Pedestrian Intention Recognition by Means of a Hidden Markov Model and Body Language

R. Quintero, I. Parra, J. Lorenzo, D. Fernández-Llorca and M. A. Sotelo

Abstract— **According to several reports published by world-wide organisations, thousands of pedestrians die on road accidents every year. Due to this fact, vehicular technologies have been evolving with the intent of reducing these fatalities. Improving these technological advances is crucial since an early recognition of pedestrian intentions can lead to much more accurate active interventions in last second automatic manoeuvres. This paper proposes a method based on a Hidden Markov Model that recognises intentions by means of 3D positions and displacements of 11 joints located along the pedestrian body. The method is able to recognise the intention with an accuracy of 95.13%. It recognises starting intentions 125 ms after gait initiation with an accuracy of 80% and stopping intentions 291.67 and 58.33 ms before the event with an accuracy of 50% and 70% respectively. In addition, an approach based on point clouds and anthropometric constraints to extract the joints in realistic environments is proposed.**

I. Introduction and Related Works

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 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 pedestrian protection systems, AEBS or other sort of ADAS. Improving these technological advances is imperative since, for example, an early braking initiation or an accurate assessment about pedestrian positions before collisions could be particularly relevant. Similarly, an early recognition of pedestrian intentions can lead to much more accurate active interventions in last second automatic manoeuvres. The assessment from a moving vehicle of whether a pedestrian will cross or stop is regularly carried out extracting motion features by means of image processing. For example, in [1], [2], augmented motion features derived from dense optical flow fields are processed for path and intention predictions. These works recognise pedestrian walking and stopping intentions about 200-230 ms in advance with an accuracy of 80%. They also provide the capacity of human experts as baseline, who reach the same accuracy about 570 ms before the event. Another example can be found in [3], where a method to recognise starting, stopping and bending in intentions is implemented by means of SVM classifiers. The motion features are gathered using the overlapping of pedestrian silhouette images which are based on depth maps at consecutive time steps. In this work, stopping intentions are detected from 500 to 125 ms before standing still with an accuracy of 80% and 100% respectively. Bending in intentions are recognised from 320 to 570 ms after the first visible lateral body motion with the same accuracy. Finally, starting intentions are detected from 125 to 250 ms after the event with an accuracy of 75% and 100%.

On the other hand, many dangerous situations arise from the fact that the driver's view of the road scene may be obstructed by objects. For this 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 [4], [5], [6] 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. In these works, linear 2-class SVM classifiers are used to determine whether a motion-based descriptor belongs to a pedestrian which is starting to walk or not. The approach developed in [4] recognises starting intentions 120 and 340 ms after the gait initiation with an accuracy of 80% and 99% respectively in a controlled scenario. Similar results are obtained in [5], [6] despite more realistic scenarios are tested. Finally, the method developed in [7], which is focused on the early recognition of the gait initiation, is also evaluated and compared with the approach developed in [6]. The first work outperforms the second one achieving a precision of 80% at the moment of the event.

This paper proposes a method based on a Hidden Markov Model (HMM) that recognises pedestrian intentions, i.e. walking, stopping, starting and standing, by means of 3D positions and displacements of 11 joints located along the bodies. These features are extracted from a high frequency and low noise dataset published by Carnegie Mellon University (CMU) [8]. Besides, a single-frame pedestrian skeleton estimation algorithm is proposed to extract the joints from video sequences.

The paper is organized as follows: Section II describes the CMU dataset. The pedestrian skeleton estimation algorithm is detailed in Section III. Section IV describes the intention recognition algorithm. Experimental results are presented in Section V. Finally, the main contributions and future works are briefly outlined in Section VI.

R. Quintero, I. Parra, J. Lorenzo, D. Fernández-Llorca and M.A Sotelo are with the Computer Engineering Department, University of Alcalá, Alcalá de Henares, Spain raul.quintero@uah.es

978-1-5386-1526-3/17/$31.00 ©2017 IEEE

II. Dataset Description

One of the goals of this paper 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. This dataset of pedestrian motion sequences were described in our previous works [9], [10]. However, in this paper, the dataset were extended. In such a way, 490 sequences composed of 302470 pedestrian poses from 31 subjects was extracted. Hereafter, this set of sequences will be named as UAH dataset.

In the sequences, people are simulating typical pedestrian activities at the same time that 3D coordinates of 41 joints located along their bodies are being gathered. Nonetheless, not all these joints offer discriminative information about the current and future pedestrian actions. In fact, joints located along the arms do not contribute to distinguish walking, starting, stopping or standing intentions. For that reason, a subset of 11 joints has been selected to determine if the detection of only shoulder and leg motions is enough to infer intentions. An example of a pedestrian pose from different points of view is shown in Figure 1.

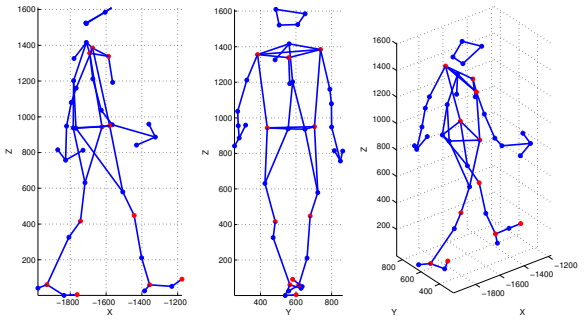


Fig. 1. Example of pedestrian pose extracted from the dataset published by CMU in which 41 joints (blue markers) and a subset of 11 joints (red markers) are shown.

As in the previous works, the UAH dataset was hierarchically divided into 8 subsets. The first division was based on the orientation, i.e. left-to-right and right-to-left, and the second one was based on the type of intention, i.e. walking, starting, stopping and standing. For this last division, a new guideline of event-labelling orientated to typical pedestrian intentions is proposed in this paper. This guideline allows to objectively identify the instant that a pedestrian starts or finishes an event such as starting or stopping. Specifically, a starting intention 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. A stopping intention 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. A breakdown of the UAH dataset based on the number of sequences and pedestrian poses is shown in Table I .

	Orientation	Walking	Starting	Stopping	Standing	Total
Seq	L-R	240	142	56	224	662
Seq	R-L	191	121	27	156	495
Total		431	263	83	380	1157
Poses	L-R	107324	10732	2522	43151	163729
Poses	R-L	95113	10940	1276	31412	138741
Total		202437	21672	3798	74563	302470

TABLE I

III. Pedestrian Skeleton Estimation

To test the proposed method with noisy observations, a single-frame pedestrian skeleton estimation algorithm based on point clouds extracted from a stereo pair and geometrical constraints was implemented. This algorithm is a variation of the method proposed in [11], [12]. The stereo pair is composed of two cameras with a resolution of 1920x1200 pixels and a focal length of 12.5 mm which captures images at 120 Hz. A baseline of 40 cm is set to detect pedestrians in a range from 5 to 15 m. The estimated skeletons are composed of 11 points placed along the pedestrian body which represent the shoulders, hips, knees, ankles and toes. This set of points is the same set described in Section II. The algorithm assumes that a pedestrian is standing and his highest point corresponds to the head.

A. Pedestrian 3D Point Cloud Extraction

Although the motivation of this paper 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 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. However, some errors could arise when pedestrians are close enough to objects from the background and their feet could not be correctly segmented since their depth values are similar to the values corresponding to the ground floor.

The vision-based pedestrian segmentation algorithm works as follows. Firstly, the depth map is computed by means of the SGM algorithm. Then, the pixels that represent the ground floor on the 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 with 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 illustrated in Figure 2.

B. Skeleton Estimation

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

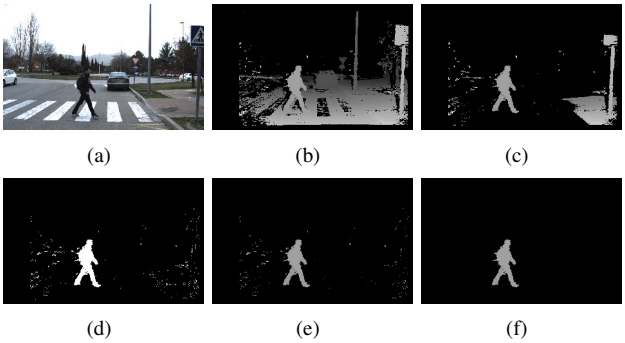


Fig. 2. Pedestrian segmentation algorithm. (a) Original colour image captured by the stereo system. (b) Depth map. (c) Depth map where values which correspond to the ground plane were removed. (d) Foreground mask of moving objects.(e) Foreground mask of moving objects with depth map values. (f) Filtered foreground mask.

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

1) Head: Firstly, the point cloud belonging to the pedestrian head is extracted and its centroid, c_{head}, computed. A Linear Least Squares (LLS) fitting of $t \in \{2,3,\ldots,N\}$ consecutive head positions, c_{head}, allows to compute the pedestrian heading line, l_{head}, whose projection onto the ground plane, l'_{head}, is represented by the red line in Figure 5. This fitting is only performed when pedestrians are moving since, in any other case, it could produce noisy measurements.

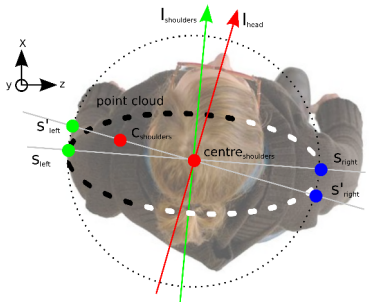


Fig. 3. Diagram of pedestrian shoulders estimation.

2) Shoulders and Hips: In the next step, the shoulders positions are estimated. A diagram of this process is shown in Figure 3. 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 correspond 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 5.

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 hips-based heading is represented by the purple line in Figure 5.

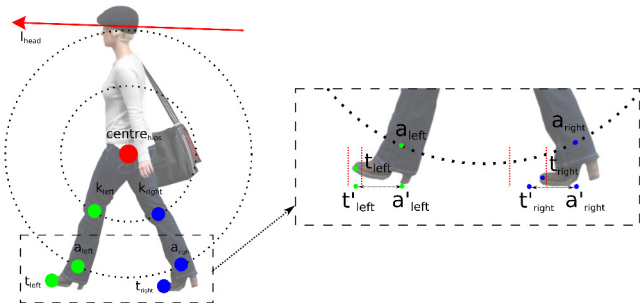
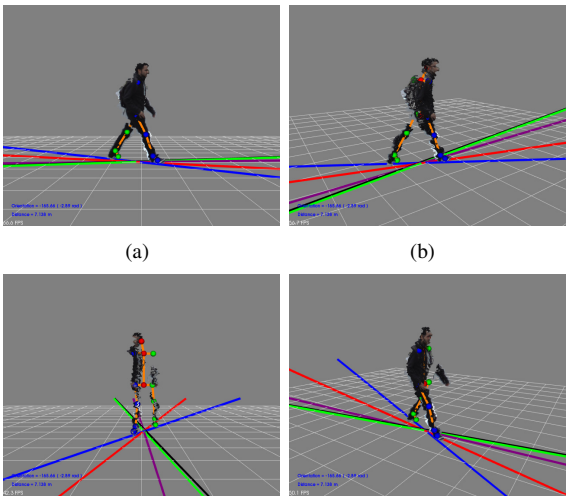


Fig. 4. Diagram of pedestrian limbs estimation.

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. 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 $centre_{hips_y}$. 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 most appropriate 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 5. 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 5. After that, the centroids of each new cluster, k_{left} and k_{right}, are computed. 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.



(a) (b) (c) (d)

Fig. 5. Example of a pedestrian skeleton estimation. Green markers correspond to left joints, blue markers to right joints and red markers to head, centre of shoulders and centre of hips.

In a similar way, the pedestrian ankles are estimated. In this case, a sphere, whose centre is also $centre_{hips}$ but the value of the radius 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.

IV. Pedestrian Intention Recognition

In this paper, a variation of the intention recognition proposed in [9], [10] is described. The maximum similarity between the current observation and each observation of the UAH dataset may determine the intention. Nevertheless, if this maximum similarity were applied directly, that is, without modelling the evolution of the pedestrian intention, higher errors would be achieved 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 action or to the end of a starting intention. Thus, if the previous observation were recognised as walking, then the next dynamics should 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. At any time, the pedestrian can do one of a set of 4 distinct actions $\mathbf{s} = \{Standing, Starting, Stopping, Walking\}$. These states are not observable since only 3D information from joints is available. Therefore, the states can be only inferred through the observations $\mathbf{x}$. Hence, the implementation of a first-order HMM allows to model the transitions between intentions and to recognise the correct one taking into account the previous dynamics.

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

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

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:

$$P(\mathbf{s}_j^t) \propto \max_{i=1}^{4}[P(\mathbf{s}_j^t|\mathbf{s}_i^{t-1})P(\mathbf{s}_i^{t-1}|\mathbf{x}^{t-1})], \quad t > 1 \quad (2)$$

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. The values of transitions between states were experimentally fixed maximising the success rate. $P(\mathbf{s}_i^{t-1}|\mathbf{x}^{t-1})$ 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 intention is unknown in $t = 1$.

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

$$P(\mathbf{x}^t|\mathbf{s}_j^t) \propto \max_{i=1}^{N}\left(\frac{1}{1+\alpha_i} + \frac{1}{1+\beta_i}\right) \quad (3)$$

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 joint displacements respectively. The SSE are computed between the current pedestrian observation $\mathbf{x}^t$ 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 to find pedestrians with similar joint velocities.

V. Results

The intention recognition algorithm was tested using the UAH dataset 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. The intention recognition results are summarised on a confusion matrix shown in Table II. Additionally, a more exhaustive evaluation is carried out to test the algorithm in a more real environments. Thereby, it was also tested using data extracted by the skeleton estimation algorithm. An exhaustive data assessment of the confusion matrix is represented in Table III where it is taken into account each pedestrian feature and intention.

TABLE II

		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

TABLE III

Features		**Pose + Disp**	**Pose**	**Disp**
Accuracy		**95.13%**	91.28%	94.23%
Precision	**Standing**	97.51%	95.54%	**98.04%**
	Starting	**87.57%**	79.38%	83.96%
	Stopping	**53.51%**	40.06%	35.72%
	Walking	**95.57%**	91.35%	94.81%
Recall	**Standing**	96.97%	87.86%	**97.19%**
	Starting	**61.49%**	39.14%	54.90%
	Stopping	**51.38%**	37.11%	40.90%
	Walking	98.88%	**99.13%**	98.36%
F1-Score	**Standing**	97.24%	91.54%	**97.61%**
	Starting	**72.25%**	52.43%	66.39%
	Stopping	**52.42%**	38.53%	38.13%
	Walking	**97.20%**	95.08%	96.55%

4) Joints Influence on the Intention Recognition Performance: The previous results verify that shoulder and leg motions, which are associated with the set of 11 joints, are valuable sources of information to recognise the current pedestrian action. Specifically, the maximum accuracy, 95.13%, is achieved when the observations are composed of poses and displacements. However, considering only body poses or displacements, the maximum accuracy falls to 91.28% and 94.23% respectively.

5) Features Influence on the Intention Recognition Performance: Regarding the distinction among intentions, the pedestrian displacements perform a better recognition of standing actions from the rest of intentions. However, with respect to starting and stopping actions, a higher 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. The body poses along with the displacements offer a more discriminative information in these cases.

Considering the body pose as the only feature, standing actions are repeatedly recognised as walking intentions 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 these cases.

When observations composed of body poses and displacements are analysed, the most frequent missclassifications are produced by delays or pedestrians with low-speed motions. 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 is normally recognised as walking action. On the other hand, walking intentions are recognised as starting or stopping actions when pedestrians with low-speed motions are tested. Likewise, the beginning of a starting action and the ending of a stopping motion contains body poses which are equivalent to poses labelled as standing actions. Hence, a significant number of missclassifications are also produced between these intentions.

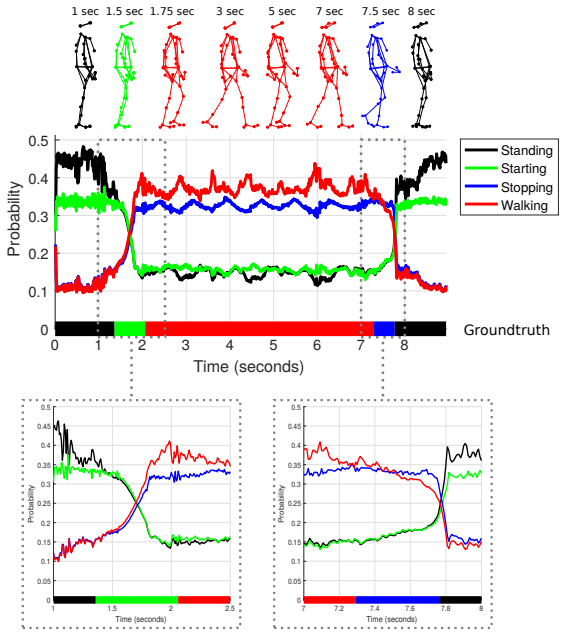


Fig. 6. Example of intention recognition probabilities using poses and displacements. Black represents standing, green starting, red walking and blue stopping. Top: pedestrian poses at significant instants of time. Middle: probabilities for each intention. Bottom: zoom in of the transitions.

A graphical example of the previous statements is shown in Figure 6 where the classification probabilities along with the groundtruth are illustrated. Several examples of pedestrian poses at different instants of time are illustrated on the top of the figures. These poses are represented in different colours according to the classification result. Black represents standing, green starting, red walking and blue stopping. In the middle, the probabilities of each intention at each instant of time are shown. Finally, at the bottom, a zoom in of the transitions are illustrated. The figure

shows that starting-walking and walking-stopping transitions usually happen in the middle of the first and last steps, thus obtaining non-critical missclassifications. Additionally, short delays appear in the standing-starting and stopping-standing transitions. 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.

6) Labelling Influence on the Intention Recognition: In Tables IV and V, the transitions from standing to starting, starting to walking, walking to stopping and stopping to standing are analysed in detail. This analysis is focused on the number of detected and non-detected transitions and its delays. The evaluation criteria fixes a range of $[-500, 500]$ ms around the event labelled by the human expert. Within this range, a multiframe validation algorithm is applied to ensure the transition detection and reduce false positive changes produced by missclassifications. The number of frames is fixed to 6 (50 ms). Thereby, the algorithm detects a transition when 6 consecutive pedestrian observations are recognised as the same intention but this is different to the action classified in $t-6$. Finally, the intention 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 was detected by the algorithm.

TABLE IV

Transition	**Detected**	**Non-Detected**	**Accuracy**
Standing - Starting	238	5	97.94%
Starting - Walking	250	12	95.42%
Walking - Stopping	61	21	74.39%
Stopping - Standing	73	7	91.25%
Overall	622	45	93.25%

TABLE V

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

The number of transitions correctly and incorrectly detected are 622 and 45 respectively, i.e. the accuracy is 93.25%. Most of the transitions which are not detected corresponds to walking-stopping changes. This could be due to the fact that the number of stopping observations in the dataset 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 ms respectively. It is worth mentioning that missclassifications produced in a transition negatively influence in the non-detection of future transitions.

Regarding the delays of the detected transitions, the results show that starting-walking transitions have negative delays since the first half of the first step contains the most perceptible information to determine starting actions. A more comprehensive assessment can be addressed comparing the results with the delays accomplished in other works. The method proposed in this document recognises starting intentions 125 ms after the gait initiation with an accuracy of 80%. These results are similar to the delays achieved in [4], [3]. Nonetheless, a multiframe validation of 50 ms is carried out 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.

Additionally, an analysis of delays from walking-stopping transitions to the standing events labelled by the human expert can be done. This analysis is important to know the delay from a stopping detection until the real standing event. Most of standing events can be predicted a few tens of ms in advance. Specifically, the method proposed in this document recognises stopping intentions 291.67 and 58.33 ms before the event with an accuracy of 50-70% respectively. These data are slightly worse than the results accomplished in [1], [2], [3] due to the non-detection of walking-stopping transitions previously discussed. However, it should be pointed out that a multiframe validation over 50 ms is carried out 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.

A. Intention Recognition using Skeleton Estimation

The intention recognition was also examined using a sequence example of noisy observations extracted by the single-frame pedestrian skeleton estimation algorithm previously described. In Figure 7, images extracted from the sequence are represented. 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.

In Figure 8, the intention recognition probabilities for the skeleton estimated are represented. The black line represents the probability of standing intention, 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 intention. At top of the figure, the pedestrian point clouds extracted by the pedestrian segmentation algorithm and the skeleton estimation at different instants of time are shown. These skeletons correspond to the scenes of the third column in Figure 7. The graph

Fig. 7. Images extracted from the sequence example.

shows that the intention has been correctly recognised in the whole sequence and the probability values for each intention are similar to the values shown in Figure 6.

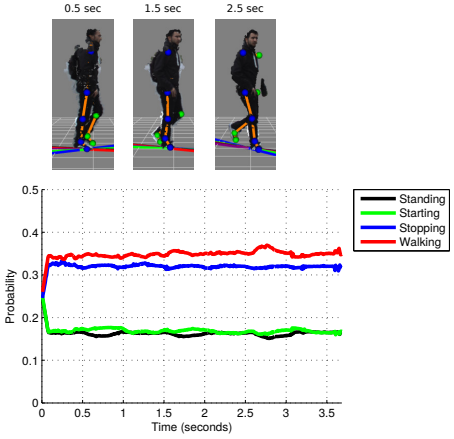


Fig. 8. Activity recognition probabilities when poses and displacements extracted from the skeleton estimation algorithm are used. Top: pedestrian poses at significant instants of time. Bottom: probabilities for each intention.

VI. Conclusions and Future Works

This paper proposes a method based on a HMM that recognises pedestrian intentions, i.e. walking, stopping, starting and standing, by means of 3D positions and displacements of 11 joints located along the pedestrian bodies. The features were extracted from a high frequency and low noise dataset published by CMU. Additionally, a single-frame pedestrian skeleton estimation algorithm is proposed in to extract the same set of joint is real-world environments. 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. For example, children or elderly people should be considered since their dynamics are not included in the CMU dataset.

The method was tested in an extensive way under ideal conditions. The high frequency of the dataset 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 method correctly recognises intentions with an accuracy of 95.13%. It recognises starting intentions 125 ms after the gait initiation with an accuracy of 80% and recognises stopping intentions 291.67 and 58.33 ms before the event with an accuracy of 50-70% respectively. The algorithm were also tested using noisy observations extracted by a single-frame pedestrian skeleton estimation algorithm. To obtain more accurate pedestrian skeletons, markerless motion capture approaches based on CNNs such as the algorithm proposed in [13] could be developed instead of algorithms based on geometrical constraints.

VII. Acknowledgements

This work was funded by Research Grants SEGVAUTO S2013/MIT-2713 (CAM), DPI2014-59276-R (Spanish Min. of Economy), BRAVE Project, H2020, Contract #723021. This project has received funding from the Electronic Component Systems for European Leadership Joint Undertaking under grant agreement No 737469. This Joint Undertaking receives support from the European Unions Horizon 2020 research and innovation programme and Germany, Austria, Spain, Italy, Latvia, Belgium,Netherlands, Sweden, Finland, Lithuania, Czech Republic, Romania, Norway.

References

[1] C. G. Keller and D. M. Gavrila, "Will the pedestrian cross? a study on pedestrian path prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 2, pp. 494–506, April 2014.

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

[3] S. Köhler, M. Goldhammer, K. Zindler, K. Doll, and K. Dietmeyer, "Stereo-vision-based pedestrian's intention detection in a moving vehicle," in *2015 IEEE 18th International Conference on ITS*, Sept 2015, pp. 2317–2322.

[4] S. Köhler, M. Goldhammer, S. Bauer, K. Doll, U. Brunsmann, and K. Dietmayer, "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.

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

[6] S. Köehler, M. Goldhammer, S. Bauer, S. Zecha, K. Doll, U. Brunsmann, and K. Dietmayer, "Stationary detection of the pedestrian?s intention at intersections," *IEEE ITSM*, vol. 5, no. 4, pp. 87–99, 2013.

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

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

[9] R. Quintero, I. Parra, D. F. Llorca, and M. A. Sotelo, "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.

[10] ——, "Pedestrian intention and pose prediction through dynamical models and behaviour classification," in *2015 IEEE 18th International Conference on ITS*, Sept 2015, pp. 83–88.

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

[12] S. Worrall, R. Quintero, Z. Xian, A. Zyner, J. Philips, J. Ward, A. Bender, and E. Nebot, "Multi-sensor detection of pedestrian position and behaviour," in *Proceedings of the 23rd ITS World Congress*, 2016.

[13] A. Elhayek, E. de Aguiar, A. Jain, J. Thompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt, "Marconi-convnet-based marker-less motion capture in outdoor and indoor scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 3, pp. 501–514, March 2017.