trajectory-based methods proposed in [7–9, 14] can also deal with long-term predictions. However, they are applied in a suitable way when a low number of motion patterns are only required to predict all possible pedestrian paths. Besides, a pedestrian motion history should be extracted previous to the prediction. Finally, changes in the pedestrian dynamics are not considered so that only walking activities are normally contemplated. Hence, when all these factors are assumed, low errors are normally obtained for long-term predictions.

2.4. Discussion

In this chapter, different studies focused on pedestrian behaviours at crosswalks and intersections inspect important variables, i.e. the pedestrian-vehicle distance, step frequency, environment, pedestrian gender and age or head-turning, that may be effective for innovative pedestrian protection systems. Unfortunately, not all these variables can be extracted using the sensors which are commonly set up in intelligent vehicles. For this reason, only positioning information, motion features,

orientation, head poses and context features are regularly considered by vision-based systems destined to predict future pedestrian states. These variables can be mainly extracted from two sources of information. The first one is directly obtained from pedestrians and the second one emerges from the situation criticality and the environment. Despite the combination of both sources of information may accomplish more accurate estimations, this approach is not often addressed.

On the other hand, no system applies pedestrian skeleton estimation to predict paths and intentions. Given that humans are not rigid objects, the motion analysis of each body part should be taken into account for these tasks. For example, whereas the motion of the head may not be relevant in starting intentions, a slightly motion of a knee could be indicative of that action. Likewise, before stopping activities, pedestrian gap steps are usually shorter than in walking activities. Determining the distance between feet could distinguish the beginning of a stopping intention.

Furthermore, infrastructural sensors in combination with roadside units could deal with many dangerous situations. Although these systems offer good results, the implementation is unfeasible since all vehicles should include the devices that allow to establish the connection with the system. Besides, the roadside units must be extensively located along roads and streets. For these reasons, with the exception of hazard spots, this approach should not be considered to improve the road safety. Hence, the improvements of pedestrian protection systems should be developed from a vehicle perspective instead of from an infrastructural point of view.

Concerning modelling approaches, switching between models with different dynamics are the best option to achieve accurate path and intention predictions. However, extensive experiments have not been carried out so far in order to fix the number of different pedestrian dynamics that could emerge in urban environments. For example, pedestrians with disabilities could have dynamics that do not correspond to any trained model. On the other hand, approaches which take into account past motion histories to accomplish accurate future paths may not be effective in situations where pedestrians suddenly appear in the vehicle trajectory.

Finally, as mentioned in previous sections, none of the works reviewed in this thesis offers a discussion about the best method of event-labelling. Establishing a standard criterion would allow to compare approaches in similar conditions. Whereas the last placement of the foot on the ground at the curbside is usually labelled as non-crossing intention when pedestrians are stopping, the closest point to the curbside and the first moment of visually recognisable body turning or leg

movements are generally selected in walking, bending in and starting intentions.

Moreover, despite most works reviewed above are focused on predicting pedestrian intentions, providing the probability of crossing with high confidence is not enough to avoid pedestrian-vehicle crashes. For example, future pedestrian positions could be decisive in the computation of the best collision avoidance trajectory for an automatic steering system. Thus, a good estimation of pedestrian paths should be also computed by innovative ADAS. Hence, although context information does not allow to estimate accurate future paths as some approaches based on pedestrian features do, the combination of both sources of information may provide longer and more accurate predictions about intentions and trajectories.

Finally, regarding the path evaluation, the RMSE and MED between estimated pedestrian positions and the groundtruth are often chosen as measure of accuracy. However, it is not clear what methodology is the most appropriate in order to standardise the path evaluation. Hence, working in these aspects is essential in order to obtain reliable comparisons between approaches.

2.5. Objectives

After reviewing different works focused on pedestrian path and intention prediction, this thesis tries to solve several problems previously discussed. Hence, the main objectives of this thesis are:

1. To develop a single-frame method to predict pedestrian path, poses and intentions up to 1 second ahead in time applying a novel probabilistic modelling technique called B-GPDM and a HMM. The B-GPDM enables to estimate future observations from pedestrian motion sequences previously modelled. These sequences, in which different pedestrian dynamics were captured, are composed of 3D positions and displacements of several joints placed along the pedestrian body. On the other hand, an activity recognition algorithm based on a HMM makes possible to select the most accurate model to estimate future pedestrian states.

2. To measure the influence of modelling four different pedestrian dynamics, i.e. standing, starting, stopping and walking, instead of two activities as proposed in other works. These dynamics enable to appropriately define typical dynamical changes which are carried out by pedestrians in real scenarios. As mentioned before, switching between models with different dynamics are the best option to achieve accurate path and intention predictions.

3. To determine what information and body parts are more relevant to make predictions of future pedestrian states. Therefore, the method will be evaluated taking into account two different sets of joints located along the pedestrian body. One set will be composed of 41 joints and the other will be composed of 11 joints which are located in shoulders and legs.

4. To test the feasibility and limits of the proposed method in an extensive way under ideal conditions by using a high frequency and low noise dataset published by Carnegie Mellon University (CMU) (see [10] for more information). On the one side, the high frequency of the dataset will help the algorithms to properly learn the dynamics of different activities and will increase the probability of finding a similar test observation in the trained data without missing intermediate observations. On the other side, low noise models will improve the prediction when working with noisy test samples.

5. To test the proposed method with noisy observations. Thereby, a single-frame pedestrian skeleton estimation algorithm based on point clouds extracted from a stereo vision system and geometrical constraints will be described. These algorithm enables to use other pedestrian features that are not considered in other works.

6. To establish a guideline of event-labelling, i.e. when a pedestrian starts or finishes an event such as crossing, starting or stopping. Addressing this issue is imperative in order to establish a standard criterion which enables to make comparisons among approaches in similar conditions.

Chapter 3

The Gaussian Process Dynamical Model

Modelling high-dimensional datasets composed of observations of multiple variables is a widespread practice in machine learning. Nonetheless, some of these measured variables are less significant than others to understand the underlying phenomena of interest. For that reason, techniques such as PCA, Factor Analysis (FA), Probabilistic Principal Component Analysis (PPCA), GPLVM or GPDM are applied to reduce the dimensionality of the original data in order to extract the most relevant information and represent them as a set of new variables called latent or hidden variables. The problem can be stated as follows: given the d-dimensional vector of observed variables $\mathbf{y} = (y_1, ..., y_d)^T$, a lower dimensional representation is obtained through the q-dimensional vector of latent variables $\mathbf{x} = (x_1, ..., x_q)^T$ with $q \leq d$.

These dimensionality reduction techniques can be classified into two groups: linear and non-linear. The former computes the latent variables as a linear combination of the original variables such as:

$$\mathbf{x} = \mathbf{W}^T \mathbf{y} \tag{3.1}$$

where $\mathbf{W}$ specifies the linear transformation between the data space and the latent space. PCA, FA, Independent Component Analysis (ICA) or Linear Discriminant Analysis (LDA) are some representative techniques of this group. On the other hand, non-linear methods, also referred to as manifolds learning algorithms, are mainly based on the idea of a dataset lying along a low-dimensional manifold em-

bedded in a high-dimensional space. Whereas the low-dimensional space reflects the underlying parameters, the high-dimensional space corresponds to the feature space so that the Euclidean distance in the new space is a meaningful measure of distance between any pair of points. The GPLVM, GPDM, Isomap and Locally Linear Embedding (LLE) are representative methods of this category. More information can be found in [6, 28].

This chapter presents the theoretical basis of the GPDM to reduce the dimensionality of a dataset in a non-linear way taking into account the dynamics of the data. For the sake of a better understanding of the process, the chapter starts explaining the most familiar linear method of dimensionality reduction, PCA. Then, how this technique can be developed under a probabilistic framework is explained. In the next section, the GPLVM is illustrated. Then, the theoretical development of the GPDM and B-GPDM is outlined. Finally, the main conclusions of this chapter are described in Section 3.5.

3.1. Principal Component Analysis

PCA is a well-known multivariate analysis technique to describe the structure of datasets with a large number of correlated variables or features, and observations. This method projects the dataset to a new orthonormal coordinate system, which is determined by the eigenvectors and eigenvalues of the covariance matrix, maximising the retained variance and minimising the least square reconstruction error. Thereby, it converts the original variables into a set of uncorrelated variables, called principal components, through linear transformations. Many applications of data compression, image processing, visualisation or pattern recognition apply PCA on large datasets to reduce the dimensionality while retaining as much as possible of the variation present in them. Multiple works about PCA can be found in the literature such as [29, 30], but, in this section, the method is briefly described.

Given a set of observations of multiple variables $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)^T$, where d is the number of variables and n the number of observations, PCA works determining the following eigen-decomposition of its covariance matrix $\mathbf{S}$ to obtain the principal components:

$$\mathbf{S} = \mathbf{U}\boldsymbol{\Delta}\mathbf{U}^T \tag{3.2}$$

where $\mathbf{U}$ and $\boldsymbol{\Delta}$ represents, respectively, the orthonormal matrix whose columns are the eigenvectors and the diagonal matrix whose elements are the eigenvalues of $\mathbf{S}$. Before applying PCA, it is important to analyse the nature of the dataset

since, when the variables are measured in different units, those whose variances are largest will tend to dominate the first principal components. For that reason, it is appropriate to scale the variables by subtracting the mean and dividing each one by its standard deviation in order to standardise them to have zero-mean and unit-variance. It is worth remarking that the covariance matrix is a positive semi-definite matrix so that the eigen-decomposition always exists, the eigenvalues are real positive or nulls and the eigenvectors are pairwise orthonormal when their eigenvalues are different. Therefore, it is possible to create the orthonormal matrix with the eigenvectors which define the principal component axes.

In general, once the eigenvectors are found from $\mathbf{S}$, the next step is to sort them in descending order according to their associated eigenvalues in order to form the orthonormal matrix $\mathbf{W}$. Since the eigenvalues indicate the variances of the principal components, the maximal variance is achieved by selecting the eigenvectors with the highest eigenvalues. Thus, the first principal component is the linear combination with the largest variance. The second principal component is the linear combination with the second largest variance and orthonormal to the first principal component, and so on. This property allows to select the $q \leq d$ principal components to map a high-dimensional dataset to a lower dimensional space with minimal loss of information.

The eigen-decomposition can also be done in a similar way applying the Singular Value Decomposition (SVD) technique to the zero-mean set of observations $\mathbf{Y}'$ computed from $\mathbf{Y}$:

$$\mathbf{Y}' = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{3.3}$$

where $\mathbf{U}$ is the matrix of left singular vectors corresponding with the eigenvectors of the matrix $\mathbf{Y}'\mathbf{Y}'^T$, $\mathbf{V}$ is the matrix of right singular vectors corresponding with the eigenvectors of the matrix $\mathbf{Y}'^T\mathbf{Y}'$ and $\mathbf{\Sigma}$ is the diagonal matrix whose elements are the singular values in descending order and corresponds to $((n-1)\mathbf{\Delta})^{\frac{1}{2}}$. As consequence, the matrix $\mathbf{V}$ conforms to the previous orthonormal matrix $\mathbf{W}$.

The values of the latent variables in the low-dimensional space are called factor scores. These can be interpreted geometrically as the projections of the observations onto the principal component axes and can be computed by:

$$\mathbf{x} = \mathbf{W}^T(\mathbf{y} - \bar{\mathbf{y}}) \tag{3.4}$$

where $\mathbf{x}$ corresponds to the factor scores, $\mathbf{y}$ to the original $d-$dimensional observation and $\bar{\mathbf{y}}$ to the $d-$dimensional mean vector of $\mathbf{Y}$. Because of the principal components are uncorrelated, the covariance matrix of the factor scores emerges as

a diagonal matrix whose elements are the eigenvalues of the covariance matrix $\mathbf{S}$.

In addition, given $\mathbf{W}$ and $\mathbf{x}$, an observation $\hat{\mathbf{y}}$ can be reconstructed from:

$$\hat{\mathbf{y}} = \mathbf{W}\mathbf{x} + \bar{\mathbf{y}} \tag{3.5}$$

where $\hat{\mathbf{y}}$ represents the reconstruction of the observation, $\bar{\mathbf{y}}$ the $d-$dimensional mean vector of $\mathbf{Y}$ and $\mathbf{x}$ the factor scores. As expected, the least square reconstruction error between $\hat{\mathbf{y}}$ and $\mathbf{y}$ is minimum.

3.2. Probabilistic Principal Component Analysis

The formulation of PCA presented in the previous section was based on a linear projection of the data onto a subspace of lower dimensionality than the original data space. However, PCA can also be expressed as the maximum likelihood solution of a probabilistic latent variable model, as described in [57, 58]. This probabilistic formulation for PCA, called PPCA, offers important advantages compared with the conventional PCA such as dealing with missing values of data or applying Bayesian inference methods.

PPCA is based on FA, which is one of the most common latent variable models for dimensionality reduction, where the relationship between a set of centred observations $\mathbf{Y} = [\mathbf{y_1}, ..., \mathbf{y_n}]^T$ and a set of latent variables $\mathbf{X} = [\mathbf{x_1}, ..., \mathbf{x_n}]^T$ is linear and corrupted by noise:

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \mathbf{u} \tag{3.6}$$

where the $d \times q$ matrix $\mathbf{W}$ relates a $d-$dimensional vector of observed variables $\mathbf{y}$ and an $q-$dimensional vector of latent variables or common factors $\mathbf{x}$, while $\mathbf{u}$ represents the error model or specific factors. The motivation is that, with $q < d$, the latent variables will offer a more concise explanation of the dependencies between the observed variables. Conventionally, given that there is no analytic solution, the common factors are defined to be independent and Gaussian with unit variance, $p(\mathbf{x}) = N(\mathbf{x}|\mathbf{0}, \mathbf{I})$. Hence, the $\mathbf{W}$ contains the correlations between the observed variables and the common factors. Likewise, the error model is also specified Gaussian, $p(\mathbf{u}) = N(\mathbf{u}|\mathbf{0}, \mathbf{\Psi})$, with a $d \times d$ diagonal matrix $\mathbf{\Psi}$. In this way, the covariance matrix of the set of observations $\mathbf{Y}$ can be divided into two parts, $\mathbf{S} = \mathbf{W}\mathbf{W}^T + \mathbf{\Psi}$, where $\mathbf{W}\mathbf{W}^T$ represents the variances of each observed variable that are shared with the other variables and $\mathbf{\Psi}$ represents the variances of the specific factors, that is, the variances of each observed variable that are not shared with the other variables. The idea of the algorithm is to determine the

values of $\mathbf{W}$ and $\mathbf{\Psi}$. Since, as mentioned above, there is no analytic solution, their values have to be obtained via an iterative procedure, mainly applying a maximum likelihood estimation. Thus, the corresponding Gaussian marginal distribution for the observed variables, i.e. the marginal likelihood, $p(\mathbf{y}|\mathbf{W}, \mathbf{\Psi}) = N(\mathbf{y}|\mathbf{0}, \mathbf{W}\mathbf{W}^T + \mathbf{\Psi})$ is derived to estimate the values of $\mathbf{W}$ and $\mathbf{\Psi}$.

In essence, FA explains the observed covariance structure of the data by representing the independent variance associated with each variable in the matrix $\mathbf{\Psi}$ and capturing the covariance between variables in the matrix $\mathbf{W}$. This is in contrast to PCA which treats the inter-variables dependencies and the independent noise identically. Additionally, unlike PCA, in FA the subspace defined by the maximum-likelihood estimates of $\mathbf{W}$ will generally not correspond to the principal subspace of the observed data.

Regarding PPCA, it differs from FA in that the conditional distribution of the observed variables $\mathbf{y}$ given the latent variables $\mathbf{x}$ is taken to have an isotropic covariance rather than a diagonal matrix $\mathbf{\Psi}$. The use of the isotropic Gaussian noise model $p(\mathbf{u}) = N(\mathbf{u}|\mathbf{0}, \sigma^2\mathbf{I})$ in conjunction with Equation 3.6 implies that the likelihood for a data point can be written as:

$$p(\mathbf{y}|\mathbf{x}, \mathbf{W}, \sigma^2) = N(\mathbf{y}|\mathbf{W}\mathbf{x}, \sigma^2\mathbf{I}) \tag{3.7}$$

and if it is assumed independence across data point then:

$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}, \sigma^2) = \prod_{i=1}^{n} N(\mathbf{y}_i|\mathbf{W}\mathbf{x}_i, \sigma^2\mathbf{I}) \tag{3.8}$$

The marginal likelihood for the observed data is obtained by integrating out the latent variables such as:

$$p(\mathbf{Y}|\mathbf{W}, \sigma^2) = \prod_{i=1}^{n} \int p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{W}, \sigma^2) p(\mathbf{x}_i) d\mathbf{x} = \prod_{i=1}^{n} N(\mathbf{y}_i|\mathbf{0}, \mathbf{C}) \tag{3.9}$$

where the covariance model is specified by $\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$ and the Gaussian prior over the latent variables is defined as:

$$p(\mathbf{X}) = \prod_{i=1}^{n} p(\mathbf{x}_i) = \prod_{i=1}^{n} N(\mathbf{x}_i|\mathbf{0}, \mathbf{I}) \tag{3.10}$$

The maximum-likelihood estimator for $\mathbf{W}$ and σ^2 can be obtained by an itera-

tive minimisation of the negative log-likelihood function defined as:

$$L = \frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln|\mathbf{C}| + \frac{1}{2}\mathrm{tr}(\mathbf{C}^{-1}\mathbf{S}) \tag{3.11}$$

where $\mathbf{S}$ corresponds to the covariance matrix, $n^{-1}\mathbf{Y}^T\mathbf{Y}$, of the set of centred observations $\mathbf{Y}$. However, in contrast to FA, the maximum-likelihood estimator for $\mathbf{W}$ and σ^2 can be obtained explicitly in PPCA. In fact, stationary points of the marginal likelihood function occur where $\mathbf{W}$ is a matrix whose columns are scaled eigenvectors of the covariance matrix $\mathbf{S}$, and σ^2 is the average variance in the discarded dimensions. In particular, the maximum-likelihood estimators for $\mathbf{W_{ML}}$ and σ^2_{ML} can be expressed in closed form from:

$$\mathbf{W_{ML}} = \mathbf{U}_q(\mathbf{\Delta}_q - \sigma^2\mathbf{I})^{\frac{1}{2}}\mathbf{R} \tag{3.12}$$

$$\sigma^2_{ML} = \frac{1}{d-q}\sum_{j=q+1}^{d}\lambda_j \tag{3.13}$$

where the q column vectors in the $d \times q$ matrix $\mathbf{U}_q$ are the principal eigenvectors of $\mathbf{S}$, with the associated eigenvalues in the $q \times q$ diagonal matrix $\mathbf{\Delta}_q$, $\mathbf{R}$ is an arbitrary $q \times q$ orthogonal rotation matrix and λ_j corresponds to the eigenvalue associated to the eigenvector j. Thus, from Equation 3.12, the latent variable model defined by Equation 3.7 effects a mapping from the latent space into the principal subspace of the observed data.

To implement PPCA, the usual eigen-decomposition of the covariance matrix $\mathbf{S}$ is firstly computed. Then, σ^2_{ML} is estimated from Equation 3.13 and, finally, the values of $\mathbf{W_{ML}}$ are found from Equation 3.12. For simplicity, the matrix $\mathbf{R}$ is chosen as $\mathbf{R} = \mathbf{I}$.

The posterior distribution of the latent variables given the observed variables can be calculated by:

$$p(\mathbf{x}|\mathbf{y}) = N(\mathbf{M}^{-1}\mathbf{W_{ML}}^T\mathbf{y}, \sigma^2_{ML}\mathbf{M}^{-1}) \tag{3.14}$$

where $\mathbf{M} = \mathbf{W_{ML}}^T\mathbf{W_{ML}} + \sigma^2_{ML}\mathbf{I}$. Whereas $\mathbf{M}$ is of size $q \times q$, $\mathbf{C}$ is of size $d \times d$. In this way, the optimal least-squares linear reconstruction of the data can be obtained from:

$$\hat{\mathbf{y}} = \mathbf{W_{ML}}(\mathbf{W_{ML}}^T\mathbf{W_{ML}})^{-1}\mathbf{W_{ML}}^T\mathbf{x} \tag{3.15}$$

3.3. The Gaussian Process Latent Variable Model

As mentioned above, PPCA relates a vector of latent variables $\mathbf{x}$ to a vector of observed variables $\mathbf{y}$ through a linear relationship given by $\mathbf{W}$. The latent variables are then marginalised and the values of $\mathbf{W}$ are found when the marginal likelihood is maximised. However, the GPLVM, as explained in [40, 41], arises from a novel interpretation of PPCA referred to as Dual Probabilistic Principal Component Analysis (DPPCA) which, unlike PPCA, marginalises $\mathbf{W}$ and optimises the latent variables.

In DPPCA, the marginal likelihood for the observed data takes the form:

$$p(\mathbf{Y}|\mathbf{X}, \sigma^2) = \int \prod_{i=1}^{d} p(\mathbf{y}_{:,i}|\mathbf{X}, \mathbf{W}, \sigma^2) p(\mathbf{W}) d\mathbf{W} = \prod_{i=1}^{d} N(\mathbf{y}_{:,i}|\mathbf{0}, \mathbf{K}) \tag{3.16}$$

where $\mathbf{K} = \mathbf{X}\mathbf{X}^T + \sigma^2\mathbf{I}$, $\mathbf{y}_{:,i}$ represents the ith column of $\mathbf{Y}$ and the Gaussian prior over the parameters $\mathbf{W}$ is defined as:

$$p(\mathbf{W}) = \prod_{i=1}^{d} N(\mathbf{w}_i|\mathbf{0}, \mathbf{I}) \tag{3.17}$$

Moreover, the maximum-likelihood estimator for $\mathbf{X}$ and σ^2 is obtained from an iterative minimisation of the negative log-likelihood function defined as:

$$L = \frac{n}{2} \ln(2\pi) + \frac{1}{2} \ln |\mathbf{K}| + \frac{1}{2} \operatorname{tr}(\mathbf{K}^{-1}\mathbf{S}) \tag{3.18}$$

where $\mathbf{S}$ is the covariance matrix, $d^{-1}\mathbf{Y}\mathbf{Y}^T$. However, as in PPCA, a closed-form solution can be applied. In particular, the values of $\mathbf{X}$ and σ^2 which maximise the marginal likelihood are given by:

$$\mathbf{X_{ML}} = \mathbf{U}(\mathbf{\Delta} - \sigma^2\mathbf{I})^{-\frac{1}{2}}\mathbf{R} \tag{3.19}$$

$$\sigma^2_{ML} = \frac{n - q}{\sum_{j=q+1}^{n} \lambda_j} \tag{3.20}$$

where $\mathbf{U}$ is an $N \times q$ matrix whose columns are the first q eigenvectors of $\mathbf{Y}\mathbf{Y}^T$, $\mathbf{\Delta}$ is a $q \times q$ diagonal matrix with the eigenvalues associated with the q eigenvectors of $d^{-1}\mathbf{Y}\mathbf{Y}^T$, $\mathbf{R}$ is an arbitrary $q \times q$ orthogonal rotation matrix and λ_j corresponds to the eigenvalue associated to the eigenvector j.

The marginal likelihood given in Equation 3.16 can be seen as a product of d independent GPs with a linear covariance function $\mathbf{K}$ where each process is

associated with a different dimension of $\mathbf{Y}$ (see [50] for more information about GPs). Therefore, a natural extension of DPPCA is the non-linearisation of the mapping from the latent space to the data space through the introduction of a non-linear covariance function. In this way, the GPLVM emerges as a probabilistic generalisation of PCA:

$$p(\mathbf{Y}|\mathbf{X}) = \prod_{i=1}^{d} N(\mathbf{y}_{:,i}|\mathbf{0}, \mathbf{K}) \tag{3.21}$$

being $\mathbf{K}$ the kernel or covariance function that can be either linear or non-linear.

The main advantage of the GPLVM is that a smooth mapping from the latent to the data space can be obtained. That is, points that are close in the latent space are close in the data space, however, this does not imply that close points in the data space are close in the latent space. For that reason, in [42], certain constraints in the model are examined in order to figure out this problem. On the other hand, unlike PPCA and DPPCA, there is no closed-form solution for the GPLVM. Iterative procedures have to be applied to find the optimal values of the latent variables $\mathbf{X}$ and the kernel parameters through the minimisation of the negative log-likelihood function L. In order to optimise the function, these iterative procedures normally rely on gradient descent algorithms, such as Scaled Conjugate Gradient Algorithm (SCG), which is described in [44]. The gradient of the negative log-likelihood function with respect to the kernel is computed as:

$$\frac{\partial L}{\partial \mathbf{K}} = \mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^T\mathbf{K}^{-1} - d\mathbf{K}^{-1} \tag{3.22}$$

and applying the chain rule with $\frac{\partial \mathbf{K}}{\partial \mathbf{X}}$ allows to obtain the optimal values of $\mathbf{X}$. Furthermore, gradients with respect to the parameters of the kernel matrix can be computed and used to optimise the latent variables and the parameters of the kernel.

3.4. The Gaussian Process Dynamical Model

The GPDM, described in [62, 63], is directly inspired by the GPLVM. The GPDM provides a framework for transforming a sequence of feature vectors, which are related in time, into a low dimensional latent space. In order to apply this transformation, the observation and dynamics mappings are computed separately in a non-linear form as the GPLVM does, marginalising out both mappings and optimising the latent variables and the hyperparameters of the kernels. The incorporation of dynamics not only allows to make predictions about future data but

also helps to regularise the latent space for modelling temporal data. Therefore, if the dynamic process defined by the latent trajectories in the latent space is smooth, the models will tend to make good predictions. Additionally, to learn smoother models, the B-GPDM is an alternative learning approach.

In the GPDM, the conditional probability of $\mathbf{Y}$ given $\mathbf{X}$, $\boldsymbol{\theta}$ and $\mathbf{W}$ for the observation mapping is defined as:

$$p(\mathbf{Y}|\mathbf{X},\boldsymbol{\theta},\mathbf{W}) = \frac{|\mathbf{W}|^n}{\sqrt{(2\pi)^{nd}|\mathbf{K_Y}|^d}} exp\left(-\frac{1}{2} tr\left(\mathbf{K_Y}^{-1}\mathbf{Y}\mathbf{W}^2\mathbf{Y}^T\right)\right) \tag{3.23}$$

where $\mathbf{Y}$ is the centred observed dataset, $\mathbf{X}$ represents the latent positions on the model, $\mathbf{K_Y}$ is the kernel matrix with hyperparameters $\boldsymbol{\theta} = [\theta_1, \theta_2, ..., \theta_n]$, n is the number of samples, d is the dimension of the dataset and $\mathbf{W}$ is the diagonal scaling matrix which model the variance in each data dimension. The elements of the kernel matrix for the observation mapping are normally computed using:

$$\mathbf{k}(\mathbf{x}_i, \mathbf{x}_j) = \theta_1 exp\left(\frac{-\theta_2}{2}(\mathbf{x}_i - \mathbf{x}_j)^T(\mathbf{x}_i - \mathbf{x}_j)\right) + \theta_3\delta_{i,j} \tag{3.24}$$

where $\delta_{i,j}$ is the Kronecher delta function. Nonetheless, another definition of the kernel function can be specified depending on the application considered.

The dynamic mapping on the latent coordinates is defined as:

$$p(\mathbf{X}|\boldsymbol{\beta}) = \frac{p(\mathbf{x}_1)}{\sqrt{(2\pi)^{(n-1)q}|\mathbf{K_X}|^q}} exp\left(-\frac{1}{2} tr\left(\mathbf{K_X}^{-1}\mathbf{X}_{2:n}\mathbf{X}_{2:n}^T\right)\right) \tag{3.25}$$

where $\mathbf{X}_{2:n} = [\mathbf{x}_2, ..., \mathbf{x}_n]^T$, q is the model dimension, and $\mathbf{K_X}$ is the kernel matrix constructed from $\mathbf{X}_{1:n-1} = [\mathbf{x}_1, ..., \mathbf{x}_{n-1}]^T$ using the kernel function:

$$\mathbf{k}(\mathbf{x}_i, \mathbf{x}_j) = \beta_1 exp\left(\frac{-\beta_2}{2}(\mathbf{x}_i - \mathbf{x}_j)^T(\mathbf{x}_i - \mathbf{x}_j)\right) + \beta_3\mathbf{x}_i^T\mathbf{x}_j + \beta_4\delta_{i,j} \tag{3.26}$$

where $\boldsymbol{\beta} = [\beta_1, \beta_2, ..., \beta_n]$ are the kernel hyperparameters. Nonetheless, as before, another definition of the kernel function can be specified depending on the application considered.

The combination of the observation and dynamics mappings defines the model:

$$p(\mathbf{X}, \mathbf{Y}, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{W}) = p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}, \mathbf{W})p(\mathbf{X}|\boldsymbol{\beta})p(\boldsymbol{\beta})p(\boldsymbol{\theta})p(\mathbf{W}) \tag{3.27}$$

where the priors of $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ and $\mathbf{W}$ are defined as:

$$p(\boldsymbol{\theta}) \propto \prod_i \theta_i^{-1} \tag{3.28}$$

$$p(\boldsymbol{\beta}) \propto \prod_i \beta_i^{-1} \tag{3.29}$$

$$p(\mathbf{W}) = \prod_{i=1}^{d} \frac{2}{\kappa\sqrt{2\pi}} exp(-\frac{w_i^2}{2\kappa^2}) \tag{3.30}$$

where κ is a constant.

The goal of learning the GPDM is to find the latent positions $\mathbf{X}$ and the kernel hyperparameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ with respect to the observations $\mathbf{Y}$ by iteratively minimising the negative log-posterior function $-ln\, p(\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{W}|\mathbf{Y})$ that is given by:

$$L = L_{\mathbf{Y}} + L_{\mathbf{X}} + \sum_j ln\, \theta_j + \frac{1}{2\kappa^2} tr\,(\mathbf{W}^2) + \sum_j ln\, \beta_j \tag{3.31}$$

where

$$L_{\mathbf{Y}} = \frac{d}{2} ln\, |\mathbf{K_Y}| + \frac{1}{2} tr\,(\mathbf{K_Y}^{-1}\mathbf{Y}\mathbf{W}^2\mathbf{Y}^T) - n\, ln\, |\mathbf{W}| \tag{3.32}$$

$$L_{\mathbf{X}} = \frac{q}{2} ln\, |\mathbf{K_X}| + \frac{1}{2} tr\,(\mathbf{K_X}^{-1}\mathbf{X}_{2:n}\mathbf{X}_{2:n}^T) + \frac{1}{2}\mathbf{x}_1^T\mathbf{x}_1 \tag{3.33}$$

In order to increase the smoothness of the learned trajectories in the latent space, a modified version of the GPDM can be used by changing the weight of $L_{\mathbf{X}}$ by means of a λ element. A value for λ of $\frac{d}{q}$ is recommended in [59]. This modification is known as the B-GPDM.

Moreover, given a latent position, a feature vector and its variance can be reconstructed by:

$$\mu_{\mathbf{Y}}(\mathbf{x}) = \mathbf{Y}^T\mathbf{K_Y}^{-1}\mathbf{k_Y}(\mathbf{x}) \tag{3.34}$$

$$\sigma_{\mathbf{Y}}^2(\mathbf{x}) = \mathbf{k_Y}(\mathbf{x}, \mathbf{x}) - \mathbf{k_Y}(\mathbf{x})^T\mathbf{K_Y}^{-1}\mathbf{k_Y}(\mathbf{x}) \tag{3.35}$$

where $\mathbf{Y}$ is the centred dataset, $\mathbf{K_Y}^{-1}$ the inverse matrix of the kernel for the observation mapping provided by Equation 3.24, $\mathbf{k_Y}(\mathbf{x})$ is a column vector with elements $\mathbf{k_Y}(\mathbf{x}, \mathbf{x}_j)$ for all other latent position $\mathbf{x}_j$ in the model and $\mathbf{k_Y}(\mathbf{x}, \mathbf{x})$ is a value of computing the kernel function provided by Equation 3.24 for a latent position $\mathbf{x}$.

The GPDM also provides the grounds for predicting the next position in the latent space based on the current latent position. Thus, the next latent position

and its variance can be obtained by:

$$\mu_{\mathbf{X}}(\mathbf{x}) = \mathbf{X}_{2:n}^{T}\mathbf{K_X}^{-1}\mathbf{k_X}(\mathbf{x}) \tag{3.36}$$

$$\sigma_{\mathbf{X}}^{2}(\mathbf{x}) = \mathbf{k_X}(\mathbf{x},\mathbf{x}) - \mathbf{k_X}(\mathbf{x})^{T}\mathbf{K_X}^{-1}\mathbf{k_X}(\mathbf{x}) \tag{3.37}$$

where $\mathbf{X}_{2:n} = [\mathbf{x}_2, ..., \mathbf{x}_n]^T$, $\mathbf{K_X}$ is the kernel matrix constructed from $\mathbf{X}_{1:n-1} = [\mathbf{x}_1, ..., \mathbf{x}_{n-1}]^T$ using the kernel function provided by Equation 3.26, $\mathbf{k_X}(\mathbf{x})$ is a column vector with elements $\mathbf{k_X}(\mathbf{x}, \mathbf{x}_j)$ for all other latent position $\mathbf{x}_j$ in the model and $\mathbf{k_X}(\mathbf{x}, \mathbf{x})$ is the kernel function provided by Equation 3.26 for a latent position $\mathbf{x}$. Thereby, a prediction of k latent positions ahead can be obtained computing Equation 3.36 iteratively.

3.5. Conclusions

Throughout this chapter, the theoretical basis of the GPDM was presented. It starts describing how the dimensionality of a dataset can be reduced in a linear way applying PCA. Then, PPCA and DPPCA are introduced, taking into account how PCA can be expressed as the maximum likelihood solution of a probabilistic latent variable model. As mentioned in previous sections, PPCA marginalises the latent positions and optimises the linear transformation matrix, however DPPCA marginalises the matrix and optimises the latent positions. Additionally, this last approach can be extended by the non-linearisation of the mapping from the latent space to the data space through the introduction of a non-linear covariance function. In this way, the GPLVM emerges as a generalisation of PCA. However, it is not a dynamical model. To solve this problem, the GPDM, which is inspired by GPLVM, provides a framework for transforming a sequence of feature vectors, which are related in time, into a low dimensional latent space computing the observation and dynamics mappings separately in a non-linear form as the GPLVM does.

In this thesis, the B-GPDM is applied to predict pedestrian intentions and paths. It allows to reduce the dimensionality of a set of pedestrian motions, i.e. a set of feature vectors related in time, and infer future positions on the latent space given the latent position corresponding to the current observation. Unfortunately, learning a generic B-GPDM for all kind of pedestrian motions or combining them is a difficult task. In the next chapter, the implementation of an approach based on B-GPDMs to make pedestrian path, pose and intention predictions is explained in detail.

Chapter 4

Development

The main goal of research methods centred on pedestrian path, poses and intention predictions is to develop commercial systems set up in moving vehicles equipped with stereo cameras and LIDAR. These systems are mainly orientated to avoid vehicle-pedestrian collisions automatically. Nonetheless, not many works have been published so far about this field once the pedestrians are detected. Generally, all these works should tackle two challenges simultaneously. One is related with the information that could be more relevant to predict future pedestrian states and the other is concerned with the learning of that information. Modelling it properly may provide more accurate pedestrian estimations.

This thesis describes a method based on B-GPDMs which learns 3D time-related information from pedestrian joints in order to predict paths, poses and intentions up to 1 second in advance. As mentioned in Section 3.4, the GPDM and B-GPDM can reduce the dimensionality of a set of feature vectors related in time and infer future latent positions. Likewise, given a latent position from the latent space, the corresponding feature vector can also be reconstructed. However, as claimed in [63], learning a generic model for all kind of pedestrian activities or combining some of them into a single model normally provides inaccurate estimations of future observations. For that reason, the proposed method learns multiple models of each type of pedestrian activity, i.e. walking, stopping, starting and standing, and selects the most appropriate among them to estimate future pedestrian states at each instant of time. A general description of the method is shown in Figure 4.1. A training dataset of pedestrian motion sequences is split into 8 subsets based on typical crossing orientations, that is, from left to right and from right to left, and type of activity. Then, a B-GPDM is obtained for each sequence contained in

the dataset. On the other hand, given a new pedestrian observation, the current activity is determined. Thus, the selection of the most appropriate model among the trained ones is centred solely on that activity. Finally, the selected model is used to predict the future latent positions and reconstruct the future pedestrian path and poses.

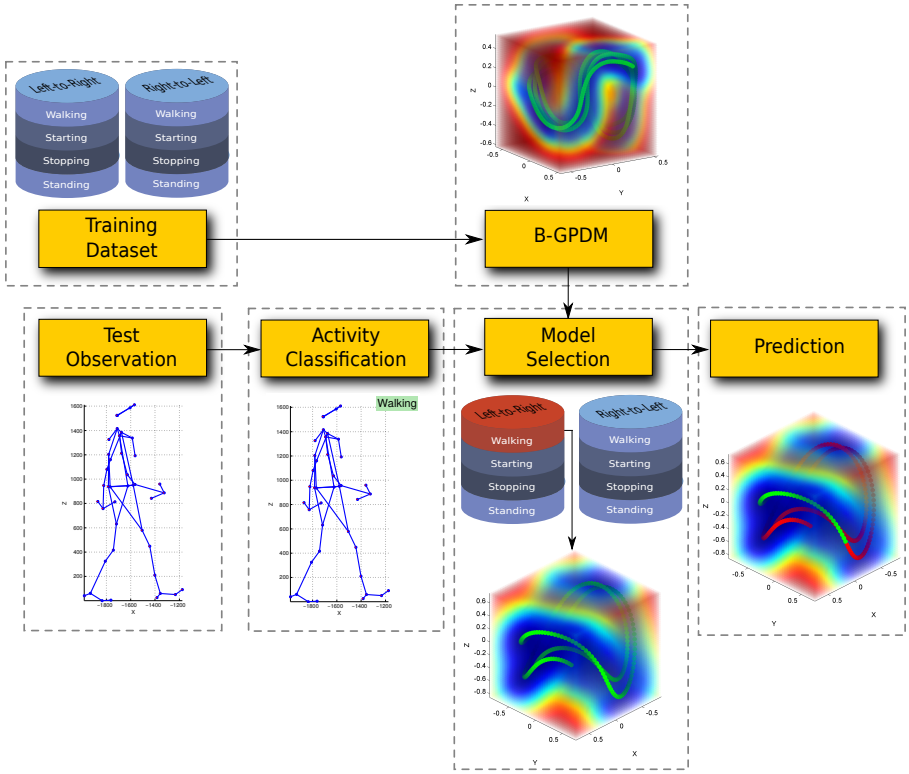


Figure 4.1: General description of the pedestrian path, pose and intention prediction method.

In this chapter, the dataset of pedestrian motion sequences and the information which is extracted and learned to create the models are firstly described in Section 4.1. Then, in Section 4.2, a pedestrian skeleton estimation algorithm is detailed. This algorithm enables to obtain noisy pedestrian observations by means of a stereo vision system. In Section 4.3, an exhaustive analysis about the learning methods is done considering different model dimensionalities, activities and pedestrian joints. After that, in Section 4.4, the algorithm to recognise pedestrian activities is explained. Then, how pedestrian pose, path and intention predictions are computed is discussed in Section 4.5. Finally, the main conclusions of this chapter are described in Section 4.6.

4.1. Dataset Description

One of the goals of this thesis is to test the feasibility and limits of the proposed method in an extensive way under ideal conditions by using a high frequency and low noise dataset published by CMU (see [10] for more information). On the one side, the high frequency of the dataset will help the algorithms to properly learn the dynamics of different activities and will increase the probability of finding a similar test observation in the trained data without missing intermediate observations. On the other side, low noise models will improve the prediction when working with noisy test samples. The CMU dataset contains sequences in which people are carrying out multiple activities captured by a Vicon motion capture system, which consists of 12 infrared MX-40 cameras, in a working volume of approximately 3 x 8 meters (see [60] for more information). In several of these sequences, people are simulating typical pedestrian activities at the same time that 3D coordinates of 41 joints along their bodies are being gathered at 120 Hz. An example of a walking pedestrian pose from different points of view is shown in Figure 4.2.

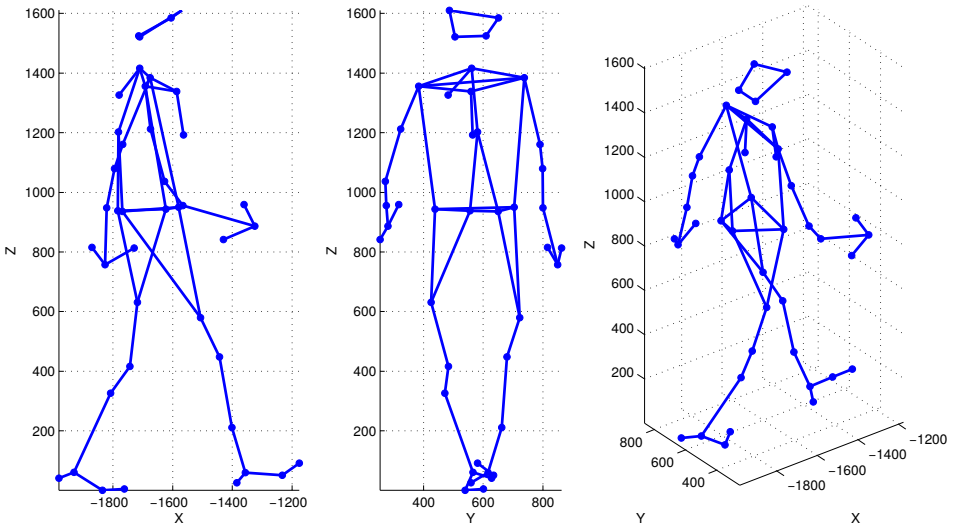


Figure 4.2: Example of pedestrian pose extracted from the dataset published by CMU in which 41 joints, represented by blue markers, are shown.

Nevertheless, not all gathered joints offer discriminative information about the current and future pedestrian activities. In fact, joints located along the arms do not contribute to distinguish walking, starting, stopping or standing activities. For that reason, a subset of 11 joints has been selected in order to determine whether the detection of only shoulder and leg motions are enough to infer future states.

In [35], head positions, centres of gravity, feet and their respective velocities are analysed during the gait initiation. The study deduces that, whereas the centres of gravity and head positions are the least sensitive information, feet position changes indicate more reliably the gait initiation. An example of a pedestrian pose of this subset from different points of view is shown in Figure 4.3.

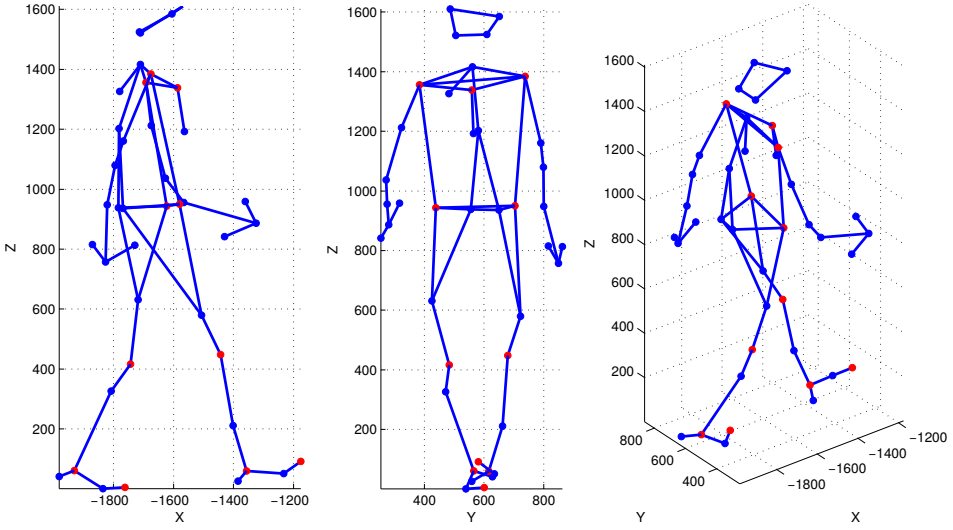


Figure 4.3: Example of pedestrian pose extracted from the dataset published by CMU in which 11 joints, represented by red markers, are shown.

Because of the large number of activities in the dataset published by CMU, an extraction of valid sequences or subsequences were performed. The criterion adopted for this process was based on the premises that a pedestrian does not change the orientation along the sequence and only carries out one or several of the activities considered in this thesis. In such a way, 490 sequences composed of 302470 pedestrian poses from 31 subjects were extracted. Hereafter, this set of sequences will be named as University of Alcalá (UAH) dataset.

After this extraction, the UAH dataset was hierarchically divided into 8 subsets. The first division was based on the orientation of typical crossing activities, i.e. left-to-right and right-to-left. The second one was based on the type of activity, i.e. walking, starting, stopping and standing. Those sequences with more than one activity were cropped into subsequences with only one action. However, as mentioned in Section 2.3, none of the works reviewed in this thesis offers a discussion about the best method of event-labelling, i.e. how to identify the instant that a pedestrian starts or finishes an event such as crossing, starting or stopping. Consequently, a guideline is proposed in Section 4.1.1.

Moreover, it is worth mentioning that each pedestrian observation is composed of 3D positions which belong to 41 or 11 joints, depending on the dataset chosen, and their displacements between two consecutive instants of time. These displacements are essential features for two main reasons. Firstly, they make possible the reconstruction of future pedestrian paths considering single-frame evaluation. And, secondly, they will help to recognise more accurately the pedestrian activity since the only consideration of the body pose would not enable to determine whether a pedestrian is moving or not.

4.1.1. Event-labelling Methodology

In this section, a guideline of event-labelling orientated to typical pedestrian activities is proposed. This guideline allows to identify the instant that a pedestrian starts or finishes an event such as starting or stopping. Thereby, a starting activity is defined as the action that begins when the pedestrian moves one knee to initiate the gait and ends when the foot of that leg touches the ground again. In addition, a stopping activity is defined as the action that begins when a foot is raised for the last step and finishes when that foot treads the ground. Examples of transitions manually labelled from standing to starting, starting to walking, walking to stopping and stopping to standing are shown in Figures 4.4, 4.5, 4.6 and 4.7 respectively. This criterion was adopted because these events happen in all UAH sequences in which starting or stopping activities are included and because they are easily labelled by human experts, thus enabling the creation of reliable groundtruths. A breakdown of the UAH dataset based on the number of sequences and pedestrian poses is shown in Table 4.1 .

	Orientation	**Walking**	**Starting**	**Stopping**	**Standing**	**Total**
Sequences	Left-to-right	240	142	56	224	662
Sequences	Right-to-left	191	121	27	156	495
Total sequences		431	263	83	380	1157
Pedestrian poses	Left-to-right	107324	10732	2522	43151	163729
Pedestrian poses	Right-to-left	95113	10940	1276	31412	138741
Total pedestrian poses		202437	21672	3798	74563	302470

Table 4.1: Breakdown of UAH dataset based on the number of sequences and pedestrian poses for each type of activity.

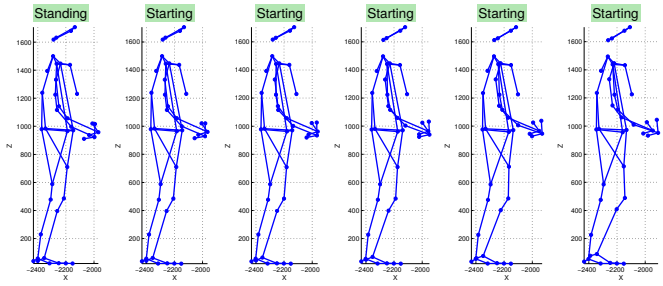


Figure 4.4: Example of transition manually labelled from standing to starting.

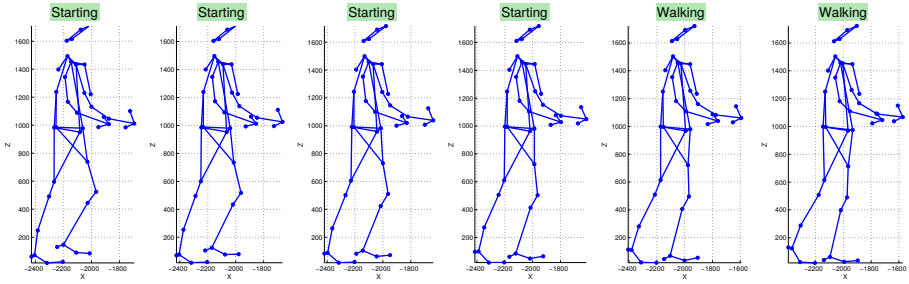


Figure 4.5: Example of transition manually labelled from starting to walking.

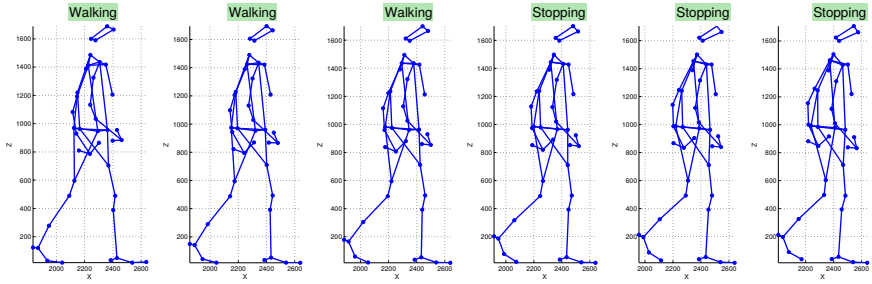


Figure 4.6: Example of transition manually labelled from walking to stopping.

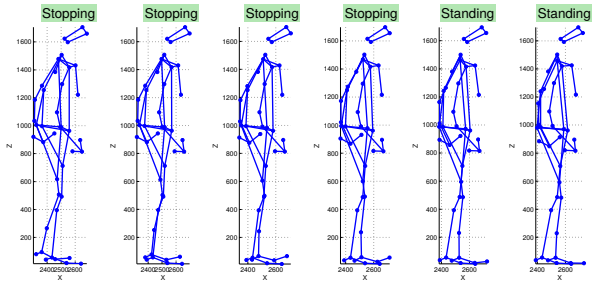


Figure 4.7: Example of transition manually labelled from stopping to standing.

4.2. Pedestrian Skeleton Estimation

Additionally, to test the proposed method with noisy observations, a single-frame pedestrian skeleton estimation algorithm based on point clouds extracted from a stereo vision system and geometrical constraints is implemented. The stereo pair is composed of two colour cameras with a resolution of 1920 x 1200 pixels and a focal length of 12.5 millimetres which captures images at a frequency of 120 Hz. A baseline of 40 centimetres was set in order to detect pedestrians in a range from 5 to 15 metres. The estimated skeletons are composed of 11 3D points placed along the pedestrian body which represent the shoulders, hips, knees, ankles and toes. It is worth mentioning that this set of points is the same set described in Section 4.1. The algorithm assumes that a pedestrian is standing and his highest point corresponds to the head.

4.2.1. Pedestrian 3D Point Cloud Extraction

Although the motivation of this thesis is not to develop a complex pedestrian detection algorithm, a good segmentation is required for the skeleton estimation. For this reason, a simple pedestrian segmentation method is implemented by applying a Gaussian mixture model background subtraction, described in [67, 68], from depth maps. This method avoids errors caused by shadows and pixels with similar values in the original images which pertain to the background and pedestrians. Nevertheless, some errors could arise if pedestrians are close enough to an object from the background and their feet could not be segmented correctly due to the fact that their values on the depth map are similar to the values corresponding to the ground floor.

Based on these considerations, the vision-based pedestrian segmentation algorithm works as follows. Firstly, the depth map is computed by means of the Semi Global Matching (SGM) algorithm. Then, the pixels that represent the ground floor on the tridimensional scene reconstruction are removed on the depth map. The intent of this step is to solve the problem related to the pedestrian feet mentioned before. After that, the background model from the filtered depth map is computed for the purpose of generating a foreground mask of moving objects. Finally, this mask is filtered by removing small clusters of pixels. An example of each pedestrian segmentation stage in a real crosswalk scenario is presented in Figure 4.8.

(a) Original colour image captured by a stereo pair system.

(b) Depth map.

(c) Depth map where values which correspond to the ground plane were removed.

(d) Foreground mask of moving objects.

(e) Foreground mask of moving objects with depth map values.

(f) Filtered foreground mask.

Figure 4.8: Pedestrian segmentation algorithm.

4.2.2. Skeleton Estimation

The skeleton estimation algorithm is based on the extraction of point clouds corresponding to different pedestrian body parts and the location of 3D joints in an hierarchical top-down search given anthropometric proportions and geometri-

cal constraints. These proportions are referred to the pedestrian height, so, with the intent of calculating this value, the coordinate system is translated from the stereo pair to the ground floor as shown in Figure 4.9. Thereby, the maximum y-coordinate point from the pedestrian point cloud provides the expected height, h. Likewise, the coordinate system translation enables to remove data which belong to the ground floor in the previous segmentation stage.

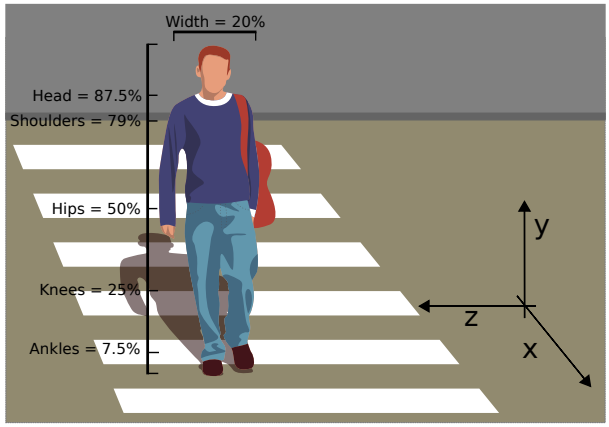


Figure 4.9: Coordinate system and anthropometric proportions with respect to pedestrian height used in the skeleton estimation algorithm.

4.2.2.1. Head

Firstly, the point cloud corresponding to the pedestrian head is extracted and its centroid, c_{head}, computed. It is important to point out that a Linear Least Squares (LLS) fitting of $t \in \{2, 3, \ldots, N\}$ consecutive head positions, c_{head}, enables to compute the pedestrian heading line, l_{head}, whose projection onto the ground plane, l'_{head}, is represented by the red line in Figure 4.12. This fitting is only carried out when pedestrians are moving since, in any other case, it could produce noisy measurements.

4.2.2.2. Shoulders and Hips

In the next step, the shoulders positions are estimated. A diagram of this process is shown in Figure 4.10. Firstly, the point cloud that belongs to the shoulders is extracted and its centroid, $c_{shoulders}$, determined. In the diagram, the point cloud that is visible is represented in black markers and the occluded body part is shown in white markers. Due to the occluded point cloud, $c_{shoulders}$ does not cor-

respond to the middle point between both shoulders. Hence, these are modelled as a circle whose centre, $centre_{shoulders}$, is the intersection of the head-based heading line, l_{head}, projected onto the plane $y = c_{shoulders_y}$ and the perpendicular line that passes through $c_{shoulders}$. The diameter of the circle corresponds to the anthropometric proportion of the pedestrian width. A prior estimation of the shoulders positions, s'_{left} and s'_{right}, assumes that they are located in this perpendicular line. Nonetheless, the final locations, s_{left} and s_{right}, are computed rotating the prior positions and getting the line that joints both shoulders and has minimum sum of point-line distance for all points in the cloud. As before, its perpendicular line, $l_{shoulders}$, could be used to compute the pedestrian heading, whose projection onto the ground plane, $l'_{shoulders}$, is represented by the green line in Figure 4.12.

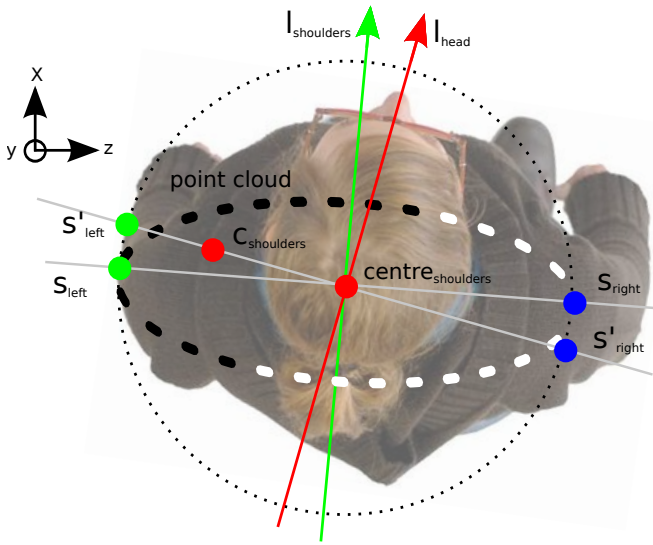


Figure 4.10: Diagram of pedestrian shoulders estimation.

The point cloud that corresponds to the pedestrian hips is also extracted using anthropometric proportions. Nonetheless, in this case, the point clouds associated with the arms and hands are removed before computing these joints. To do this, the circle that models the shoulders is projected onto the plane $y = \frac{h}{2}$. Then, the points from the pedestrian cloud which are not enclosed by this projection are removed. After that, the algorithm estimates the pedestrian hips positions in the same way as the shoulders locations. The pedestrian heading that is based on hips positions is represented by the purple line in Figure 4.12.

4.2.2.3. Lower Limbs

The lower limbs are estimated by locating the knees, ankles and toes. A diagram of this process is shown in Figure 4.11. As before, the point clouds of each body part are extracted using anthropometric proportions. Regarding the knees, a sphere, whose centre corresponds to the centre of hips, $centre_{hips}$, and radius to 25% of the pedestrian height, is used to extract the point cloud associated with these body parts. The cloud is composed of all points close to the sphere with a y-coordinate lower than $centre_{hips_y}$. In order to locate the knees positions, two methods were implemented. The first one detects clusters of points. This method works well when the pedestrian legs are separated because two clusters are clearly detected. However, in other cases, only one cluster is observed. Hence, the second method divides a point cloud into two clusters using a line. This line is selected among the heading lines previously computed and projected on the ground floor, l'_{head}, $l'_{shoulders}$ and l'_{hips}. To determine the line, the heading line based on the lower limbs, l_{legs}, is previously obtained by a LLS fitting of the point cloud extracted from the pedestrian legs. Its projection onto the ground plane, l'_{legs}, is represented by the blue line in Figure 4.12. Thus, the maximum angle between l'_{legs} and each line of the listed before determines the line that divides the original cluster. This line is represented by a black line in Figure 4.12. In this case, the line corresponds to $l'_{shoulders}$. After that, the centroids of each cluster, k_{left} and k_{right}, are computed. Nonetheless, it is assumed an occlusion when the second method detects only one cluster. To solve it, the line which joints the sensor and the non-occluded centroid is computed and used to determine the position of the occluded knee. Finally, the distances of each centroid to each hip indicate whether a knee corresponds to the left or right side of the pedestrian body.

In a similar way, the pedestrian ankles are estimated. In this case, a sphere, whose centre is also $centre_{hips}$ but its radius value is 42.5% of the pedestrian height, is modelled to extract the point clouds. Once again, the same two methods are applied to locate the ankles positions, a_{left} and a_{right}.

Finally, regarding pedestrian toes, their positions, t_{left} and t_{right}, are computed using l'_{head} and the ankles positions, a_{left} and a_{right}. Firstly, a prior positions, t'_{left} and t'_{right}, are estimated along the parallel lines to l'_{head} that passes through the ankles projections onto the ground plane, a'_{left} and a'_{right}. These prior positions are located at a distance 10% of the pedestrian height from a'_{left} and a'_{right} respectively. Then, an iterative search of the point clouds corresponding to the tiptoes is done. This search consists in extending the search radius from t'_{left} and t'_{right} until the point clouds are located. Finally, their centroids, t_{left} and t_{right}, are computed.

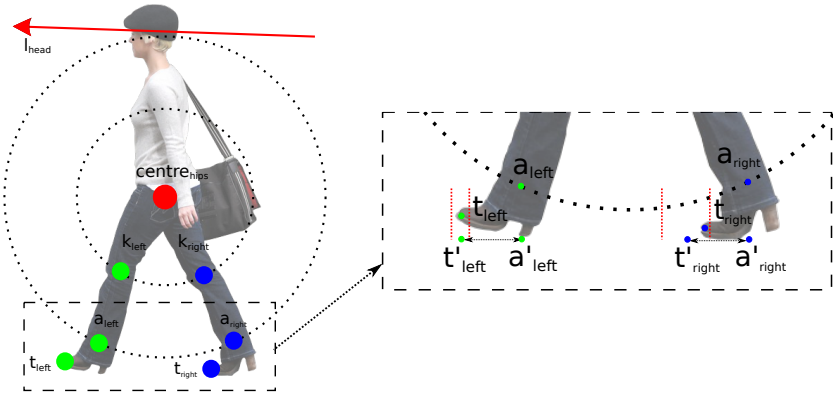


Figure 4.11: Diagram of pedestrian limbs estimation.

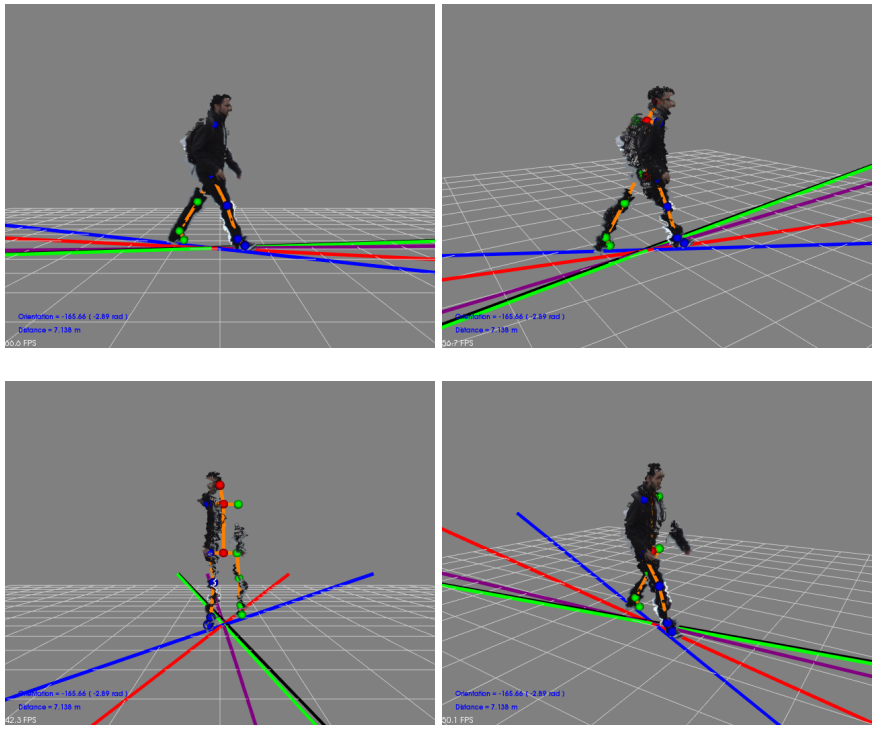


Figure 4.12: Example of a pedestrian skeleton estimation. Green markers correspond to 3D left joints, blue markers to 3D right joints and red markers to head, centre of shoulders and centre of hips. The red line indicates the pedestrian heading computed from consecutive head positions. The blue line represents the heading computed from the legs. The green line corresponds to the heading based on the shoulders positions. The purple line is associated with the heading based on the hips positions. Finally, the black line determines the line that divides the pedestrian legs.

4.3. Learning Pedestrian Motion Sequences

As mentioned above, this thesis describes a method based on the B-GPDM to learn 3D time-related information extracted from pedestrian joints in order to predict paths, poses and intentions. In contrast to PCA and the GPLVM, the GPDM and B-GPDM reduce the dimensionality of a set of feature vectors related in time and infer future latent positions. Likewise, given a latent position from the latent space, the corresponding observation can be reconstructed. Nonetheless, learning a generic model for all kind of pedestrian activities or combining some of them into a single model could produce poor dynamical predictions as claimed in [63]. For that reason, the proposed method learns multiple models for each type of pedestrian activity, i.e. walking, stopping, starting and standing, and selects the most appropriate among them to estimate future pedestrian states at each instant of time.

Throughout this section, the learning stage is described and a comparison among the methods, i.e. PCA, GPLVM, GPDM and B-GPDM, is illustrated. This comparison is done by means of models obtained from 4 sequences, which correspond to different activities, taking into account different model dimensions and pedestrian joints.

4.3.1. Learning Stage

The learning stage starts loading all cropped sequences contained in the UAH dataset. Because of the coordinate system of these sequences is referenced to the sensor, the 3D translation parameters of each observation are removed so that the origin of the reference system is relocated in the pedestrian. The deletion of these parameters let the algorithms deal with pedestrians regardless of their positions with respect to the sensors.

After that, as noted in Section 3.1, it is appropriate to scale the variables by subtracting the mean and dividing each one by its standard deviation in order to have zero-mean and unit-variance data. For this reason, this preprocessed step is applied to each sequence separately before reducing their dimensionality.

As mentioned before, the learning models based on GPs require iterative procedures. Hence, the latent positions $\boldsymbol{X}$, the hyperparameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, and the constant κ (see Equations 3.28, 3.29 and 3.30) are initialised. On the one hand, the latent coordinates are initialised by PCA and, on the other hand, the kernel parameters and κ are initialised by using the values proposed in [63]. Finally, the

GPLVMs, GPDMs and B-GPDMs are learned for each sequence. An example of a B-GPDM corresponding to a pedestrian that is walking 6 steps is shown in Figure 4.13. The green markers indicate the projection of the pedestrian observations onto the subspace. Additionally, the model variance is represented from cold to warm colours. Whereas a high variance (warm colours) indicates that illogical pedestrian observations can be reconstructed, a low variance (cold colours) indicates that pedestrian observations similar to the learned sequence may be obtained from a latent position.

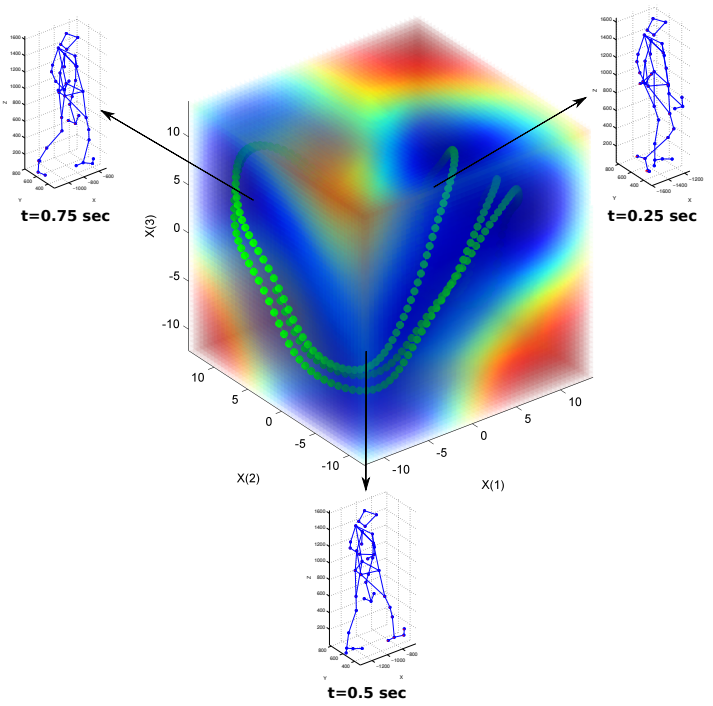


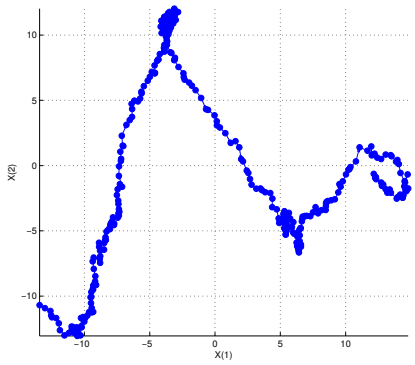
Figure 4.13: Example of a B-GPDM corresponding to a pedestrian that is walking 6 steps. The projection of the pedestrian motion sequence onto the subspace is represented by green markers. The model variance is indicated from cold to warm colours.

4.3.2. Comparison among Techniques

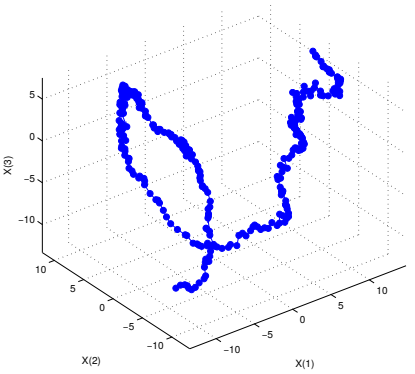
Throughout this section, a comparison among PCA, GPLVM, GPDM and B-GPDM is illustrated by means of models obtained from 4 sequences, which correspond to the activities considered in this thesis, taking into account different model dimensionalities and pedestrian joints.

4.3.2.1. Principal Component Analysis

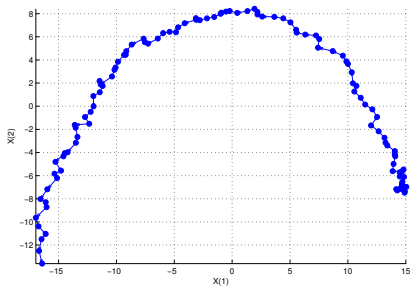
Regarding PCA, examples of 2D and 3D models are shown in Figure 4.14. These models represent observations of 3D positions and displacements of 41 joints located along the pedestrian bodies extracted from a standing, starting, stopping and walking activity respectively. Because of the high frequency and low noise sequences included in the UAH dataset, the projection of pedestrian observations related in time onto a PCA subspace emerges as well-defined trajectories. For example, walking activities generate cyclic trajectories where each cycle corresponds to two pedestrian steps. Therefore, it seems that close pedestrian observations are projected onto close positions in the subspace. An example of a cyclic model with reconstructed pedestrian observations were shown in Figure 4.13. Furthermore, starting and stopping activities generate trajectories of a half cycle since only one step was considered in the event-labelling. Finally, the models that correspond to standing sequences produce non-cyclic trajectories.



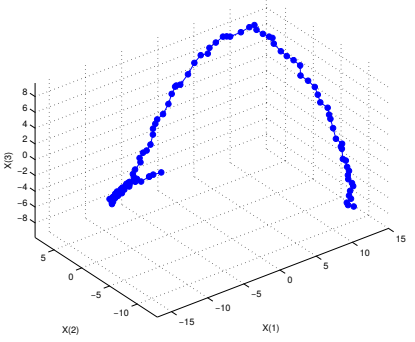
(a) 2D model of a standing activity computed by PCA.



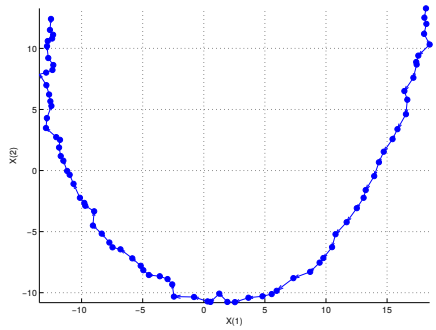
(b) 3D model of a standing activity computed by PCA.



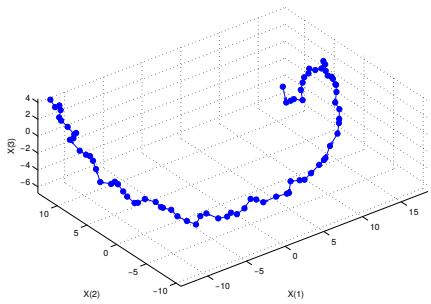
(c) 2D model of a starting activity computed by PCA.



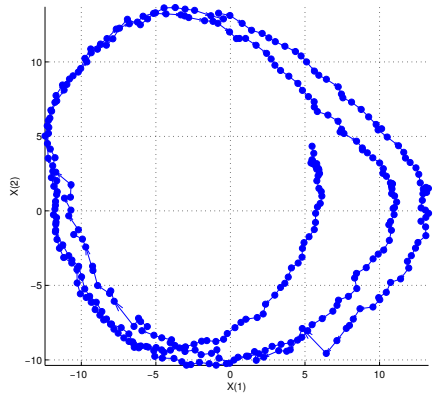
(d) 3D model of a starting activity computed by PCA.



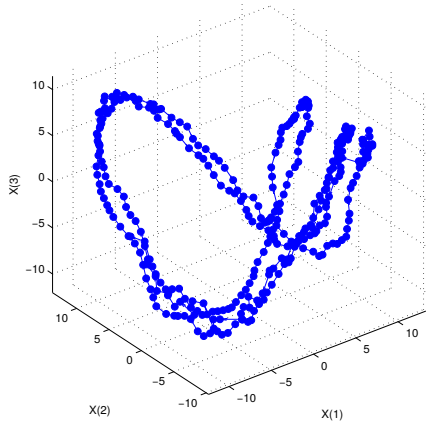
(e) 2D model of a stopping activity computed by PCA.



(f) 3D model of a stopping activity computed by PCA.



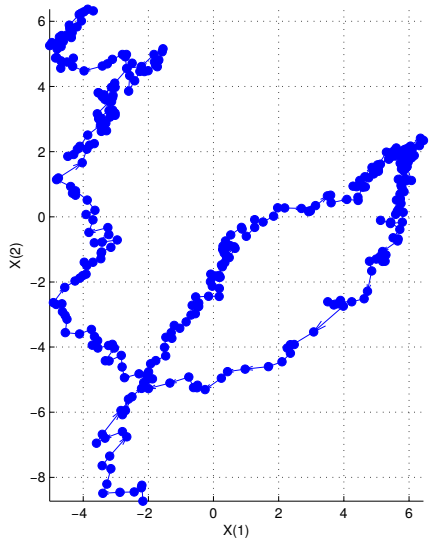
(g) 2D model of a walking activity computed by PCA.



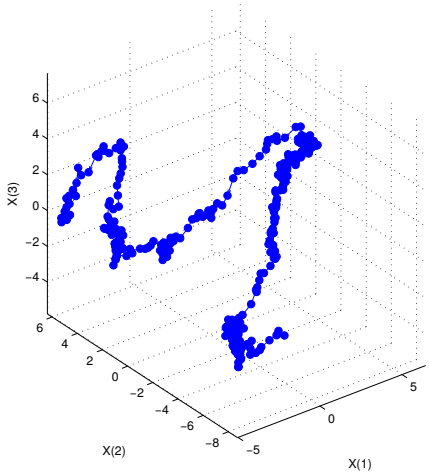
(h) 3D model of a walking activity computed by PCA.

Figure 4.14: Examples of 2D and 3D models accomplished by PCA for a standing, starting, stopping and walking activity respectively using 3D coordinates and displacements of 41 joints located along the pedestrian body.

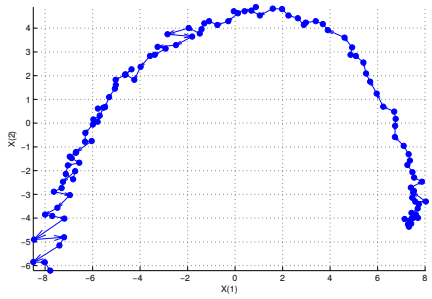
Moreover, examples of 2D and 3D models accomplished by PCA, where observations extracted from 11 pedestrian joints are only considered, are shown in Figure 4.15. Again, the projection of pedestrian observations onto a subspace emerges as well-defined trajectories. Concretely, walking activities produce cyclic trajectories where each cycle represents two pedestrian steps. Starting and stopping activities generate trajectories of a half cycle which corresponds to only one step and standing activities produce non-cyclic trajectories. As previously mentioned, it seems that close pedestrian observations are projected onto close positions in the subspace. Besides, the sequences whose observations were captured from a lower number of pedestrian joints produce noisier models than the sequences whose observations were extracted from 41 joints.



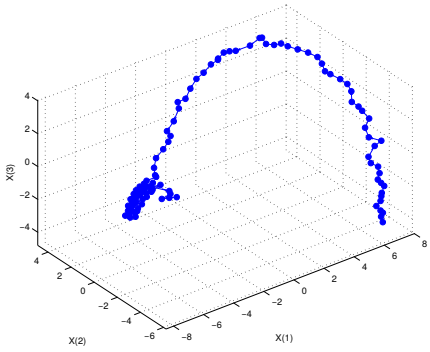
(a) 2D model of a standing activity computed by PCA.



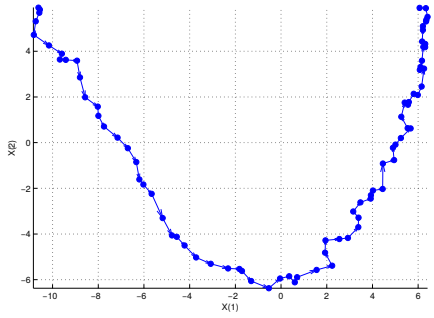
(b) 3D model of a standing activity computed by PCA.



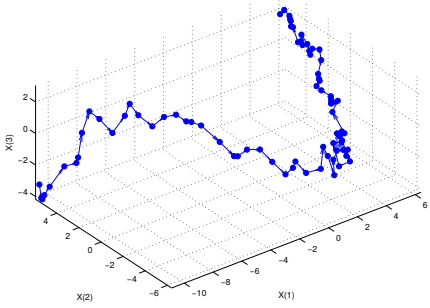
(c) 2D model of a starting activity computed by PCA.



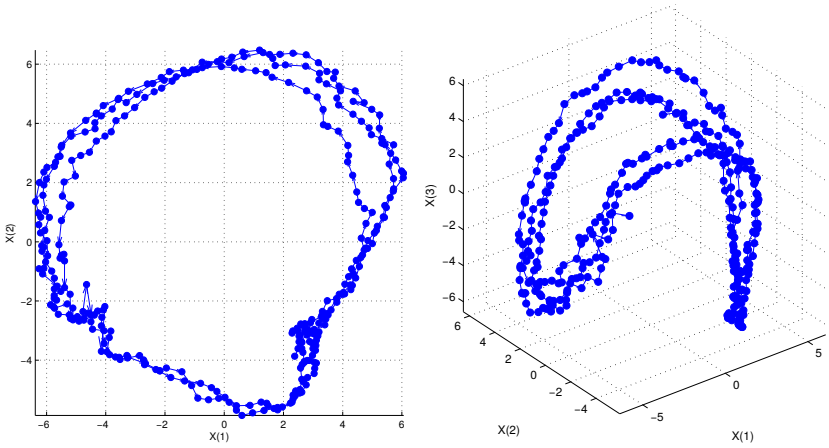
(d) 3D model of a starting activity computed by PCA.



(e) 2D model of a stopping activity computed by PCA.



(f) 3D model of a stopping activity computed by PCA.



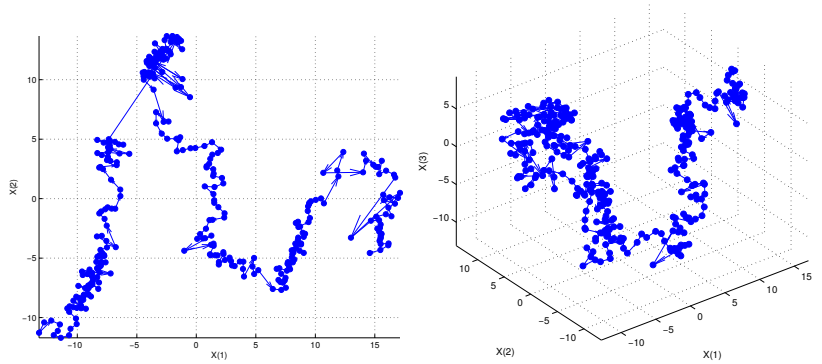
(g) 2D model of a walking activity computed by PCA.

(h) 3D model of a walking activity computed by PCA.

Figure 4.15: Examples of 2D and 3D models accomplished by PCA for a standing, starting, stopping and walking activity using 3D coordinates and displacements of 11 joints located along the pedestrian body.

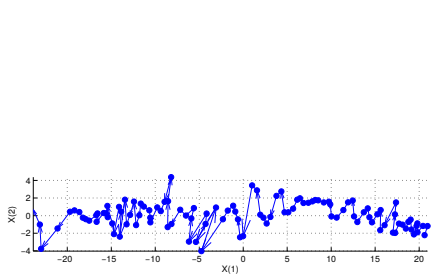
4.3.2.2. Gaussian Process Latent Variable Models

A similar analysis can be done in reference to the GPLVM. In particular, examples of 2D and 3D GPLVMs are shown in Figures 4.16 and 4.17. These figures also represent models of a standing, starting, stopping and walking activity whose observations were extracted from 3D coordinates and displacements of 41 and 11 joints respectively. As shown in the figures, very noisy trajectories are produced in the subspace. This may be caused by the fact that this modelling technique is mainly focused on pattern recognition instead of modelling time-related data.

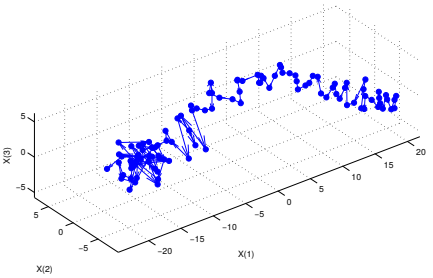


(a) 2D GPLVM of a standing activity.

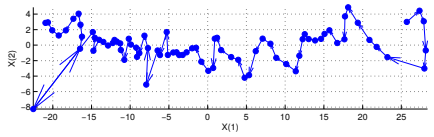
(b) 3D GPLVM of a standing activity.



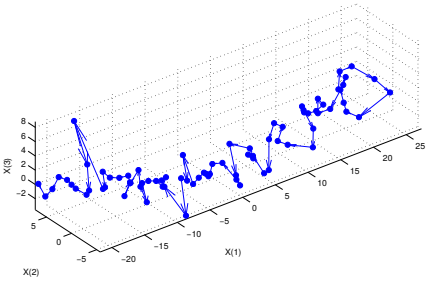
(c) 2D GPLVM of a starting activity.



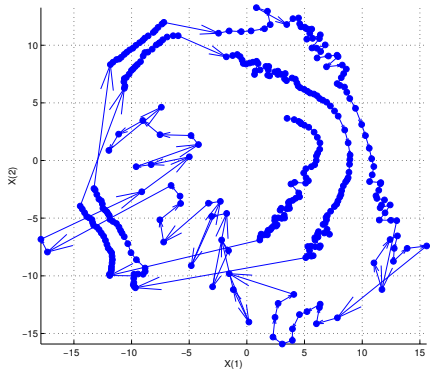
(d) 3D GPLVM of a starting activity.



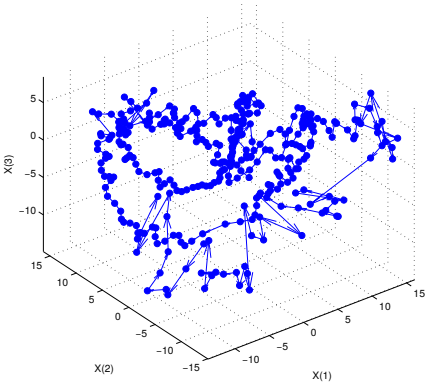
(e) 2D GPLVM of a stopping activity.



(f) 3D GPLVM of a stopping activity.



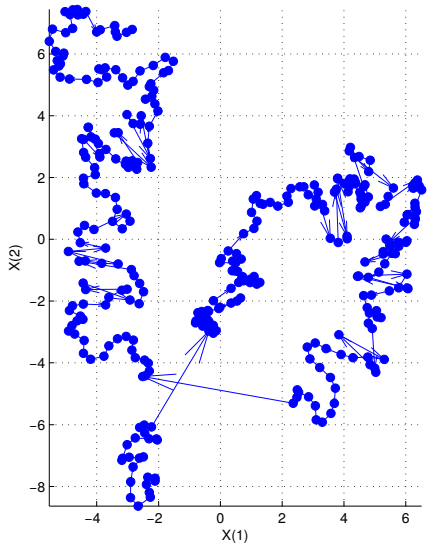
(g) 2D GPLVM of a walking activity.



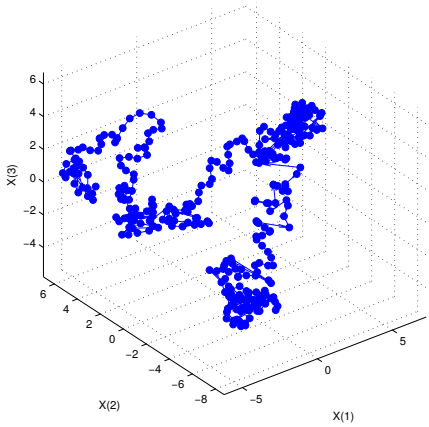
(h) 3D GPLVM of a walking activity.

Figure 4.16: Examples of 2D and 3D GPLVMs for a standing, starting, stopping and walking activity using 3D coordinates and displacements of 41 joints located along the pedestrian body.

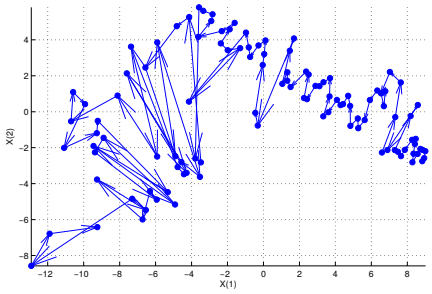
Furthermore, since the GPLVM is an iterative procedure, the trajectories created in the subspace are caused by the latent position initialisation carried out by PCA and the termination conditions chosen. These conditions are referred to the maximum number of iterations and the termination tolerance for the minimisation of the negative log-likelihood function defined in Equation 3.18.



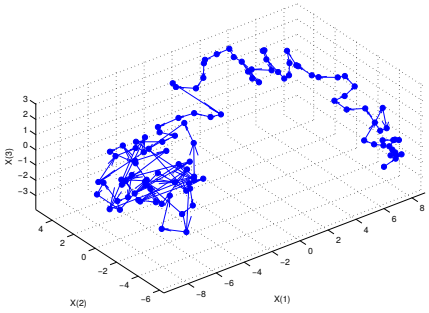
(a) 2D GPLVM of a standing activity.



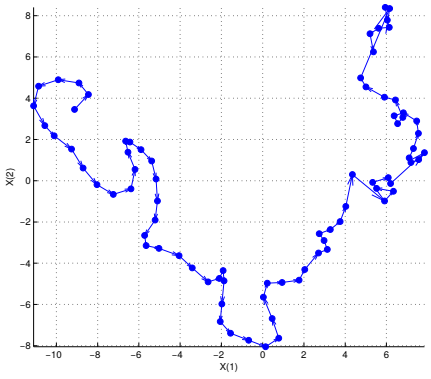
(b) 3D GPLVM of a standing activity.



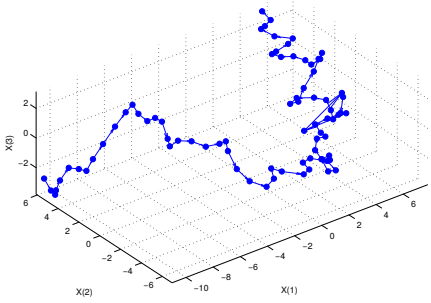
(c) 2D GPLVM of a starting activity.



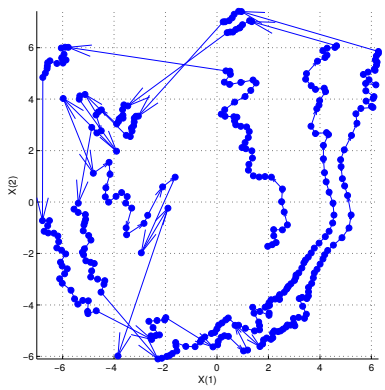
(d) 3D GPLVM of a starting activity.



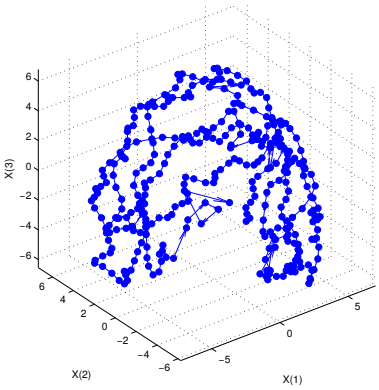
(e) 2D GPLVM of a stopping activity.



(f) 3D GPLVM of a stopping activity.



(g) 2D GPLVM of a walking activity.

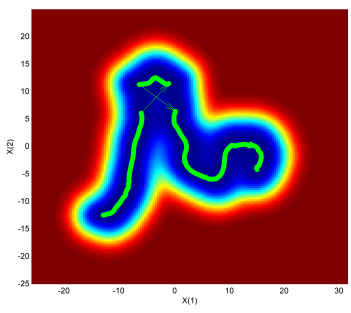


(h) 3D GPLVM of a walking activity.

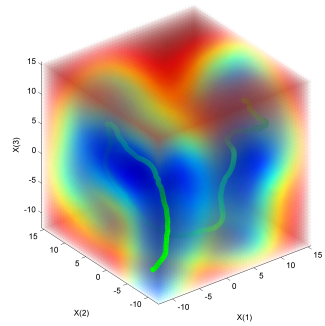
Figure 4.17: Examples of 2D and 3D GPLVMs for a standing, starting, stopping and walking activity using 3D coordinates and displacements of 11 joints located along the pedestrian body.

4.3.2.3. Gaussian Process Dynamical Models

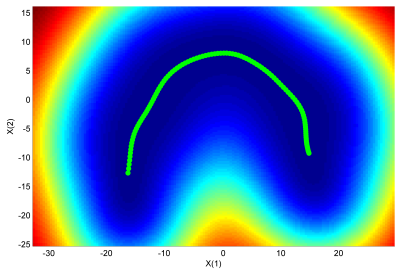
Certainly, the most interesting analysis is referred to the GPDM and B-GPDM since they are able to deal with temporal trends. As before, examples of 2D and 3D GPDMs that represent a standing, starting, stopping and walking activity are shown in Figures 4.18 and 4.19. Once again, the observations were extracted from 3D positions and displacements of 41 and 11 pedestrians joints respectively. The green markers indicates the projection of the pedestrian observations onto the subspace and the model variance is represented from cold to warm colours. A high variance (warm colours) indicates that illogical pedestrian observations can be reconstructed and a low variance (cold colours) indicates that observations similar to an observation from the learned sequence may be reconstructed.



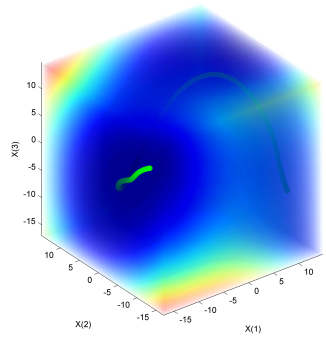
(a) 2D GPDM of a standing activity.



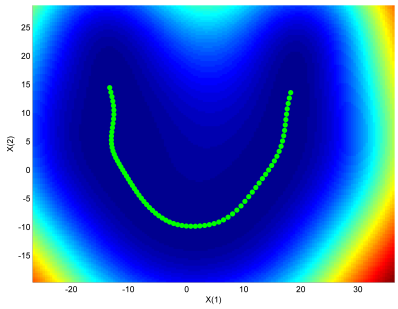
(b) 3D GPDM of a standing activity.



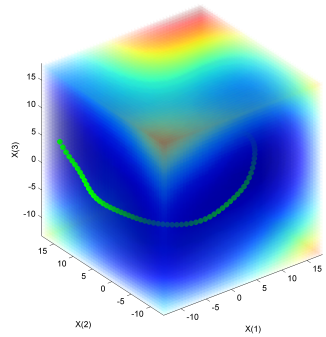
(c) 2D GPDM of a starting activity.



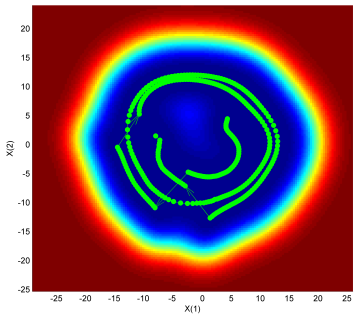
(d) 3D GPDM of a starting activity.



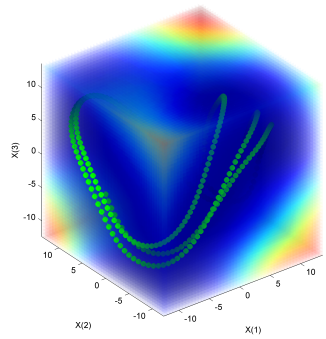
(e) 2D GPDM of a stopping activity.



(f) 3D GPDM of a stopping activity.



(g) 2D GPDM of a walking activity.

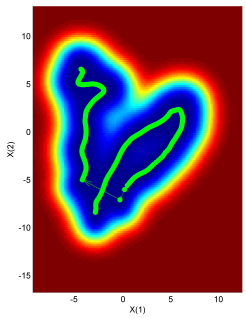


(h) 3D GPDM of a walking activity.

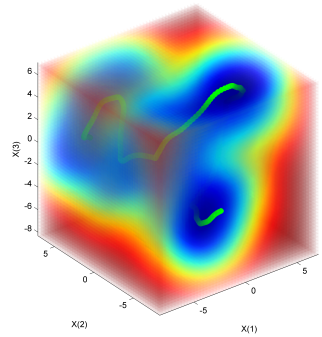
Figure 4.18: Examples of 2D and 3D GPDMs for a standing, starting, stopping and walking activity using 3D coordinates and displacements of 41 joints located along the pedestrian body.

Given that the GPDM takes into account the dynamical relationships between observations, smoother trajectories than the GPLVM are created in the subspaces. Nonetheless, several discontinuities appear in the trajectories which could produce errors in latent position predictions. Likewise, as in previous cases, walking ac-

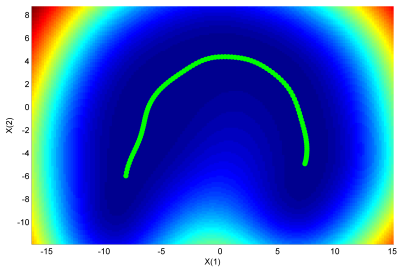
tivities generate cyclic trajectories where each cycle corresponds to two pedestrian steps. Starting and stopping activities produce half-cycle trajectories which represent only one pedestrian step. Finally, in contrast to PCA and GPLVM, standing activities generate smooth non-cyclic trajectories. Additionally, the sequences whose observations were captured from a lower number of pedestrian joints do not produce noisier models than the sequences whose observations were extracted from 41 pedestrian joints as it occurs in PCA.



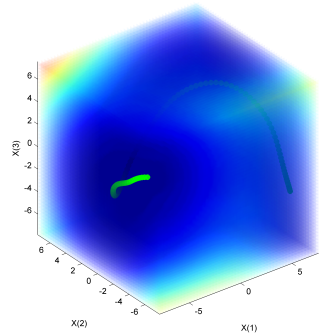
(a) 2D GPDM of a standing activity.



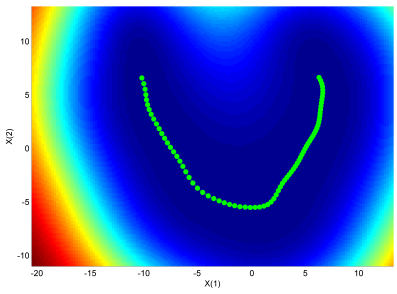
(b) 3D GPDM of a standing activity.



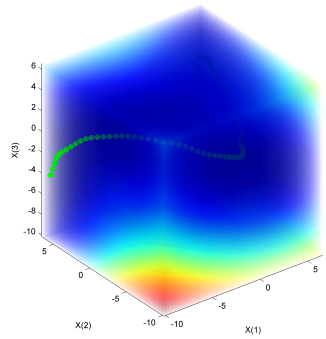
(c) 2D GPDM of a starting activity.



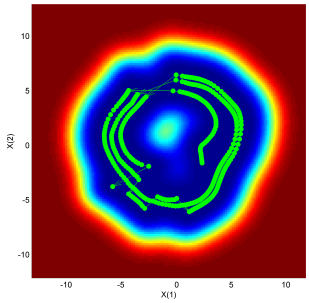
(d) 3D GPDM of a starting activity.



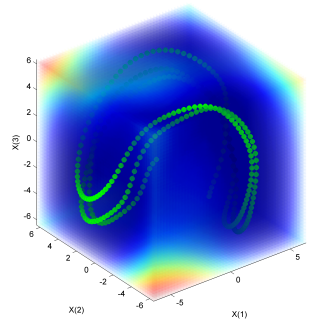
(e) 2D GPDM of a stopping activity.



(f) 3D GPDM of a stopping activity.



(g) 2D GPDM of a walking activity.

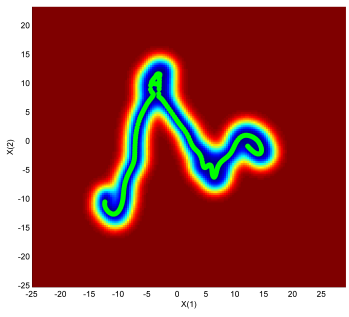


(h) 3D GPDM of a walking activity.

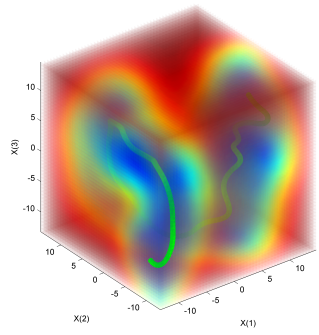
Figure 4.19: Examples of 2D and 3D GPDMs for a standing, starting, stopping and walking activity using 3D coordinates and displacements of 11 joints located along the pedestrian body.

4.3.2.4. Balanced Gaussian Process Dynamical Models

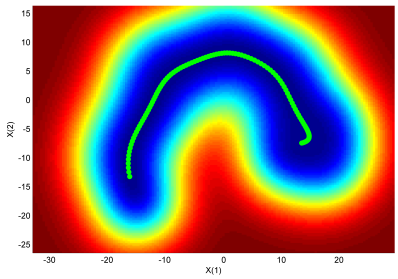
Regarding the B-GPDM, examples of 2D and 3D models that represent a standing, starting, stopping and walking activity are shown in Figure 4.20 and 4.21. Once again, the observations were extracted from 3D coordinates and displacements of 41 and 11 pedestrians joints respectively. As mentioned above, the green markers indicate the projection of the pedestrian observations onto the subspace and the model variance is represented from cold to warm colours. A high variance (warm colours) indicates that illogical pedestrian observations can be reconstructed and a low variance (cold colours) indicates that observations similar to an observation from the learned sequence may be reconstructed. Comparing the models obtained by the GPDM and B-GPDM, the latter removes discontinuities in the trajectories and increases the variance when the latent positions get further from the learned sequence.



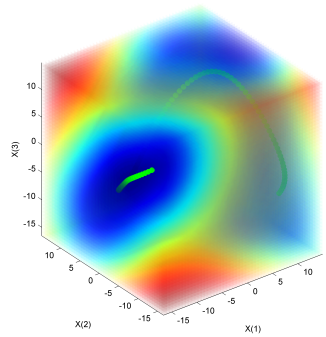
(a) 2D B-GPDM of a standing activity.



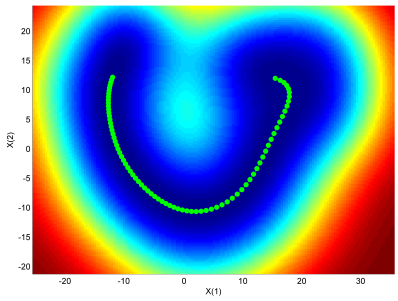
(b) 3D B-GPDM of a standing activity.



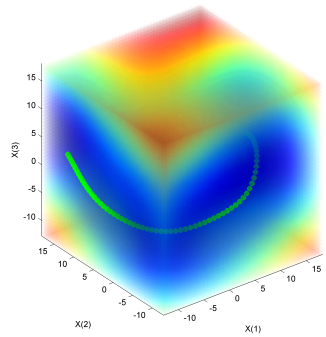
(c) 2D B-GPDM of a starting activity.



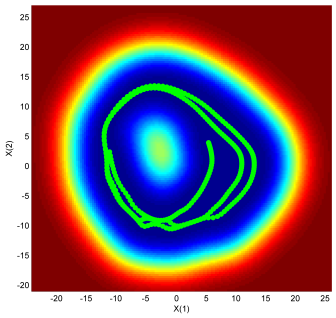
(d) 3D B-GPDM of a starting activity.



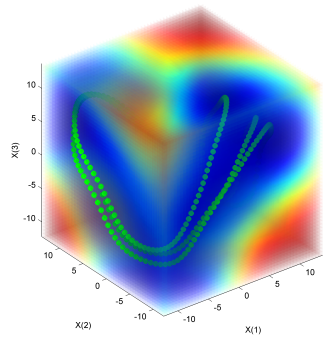
(e) 2D B-GPDM of a stopping activity.



(f) 3D B-GPDM of a stopping activity.



(g) 2D B-GPDM of a walking activity.

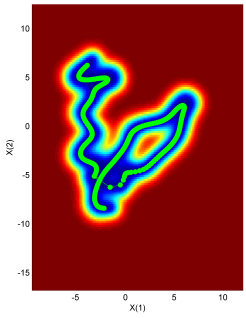


(h) 3D B-GPDM of a walking activity.

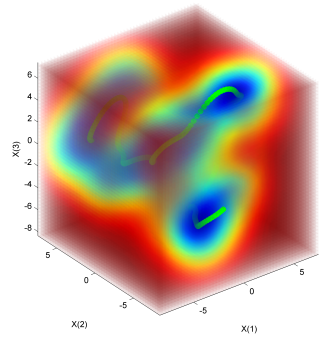
Figure 4.20: Examples of 2D and 3D B-GPDMs for a standing, starting, stopping and walking activity using 3D coordinates and displacements of 41 joints located along the pedestrian body.

In the same way as the GPDM, smoother trajectories than the GPLVM are created in the subspace. Thus, walking activities produce cyclic trajectories where each cycle corresponds to two pedestrian steps. Starting and stopping activities generate half-cycle trajectories which represent only one pedestrian step. Finally,

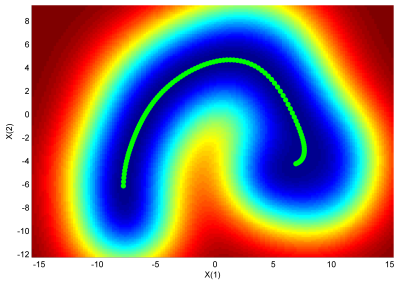
standing activities produce smooth non-cyclic trajectories. Again, the sequences whose observations were captured from a lower number of pedestrian joints do not produce noisier models than the sequences whose observations were extracted from 41 pedestrian joints as it occurs with PCA. Due to all these considerations, it seems that the B-GPDM is the most appropriate modelling technique among the methods analysed in this section in order to predict future observations of dynamical processes.



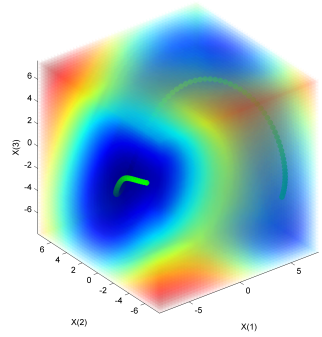
(a) 2D B-GPDM of a standing activity.



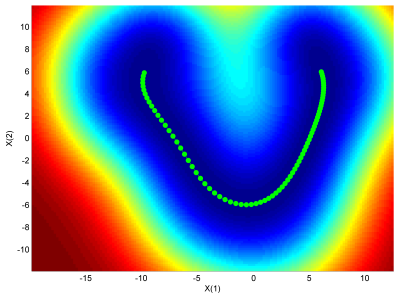
(b) 3D B-GPDM of a standing activity.



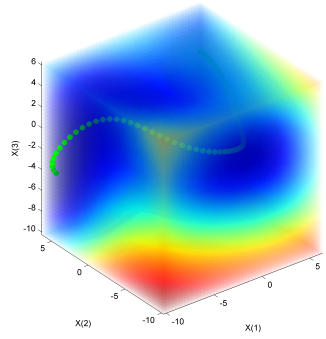
(c) 2D B-GPDM of a starting activity.



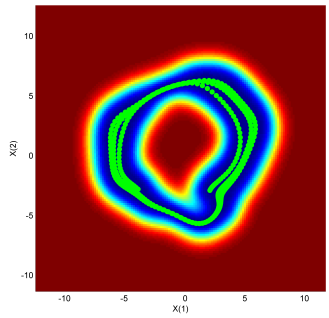
(d) 3D B-GPDM of a starting activity.



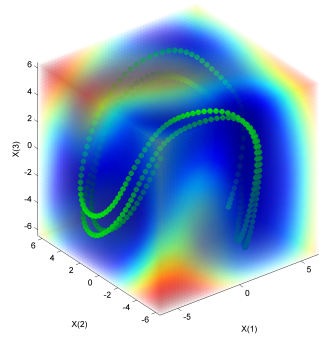
(e) 2D B-GPDM of a stopping activity.



(f) 3D B-GPDM of a stopping activity.



(g) 2D B-GPDM of a walking activity.



(h) 3D B-GPDM of a walking activity.

Figure 4.21: Examples of 2D and 3D B-GPDMs for a standing, starting, stopping and walking activity using 3D coordinates and displacements of 11 joints located along the pedestrian body.

4.4. Activity Recognition

Since several models with different dynamics were previously trained, an activity recognition from the current pedestrian observation allows to select afterwards the most accurate model to estimate future pedestrian states. The maximum similarity between the current observation and each observation of the training dataset may determine the activity. Nevertheless, if this maximum similarity is applied directly, that is, without modelling the evolution of the pedestrian activity, higher errors are achieved in selecting the most appropriate model due to the likeness between observations of different dynamics. For example, an observation of a pedestrian that is walking may be similar to an observation belonging to the beginning of a stopping sequence or to the end of a starting sequence. Thus, if the previous activity were recognised as walking, then the next dynamics would be determined as walking or stopping and not as starting. Thereby, the process of how a pedestrian changes its dynamics over time can be described by a Markov Process, which is represented in Figure 4.22. At any time, the pedestrian can do one of a set of 4 distinct actions $\mathbf{s} = \{Standing, Starting, Stopping, Walking\}$. However, these activities or states are not observable since only 3D information from joints belonging to the pedestrian is available. Therefore, the states can be only inferred through the observations $\mathbf{x}$. For this reason, the implementation of a first-order HMM allows to model the transitions between activities and to recognise the correct one taking into account the previous dynamics. A tutorial on HMM is available in [49].

The Viterbi algorithm is a dynamic programming procedure for finding the most likely state sequence given an observation sequence. That way, choosing sequences

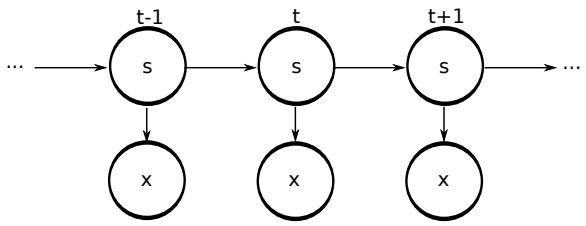


Figure 4.22: HMM graphical description.

of a single element, the probability of an observation $\mathbf{x}$ of being in the j-th state of $\mathbf{s}$ at an instant of time t is formulated as:

$$P(\mathbf{s}_j^t|\mathbf{x}^t) = \frac{P(\mathbf{x}^t|\mathbf{s}_j^t)P(\mathbf{s}_j^t)}{\sum_{i=1}^{4} P(\mathbf{x}^t|\mathbf{s}_i^t)P(\mathbf{s}_i^t)} \tag{4.1}$$

where $P(\mathbf{s}_j^t)$ represents the prior probability and $P(\mathbf{x}^t|\mathbf{s}_j^t)$ the emission probability.

The prior probability is computed as:

$$P(\mathbf{s}_j^t) \propto \max_{i=1}^{4}[P(\mathbf{s}_j^t|\mathbf{s}_i^{t-1})P(\mathbf{s}_i^{t-1}|\mathbf{x}^{t-1})], \quad t > 1 \tag{4.2}$$

where $P(\mathbf{s}_j^t|\mathbf{s}_i^{t-1})$ corresponds to the probability of changing from the i-th to the j-th state defined by means of a TPM which is graphically represented in Figure 4.23. The values of transitions between states were experimentally fixed maximising the success rate (see Section 5.1). $P(\mathbf{s}_i^{t-1}|\mathbf{x}^{t-1})$ corresponds to the probability of being in the i-th state of $\mathbf{s}$ at the previous instant. The initial probability $P(\mathbf{s}^t)$ is uniformly distributed since the pedestrian activity is unknown in $t = 1$.

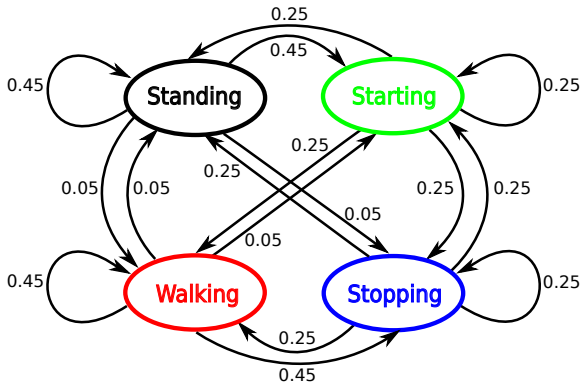


Figure 4.23: Probabilities of transitions between pedestrian activities.

The emission probability $P(\mathbf{x}^t|\mathbf{s}_j^t)$ is defined as:

$$P(\mathbf{x}^t|\mathbf{s}_j^t) \propto \max_{i=1}^{N} \left(\frac{1}{1+\alpha_i} + \frac{1}{1+\beta_i} \right) \tag{4.3}$$

where $\alpha_i \in [0, \infty]$ and $\beta_i \in [0, \infty]$ correspond to the Sum of Squared Errors (SSE) for the pedestrian pose and the displacements of the joints respectively. The SSE are computed between the current pedestrian observation and the N observations of the training data subset belonging to the j-th state of $\mathbf{s}$. Before computing α_i, the pose of the current pedestrian observation and the poses of the training observations are scaled and referenced to the same joint. The scale factor applied to each observation is obtained by the sum of ankle-knee and knee-hip distances. The displacements are not scaled with the intent of finding pedestrians with similar joint velocities.

4.5. Path, Pose and Intention Prediction

Once the pedestrian activity in t has been estimated, the selection of the most appropriate model allows to make accurate predictions about the path, poses and intentions. For this task, a search of the most similar training observation and its model is computed. This observation corresponds to the i-th element in Equation 4.3. Hence, the most appropriate model is directly selected.

Additionally, the latent position that represents the most similar observation is used as the starting point for a more accurate search in the selected model applying a gradient descent algorithm. Due to the fact that close points in the latent space are also close in the data space, it is expected that a more similar non-trained observation can be found around this starting point. The function that is minimised in the gradient descent algorithm is defined by:

$$\epsilon(\mathbf{x}) = \sum_{j=1}^{d}((\mathbf{y} - \boldsymbol{\mu})^2) + \frac{1}{2}\sum_{j=1}^{q}(\mathbf{x}^2) \tag{4.4}$$

where $\mathbf{y}$ is the current pedestrian observation and $\boldsymbol{\mu}$ represents the pedestrian observation reconstructed from the latent position $\mathbf{x}$ (see Equation 3.34). Both observations are previously scaled and referenced to the same joint. Finally, d corresponds to the dimension of the original observation and q to the dimension of the model.

Once the final latent position has been estimated, predictions of N observations

ahead are made using Equations 3.36 and 3.34 iteratively. Thereby, given the current pedestrian location with respect to the sensor, the future pedestrian path can be computed adding the consecutive N predicted displacements. It is noteworthy that the reference point to reconstruct the path is the right hip since it corresponds to a point close to the centre of gravity. Additionally, given the future pedestrian observations, the future intentions can be estimated through the application of the activity recognition algorithm explained in Section 4.4 to the N future pedestrian observations.

An example of the latent position prediction for an observation of a walking activity is shown in Figure 4.24. The most appropriate model is shown and the latent position of the most similar observation is represented by a yellow marker. This latent position corresponds to the initial point for the gradient descent algorithm. The final point is represented by the black marker. Finally, the future latent positions are shown in red markers. As expected, the predicted latent positions are close to the trained latent positions.

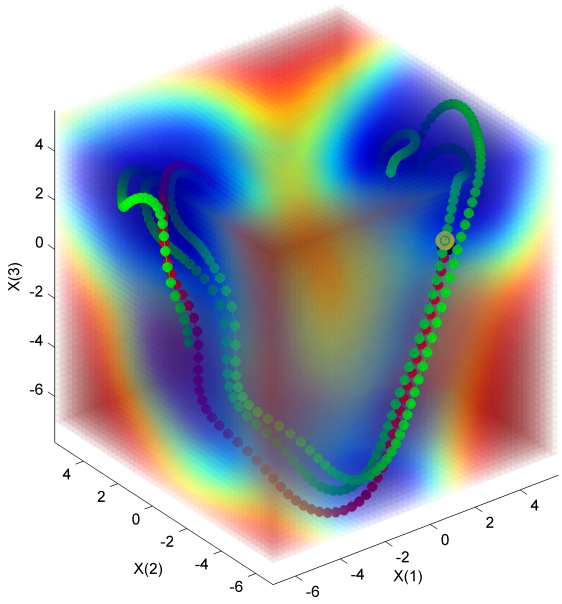


Figure 4.24: Example of latent positions prediction. The green markers indicate the projection of a walking sequence onto the subspace. The model variance is represented from cold to warm colours. The latent position of the most similar observation is represented by a yellow marker. The final point obtained by the gradient descent algorithm is represented by the black marker. The future latent positions are shown in red markers,

4.6. Conclusions

Throughout this chapter, a method based on B-GPDMs, which learns 3D time-related information from pedestrian joints, has been described with the intent of predicting paths, poses and intentions up to 1 second in advance. Given that learning a generic model for all kind of pedestrian activities or combining some of them into a single model normally provides inaccurate estimations of future observations, the method learns multiple models of each type of pedestrian activity and selects the most appropriate among them to estimate future pedestrian states at each instant of time. This strategy allows to design scalable systems in which new sequences with different dynamics can be added to the dataset without negatively impacting the performance.

Additionally, a high frequency and low noise dataset published by CMU is used in order to test the feasibility and limits of the proposed method. On the one hand, the high frequency of the dataset helps the algorithms to properly learn the dynamics of different activities and increases the probability of finding a similar test observation in the trained data without missing intermediate observations. On the other hand, low noise models improve the prediction when working with noisy test samples.

The CMU dataset is composed of sequences where people are simulating typical pedestrian activities at the same time that 3D coordinates of 41 joints along their bodies are being gathered at 120 Hz. Nonetheless, due to the fact that not all joints offer discriminative information about the current and future pedestrian activities, a subset of 11 joints is also considered. Comparing the models obtained from both set of joints, it seems that these models are not influenced by the reduction in the number of joints. In all cases, walking activities produce cyclic trajectories where each cycle corresponds to two pedestrian steps. Starting and stopping activities generate half-cycle trajectories which represent only one pedestrian step. And, finally, standing activities produce smooth non-cyclic trajectories. Due to the B-GPDM produces smoother trajectories than other models, it can be considered as the most appropriate modelling technique among the methods analysed to predict future observations of dynamical processes.

Moreover, a guideline of event-labelling is proposed in this document. A starting activity is defined as the action that begins when the pedestrian moves one knee to initiate the gait and ends when the foot of that leg touches the ground again. In addition, a stopping activity is defined as the action that begins when a foot is raised for the last step and finishes when that foot treads the ground. This

criterion was adopted because these events are easily labelled by human experts, thus enabling the creation of reliable groundtruths.

Additionally, to test the proposed method with noisy observations, a single-frame pedestrian skeleton estimation algorithm is proposed. This algorithm is based on the extraction of point clouds corresponding to different pedestrian body parts and the location of 3D joints in an hierarchical top-down search given anthropometric proportions and geometrical constraints.

Finally, since several models with different dynamics were trained, an activity recognition from the current pedestrian observation allows to select the most accurate model to estimate future pedestrian states. The maximum similarity between the current observation and each observation of the training dataset will determine the activity. Nevertheless, if this maximum similarity is applied directly, that is, without modelling the evolution of the pedestrian activity, higher errors are achieved in selecting the most appropriate model due to the likeness between observations of different dynamics. Therefore, a HMM is developed to model the pedestrian dynamics over time.

Chapter 5

Results

Throughout this chapter, the main results of the algorithms described in Chapter 4 are discussed. All algorithms were tested using the UAH dataset, which contains 490 sequences composed of 302470 pedestrian poses from 31 subjects, adopting a one vs. all strategy. This means that all the models generated by one test subject were removed from the training data before performing tests on this subject. This strategy was assumed because the number of subjects is not enough to divide the UAH dataset into two subsets, one for training and other for testing. Because of the pedestrian displacements are computed from the two initial poses, 301.980 observations are finally analysed. Additionally, the activity recognition and prediction algorithms were also tested using a sequence of noisy pedestrian data extracted by the skeleton estimation algorithm. Thereby, a more exhaustive evaluation is carried out in order to test the algorithms in a more real environment.

This chapter is structured into three main sections. Firstly, the results obtained by the activity recognition algorithm are examined in detail in Section 5.1. After that, the path prediction accuracies at several TTEs are explored in Section 5.2. Furthermore, in Section 5.3, the results of pedestrian pose prediction are examined. The processing times of each algorithm are analysed in Section 5.4. Finally, the main conclusions of this chapter are described in Section 5.5.

5.1. Activity Recognition Results

As described in Section 4.1.1, in order to test the performance of the proposed activity recognition algorithm, all pedestrian poses contained in the UAH dataset

were manually labelled by a human expert. The adopted event-labelling criteria defines a starting activity as the action that begins when the pedestrian moves one knee to initiate the gait and ends when the foot of that leg touches the ground again. Besides, a stopping activity is defined as the action that begins when a foot is raised for the last step and finishes when that foot treads the ground. Examples of transitions manually labelled are shown in Figures 4.4, 4.5, 4.6 and 4.7.

On the other hand, as described in Section 4.4, the process of how a pedestrian changes his dynamics over time is modelled by a HMM. Therefore, at any time, a pedestrian is able to carry out one of a set of 4 distinct actions, i.e. standing, starting, stopping and walking, and the probability of changing from one to another state is defined by means of the TPM illustrated in Figure 4.23. The transition values were experimentally fixed by maximising the accuracy and minimising the number of critical missclassifications, i.e. missclassifications between standing and walking, and between starting and stopping. The activity recognition results are summarised on two confusion matrices in Tables 5.1a and 5.1b. Whereas the first matrix represents the results extracted from 41 pedestrian joints, the second matrix shows the results when solely 11 joints are used. It is worth remarking that the pedestrian observations are composed of body poses and displacements.

		Predicted			
		Standing	**Starting**	**Stopping**	**Walking**
Actual	**Standing**	65979	2306	286	5692
	Starting	3171	10522	0	7959
	Stopping	181	0	1384	2232
	Walking	4163	682	1434	195989

(a) Classification results computed from 41 pedestrian joints.

		Predicted			
		Standing	**Starting**	**Stopping**	**Walking**
Actual	**Standing**	72011	1396	174	682
	Starting	1451	13313	13	6875
	Stopping	126	0	1951	1720
	Walking	262	494	1508	200004

(b) Classification results computed from 11 pedestrian joints.

Table 5.1: Classification results computed when pedestrian observations composed of body poses and displacements are used.

As claimed in Section 4.1, the pedestrian displacements help to increase the activity recognition accuracy since the only consideration of the body pose would not enable to determine whether a pedestrian is moving or not. This statement is

confirmed when these last outcomes are compared with the activity recognition results computed when observations composed solely of body poses or displacements are considered. The confusion matrices for the first case are summarised in Table 5.2 and the results based on the displacements are shown in Table 5.3.

		Predicted			
		Standing	**Starting**	**Stopping**	**Walking**
Actual	**Standing**	62614	3118	309	8222
	Starting	3304	6784	0	11564
	Stopping	270	0	1248	2279
	Walking	4221	331	1430	196286

(a) Classification results computed from 41 pedestrian joints.

		Predicted			
		Standing	**Starting**	**Stopping**	**Walking**
Actual	**Standing**	65245	2017	860	6141
	Starting	2610	8475	0	10567
	Stopping	109	2	1409	2277
	Walking	327	182	1248	200511

(b) Classification results computed from 11 pedestrian joints.

Table 5.2: Classification results computed when pedestrian observations composed of body poses are used.

		Predicted			
		Standing	**Starting**	**Stopping**	**Walking**
Actual	**Standing**	72046	1351	69	797
	Starting	1600	11393	789	7870
	Stopping	95	109	1562	2031
	Walking	326	908	1944	199090

(a) Classification results extracted from 41 pedestrian joints.

		Predicted			
		Standing	**Starting**	**Stopping**	**Walking**
Actual	**Standing**	72173	1255	29	806
	Starting	150	11887	659	7956
	Stopping	57	48	1553	2139
	Walking	237	968	2107	198956

(b) Classification results extracted from 11 pedestrian joints.

Table 5.3: Classification results computed when pedestrian observations composed of displacements are used.

5.1.1. Discussion

An exhaustive data assessment of the previous confusion matrices is represented in Table 5.4 where the activity recognition results are compared taking into account the pedestrian features, number of joints and activity.

Features		**Pose + Disp**		**Pose**		**Disp**	
Joints		**41**	**11**	**41**	**11**	**41**	**11**
Overall Accuracy		90.69%	**95.13%**	88.39%	91.28%	94.76%	94.23%
Precision	**Standing**	89.77%	97.51%	88.93%	95.54%	97.27%	**98.04%**
	Starting	77.88%	**87.57%**	66.30%	79.38%	82.79%	83.96%
	Stopping	44.59%	**53.51%**	41.78%	40.06%	35.79%	35.72%
	Walking	92.50%	**95.57%**	89.89%	91.35%	94.90%	94.81%
Recall	**Standing**	88.85%	96.97%	84.31%	87.86%	97.01%	**97.19%**
	Starting	48.60%	**61.49%**	31.33%	39.14%	52.62%	54.90%
	Stopping	36.45%	**51.38%**	32.87%	37.11%	41.14%	40.90%
	Walking	96.90%	98.88%	97.04%	**99.13%**	98.43%	98.36%
F1-Score	**Standing**	89.31%	97.24%	86.56%	91.54%	97.14%	**97.61%**
	Starting	59.85%	**72.25%**	42.55%	52.43%	64.34%	66.39%
	Stopping	40.11%	**52.42%**	36.79%	38.53%	38.28%	38.13%
	Walking	94.65%	**97.20%**	93.33%	95.08%	96.63%	96.55%

Table 5.4: Evaluation of activity recognition results based on pedestrian features, number of joints and activity.

5.1.1.1. Joints

These results verify that shoulder and leg motions are more valuable sources of information than other body parts to recognise the current pedestrian action. More specifically, the maximum accuracy rate, 95.13%, is achieved when observations composed of poses and displacements from only 11 joints are taken into consideration. However, the accuracy rate falls to 90.69% whether 41 joints are used. Likewise, by considering only body poses, a similar conclusion is drawn since the maximum accuracy rate is 91.28% and 88.39% for 11 and 41 joints respectively. Finally, when the observations are composed solely of pedestrian displacements, the activity recognition results are not significantly influenced by the number of joints. It is noteworthy that these results are in accordance with those of [35] where it is deduced that gait initiations are recognised more reliably by means of feet position changes.

5.1.1.2. Activities

Regarding the distinction among activities, the pedestrian displacements achieve a better differentiation of standing actions from the rest of activities. However, with respect to starting and stopping actions, a larger number of critical missclassifications are produced. This means that the displacements do not allow to reliably distinguish whether a pedestrian is carrying out the first or last step. Therefore, the body poses along with the displacements offer a more discriminative information in these cases. It is worth mentioning that the poses are not usually applied to predict paths and intentions, as reviewed in Section 2.1.1. Given that humans are not rigid objects, the motion analysis of each body part should be taken into account for these tasks. The non-use of body poses and, thus the use of only motion features, may be due to the fact that only two dynamical behaviours are usually considered in other works, i.e. standing and walking. However, when a large number of dynamical activities are considered, such as standing, starting, stopping and walking, the body pose is an important feature. Beyond that, considering the body pose as the only feature, standing actions are repeatedly recognised as walking activities since, when the pedestrian legs are closed, the poses from both states are very similar in those instants of time. Therefore, the displacements are valuable information in those cases. On the other hand, including the acceleration as an additional feature may improve the recognition of starting and stopping activities. However, in walking activities, when the pedestrian legs are completely opened, the acceleration is minimum and it is maximum when the legs are closed. Hence, the body pose is again an essential information to distinguish standing and walking actions. As a conclusion, at least two types of features are needed in the activity recognition when more than two states are considered, either body poses and displacements or body poses and accelerations. The advantage of using body poses along with displacements is that only two pedestrian observations are needed for the activity recognition.

Considering observations composed of body poses and displacements, the most frequent missclassifications are produced by delays or pedestrians with low-speed movements. The first cause is related to the event-labelling methodology selected by the human expert. It seems that the first half of the first step and the second half of the last step contain the most perceptible information to determine starting and stopping actions respectively. Hence, the rest of these steps are normally recognised as walking action. It is worth mentioning that, unlike other event-labelling methodologies discussed in Section 2.3, the event-labelling method proposed in this document takes into account four transitions among activities in-

stead of only two dynamical changes. Thereby, the labelling of starting-walking and walking-stopping transitions can be objectively done when the first and last steps are completely marked. In any other cases, the event-labelling depends on the human expert. On the other hand, when body poses and displacements are used, walking activities are recognised as starting or stopping actions when pedestrians with low-speed movements are tested. Nonetheless, all these last missclassifications are not critical from the point of view of the path estimation since these actions have similar dynamics. Likewise, the beginning of a starting action and the ending of a stopping movement contains body poses which are equivalent to poses labelled as standing actions. Hence, a significant number of missclassifications are also produced between these activities. Recognising all these dynamical changes as soon as possible is a major challenge in order to increase the effectiveness of AEBSs and pedestrian protection systems. As will be discussed later, the delays obtained by the proposed method are in accordance with the results analysed from other significant works in Section 2.3.1.

In order to graphically show the previous statements, a sequence example is analysed. The classification probabilities using 41 and 11 joints along with the groundtruth are shown in Figures 5.1 and 5.2. Several examples of pedestrian poses at different instants of time are illustrated in the top of the figures. These poses are represented in different colours according to the classification result. Black represents a standing activity, green a starting action, red a walking action and blue a stopping activity. In the middle, the probabilities of each activity at each instant of time are shown. Finally, at the bottom, a zoom in of the transitions are illustrated.

As mentioned above, the figures show that starting-walking and walking-stopping transitions usually happen in the middle of the first and last steps, thus obtaining non-critical missclassifications. Additionally, in Figure 5.1, an example of missclassifications between standing and starting is illustrated in the standing-starting transition. Likewise, short delays appear in the standing-starting and stopping-standing transitions. These delays will be discussed later. On the other hand, throughout walking actions, local maxima and local minima of walking probabilities appear in the graph when the pedestrian legs are open and closed respectively. This is due to the fact that, when the legs are open, these observations are totally distinguishable from others contained in the rest of states. However, an observation from a pedestrian whose legs are closed may be similar to observations from any other state.

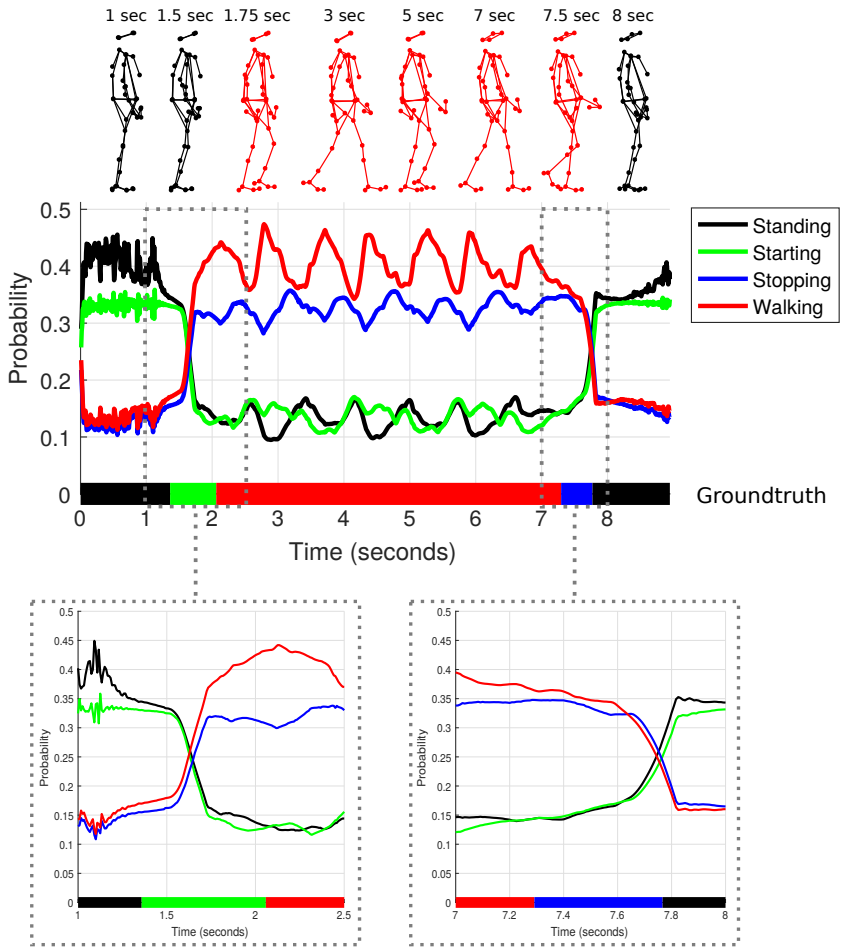


Figure 5.1: Example of activity recognition probabilities when poses and displacements extracted from 41 joints are used. Black represents a standing activity, green a starting action, red a walking action and blue a stopping activity. Top: pedestrian poses at significant instants of time. Middle: probabilities for each activity. Bottom: zoom in of the transitions.

5.1.1.3. Transitions and Delays

In Tables 5.5, 5.6 and 5.7, and Figures 5.3 and 5.4, the transitions from standing to starting, starting to walking, walking to stopping and stopping to standing are analysed in detail. This assessment is focused on the number of detected and non-detected transitions and delays, where the mean, standard deviation, median, maximum and minimum values using 41 and 11 joints are exposed. The evaluation criteria fixes a range of $[-500, 500]$ milliseconds around the event labelled by the human expert. Within this range, a multiframe validation algorithm is applied in

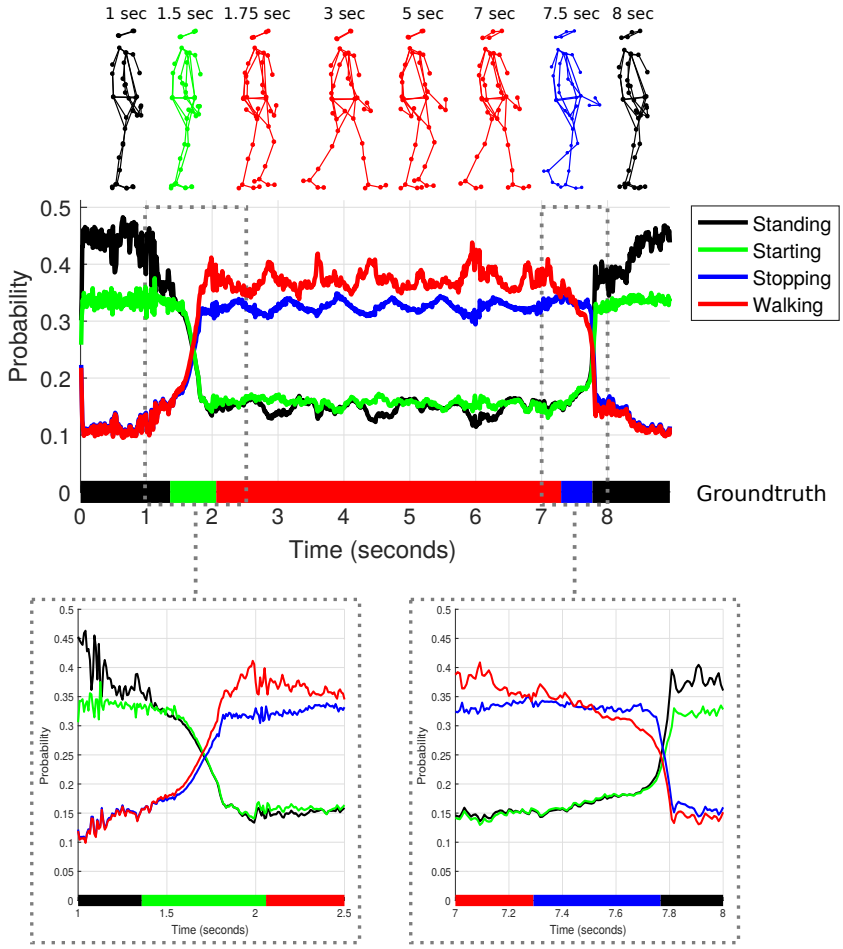


Figure 5.2: Example of activity recognition probabilities when poses and displacements extracted from 11 joints are used. Black represents a standing activity, green a starting action, red a walking action and blue a stopping activity. Top: pedestrian poses at significant instants of time. Middle: probabilities for each activity. Bottom: zoom in of the transitions.

order to ensure the transition detection and reduce false positive changes produced by missclassifications. The number of frames is fixed to 6, which corresponds to 50 milliseconds. Thereby, the algorithm detects a transition when 6 consecutive pedestrian observations are recognised as the same activity but this is different to the action classified in $t - 6$. Finally, the activity detection delay is computed from the instant of time where the event was marked by the human expert and the instant of time where the transition were detected by the algorithm.

Regarding the number of detected and non-detected transitions, a breakdown

is shown in Table 5.5. When 41 joints are used, the number of transitions correctly and incorrectly detected is 508 and 159 respectively, i.e. the accuracy rate is 76.16%. This low accuracy rate is mainly produced by a large number of non-detected standing-starting transitions (one example is shown in Figure 5.1) since a high uncertainty is obtained in the recognition of these activities. Another reason is the large number of non-detected starting-walking transitions due to the range chosen for the evaluation criteria. That is, when longer starting steps are tested, delays in starting-walking transitions are longer than 500 milliseconds. On the other hand, the number of transitions correctly and incorrectly detected when 11 joints are used is 622 and 45 respectively, i.e. the accuracy rate is 93.25%. In this case, most of the transitions which are not detected corresponds to walking-stopping changes. This could be due to the fact that the number of observations in the dataset belonging to a stopping activity is significantly smaller than other actions and stopping steps are usually faster than starting steps. An analysis of the starting and stopping steps in the groundtruth confirms this last hypothesis. The mean lengths of both steps along with their standard deviations are 686.06 ± 202.91 and 381.22 ± 78.92 milliseconds respectively. It is worth mentioning that missclassifications produced in a transition negatively influence in the non-detection of future transitions. This does not happen when only one transition, such as standing-walking or walking-standing, is considered, as several works reviewed in Section 2.3.1 do. Nonetheless, the selection of four pedestrian dynamics influences positively in the path estimation.

Transition	**Detected**		**Non-Detected**		**Accuracy**	
	41 Joints	**11 Joints**	**41 Joints**	**11 Joints**	**41 Joints**	**11 Joints**
Standing - Starting	174	238	69	5	71.60%	97.94%
Starting - Walking	220	250	42	12	83.97%	95.42%
Walking - Stopping	51	61	31	21	62.20%	74.39%
Stopping - Standing	63	73	17	7	78.85%	91.25%
Overall	508	622	159	45	76.16%	93.25%

Table 5.5: Breakdown of detected and non-detected transitions for a different number of joints. The pedestrian observations are composed of body poses and displacements.

Regarding the delays of the detected transitions, the results show that these are not significantly influenced by the number of joints since the multiframe validation algorithm filters most of the missclassifications. Moreover, it should be pointed out that starting-walking transitions have negative delays since the first half of the first step contains the most perceptible information to determine starting actions. However, as shown in Figures 5.3d and 5.4d, a bimodal distribution clearly arises

Transition	Mean	Std	Median	Max	Min
Standing - Starting	63.94 ms	147.73 ms	50.00 ms	525.00 ms	-450.00 ms
Starting - Walking	-140.42 ms	188.09 ms	-154.17 ms	283.33 ms	-466.67 ms
Walking - Stopping	33.50 ms	206.15 ms	50.00 ms	525.00 ms	-458.33 ms
Stopping - Standing	99.47 ms	142.82 ms	66.67 ms	358.33 ms	-366.67 ms

Table 5.6: Delays in milliseconds of detected transitions when 41 joints are used. The pedestrian observations are composed of body poses and displacements.

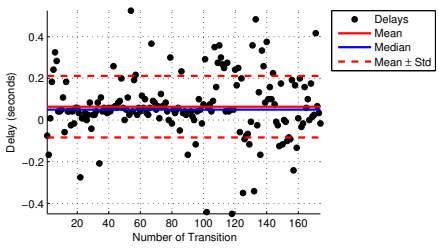
Transition	Mean	Std	Median	Max	Min
Standing - Starting	57.98 ms	120.87 ms	50.00 ms	525.00 ms	-441.67 ms
Starting - Walking	-154.30 ms	183.66 ms	-208.33 ms	341.67 ms	-446.67 ms
Walking - Stopping	102.05 ms	157.86 ms	66.67 ms	416.67 ms	-450.00 ms
Stopping - Standing	89.84 ms	131.48 ms	58.33 ms	450.00 ms	-466.67 ms

Table 5.7: Delays in milliseconds of detected transitions when 11 joints are used. The pedestrian observations are composed of body poses and displacements.

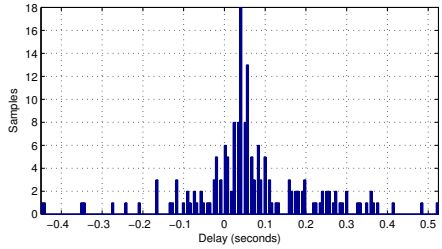
in this transition due to the fact that the walking actions are detected before and after the events. The delays of each detected transition along with the histogram, mean, median and standard deviation values are illustrated in Figures 5.3 and 5.4.

Additionally, a more comprehensive assessment can be addressed comparing the results with the delays accomplished in other works that were reviewed in Section 2.3.1. The method proposed in this document recognises starting intentions 125 milliseconds after the gait initiation with an accuracy rate of 80% when 11 joints are considered. These results are similar to the delays achieved in [35, 36]. Nonetheless, it should be pointed out that a multiframe validation of 50 milliseconds is carried out in order to filter missclassifications and a higher number of different dynamics are modelled in the proposed method. This means that the consideration of only one transition, i.e. standing-walking, instead of two dynamical changes, i.e. standing-starting and starting-walking, could accomplish better results. However, if only two states are taken into account, as several works reviewed in Section 2.3.1 do, the path prediction could be negatively influenced by this decision.

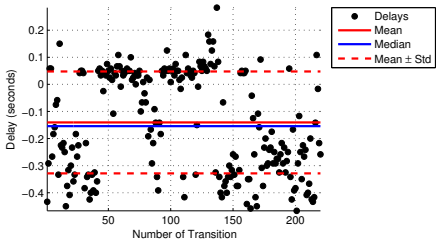
On the other hand, an analysis of delays from walking-stopping transitions to the standing events labelled by the human expert is shown in Table 5.8 and Figure 5.5. This analysis is important in order to estimate the prediction time before a standing event carried out by a pedestrian. As shown, most standing events can be predicted several tens of milliseconds in advance. More specifically, the method proposed in this document recognises stopping intentions 58.33 milliseconds before the event with an accuracy rate of 70% when 11 joints are considered. This data is



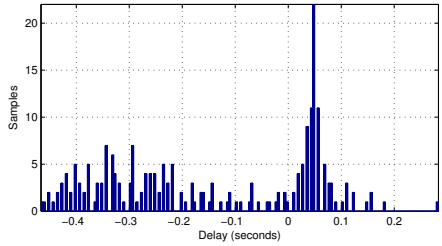
(a) Delays of standing-starting transitions.



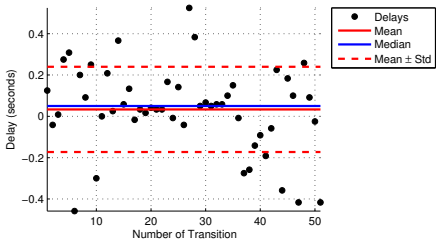
(b) Histogram of delays for standing-starting transitions.



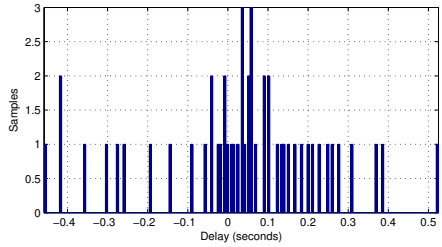
(c) Delays of starting-walking transitions.



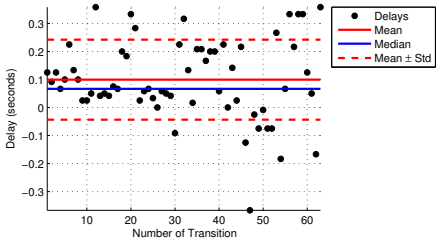
(d) Histogram of delays for starting-walking transitions.



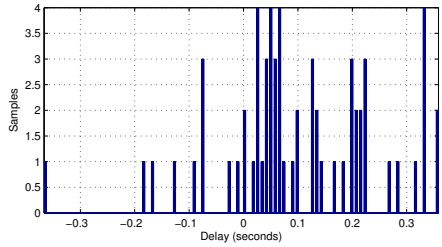
(e) Delays of walking-stopping transitions.



(f) Histogram of delays for walking-stopping transitions.

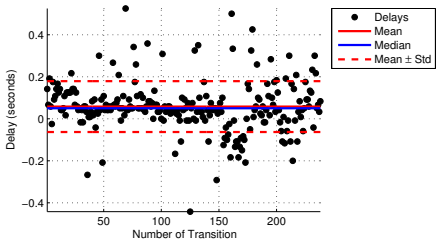


(g) Delays of stopping-standing transitions.

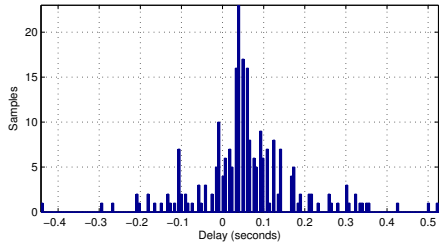


(h) Histogram of delays for stopping-standing transitions.

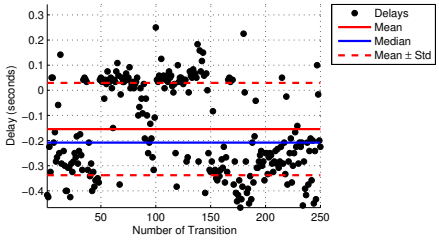
Figure 5.3: Delays in seconds of detected transitions when 41 joints are used. Left graphs show the delays of each transition along with the mean, median and standard deviation values. Right images show the corresponding histograms. The pedestrian observations are composed of body poses and displacements.



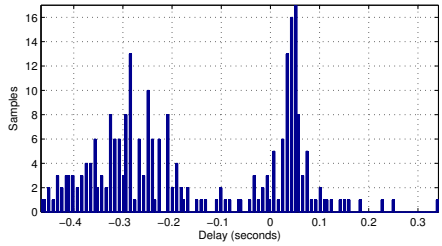
(a) Delays for standing-starting transitions.



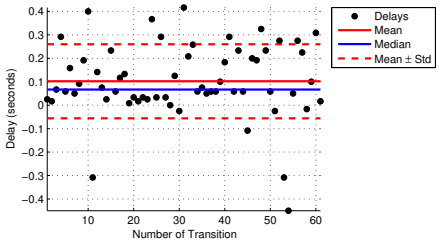
(b) Histogram of delays for delays for standing-starting transitions.



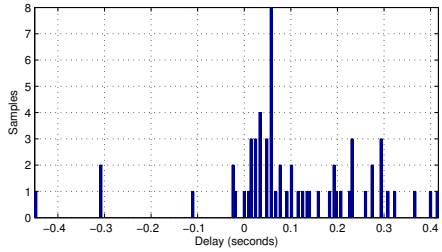
(c) Delays for starting-walking transitions.



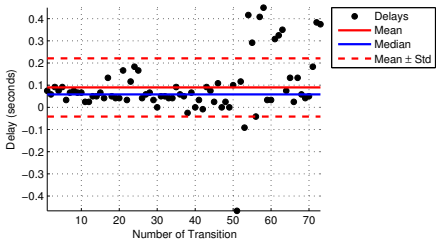
(d) Histogram of delays for starting-walking transitions.



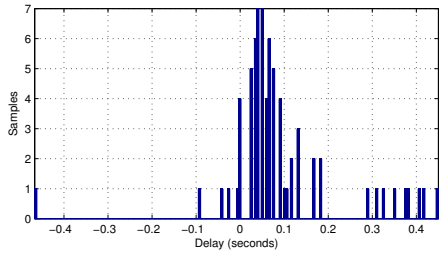
(e) Delays of walking-stopping transitions.



(f) Histogram of delays for walking-stopping transitions.



(g) Delays of stopping-standing transitions.



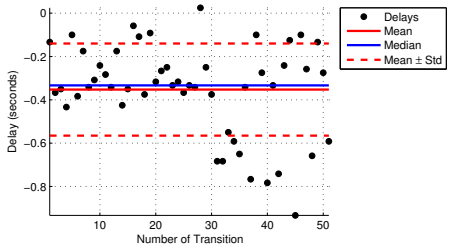
(h) Histogram of delays for stopping-standing transitions.

Figure 5.4: Delays in seconds of detected transitions when 11 joints are used. Left graphs show the delays of each transition along with the mean, median and standard deviation values. Right images show the corresponding histograms. The pedestrian observations are composed of body poses and displacements.

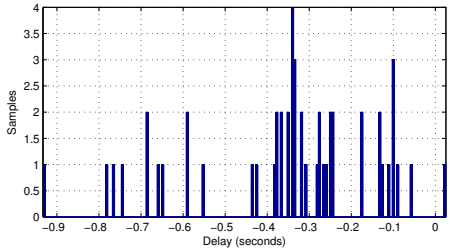
slightly worse than the results accomplished in [31,32,36] due to the non-detection of walking-stopping transitions previously discussed. However, once again, it should be pointed out that a multiframe validation over 50 milliseconds is carried out in order to filter missclassifications and a larger number of different dynamics are considered in the proposed method. Likewise, the smaller number of stopping sequences with respect to other states and the lengths of the last steps, which were previously analysed, explain the data difference. As before, if only two states are taken into account, as several works propose, the path prediction could be negatively influenced as well.

Joints	Mean	Std	Median	Max	Min
41	-352.61 ms	212.51 ms	-333.33 ms	25.00 ms	-933.33 ms
11	-279.92 ms	158.59 ms	-291.67 ms	66.67 ms	-875.00 ms

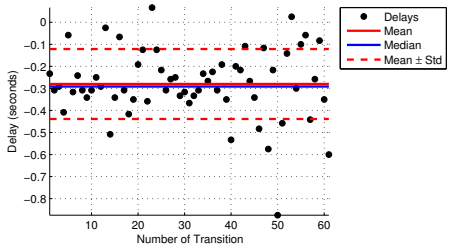
Table 5.8: Analysis of delays from walking-stopping transitions to the standing events labelled by the human expert. The pedestrian observations are composed of body poses and displacements.



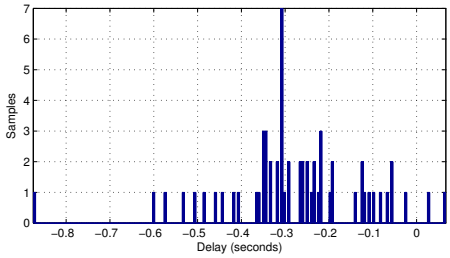
(a) Delays from walking-stopping transitions to standing events when 41 joints are used.



(b) Histogram of delays from walking-stopping transitions to standing events when 41 joints are used.



(c) Delays from walking-stopping transitions to standing events when 11 joints are used.



(d) Histogram of delays from walking-stopping transitions to standing events when 11 joints are used.

Figure 5.5: Delays from walking-stopping transitions to standing events labelled by the human expert. The pedestrian observations are composed of body poses and displacements.

5.1.2. Noisy Observations

In this section, the activity recognition is examined using a sequence example of noisy observations extracted by the single-frame pedestrian skeleton estimation algorithm described in Section 4.2. In Figure 5.6, images extracted from the sequence are presented. The sequence length is around 3.75 seconds and the time step value between each image is 0.25 seconds. As shown, the sequence corresponds to a pedestrian that is walking on a zebra crossing from the left to right.

Figure 5.6: Images extracted from the sequence example. The sequence length is around 3.75 seconds and the time step value between each image is 0.25 seconds.

Furthermore, in Figure 5.7, the tridimensional reconstructions of the scenes along with the skeleton estimations and the pedestrian headings extracted from consecutive head locations are illustrated from two different points of view. These reconstructions correspond to the scenes of the third column in Figure 5.6. As shown, all joints are correctly extracted despite the fact that the left knee is occluded in the last scenario.

In Figure 5.8, the activity recognition probabilities when poses and displacements computed from the sequence by using the skeleton estimation algorithm are represented. As in Figures 5.1 and 5.2, where an example of activity recognition by means of poses and displacements extracted from the UAH dataset is illustrated, the black line represents the probability of standing activity, the green line corresponds to the probability of starting action, the red line to the probability of walking action and, finally, the blue line represents the probability of stopping activity. On the top of the figure, the pedestrian point clouds extracted by the pedestrian segmentation algorithm and the skeleton estimations at different in-

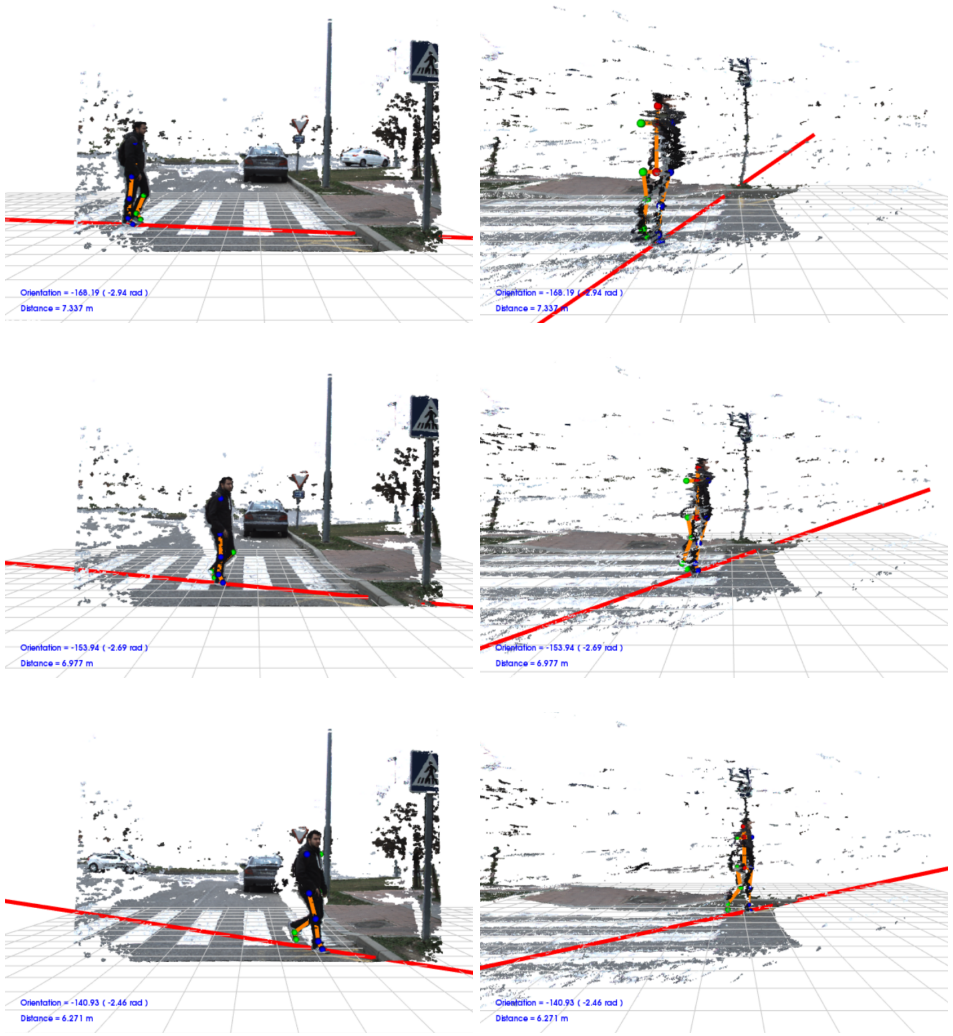


Figure 5.7: Tridimensional reconstructions of the scenes along with the skeleton estimations and the pedestrian headings extracted by means of consecutive head positions. The reconstructions are shown from two different points of view.

stants of time are shown. These skeletons correspond to the scenes of the third column in Figure 5.6 and the reconstructions of the scenes illustrated in Figure 5.7. The graph shows that the activity has been correctly recognised in the whole sequence and the probability values for each activity are similar to the values shown in Figures 5.1 and 5.2.

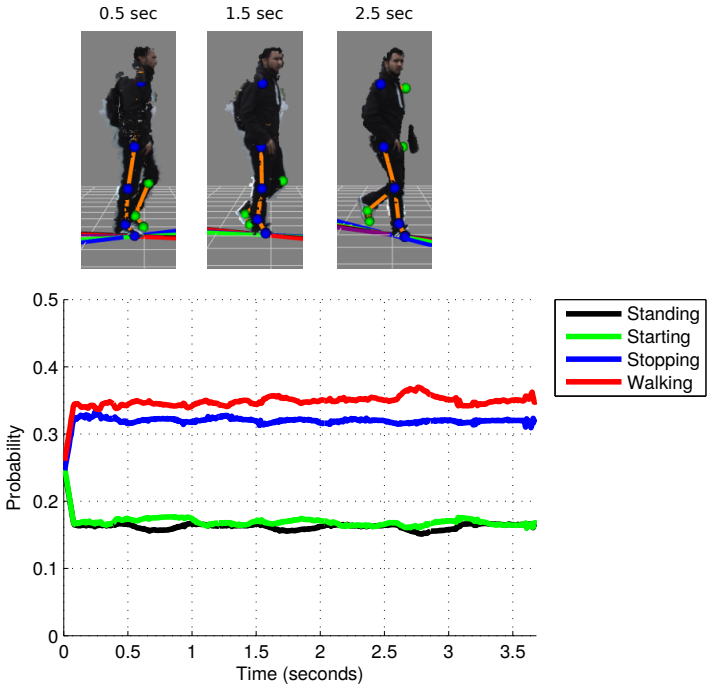


Figure 5.8: Activity recognition probabilities when poses and displacements extracted from the skeleton estimation algorithm are used. The black line represents the probability of standing activity, the green line corresponds to the probability of starting action, the red line to the probability of walking action and the blue line represents the probability of stopping activity. Top: pedestrian poses at significant instants of time. Bottom: probabilities for each activity.

5.2. Pedestrian Path Prediction Results

As mentioned in Section 2.3.1, the RMSE and MED between estimated pedestrian positions and the groundtruth are often chosen as measure of accuracy for pedestrian path evaluations. Nonetheless, some measures provide a better idea of how well a system works than others. For example, the MED used in [19, 20, 55] gives a more precise physical interpretation of the predicted pedestrian positions with respect to a groundtruth than the RMSE used in [22]. Likewise, the mean and standard deviation of the per-sequence RMSE used in [31, 32] provide vague information of the system performance since the RMSE for each sequence does not offer information about the similarity between predicted positions and the groundtruth at discrete time steps. Besides, although most of the works reviewed in Section 2.3.1 consider that the evaluation should be done for each type of activity separately, it is not clear what methodology is the most appropriate in order to standardise the

path evaluation. Thereby, a reliable comparison of path prediction approaches has not been done for the moment. For all these reasons, in this thesis, the measure of accuracy chosen for the path evaluation is the MED at different TTEs.

Throughout this section, the evaluation of path prediction results is performed considering 41 and 11 joints. Firstly, the outcomes of this task are shown assuming the best activity recognition results which were examined before. That is, the activity recognition is performed by means of 11 joints. After that, in Section 5.2.2, the path prediction results are shown assuming that the activity recognition has an accurate rate of 100%. This assessment enables to estimate the influence of the activity recognition in the path prediction task. Finally, in Section 5.2.3, the path prediction performed by means of noisy observations extracted by the skeleton estimation algorithm are analysed.

5.2.1. Pedestrian Path Prediction Results with Activity Recognition

As explained in Section 4.5, once the pedestrian activity is estimated, the most appropriate model is selected and the prediction of future observations is iteratively performed using that model. Accordingly, a good path prediction depends strongly on a good activity recognition. In this section, a path prediction evaluation is performed considering the activity recognition results previously discussed. This evaluation makes reference to MEDs between the predicted pedestrian locations and the groundtruth for time horizon values up to 1 second. Due to the fact that the most dangerous traffic situations related to pedestrians usually happen when they start to cross or whether they will stop before crossing, the evaluation is done around these situations. Thereby, the MEDs are computed at different TTEs, i.e. time to start walking and time to stop walking. It is noteworthy that positive TTE values make reference to instants of time before the event and negative values to instants of time after the event. Moreover, as mentioned in Section 4.5, the reference point to reconstruct the pedestrian path corresponds to the right hip.

In Tables 5.9 and 5.10, and Figure 5.9, the combined longitudinal and lateral MEDs along with the standard deviation are shown. Regarding starting activities, the errors before the event are mainly produced due to to the fact the algorithm assumes zero displacements when the pedestrian activity is recognised as standing, however, this is not the case in the groundtruth since small movements were gathered. On the other hand, the errors after the event exponentially grows up since, as explained in Section 5.1.1.3, the recognition of a starting activity has a

mean delay around 60 milliseconds and the pedestrian is accelerating. However, it seems that, when the pedestrian finishes to speed up, the MEDs tend to be linear. Additionally, due to the fact that the B-GPDM is a dimensionality reduction technique, the errors are not significantly influenced by the number of joints. In order to contextualise the errors, the mean displacement for starting activities belonging to the UAH dataset was computed. Throughout a starting activity, the pedestrian has a mean displacement value of 193.98±78.52 millimetres. Likewise, the mean displacement at 1 second after and before the event is 467.92±264.97 and 41.24±67.91 millimetres respectively. It is worth mentioning that other dynamical changes could happen within the TTE range of [1-0] seconds. For example, a stopping-standing transition could be carried out by the pedestrian a few hundreds of milliseconds before the event.

These results, focused on starting activities, are similar to the results achieved in other works which were reviewed in Section 2.3.1. More specifically, in [19] a MED value of 315 millimetres is accomplished for a time horizon of 1.2 seconds. This value is similar to the value obtained by the approach described in this thesis for a TTE of 0 seconds and a time horizon of 1 second (331.93 millimetres when 11 joints are used). Nonetheless, the event-labelling methodology proposed in that work changes with respect to the described in this document. The authors define a starting activity from 1 second before the initial movement to approximately 3 seconds after reaching the steady state velocity. Besides, the predictions are evaluated for all time steps instead of being assessed at different TTEs. In [20], the MED at a starting event for a time horizon of 0.6 seconds is 80 millimetres. However, the method described in this thesis achieves a MED value of 88.07 millimetres (whether 41 joints are used) at the instant of a starting event for a time horizon of 0.5 seconds. In [22], a RMSE value of 334 millimetres at 1 second is obtained, this value is slightly lower than the RMSE obtained by the approach described in this document for a TTE of 0 seconds and a time horizon of 1 second (418.09 millimetres when 11 joints are used). However, the predictions of this work need a temporal windows of n trajectory points to be performed instead of using two observations as the method described in this thesis do. Moreover, the predictions are evaluated for all time steps instead of being assessed at different TTEs.

Regarding stopping activities, the errors before the event tend to be linear since, as mentioned in Section 5.1.1.3, the mean length of stopping steps are 381.22±78.92 milliseconds and the second half of the last step contain the most perceptible information to determine stopping actions. Thereby, an appropriate model could not be chosen until a few hundreds of milliseconds before the event. Moreover, as

	Transition	**Standing-Starting**				**Stopping-Standing**			
TTE (sec)	*Horizon (sec)*	**0.25**	**0.5**	**0.75**	**1**	**0.25**	**0.5**	**0.75**	**1**
1	*MED*	8.69	15.33	20.79	**38.86**	49.05	90.94	152.29	**238.01**
	±*Std*	±10.77	±17.55	±27.16	±**54.07**	±59.41	±103.91	±154.67	±**206.93**
0.75	*MED*	11.11	19.33	**39.94**	72.96	43.97	91.98	**175.41**	289.33
	±*Std*	±17.40	±31.51	±**60.86**	±94.61	±58.86	±77.68	±**233.19**	±239.60
0.5	*MED*	10.42	**28.33**	61.07	141.79	50.54	**150.83**	294.00	462.06
	±*Std*	±13.60	±**33.52**	±71.68	±140.89	±72.14	±**223.89**	±397.80	±567.53
0.25	*MED*	**21.25**	55.45	137.62	290.77	**38.41**	105.71	241.39	333.61
	±*Std*	±**27.73**	±65.14	±129.87	±207.63	±**38.22**	±85.09	±142.72	±202.10
0	*MED*	31.22	89.10	192.82	331.93	48.51	100.66	162.78	244.23
	±*Std*	±35.77	±88.38	±164.96	±254.73	±36.80	±88.64	±155.60	±250.99
-0.25	*MED*	46.24	111.90	202.03	302.40	44.68	83.11	129.60	189.73
	±*Std*	±59.76	±119.04	±184.70	±247.43	±61.04	±127.20	±202.91	±292.71
-0.5	*MED*	48.69	116.00	202.58	296.23	11.54	20.16	40.69	64.34
	±*Std*	±59.26	±113.39	±168.16	±228.83	±9.09	±19.49	±52.16	±95.74
-0.75	*MED*	42.39	89.43	145.89	206.80	9.92	39.17	67.47	121.64
	±*Std*	±50.36	±96.36	±144.80	±210.75	±10.60	±46.70	±97.17	±163.10
-1	*MED*	35.56	79.17	120.19	161.14	21.63	51.24	100.53	183.66
	±*Std*	±46.90	±93.07	±144.91	±186.36	±29.10	±67.85	±120.69	±183.17

Table 5.9: Combined longitudinal and lateral MED±Standard Deviation in millimetres at different TTEs for predictions up to 1 second when 11 joints are solely considered.

mentioned before, delays in the transition detection could negatively influence in the path estimation. On the other hand, after the event, the error decreases and tend to be logarithmic. However, at a TTE value of -1 second, the errors grow up due to the fact that a new pedestrian dynamical change could happen. Once again, in order to contextualise the errors, the mean displacement for stopping activities belonging to the UAH dataset was computed. Throughout these activities, the pedestrian has a mean displacement value of 164.37±63.33 millimetres. Likewise, the mean displacement at 1 second after and before the event is 102.15±63.50 and 679.15.37±306.77 millimetres respectively.

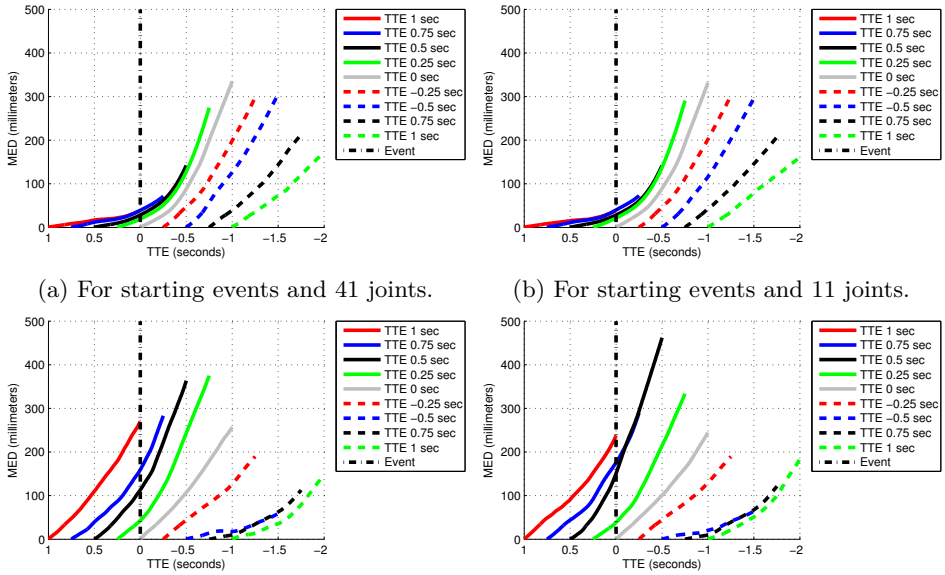
Comparing the results with the outcomes achieved by other works, these are similar. In particular, in [19], a MED value of 224 millimetres is accomplished for stopping activities at 1.2 seconds. The method proposed in this thesis achieves a MED value of 238.01 millimetres for a TTE of 1 second and a time horizon of 1 second when solely 11 joints are used. In [22], a RMSE value of 292 millimetres at 1 second is obtained, this value is slightly lower than the RMSE obtained by the

	Transition	**Standing-Starting**				**Stopping-Standing**			
TTE (sec)	*Horizon (sec)*	**0.25**	**0.5**	**0.75**	**1**	**0.25**	**0.5**	**0.75**	**1**
1	*MED*	8.75	16.91	21.95	**38.61**	47.78	109.83	179.06	**270.04**
	±*Std*	±10.95	±25.61	±34.98	±**57.77**	±62.53	±126.24	±200.33	±**258.83**
0.75	*MED*	12.10	19.78	**39.20**	71.60	39.49	86.70	**159.45**	282.92
	±*Std*	±22.62	±35.04	±**62.18**	±96.10	±34.14	±64.18	±**97.29**	±146.71
0.5	*MED*	9.84	**28.43**	62.22	142.12	47.69	**113.06**	222.37	363.58
	±*Std*	±9.72	±**33.62**	±73.90	±141.09	±45.90	±**89.04**	±160.40	±257.84
0.25	*MED*	**19.22**	51.20	127.64	274.49	**42.84**	118.84	243.12	374.97
	±*Std*	±**27.77**	±60.00	±119.99	±198.19	±**33.98**	±92.40	±172.19	±255.41
0	*MED*	28.36	88.07	200.23	334.97	49.48	106.93	179.31	256.95
	±*Std*	±29.99	±97.03	±177.98	±262.16	±35.74	±99.45	±183.19	±275.81
-0.25	*MED*	47.07	109.52	198.80	298.44	44.19	83.02	126.38	189.68
	±*Std*	±62.48	±118.79	±183.59	±253.08	±56.91	±118.75	±191.16	±281.58
-0.5	*MED*	57.31	126.35	207.71	305.50	14.20	17.83	35.33	58.73
	±*Std*	±105.49	±155.67	±206.81	±282.90	±20.28	±21.08	±42.62	±88.89
-0.75	*MED*	38.59	85.56	143.41	212.88	10.07	34.47	62.51	112.76
	±*Std*	±54.03	±104.30	±167.24	±252.28	±10.81	±43.50	±93.63	±161.74
-1	*MED*	35.58	73.94	120.22	170.60	14.31	35.55	79.05	148.71
	±*Std*	±63.23	±108.35	±181.21	±244.03	±15.22	±46.40	±97.14	±172.87

Table 5.10: Combined longitudinal and lateral MED±Standard Deviation in millimetres at different TTEs for predictions up to 1 second when 41 joints are solely considered.

approach described in this document for a TTE of 1 second and a time horizon of 1 second (314.5 millimetres when 11 joints are used). However, the algorithm described in that work needs a temporal windows of n trajectory points to performed the predictions instead of using two observations as the method described in this thesis do. Moreover, the predictions are evaluated for all time steps instead of being assessed at different TTEs. In [55], the lateral MED for a time horizon of 1 second at 1 second before the event is 140±180 millimetres. The method described in this document achieves a lateral MED value of 226.99±208.01 millimetres when solely 11 joints are considered.

Additionally, walking activities were also analysed at different time horizons. The MEDs achieved by the method described in this thesis at 0.25, 0.5, 0.75 and 1 second are 33.03±43.84, 70.87±89.69, 113.34±140.64 and 159.48±196.19 millimetres respectively when 11 joints are used. These errors are similar or lower than the outcomes obtained in other works. For example, in [19], a MED value of 230 millimetres is accomplished for walking activities at 1.2 second. Furthermore, in [22],



(a) For starting events and 41 joints.

(b) For starting events and 11 joints.

(c) For stopping events and 41 joints.

(d) For stopping events and 11 joints.

Figure 5.9: Combined longitudinal and lateral MED in millimetres at different TTEs for predictions up to 1 second.

a RMSE value of 250 millimetres at 1 second is achieved, however, the proposed algorithm accomplishes a value of 252.83 millimetres. Finally, in [55], the lateral MED value of 190±220 is obtained. The algorithm developed in this thesis achieves a lateral MED value of 149.88±194.75. Once again, in order to contextualise the errors, the mean displacement for walking activities from the UAH dataset at 1 second is 816.47±315.45 millimetres.

5.2.2. Pedestrian Path Prediction Results without Activity Recognition

With the motivation of determining the influence of the activity recognition algorithm into the path prediction, the method is also tested assuming that the activity recognition has an accurate rate of 100%. In Tables 5.11 and 5.12, and Figure 5.10, the combined longitudinal and lateral MEDs along with the standard deviations are shown. Regarding starting activities, similar to the previous case, the errors before the event are mainly produced due to to the fact the algorithm assumes zero displacements when the pedestrian activity is recognised as standing, however, small movements were gathered in the groundtruth. On the other hand, the errors after the event exponentially grows up since the pedestrian is acceler-

ating. Once again, it seems that, when the pedestrian finishes to speed up, the MEDs tend to be linear. Additionally, because of the B-GPDM is a dimensionality reduction technique, the errors are not significantly influenced by the number of joints. Likewise, as before, other dynamical changes could happen within the TTE range of [1-0] seconds. It is worth remarking that, throughout a starting activity, the pedestrian has a mean displacement value of 193.98±78.52 millimetres and the mean displacement at 1 second after and before the event is 467.92±264.97 and 41.24±67.91 millimetres respectively

	Transition	**Standing-Starting**				**Stopping-Standing**			
TTE (sec)	*Horizon (sec)*	**0.25**	**0.5**	**0.75**	**1**	**0.25**	**0.5**	**0.75**	**1**
1	*MED*	8.60	15.25	20.50	**37.48**	49.93	92.74	154.94	**240.67**
	±*Std*	±10.67	±17.12	±25.46	±**47.52**	±59.54	±105.63	±159.89	±**215.22**
0.75	*MED*	9.67	16.44	**34.53**	67.14	44.37	92.02	**175.07**	289.12
	±*Std*	±10.15	±18.97	±**40.87**	±81.03	±58.93	±77.68	±**233.25**	±239.69
0.5	*MED*	9.40	**27.47**	60.52	142.00	52.26	**157.56**	309.96	484.63
	±*Std*	±8.96	±**32.70**	±72.29	±142.13	±71.57	±**222.07**	±395.41	±562.94
0.25	*MED*	**19.79**	53.95	137.49	292.64	**25.65**	77.62	156.91	265.14
	±*Std*	±**24.66**	±62.21	±130.46	±210.86	±**27.73**	±62.66	±101.30	±157.27
0	*MED*	36.46	82.71	151.46	247.82	57.78	80.88	90.62	100.75
	±*Std*	±34.66	±70.43	±119.88	±195.70	±28.41	±41.14	±44.63	±63.22
-0.25	*MED*	40.83	109.06	207.49	320.32	25.43	38.56	53.46	79.28
	±*Std*	±48.92	±113.78	±185.12	±258.63	±20.07	±29.68	±64.24	±134.63
-0.5	*MED*	54.25	139.35	241.31	352.48	10.87	21.35	35.94	62.77
	±*Std*	±60.56	±117.57	±173.72	±241.57	±5.97	±18.09	±39.91	±91.09
-0.75	*MED*	44.07	93.53	151.00	212.78	9.84	82.53	129.33	183.77
	±*Std*	±50.46	±97.45	±146.33	±213.55	±10.63	±148.90	±211.86	±248.93
-1	*MED*	35.46	78.92	118.92	159.44	32.98	79.07	136.84	230.54
	±*Std*	±46.72	±93.24	±144.49	±185.10	±39.26	±80.22	±128.23	±196.37

Table 5.11: Combined longitudinal and lateral MED±Standard Deviation in millimetres at different TTEs when 11 joints are solely considered.

These results, focused on starting activities, are similar or even slightly better than the results achieved in other works which were reviewed in Section 2.3.1. More specifically, in [19], a MED value of 315 millimetres is accomplished for a time horizon of 1.2 seconds. This value is higher than the value obtained by the approach described in this thesis for a TTE of 0 seconds and a time horizon of 1 second (247.82 millimetres when 11 joints are used). Nonetheless, as mentioned before, the event-labelling methodology proposed in that work changes with respect

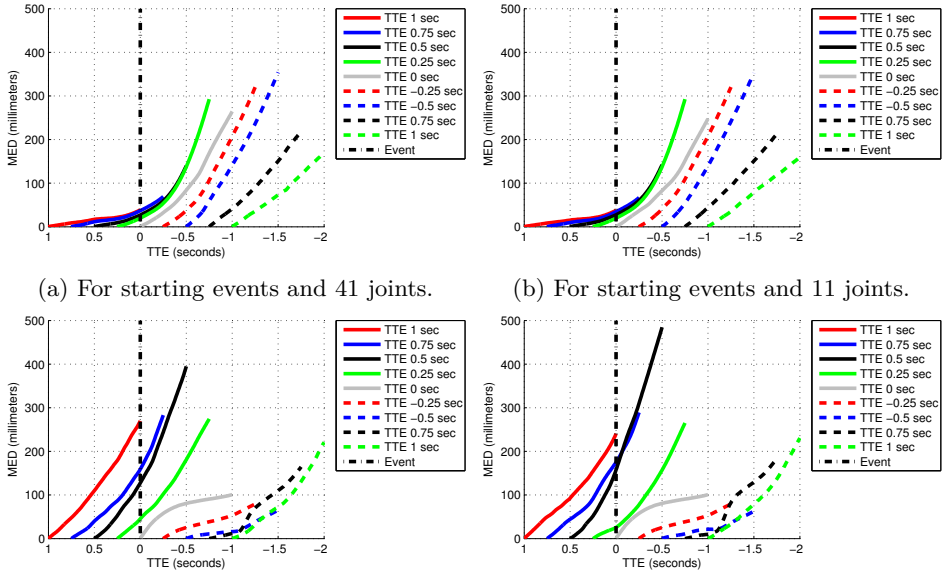
to the described in this document. Besides, the predictions are not assessed at different TTEs. In [20], the MED value at a starting event for a time horizon of 0.6 seconds is 80 millimetres. However, the method described in this thesis achieves a MED value of 82.71 millimetres (whether 11 joints are used) at the instant of a starting event for a time horizon of 0.5 seconds. In [22], a RMSE value of 334 millimetres at 1 second is obtained. This value is slightly higher than the RMSE obtained by the approach described in this document for a TTE of 0 seconds and a time horizon of 1 second (315.52 millimetres when 11 joints are used). However, in that work, the predictions need a temporal windows of n trajectory points to be performed instead of using two observations as the method described in this thesis do. Moreover, the predictions are evaluated for all time steps instead of being assessed at different TTEs.

	Transition	**Standing-Starting**				**Stopping-Standing**			
TTE (sec)	*Horizon (sec)*	**0.25**	**0.5**	**0.75**	**1**	**0.25**	**0.5**	**0.75**	**1**
1	*MED*	8.95	17.04	21.76	**38.52**	47.78	109.83	179.06	**270.04**
	±*Std*	±11.57	±26.13	±33.84	±**57.30**	±62.53	±126.24	±200.33	±**258.83**
0.75	*MED*	11.47	18.45	**36.79**	69.40	39.94	86.77	**159.17**	282.90
	±*Std*	±20.53	±29.71	±**53.03**	±90.73	±34.37	±64.20	±**97.34**	±146.71
0.5	*MED*	9.40	**27.03**	59.64	139.69	52.62	**127.53**	244.14	394.92
	±*Std*	±8.78	±**30.82**	±70.08	±140.13	±46.08	±**93.91**	±165.14	±259.51
0.25	*MED*	**19.79**	53.95	137.49	292.64	**45.32**	102.18	180.84	274.58
	±*Std*	±**24.66**	±62.21	±130.46	±210.86	±**45.09**	±92.04	±133.85	±193.52
0	*MED*	32.46	83.68	168.94	264.42	57.78	80.88	90.62	100.75
	±*Std*	±30.97	±83.31	±144.48	±218.94	±28.41	±41.14	±44.63	±63.22
-0.25	*MED*	37.55	106.64	207.44	320.16	25.43	38.56	53.46	79.28
	±*Std*	±39.98	±109.40	±184.46	±265.96	±20.07	±29.68	±64.24	±134.63
-0.5	*MED*	59.08	141.30	236.42	354.17	10.69	15.84	38.08	65.24
	±*Std*	±92.94	±146.71	±206.58	±298.25	±5.24	±10.60	±44.29	±95.49
-0.75	*MED*	40.98	90.88	150.61	218.68	11.65	70.14	110.07	164.38
	±*Std*	±55.00	±106.13	±174.09	±239.69	±10.77	±121.18	±165.81	±210.11
-1	*MED*	35.96	73.83	119.50	168.03	32.47	72.46	130.83	220.78
	±*Std*	±63.22	±107.33	±178.52	±237.19	±38.57	±87.07	±122.56	±196.17

Table 5.12: Combined longitudinal and lateral MED±Standard Deviation in millimetres at different TTE when 41 joints are solely considered.

Regarding stopping activities, the errors before the event tend to be linear since, as mentioned before, the mean length of stopping steps are 381.22±78.92 milliseconds and the second half of the last step contain the most perceptible information

to determine stopping actions. Hence, an appropriate model could not be chosen up to a few hundreds of milliseconds before the event. Likewise, after the event, the error decreases and tend to be logarithmic. However, at a TTE value of -1 second, the errors grow up due to the fact that a new pedestrian dynamical change could happen. Once again, in order to contextualise the errors, it is worth remarking that the mean displacement for stopping activities is 164.37±63.33 millimetres. In addition, the mean displacement at 1 second after and before the event is 102.15±63.50 and 679.15.37±306.77 millimetres respectively.



(a) For starting events and 41 joints.

(b) For starting events and 11 joints.

(c) For stopping events and 41 joints.

(d) For stopping events and 11 joints.

Figure 5.10: Combined longitudinal and lateral MED in millimetres at different TTEs.

Comparing the results with the outcomes achieved by other works, these are similar. In particular, in [19], a MED value of 224 millimetres for stopping activities at 1.2 second is obtained. The method proposed in this thesis achieves a MED value of 240.67 millimetres when 11 joints are used for a TTE of 1 second and a time horizon of 1 second. Moreover, in [22], the RMSE value obtained at 1 second is 292 millimetres. This value is slightly lower than the RMSE obtained by the approach described in this document for a TTE of 1 second and a time horizon of 1 second (321.94 millimetres when 11 joints are used). However, the algorithm described in that work needs a temporal windows of n trajectory points to performed the predictions instead of using two observations as the method described in this thesis do. Finally, in [55], the lateral MED for a time horizon of 1 second and at 1 second

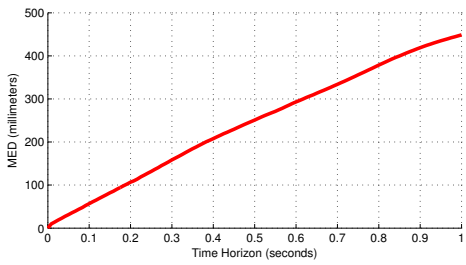
before the event is 140±180 millimetres. The method described in this document achieves a lateral MED value of 240.67±215.22 millimetres.

Furthermore, walking activities were also analysed at different time horizons. The MEDs achieved by the method described in this thesis at 0.25, 0.5, 0.75 and 1 second are 32.12±42.95, 68.74±87.14, 109.54±137.65 and 153.62±192.40 millimetres respectively. These errors are similar or lower than the outcomes obtained in other works. For example, in [19], a MED value of 230 millimetres for walking activities at 1.2 second is achieved. Additionally, in [22], a RMSE value of 250 millimetres at 1 second is obtained, however, the algorithm described here accomplishes a value of 246.20 millimetres. Finally, in [55], the lateral MED value of 190±220 millimetres is obtained. The algorithm developed in this thesis achieves a lateral MED value of 143.83±190.82. Once again, in order to contextualise the errors, the mean displacement for walking activities from the UAH dataset at 1 second is 816.47±315.45 millimetres.

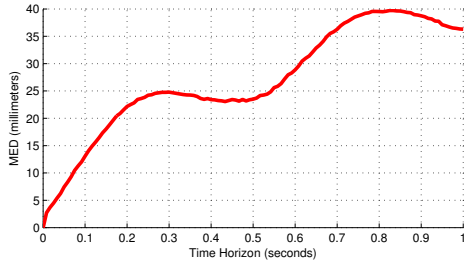
5.2.3. Noisy Observations

In this section, the path prediction algorithm is examined using a sequence example of noisy observations extracted by the single-frame pedestrian skeleton estimation algorithm described in Section 4.2. In Figure 5.6, images extracted from the sequence were presented. As shown, the sequence corresponds to a pedestrian that is walking on a zebra crossing from the left to right.

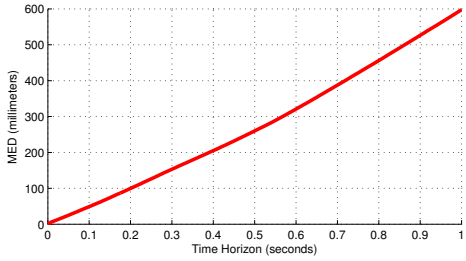
In Figure 5.11, the MEDs in millimetres for predictions up to 1 second when poses and displacements computed from the sequence by using the skeleton estimation algorithm are represented. The method achieves lateral MEDs values of 131.71±57.89, 250.95±89.00, 355.80±123.37 and 448.84±157.39 millimetres at 0.25, 0.5, 0.75 and 1 second respectively. However, the combined lateral and longitudinal MEDs are significantly longer. This is due to the fact that the pedestrian is not walking perpendicular to the sensor. As explained in Section 4.1, the training dataset is composed of people with left-to-right and right-to-left heading with a variance in the longitudinal component close to zero. Hence, the future path reconstruction is corrupted by the predicted displacement vectors. To solve this problem, the observations in the training set and test set should be normalised by means of rotations to have the same orientation with respect to the sensor. In this way, the method could predict future paths regardless the pedestrian direction.



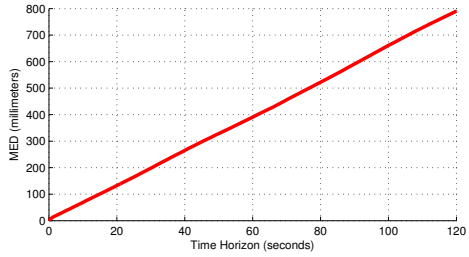
(a) Lateral MEDs.



(b) Height MEDs.



(c) Longitudinal MEDs



(d) Combined lateral and longitudinal MEDs.

Figure 5.11: MEDs in millimetres for predictions up to 1 second in the sequence example.

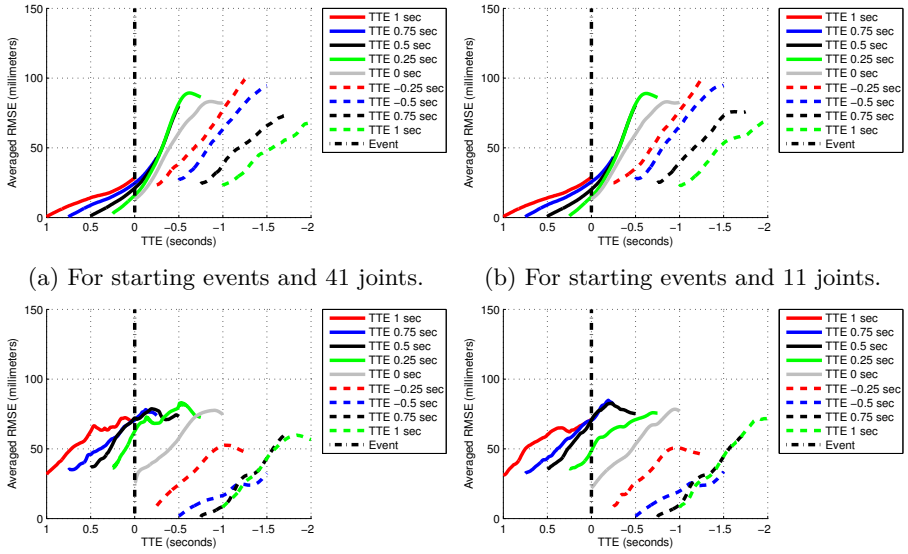
5.3. Pedestrian Pose Prediction Results

Throughout this section, the evaluation of pose prediction results is performed considering 41 and 11 joints. Firstly, as in other sections, the assessment is performed assuming the activity recognition with 11 joints. After that, in Section 5.3.2, the pose prediction results assuming that the activity recognition has an accurate rate of 100% are analysed with the motivation of estimating the influence of this task in the results. Finally, in Section 5.3.3, the pose prediction performed by means of noisy observations extracted by the skeleton estimation algorithm are examined.

5.3.1. Pedestrian Pose Prediction Results with Activity Recognition

In this section, the evaluation of pose prediction results is performed assuming an activity recognition with 11 joints. In Figure 5.12, the averaged RMSEs of pedestrian joints, i.e. the pedestrian posture, for different time horizons and TTEs are shown. As expected, when the pedestrian is standing, low errors are

obtained since all postures are similar for this activity. In fact, the low pose reconstruction error in the prediction $t = 0$ is especially significant since it denotes the low variability in the pedestrian poses. Thereby, a similar training posture to the test pedestrian pose is usually found when the most appropriate model is selected. However, higher errors in the reconstruction of the future poses are achieved when the pedestrian is moving. In this case, the pose reconstruction error in the prediction $t = 0$ denotes that a higher number of pedestrians should be included in the dataset in order to find similar pedestrian postures.



(a) For starting events and 41 joints.

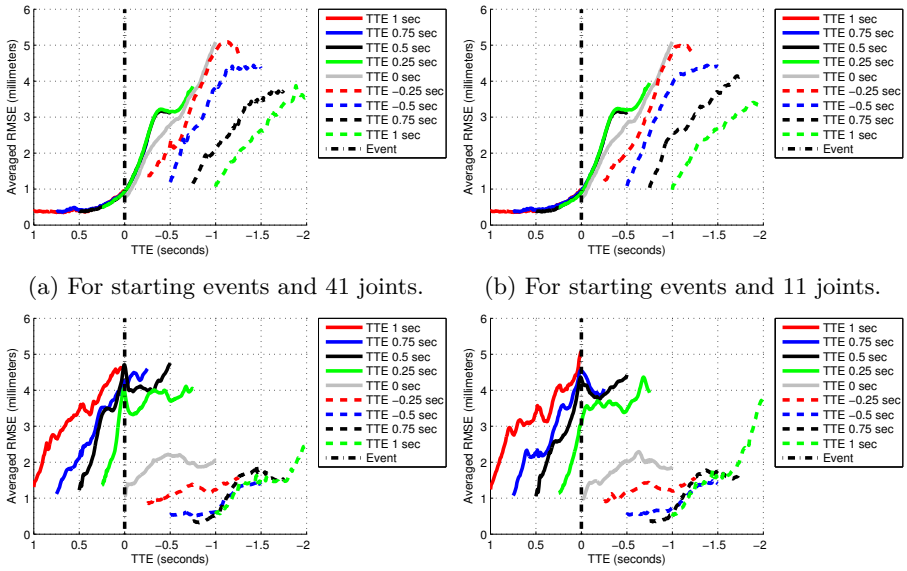
(b) For starting events and 11 joints.

(c) For stopping events and 41 joints.

(d) For stopping events and 11 joints.

Figure 5.12: Averaged RMSEs of pedestrian joints for time horizons up to 1 second and different TTEs.

Furthermore, in Figure 5.13, the averaged RMSEs of pedestrian displacements, i.e. the joint displacements between samples, for different time horizons and TTEs are illustrated. A similar analysis to the previous one can be done. As expected, when the pedestrian is standing, low errors are obtained since low displacements are gathered for this activity. Again, the low displacement reconstruction error in the prediction $t = 0$ denotes the low variability in the pedestrian displacements. However, higher errors in the reconstruction of the future displacement are achieved when the pedestrian is moving. In this case, the reconstruction error in the prediction $t = 0$ denotes that a higher number of pedestrians should be included in the dataset in order to find similar pedestrian displacements. Likewise, at a TTE value of -1 second, the errors grow up due to the fact that a new pedestrian dynam-



(a) For starting events and 41 joints.

(b) For starting events and 11 joints.

(c) For stopping events and 41 joints.

(d) For stopping events and 11 joints.

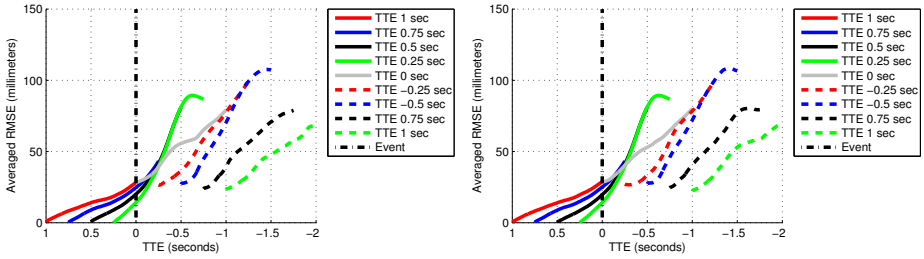
Figure 5.13: Averaged RMSEs of pedestrian displacements for time horizons up to 1 second and different TTEs.

ical change could happen. It is worth remarking that the path errors are directly influenced by the displacement reconstruction errors.

5.3.2. Pedestrian Pose Prediction Results without Activity Recognition

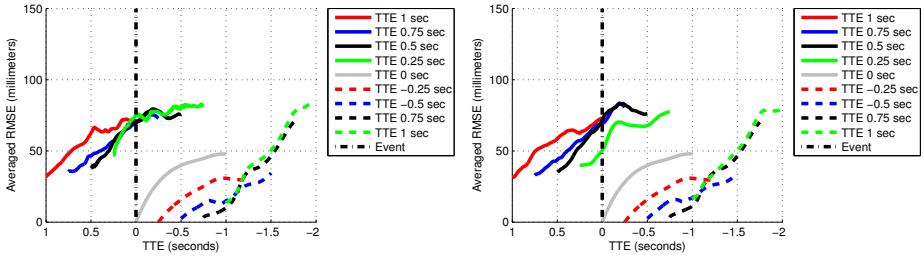
In this section, the evaluation of pose prediction results is performed assuming an activity recognition accurate rate of 100%. In Figure 5.14, the averaged RMSEs of pedestrian joints, i.e. the pedestrian posture, for different time horizons and TTEs are shown. As expected, when the pedestrian is standing, low errors are obtained since all postures are similar for this activity. As before, the low pose reconstruction error in the prediction $t = 0$ is especially significant since it denotes the low variability in the pedestrian poses. Thereby, a similar training posture to the test pedestrian pose is usually found when the most appropriate model is selected. However, higher errors in the reconstruction of the future poses are achieved when the pedestrian is moving. In this case, the reconstruction error in the prediction $t = 0$ denotes that a higher number of pedestrians should be included in the dataset in order to find similar pedestrian postures.

Additionally, in Figure 5.15, the averaged RMSEs of pedestrian displacements,



(a) For starting events and 41 joints.

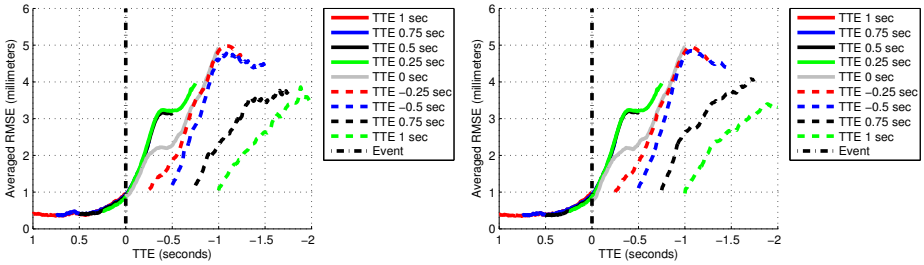
(b) For starting events and 11 joints.



(c) For stopping events and 41 joints.

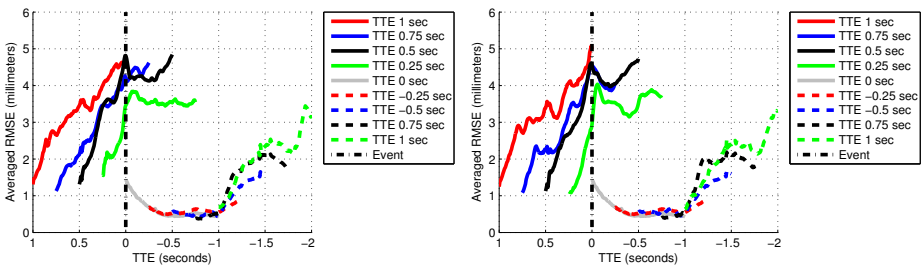
(d) For stopping events and 11 joints.

Figure 5.14: Averaged RMSEs of pedestrian joints for time horizons up to 1 second and different TTEs.



(a) For starting events and 41 joints.

(b) For starting events and 11 joints.



(c) For stopping events and 41 joints.

(d) For stopping events and 11 joints.

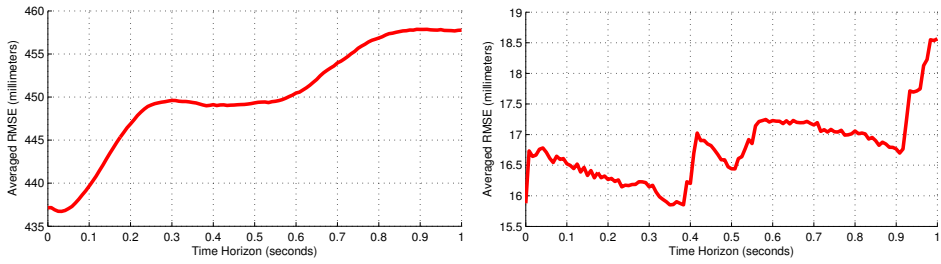
Figure 5.15: Averaged RMSEs of pedestrian displacements for time horizons up to 1 second and different TTEs.

i.e. the joint displacements between samples, for different time horizons and TTEs are shown. A similar analysis to the previous cases can be done. As expected, when the pedestrian is standing, low errors are obtained since low displacements are gathered for this activity. Again, the low displacement reconstruction error in the prediction $t = 0$ denotes the low variability in the pedestrian displacements. However, higher errors in the reconstruction of the future displacement are achieved when the pedestrian is moving. In this case, the reconstruction error in the prediction $t = 0$ denotes that a higher number of pedestrians should be included in the dataset in order to find similar pedestrian displacements.

When the results with and without the application of the activity recognition algorithm are compared, the influence of the transition delays in the reconstruction of the observations comes to light. The case especially significant is the stopping activity. After the event, the displacements, when the activity recognition has an accurate rate of 100%, have lower errors than the displacements when the activity recognition is applied. This difference in the errors explains the difference in the path prediction since, as mentioned in Section 4.5, the future pedestrian paths are computed adding N consecutive displacements.

5.3.3. Noisy Observations

In this section, the pose prediction algorithm is examined using a sequence example of noisy observations extracted by the single-frame pedestrian skeleton estimation algorithm described in Section 4.2. In Figure 5.6, images extracted from the sequence were presented. As shown, the sequence corresponds to a pedestrian that is walking on a zebra crossing from the left to right.



(a) Pose. (b) Displacement.

Figure 5.16: Averaged RMSE in the observation reconstruction for predictions up to 1 second.

In Figure 5.16, the averaged RMSEs in the pose and displacement reconstructions for predictions up to 1 second are illustrated. As shown, due to the fact

that noisy test observations are analysed, the reconstruction errors are higher than the errors with less noisy observations which were analysed in Section 5.3.1. It is worth remarking that the motivation of this thesis is not to develop a complex pedestrian skeleton estimation algorithm. Hence, it is expected that strong gains could be made in the performance of the method described in this document if more sophisticated systems are applied in the pedestrian pose extraction.

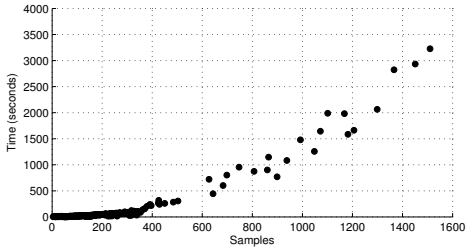
5.4. Processing Time

This section resumes the processing times of each step carried out by the method described in Chapter 4. In Figure 5.13, the processing times in milliseconds of the training step are represented for each activity. As mentioned in [63], the computational bottleneck for the B-GPDM is the inversion of the kernel matrices, which is necessary to evaluate the likelihood function and its gradient. As expected, the longer the sequence, the higher the processing time due to the fact that the dimensions of the kernel matrices depends on the number of samples in the sequence. For this reason, the processing time tends to be exponential with the number of samples in the sequences. Moreover, the SCG algorithm is sometimes unable to correctly optimise the models, thus accomplishing short processing times. It is worth noting that the training has been performed using MATLAB 2014 64-bits with a processor Intel i7-2600K 3.40GHz.

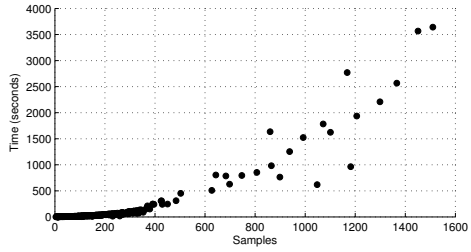
	Joints	**41**	**11**
Activity Recognition	**Mean**	85.0	43.6
	Std	29.7	21.0
	Min	33.5	23.9
	Max	858.5	242.2
Path Prediction	**Mean**	868.3	829.7
	Std	1284.4	1232.6
	Min	11.4	10.6
	Max	117685.7	129173.4
Total	**Mean**	741.5	670.5
	Std	1183.2	1129.8
	Min	34.8	25.1
	Max	117766.0	129215.5

Table 5.13: Processing times in milliseconds of each prediction step per pedestrian observation.

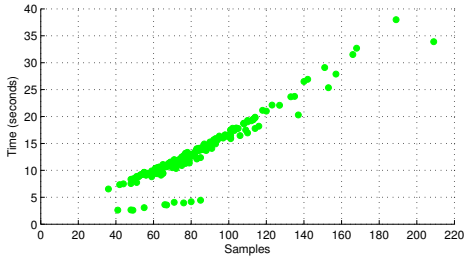
On the other hand, the path prediction and activity recognition has been performed by means of MATLAB 2016 64-bits with a processor Intel i7-7700K



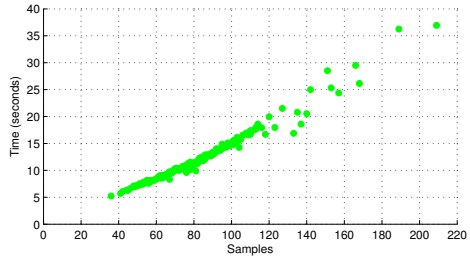
(a) Sequences of standing activities for 41 joints.



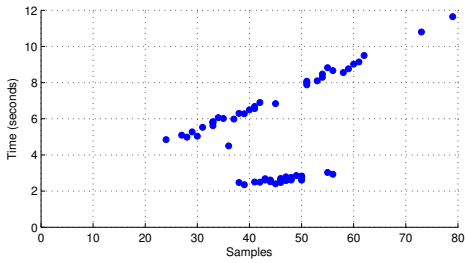
(b) Sequences of standing activities for 11 joints.



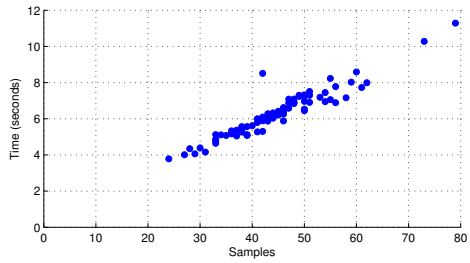
(c) Sequences of starting activities for 41 joints.



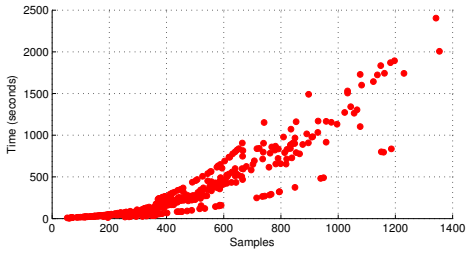
(d) Sequences of starting activities for 11 joints.



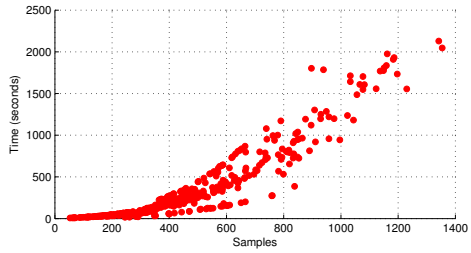
(e) Sequences of stopping activities for 41 joints.



(f) Sequences of stopping activities for 11 joints.



(g) Sequences of walking activities for 41 joints.



(h) Sequences of walking activities for 11 joints.

Figure 5.17: Processing times in milliseconds of the training step. The data are shown by pedestrian activity and number of joints.

4.20GHz. The processing times are showed in Table 5.13. The path prediction depends on the model selected in order to estimate the future pedestrian trajectory. If this model corresponds to a long sequence, the processing time is higher because the path prediction compute the inversion of the kernel matrix, which is necessary to evaluate the likelihood function and its gradient between the test observation and the reconstructed observation from the model (see Equation 4.4). Besides, the mean total processing time is shorter than the the mean path prediction time due to the fact that when the activity is recognised as standing, the path prediction is not performed.

5.5. Conclusions

An exhaustive assessment about activity recognition and path prediction algorithms has been performed throughout this chapter. Concerning activity recognition, the results verify that shoulder and leg motions are more valuable sources of information than other body parts to recognise the current pedestrian action. More specifically, the maximum accuracy rate, 95.13%, is achieved when observations composed of a few joints placed along the legs and shoulders are taken into consideration. However, the accuracy rate falls to 90.69% whether a higher number of joints located along the whole body are used. Additionally, at least two types of features are needed in the action recognition when more than two dynamical behaviours are considered, either body poses and displacements, or displacements and accelerations. The advantage of using the former is that only two pedestrian observations are needed for the activity recognition. Regarding this task, the method proposed in this document detects starting intentions 125 milliseconds after the gait initiation with an accuracy rate of 80% and recognises stopping intentions 58.33 milliseconds before the event with an accuracy rate of 70% when joints from shoulders and legs are considered.

Concerning the path prediction results, similar errors are obtained with respect to other works. However, some measures of accuracy used by other methods provide a vague idea of how well a system works. For example, the MED gives a more precise physical interpretation of the predicted pedestrian positions with respect to a groundtruth than the RMSE or the mean and standard deviation of the per-sequence RMSE. Hence, in this thesis, the measure of accuracy chosen for the path evaluation is the MED at different TTEs that gives objective information of the path prediction performance. Besides, although other works accomplished slightly errors than the method proposed in this document, their prediction algorithms

need a temporal window of n trajectory points instead of using two observations and the errors are evaluated for all time steps instead of being assessed at different TTEs.

On the other hand, the algorithms have been also tested using noisy observations extracted by a single-frame pedestrian skeleton estimation algorithm. Although the motivation of this thesis is not to develop a complex procedure for this task it is expected that strong gains could be made in the performance of the method described in this document if more sophisticated systems are applied in the pedestrian pose extraction.

Chapter 6

Main Contributions and Future Work

This chapter presents the global conclusions and discusses the main contributions introduced and developed along the chapters of this thesis. Finally, in Section 6.2, several futures lines of research which this thesis leaves open are drawn.

6.1. Main Contributions

This thesis proposes a single-frame method to predict pedestrian path, poses and intentions up to 1 second ahead in time by means of the B-GPDM and a HMM. The B-GPDM reduces the dimensionality of a set of feature vectors related in time and infers future latent positions. Likewise, given a latent position from the latent space, the corresponding feature vector can also be reconstructed. However, as claimed in [63], learning a generic model for all kind of pedestrian activities or combining some of them into a single model normally provides inaccurate estimations of future observations. For that reason, the method proposed in this thesis learns multiple models of each type of pedestrian activities, i.e. walking, stopping, starting and standing, and selects the most appropriate among them to estimate future pedestrian states at each instant of time. This strategy allows to design scalable systems in which new sequences with different dynamics can be added to the dataset without negatively impacting the performance.

Additionally, a event-labelling methodology was proposed. This methodology allows to identify the instant of time that a pedestrian starts or finishes an event

such as starting or stopping. Thereby, a starting activity was defined as the action that begins when the pedestrian moves one knee to initiate the gait and ends when the foot of that leg touches the ground again. Besides, a stopping activity was defined as the action that begins when a foot is raised for the last step and finishes when that foot treads the ground. This criterion was adopted because these events happen in all sequences in which starting or stopping activities are included and because they are easily labelled by human experts, thus enabling the creation of reliable groundtruths. Moreover, to test the proposed method with noisy observations, a single-frame pedestrian skeleton estimation algorithm was proposed. This algorithm is based on the extraction of point clouds corresponding to different pedestrian body parts and the location of 3D joints in an hierarchical top-down search given anthropometric proportions and geometrical constraints.

On the other hand, one of the goals of this thesis was to test the feasibility and limits of the proposed method in an extensive way under ideal conditions by using a high frequency and low noise dataset published by CMU. The high frequency of the dataset helps the algorithms to properly learn the dynamics of different activities and increases the probability of finding a similar test observation in the trained data without missing intermediate observations. Besides, low noise models improve the prediction when working with noisy test samples. The CMU dataset is composed of sequences where people are simulating typical pedestrian activities at the same time that 3D coordinates of 41 joints along their bodies are being gathered at 120 Hz. Because of the high frequency and low noise sequences included in the dataset and the the event-labelling methodology chosen, the projection of pedestrian observations related in time onto the different subspaces compared in this thesis emerges as well-defined trajectories. For example, walking activities generate cyclic trajectories where each cycle corresponds to two pedestrian steps, starting and stopping activities generate trajectories of a half cycle since only one step was considered in the event-labelling. Finally, the models that correspond to standing sequences produce non-cyclic trajectories. Unlike other dimensionality reduction techniques such as PCA, PPCA, GPLVM or GPDM, B-GPDM obtains smoother trajectories onto the learned subspaces which provide more accurate estimations of future pedestrian states. It is worth mentioning that GPLVM produces very noisy trajectories in the subspace caused by the fact that this modelling technique is mainly focused on pattern recognition instead of modelling time-related data.

Moreover, due to the fact that not all gathered joints in the CMU dataset offer discriminative information about the current and future pedestrian activities, two

different set of joints are compared in order to determine whether the detection of only shoulder and leg motions are enough to infer future states. It seems that the models are not influenced by the reduction in the number of joints. However, with respect to the activity recognition, using a less number of joint provides more accurate results. Therefore, the results verify that shoulder and leg motions are more valuable sources of information than other body parts to recognise the current pedestrian action. More specifically, the maximum accuracy rate, 95.13%, is achieved when observations composed of poses and displacements from only 11 joints were taken into consideration. However, the accuracy rate falls to 90.69% whether 41 joints are used. Likewise, by considering only body poses, a similar conclusion is drawn since the maximum accuracy rate is 91.28% and 88.39% for 11 and 41 joints respectively. Finally, when the observations are composed solely of pedestrian displacements, the activity recognition results are not significantly influenced by the number of joints.

Regarding the distinction among activities, the pedestrian displacements achieve a better differentiation of standing actions from the rest of activities. However, with respect to starting and stopping actions, a larger number of critical missclassifications are produced. This means that the displacements do not allow to reliably distinguish whether a pedestrian is carrying out the first or last step. Therefore, the body poses along with the displacements offer a more discriminative information in these cases. Besides, it seems that the first half of the first step and the second half of the last step contain the most perceptible information to determine starting and stopping actions respectively. Beyond that, considering the body pose as the only feature, standing actions are repeatedly recognised as walking activities since, when the pedestrian legs are closed, the poses from both states are very similar in those instants of time. Therefore, the displacements are valuable information in those cases. Thereby, when a large number of dynamical activities are considered, such as standing, starting, stopping and walking, the body poses and displacements are important features. Moreover, including the acceleration as an additional feature may improve the recognition of starting and stopping activities. However, when the pedestrian legs are completely opened, the acceleration is minimum and it is maximum when the legs are closed. Hence, the body pose is again an essential information to distinguish standing and walking actions. As a conclusion, at least two types of features are needed in the activity recognition when more than two state are considered, either body poses and displacements, or displacements and accelerations. The advantage of using body poses and displacements is that only two pedestrian observations are needed for the activity recognition.

Regarding the delays of the transitions between activities, the results show that these are not significantly influenced by the number of joints. Moreover, it should be pointed out that starting-walking transitions have negative delays due to the fact that the first half of the first step contains the most perceptible information to determine starting actions. The method proposed in this document recognises starting intentions 125 milliseconds after the gait initiation with an accuracy rate of 80% when 11 joints are considered. These results are similar to the delays achieved in other works. On the other hand, standing actions are recognised 58.33 milliseconds before the event with an accuracy rate of 70% when 11 joints are considered.

Concerning the path prediction results, similar errors are obtained with respect to other works. However, some measures of accuracy used by other methods provide a vague idea of how well a system works. For example, the MED gives a more precise physical interpretation of the predicted pedestrian positions with respect to a groundtruth than the RMSE or the mean and standard deviation of the per-sequence RMSE. Hence, in this thesis, the measure of accuracy chosen for the path evaluation is the MED at different TTEs that gives objective information of the path prediction performance. Besides, although other works accomplished slightly errors than the method proposed in this document, their prediction algorithms need a temporal window of n trajectory points instead of using two observations and the errors are evaluated for all time steps instead of being assessed at different TTEs.

On the other hand, the algorithms have been also tested using noisy observations extracted by a single-frame pedestrian skeleton estimation algorithm. Although the motivation of this thesis is not to develop a complex procedure for this task it is expected that strong gains could be made in the performance of the method described in this document if more sophisticated systems are applied in the pedestrian pose extraction.

Finally, four publications were presented from this thesis in different international conferences about ITS, i.e. [46–48, 64]. It is worth mentioning that the [48] were awarded with the *Best Paper of Workshop on 18th IEEE International Conference on Intelligent Transportation Systems 2015.*

6.2. Future Work

From the results and conclusions of the present work, several lines of work can be proposed. They correspond to different aspect that have not been solved or

need a further analysis to improved the performance:

1. A higher number of sequences should be considered since children or elderly people are not included in the CMU dataset. As claimed in [45], elderly pedestrians select more dangerous decisions than younger people despite the fact that they normally take more time to make them.

2. Testing all algorithms with different type of features or combining them may improve the performance of the method proposed in this thesis. For example, motion features obtained by means of optical flow or motion history images instead of pedestrian displacements extracted from body poses can be used as well. Additionally, in a higher level, the combination of context-based information along with a situation criticality evaluation and a pedestrian body language analysis would allow to develop more reliable AEBSs. Thus, scene understanding, pedestrian detection and prediction algorithms are interesting lines of research in the ITS field.

3. In order to obtain more accurate pedestrian skeletons, markerless motion capture approaches based on Convolutional Neural Networks (CNNs) such as the algorithm proposed in [13] could be developed instead of algorithm based on geometrical constrains.

4. Comparing the B-GPDM and other modelling technique which are able to predict future observation such as ANNs and KFs using high frequency and low noise datasets and body pose features.

5. Creating a extensive dataset of real pedestrian situations would make possible to compare different approaches in similar conditions. The event-labelling methodology proposed in this thesis would help to human experts determine the different pedestrian activities.

6. Testing the algorithms in moving vehicles. To do that, the ego-motion should be compensated every instant of time.

Bibliography

[1] Abramson, Y., and Steux, B. Hardware-friendly pedestrian detection and impact prediction. In *Intelligent Vehicles Symposium, 2004 IEEE* (June 2004), pp. 590–595.

[2] Asahara, A., Maruyama, K., Sato, A., and Seto, K. Pedestrian-movement prediction based on mixed markov-chain model. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (New York, NY, USA, 2011), GIS '11, ACM, pp. 25–33.

[3] Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[4] Bohannon, R. W. Comfortable and maximum walking speed of adults aged 20-79 years: reference values and determinants. *Age and ageing 26*, 1 (1997), 15–19.

[5] Bonnin, S., Weisswange, T. H., Kummert, F., and Schmuedderich, J. Pedestrian crossing prediction using multiple context-based models. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)* (Oct 2014), pp. 378–385.

[6] Cayton, L. Algorithms for manifold learning. *Univ. of California at San Diego Tech. Rep* (2005), 1–17.

[7] Chen, Z., Ngai, D. C. K., and Yung, N. H. C. Pedestrian behavior prediction based on motion patterns for vehicle-to-pedestrian collision avoidance. In *2008 11th International IEEE Conference on Intelligent Transportation Systems* (Oct 2008), pp. 316–321.

[8] Chen, Z., Wang, L., and Yung, N. H. Adaptive human motion analysis and prediction. *Pattern Recognition 44*, 12 (2011), 2902 – 2914.

[9] Chen, Z., and Yung, N. H. C. Improved multi-level pedestrian behavior prediction based on matching with classified motion patterns. In *2009 12th International IEEE Conference on Intelligent Transportation Systems* (Oct 2009), pp. 1–6.

[10] CMU. Cmu graphics lab motion capture database. http://mocap.cs.cmu.edu/.

[11] Coelingh, E., Eidehall, A., and Bengtsson, M. Collision warning with full auto brake and pedestrian detection - a practical example of automatic emergency braking. In *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on* (Sept 2010), pp. 155–160.

[12] Dollar, P., Wojek, C., Schiele, B., and Perona, P. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence 34*, 4 (April 2012), 743–761.

[13] Elhayek, A., de Aguiar, E., Jain, A., Thompson, J., Pishchulin, L., Andriluka, M., Bregler, C., Schiele, B., and Theobalt, C. Marconi-convnet-based marker-less motion capture in outdoor and indoor scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence 39*, 3 (March 2017), 501–514.

[14] Ellis, D., Sommerlade, E., and Reid, I. Modelling pedestrian trajectory patterns with gaussian processes. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on* (Sept 2009), pp. 1229–1234.

[15] Enzweiler, M., and Gavrila, D. M. Monocular pedestrian detection: Survey and experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence 31*, 12 (2009), 2179–2195.

[16] Fugger, T., Randles, B., Stein, A., Whiting, W., and Gallagher, B. Analysis of pedestrian gait and perception-reaction at signal-controlled crosswalk intersections. *Transportation Research Record: Journal of the Transportation Research Board*, 1705 (2000), 20–25.

[17] Gandhi, T., and Trivedi, M. M. Image based estimation of pedestrian orientation for improving path prediction. In *Intelligent Vehicles Symposium, 2008 IEEE* (June 2008), pp. 506–511.

[18] Geronimo, D., Lopez, A. M., Sappa, A. D., and Graf, T. Survey of pedestrian detection for advanced driver assistance systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence 32*, 7 (July 2010), 1239–1258.

[19] Goldhammer, M., Doll, K., Brunsmann, U., Gensler, A., and Sick, B. Pedestrian's trajectory forecast in public traffic with artificial neural networks. In *Pattern Recognition (ICPR), 2014 22nd International Conference on* (Aug 2014), pp. 4110–4115.

[20] Goldhammer, M., Gerhard, M., Zernetsch, S., Doll, K., and Brunsmann, U. Early prediction of a pedestrian's trajectory at intersections. In *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)* (Oct 2013), pp. 237–242.

[21] Goldhammer, M., Hubert, A., Koehler, S., Zindler, K., Brunsmann, U., Doll, K., and Sick, B. Analysis on termination of pedestrians' gait at urban intersections. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)* (2014), IEEE, pp. 1758–1763.

[22] Goldhammer, M., Köhler, S., Doll, K., and Sick, B. Camera based pedestrian path prediction by means of polynomial least-squares approximation and multilayer perceptron neural networks. In *SAI Intelligent Systems Conference (IntelliSys), 2015* (Nov 2015), pp. 390–399.

[23] Hamaoka, H., Hagiwara, T., Tada, M., and Munehiro, K. A study on the behavior of pedestrians when confirming approach of right/left-turning vehicle while crossing a crosswalk. In *Intelligent Vehicles Symposium* (2013), IEEE, pp. 106–110.

[24] Hamdane, H., Serre, T., Masson, C., and Anderson, R. Issues and challenges for pedestrian active safety systems based on real world accidents. *Accident Analysis and Prevention*, 82 (Jan 2015), pp. 53–60.

[25] Helbing, D., and Molnar, P. Social force model for pedestrian dynamics. *Physical review E 51*, 5 (1995), 4282.

[26] Hermes, C., Wohler, C., Schenk, K., and Kummert, F. Long-term vehicle motion prediction. In *Intelligent Vehicles Symposium, 2009 IEEE* (June 2009), pp. 652–657.

[27] Huang, Y., Cui, J., Davoine, F., Zhao, H., and Zha, H. Head pose based intention prediction using discrete dynamic bayesian network. In *Distributed*

Smart Cameras (ICDSC), 2013 Seventh International Conference on (Oct 2013), pp. 1–6.

[28] Izenman, A. J. Introduction to manifold learning. *Wiley Interdisciplinary Reviews: Computational Statistics 4*, 5 (2012), 439–446.

[29] Jackson, J. *A User's Guide to Principal Components*. Wiley Series in Probability and Statistics. Wiley, 2005.

[30] Jolliffe, I. *Principal Component Analysis*. Springer Series in Statistics. Springer, 2002.

[31] Keller, C. G., and Gavrila, D. M. Will the pedestrian cross? a study on pedestrian path prediction. *IEEE Transactions on Intelligent Transportation Systems 15*, 2 (April 2014), 494–506.

[32] Keller, C. G., Hermes, C., and Gavrila, D. M. *Will the Pedestrian Cross? Probabilistic Path Prediction Based on Learned Motion Features*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 386–395.

[33] Kitani, K. M., Ziebart, B. D., Bagnell, J. A., and Hebert, M. *Activity Forecasting*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 201–214.

[34] Köehler, S., Goldhammer, M., Bauer, S., Zecha, S., Doll, K., Brunsmann, U., and Dietmayer, K. Stationary detection of the pedestrian?s intention at intersections. *IEEE Intelligent Transportation Systems Magazine 5*, 4 (winter 2013), 87–99.

[35] Köhler, S., Goldhammer, M., Bauer, S., Doll, K., Brunsmann, U., and Dietmayer, K. Early detection of the pedestrian's intention to cross the street. In *2012 15th International IEEE Conference on Intelligent Transportation Systems* (Sept 2012), pp. 1759–1764.

[36] Köhler, S., Goldhammer, M., Zindler, K., Doll, K., and Dietmeyer, K. Stereo-vision-based pedestrian's intention detection in a moving vehicle. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems* (Sept 2015), pp. 2317–2322.

[37] Köhler, S., Schreiner, B., Ronalter, S., Doll, K., Brunsmann, U., and Zindler, K. Autonomous evasive maneuvers triggered by infrastructure-based detection of pedestrian intentions. In *Intelligent Vehicles Symposium (IV), 2013 IEEE* (June 2013), pp. 519–526.

[38] Kooij, J. F. P., Schneider, N., Flohr, F., and Gavrila, D. M. Context-based pedestrian path prediction. In *ECCV (6)* (2014), vol. 8694 of *Lecture Notes in Computer Science*, Springer, pp. 618–633.

[39] Kwak, J.-Y., Lee, E.-J., Ko, B., and Jeong, M. Pedestrian's intention prediction based on fuzzy finite automata and spatial-temporal features. *Electronic Imaging 2016*, 3 (2016), 1–6.

[40] Lawrence, N. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of Machine Learning Research 6*, Nov (2005), 1783–1816.

[41] Lawrence, N. D. Gaussian process latent variable models for visualisation of high dimensional data. *Advances in neural information processing systems 16*, 3 (2004), 329–336.

[42] Lawrence, N. D., and Quiñonero Candela, J. Local distance preservation in the gp-lvm through back constraints. In *Proceedings of the 23rd International Conference on Machine Learning* (New York, NY, USA, 2006), ICML '06, ACM, pp. 513–520.

[43] Lindman, M., Ödblom, A., Bergvall, E., Eidehall, A., Svanberg, B., and Lukaszewicz, T. Benefit estimation model for pedestrian auto brake functionality. *Expert Symposium on Accident Research, 4th International Conference on*, 77 (2010).

[44] Møller, M. F. A scaled conjugate gradient algorithm for fast supervised learning. *Neural networks 6*, 4 (1993), 525–533.

[45] Oxley, J. A., Ihsen, E., Fildes, B. N., Charlton, J. L., and Day, R. H. Crossing roads safely: an experimental study of age differences in gap selection by pedestrians. *Accident; analysis and prevention 37*, 5 (September 2005), 962–971.

[46] Quintero, R., Almeida, J., Llorca, D. F., and Sotelo, M. A. Pedestrian path prediction using body language traits. In *2014 IEEE Intelligent Vehicles Symposium Proceedings* (June 2014), pp. 317–323.

[47] Quintero, R., Parra, I., Llorca, D. F., and Sotelo, M. A. Pedestrian path prediction based on body language and action classification. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)* (Oct 2014), pp. 679–684.

[48] Quintero, R., Parra, I., Llorca, D. F., and Sotelo, M. A. Pedestrian intention and pose prediction through dynamical models and behaviour classification. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems* (Sept 2015), pp. 83–88.

[49] Rabiner, L. R. Readings in speech recognition. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990, ch. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, pp. 267–296.

[50] Rasmussen, C. E., and Williams, C. K. I. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[51] Ratsamee, P., Mae, Y., Ohara, K., Takubo, T., and Arai, T. People tracking with body pose estimation for human path prediction. In *2012 IEEE International Conference on Mechatronics and Automation* (Aug 2012), pp. 1915–1920.

[52] Rosén, E., Källhammer, J.-E., Eriksson, D., Nentwich, M., Fredriksson, R., and Smith, K. Pedestrian injury mitigation by autonomous braking. *Accident Analysis & Prevention 42*, 6 (2010), 1949–1957.

[53] Schmidt, S., and Färber, B. Pedestrians at the kerb - recognising the action intentions of humans. *Transportation Research Part F: Traffic Psychology and Behaviour 12*, 4 (2009), 300 – 310.

[54] Schneider, N., and Gavrila, D. M. *Pedestrian Path Prediction with Recursive Bayesian Filters: A Comparative Study*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 174–183.

[55] Schulz, A. T., and Stiefelhagen, R. A controlled interactive multiple model filter for combined pedestrian intention recognition and path prediction. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems* (Sept 2015), pp. 173–178.

[56] Tian, R., Du, E. Y., Yang, K., Jiang, P., Jiang, F., Chen, Y., Sherony, R., and Takahashi, H. Pilot study on pedestrian step frequency in naturalistic driving environment. In *Intelligent Vehicles Symposium (IV), 2013 IEEE* (June 2013), pp. 1215–1220.

[57] Tipping, M. E., and Bishop, C. M. Mixtures of probabilistic principal component analyzers. *Neural Comput. 11*, 2 (Feb. 1999), 443–482.

[58] Tipping, M. E., and Bishop, C. M. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B 61* (1999), 611–622.

[59] Urtasun, R., Fleet, D. J., and Fua, P. 3D People Tracking with Gaussian Process Dynamical Models. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Washington, DC, USA, 2006), IEEE Computer Society, pp. 238–245.

[60] Vicon. Vicon motion capture system. https://www.vicon.com/.

[61] Völz, B., Mielenz, H., Agamennoni, G., and Siegwart, R. Feature relevance estimation for learning pedestrian behavior at crosswalks. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems* (Sept 2015), pp. 854–860.

[62] Wang, J. M., Fleet, D. J., and Hertzmann, A. Gaussian process dynamical models. In *In NIPS* (2006), MIT Press, pp. 1441–1448.

[63] Wang, J. M., Fleet, D. J., and Hertzmann, A. Gaussian process dynamical models for human motion. *IEEE transactions on pattern analysis and machine intelligence 30*, 2 (2008), 283–298.

[64] Worrall, S., Quintero, R., Xian, Z., Zyner, A., Philips, J., Ward, J., Bender, A., and Nebot, E. Multi-sensor detection of pedestrian position and behaviour. In *Proceedings of the 23rd World Congress on Intelligent Transport Systems* (2016).

[65] Yannis, G., Papadimitriou, E., and Theofilatos, A. Pedestrian gap acceptance for mid-block street crossing. *Transportation Planning and Technology 36*, 5 (2013), 450–462.

[66] Zebala, J., Ciepka, P., and Reza, A. Pedestrian acceleration and speeds. *Problems of Forensic Sciences*, 91 (2012), 227–234.

[67] Zivkovic, Z. Improved adaptive gaussian mixture model for background subtraction. In *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 2 - Volume 02* (Washington, DC, USA, 2004), ICPR '04, IEEE Computer Society, pp. 28–31.

[68] Zivkovic, Z., and van der Heijden, F. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recogn. Lett. 27*, 7 (May 2006), 773–780.