Pedestrian Intention and Pose Prediction through Dynamical Models and Behaviour Classification

R. Quintero, I. Parra, D. F. Llorca and M. A. Sotelo

***Abstract*— Pedestrian protection systems are being included by many automobile manufacturers in their commercial vehicles. However, improving the accuracy of these systems is imperative since the difference between an effective and a non-effective intervention can depend only on a few centimeters or on a fraction of a second. In this paper, we describe a method to carry out the prediction of pedestrian locations and pose and to classify intentions up to 1 s ahead in time applying Balanced Gaussian Process Dynamical Models (B-GPDM) and naïve-Bayes classifiers. These classifiers are combined in order to increase the action classification precision. The system provides accurate path predictions with mean errors of 24.4 cm, for walking trajectories, 26.67 cm, for stopping trajectories and 37.36 cm for starting trajectories, at a time horizon of 1 second.**

I. Introduction and Related Works

The effective interaction with other traffic participants is an open challenge for automated vehicles. This is particularly true for urban environments that are not primarily dedicated to traffic and are populated with vulnerable road users like pedestrians and bicyclists. In order to cope with the wide variations in traffic situations and behaviour of traffic participants scientific progress is required in perception, prediction and interaction techniques.

In the context of pedestrian protection, Toyota recently developed the Pre-Collision System with Pedestrian-avoidance Steer Assist that warns the driver when a pedestrian is in front of the vehicle and, if the driver does not take action to avoid the collision, an automatic emergency braking in addition to automatic steering is activated. Improving the accuracy of these systems is imperative since the lateral component of the pedestrian localization could be particularly relevant. Thereby a precise assessment about the current and future pedestrian locations is required. A difference of only 30 cm in the estimated lateral position can make the difference for a successful collision avoidance maneuver [1]. Moreover, accident analysis in [2] demostrated that initiating an emergency braking 0.16 s in advance reduces the severity of accident injuries up to 50% given an initial vehicle speed of 50 km/h. As a consequence, over the last few years, a lot of effort has been put into understanding the pedestrian intentions and predicting their trajectories.

Early approaches to perform path prediction and tracking used Kalman Filters in a trajectory-based framework [3] for walking motions, applying the current pedestrian position and velocity to estimate the next location. Nonetheless, the sole consideration of using trajectory, assuming entirely walking intentions, is clearly insufficient to predict the pedestrian path due to the highly dynamic behavior of humans, since changes in their walking direction or intentions can happen in an instant. For this reason, some intentions such as start walking could be hard to predict in advance, even for a human expert.

R. Quintero, I. Parra, D. F. Llorca and M.A Sotelo are with the Computer Engineering Department, University of Alcalá, Alcalá de Henares, Spain {raul,ignacio,llorca,sotelo}@aut.uah.es

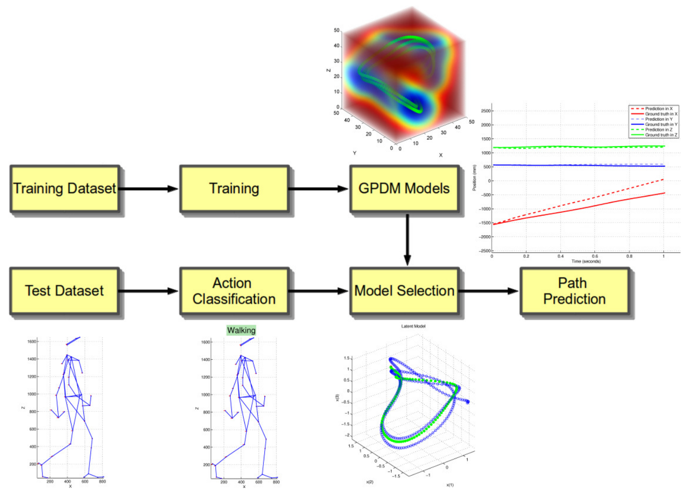

Fig. 1. Pedestrian intention and pose prediction algorithm.

Moreover, other approaches find similarities between observed and learned pedestrian trajectories in order to predict future states. These trajectories can be composed of varied features such as positional information, motion vectors or dense optical flow. In [4] a trajectory matching algorithm is applied to measure the similarity between trajectories in order to classify walking and stopping actions and predict pedestrian paths at short intervals, combining positional and optical flow features.

More advanced methods are based on human motion features or body language of different actions using a low-dimensional nonlinear manifold that reduces the dimensionality of the input data, considering its dependence over time, in the so-called latent space. In [1] two Gaussian process dynamical models (GPDM) [5] are separately trained using augmented features derived from dense optical flow of different sequences of stopping and walking pedestrian motions. A particle filter allows to combine both models with the purpose of computing the probability of the pedestrian state. Their proposed method can achieve more accurate path prediction than basic approaches mostly for stopping actions. However, learning long sequences with different actions could result in degenerated GPDMs. To avoid this problem the perspectives of trajectory-based and GPDM-based approaches can be mixed. In [6] the action is classified comparing observed sequences with GPDM-trained sequences. In [7] a large dataset of typical human behaviours is learned and the most similar trained sequence on the dataset to the observed sequence is selected in order to predict the pedestrian path.

In the latest years, context-based pedestrian behaviour prediction systems have been developed in a succesful way. They analyze the current situation infering what the pedestrian will do in advance. These approaches have longer prediction horizons than the above mentioned methods, especially for walking motions, although they can not deal with starting or stopping actions correctly because the information about these actions is extracted better from the pedestrian pose, not from the context. In [8] a generic context-based model to predict crossing behaviours of pedestrians in inner-city and an additional model to the context of zebra crossings are proposed. Both models are learned computing features such as the lateral distance between the pedestrian and the collision point, the time for the pedestrian to reach the collision point, the distance to curbstone, etc. Finally those models are hierarchically combined applying a “Context Model Tree” framework.

This paper describes a method for predicting the pedestrian locations and pose and classifying intentions up to 1 s ahead in time applying a novel approach for pedestrian path and pose prediction for walking, starting, stopping and standing behaviours based on Balanced Gaussian Process Dynamical Models (B-GPDM) and naïve-Bayes classifiers. This approach is described in our previous works [7], [9]. In [9] a classifier based on the similarity between consecutive pedestrian poses and the sum of absolute joint velocities was developed. The drawback of this classifier is that a history of the previous features have to be taken into account for distinguishing between starting and stopping behaviours since those features are noisy and the poses in these actions are similar each other. However, in this paper, we propose two new single-frame action classifiers, the first one is based on joint positions in lateral direction and the second one is based on their displacement in the same direction. The lateral direction is selected due to all sequences simulate a pedestrian crossing in front of a vehicle so that longitudinal direction and height are not discriminative among actions. A prior, computed from a transition matrix, allows us to solve the drawback of the previous classifier. Finally, the overall action probability is chosen depending on the confidence of each classifier in each instant.

The paper is organized as follows: Section II describes the goal of our method and the data-sets used for learning and testing. In section II-A we briefly resume how GPDM works with the purpose of making easier the understanding of the next sections. The sections II-B describes the new naïve-Bayes classifiers that perform the action classification. Experimental results from long sequences where pedestrians do different actions are presented in section III. Finally, we discuss our conclusions and future works in section IV.

II. System Description

Our future goal is to develop a pedestrian path and pose prediction system set up in a moving vehicle equipped with stereo cameras and LIDAR. In this paper, we will test the feasibility and limits of our method in an extensive way under ideal conditions by using the high frequency and low noise data-set from CMU [10]. The CMU data-set contains different pedestrian sequences captured from a Vicon motion capture system, consisting of 12 infrared MX-40 cameras. Motions are captured in a working volume of approximately 3 m x 8 m. Each pedestrian pose is composed of the 3D coordinates of 41 joints along the body (see Fig. 2). The accuracy of pedestrian path and pose prediction and action classification algorithms will be tested with 129 sequences in which different subjects are simulating pedestrian behaviours. The processing time of each step will be analyzed as well. All results have been obtained in MATLAB 2009 64-bits with a processor Intel i7-2600K 3.40GHz.

As we mentioned above, we learn high frequency and low noise sequences to get high quality individual models, reducing the dimensionality of a feature vector using the B-GPDM algorithm to construct a latent space. Our feature vector is composed of the 3D positions and displacement of the pedestrian joints, removing the 3D body translation parameters. The displacements are included in the model because it was observed to increase the accuracy in the prediction of the pedestrian path. The high frequency will help the B-GPDM to properly learn the dynamics of the different actions and will increase the probability of finding a similar test pose in the trained data without missing intermediate poses. In addition, these low noise models will improve the prediction when working with noisy test samples.

In the learning step, the pedestrian motions from the CMU data-set are hierarchically divided into eight sub-sets. The first division is based on the direction, left-to-right and right-to-left. The second one is based on the action (standing, starting, stopping and walking). To capture the dynamics of the different actions, the beginning and end of the sequences were cropped manually trying that all the poses in a sequence were representative of their action.

On the other hand, in the prediction step, the original sequences are used since variations in the pedestrian behaviours were captured. Table I shows the overall number of poses for the learning models. The data-set is composed of 187 sequences (29 of standing actions, 45 of starting actions, 16 of stopping actions and 97 of walking actions) from 26 different subjects divided according to the action and direction.

TABLE I
NUMBER OF PEDESTRIAN POSES IN LEARNING STEP.

	Standing	Starting	Stopping	Walking
Left-to-Right	16963	1752	1181	25397
Right-to-Left	2512	1877	1147	11056
Total	19475	3629	2328	36453

A. GPDM

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 the dynamics mapping are computed separately in a non-linear form, marginalizing out both mappings and optimizing the latent variables and the hyper-parameters of the kernels. The conditional probability of Y given X, θ and W for the observation mapping is defined in (1)

$$p(Y|X,\theta,W) = \frac{|W|^N}{\sqrt{(2\pi)^{ND}|K_Y|^D}} exp(-\frac{1}{2}tr(K_Y^{-1}YW^2Y^T)) \tag{1}$$

where Y is the centred observed data-set, X represents the latent positions on the model, K_Y is the kernel matrix, $\theta = [\theta_1, \theta_2, ..., \theta_N]$ contains the kernel hyper-parameters, N is the number of samples, D is the dimension of the data-set, and W is the scaling matrix (to account for different variances in the different data dimension). The elements of kernel matrix for the observation mapping are computed using (2).

$$k(x_i,x_j) = \theta_1 exp(\frac{-\theta_2}{2}(x_i-x_j)^T(x_i-x_j)) + \theta_3\delta_{i,j} \tag{2}$$

where $\delta_{i,j}$ is the Kronecher delta function.

The dynamic mapping from the latent coordinates is defined in (3),

$$p(X|\beta) = \frac{p(x_1)}{\sqrt{(2\pi)^{(N-1)d}|K_X|^d}} exp(-\frac{1}{2}tr(K_X^{-1}X_{out}X_{out}^T)) \tag{3}$$

where $X_{out} = [x_2, ..., x_N]^T$, d is the model dimension, and K_X is the kernel matrix constructed from $\{x_1, ..., x_{N-1}\}$ using the kernel function provided in (4)

$$k(x_i,x_j) = \beta_1 exp(\frac{-\beta_2}{2}(x_i-x_j)^T(x_i-x_j)) + \beta_3 x_i^T x_j + \beta_4\delta_{i,j} \tag{4}$$

where β_1 to β_4 are the kernel hyper-parameters.

The goal is to minimize the negative log-likelihood function $-ln\, p(X,\theta,\beta,W|Y)$ that is given in (5)

$$\mathscr{L} = \mathscr{L}_Y + \mathscr{L}_X + \sum_j ln\,\theta_j + \frac{1}{2\kappa^2}tr(W^2) + \sum_j ln\,\beta_j \tag{5}$$

where

$$\mathscr{L}_Y = \frac{D}{2}ln|K_Y| + \frac{1}{2}tr(K_Y^{-1}YW^2Y^T) - N\,ln|W| \tag{6}$$

$$\mathscr{L}_X = \frac{d}{2}ln|K_X| + \frac{1}{2}tr(K_X^{-1}X_{out}X_{out}^T) + \frac{1}{2}x_1^T x_1 \tag{7}$$

In order to increase the smoothness of the learned trajectories in the latent space, a modified version of GPDM can be used by changing the weight of $\mathscr{L}_X$ by means of a λ element. A value for λ of $\frac{D}{d}$ is recommended in [11]. This modification is known as Balanced GPDM (B-GPDM).

Given a latent position the original feature vector can be recovered as described in (8).

$$\mu = Y^T K_Y^{-1} k_Y(x) \tag{8}$$

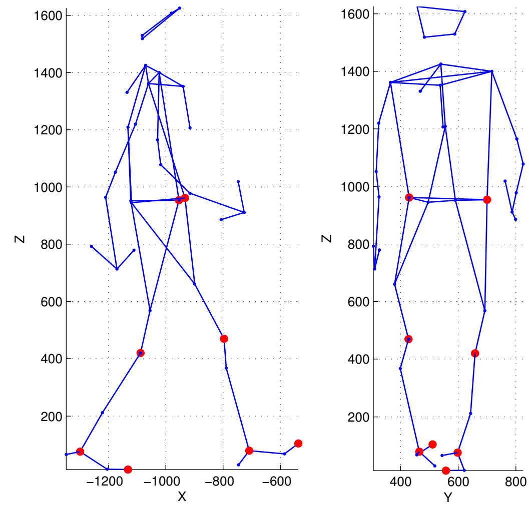

Fig. 2. Pedestrian joints. Red markers stand for the action classification joints.

where Y is the centred data-set, K_Y^{-1} the inverse matrix of the kernel for the observation mapping (see 2) and $k_Y(x)$ is a column vector with elements $k_Y(x,x_j)$ for all other latent position x_j in the model.

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 can be obtained as described in (9)

$$\mu_X(x) = X_{out}^T K_X^{-1} k_X(x) \tag{9}$$

where $X_{out} = [x_2, ..., x_N]^T$, K_X is the kernel matrix constructed from $\{x_1, ..., x_{N-1}\}$ using the kernel function provided in (4) and $k_X(x)$ is a column vector with elements $k_X(x,x_j)$ for all other latent position x_j in the model. A prediction at a time horizon of N latent positions ahead can be obtained computing (9) iteratively.

B. Action classification

In this paper, we propose two new single frame classifiers to estimate the pedestrian action separately. The first one is based on 3D joint positions in lateral direction and the second one is based on their displacements in the same direction. At this point we should wonder what joints are more relevant to the action classification algorithm. Some experiments demostrated that a few joints in the legs are sufficent, so that, the feature vector for this purpose is composed of 8 points: hips, knees, anckles and tiptoes. In Fig. 2 it is shown all pedestrian joints (blue markers) and the selected joints for action classification (red markers). Other joints in the legs are correlated with the outlined before, consequently their information is redundant. On the other hand, adding joints from the upper body in the feature vector could increase the error rate of the classifiers since a pedestrian could move the arms in a similar way when it is walking and standing.

Both classifiers only need to considerer four actions (walking, stopping, starting and standing). However, during the learning step, the data-set was hierarchically divided into eight sub-sets. The first division was based on the direction, left-to-right and right-to-left, and the second one was based on the action. Hence, a mirror rotation is applied to all right-to-left sequences in order to get only pedestrians moving from left to right and reduce the number of classes to classify from eight to four.

The classifiers are trained getting the mean and the variance from the feature vectors of each considered action. Given a new feature vector, the posterior probability for each class is computed as:

$$P(C|X) = \prod_{j=1}^{n} P(X_j|C)P(C) \tag{10}$$

where X means the feature vector, C is the class and n is the feature vector length. For each classifier, a Maximum A Posteriori (MAP) estimation is computed to obtain the pedestrian action. The initial prior $P(C)$ is defined in such a way that all actions probabilities are identical.

The overall action probability is chosen depending on the confidence of each classifier. If the displacement-based classifier obtains a high confidence on walking or standing action then the overall action probability corresponds to the computed with this classifier, otherwise the overall probability is the result from the position-based classifier.

At later instances, a transition matrix M, given the overall action probability $P(C|X)$, allows us to compute the prior $P(C)$ as:

$$P(C) = P(C|X)M \tag{11}$$

This transition matrix takes into account how a pedestrian can change its intentions, i.e, if a pedestrian is standing, it will only change to a starting behaviour. Therefore, the transitions between actions is a Finite Markov chain with stationary transition probabilities given an initial vector of probabilities.

Once we have estimated the pedestrian action we focus on selecting the appropiate model. To select it a search of the most similar 3D pose (joint positions and displacements) in the corresponding action training sub-set is computed, this pose and its latent position is used as starting point for a more accurate search in the latent space applying a gradient descent algorithm. Once the latent position has been estimated, a prediction at a time horizon of N poses ahead can be done using (8) and (9) iteratively.

III. Experimental results

The described method was tested using the CMU dataset with 129 sequences (63508 poses) from 24 subjects adopting a one vs. all strategy. This means that all the models generated by one test subject were removed from the training data while performing tests on this subject. This strategy was chosen due to the number of subjects is not enough to divide them into two sets, one for training and other for testing.

A. *Results on action classification*

To test the performance of the proposed action classification algorithm all pedestrian poses were manually labelled on the sequences by a human expert. The adopted criteria of labelling for a starting action is defined as the movement that begins when the pedestrian moves one knee and ends when its knee and anckle are aligned in the lateral axis. In addition, a stopping action is defined as the movement that begins in the middle of the last step and finishes when the foot treads the ground. Table II summarizes the classification results on a confusion matrix for each classifier. The joint-based classifier and the displacement-based classifier have a precision of 78.89% and 72.90% respectively. The overall precision is 85.90% for the four different actions. Missclassifications such as standing movements as walking actions and viceversa or starting movements as stopping behaviours and viceversa (4.16%) are produced by classification errors at the beginning of the sequences. Other missclassifications are produced by delays. However these last missclassificatios are not critical from the point of view of the path estimation as both actions have similar dynamics and the path predictions will be also very similar.

TABLE II
CONFUSION MATRICES FOR ACTION CLASSIFICATION ALGORITHM

(a) Joint-based classifier

		Classification			
		Standing	**Starting**	**Stopping**	**Walking**
Actual	**Standing**	21556	2594	384	1320
	Starting	705	677	353	977
	Stopping	0	49	217	181
	Walking	531	1816	4495	27653

(b) Displacement-based classifier

		Classification			
		Standing	**Starting**	**Stopping**	**Walking**
Actual	**Standing**	21992	2257	1380	295
	Starting	188	1231	851	442
	Stopping	0	125	142	180
	Walking	908	836	9814	22937

(c) Overall classification

		Classification			
		Standing	**Starting**	**Stopping**	**Walking**
Actual	**Standing**	23596	591	244	1423
	Starting	739	633	316	1024
	Stopping	0	49	186	212
	Walking	935	1395	2092	30073

Figures 3 and 4 show the action probabilities for a stopping and starting sequence respectively. In the top of each figure, the probabilities from displacement-based classifier is represented. During the walking actions, some peaks of stopping probabilities appears due to the pedestrian legs are opened and the displacement in that instant is lower than when the legs are closed. In the middle of each figure, the probabilities from joint-based classifier are shown. In this case, each peak of stopping probabilities corresponds with closed legs. Finally, in the bottom, the overall probabilities are represented. This combination of classifiers allows solving the peaks of stopping probabilities and missclassifications and avoiding continuous changes in the transitions between actions, specially from walking and starting to walking and

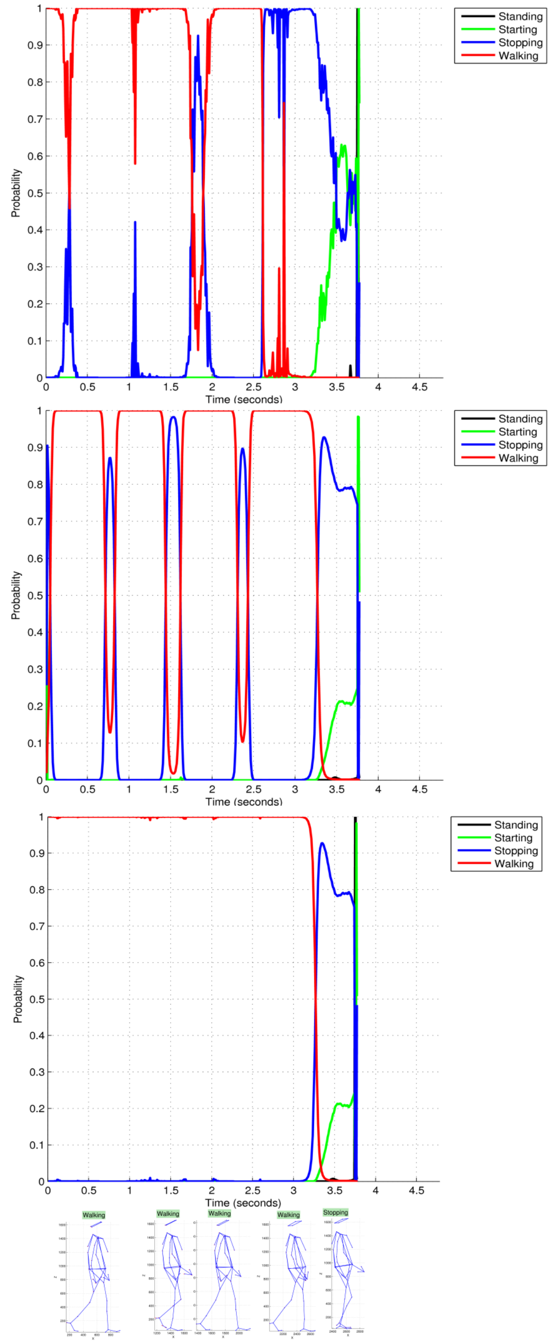

Fig. 3. Action classification probabilities for a stopping sequence. Top: Displacement-based classifier. Middle: Joint-based classifier. Bottom: Overall classification.

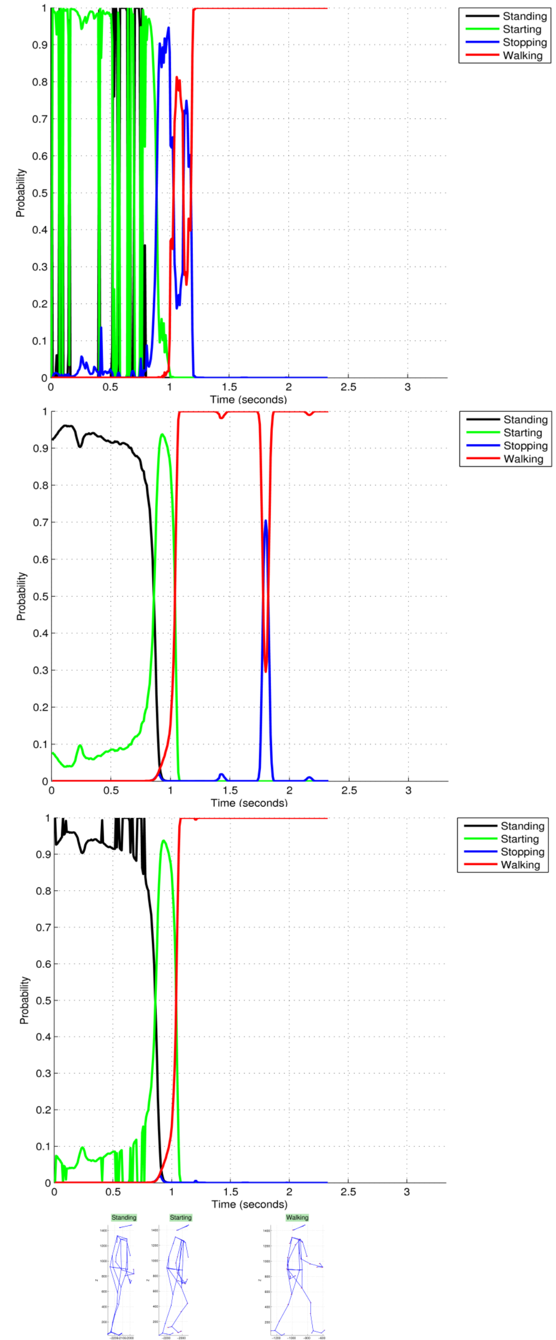

Fig. 4. Action classification probabilities for a starting sequence. Top: Displacement-based classifier. Middle: Joint-based classifier. Bottom: Overall classification.

stopping respectively.

B. Results on pedestrian path prediction

As explained before, once the pedestrian action is estimated, the model is first selected from each one of the action data-sets and then a path prediction estimation is performed using the selected model. Accordingly, a good path prediction strongly depends on a good classification. Table III shows the mean combined longitudinal and lateral path prediction error and standard deviation (cm) for different prediction horizons. As can be observed, prediction accuracy at 1 second is higher for walking sequences (24.4 cm) than for stopping (26.67 cm) or starting (37.36 cm). Compared to our previous results in [9] mean errors for walking, stopping and starting are much more similar to each other, probably due to the fact that one second is a too long time horizon for our action classifier to anticipate stopping actions from walking poses. Although a much more detailed analysis of the classifier is required we estimate that, in average, we are detecting stopping actions 0.5 seconds in advance, and this delay in the detection is introducing prediction errors that close the gap with the "change" actions (stopping and starting). This indicates that the predictive power of the B-GPDM if far larger than that of our action classifiers that are limiting our prediction time horizon.

TABLE III
MEAN COMBINED LONGITUDINAL AND LATERAL PREDICTION ERROR±STD (CM) FOR DIFFERENT PREDICTION HORIZONS (SECONDS)

	0 sec.	0.25 sec.	0.5 sec.	0.75 sec.	1 sec.
Walking	2.16 ±2.78	6.95 ±7.72	12.70 ±13.83	18.52 ±19.99	24.40 ±26.31
Stopping	2.99 ±3.06	6.34 ±5.78	12.40 ±9.67	19.85 ±14.67	26.67 ±19.61
Starting	3.25 ±3.32	7.60 ±5.56	17.75 ±10.43	27.67 ±14.92	37.36 ±21.38

C. *Processing time*

Table IV resumes the processing time of each step. All the results have been obtained using MATLAB 2009 64-bits with a processor Intel i7-2600K 3.40GHz. As can be seen, B-GPDM is the limiting section for a real time implementation because the most expensive operation is the inversion of kernel matrices, especially when the number of training data is large. However, we believe there is great margin for improvement with a GPU implementation of the Matlab code.

TABLE IV
PROCESSING TIMES

	Milliseconds
Action Classification	0.23
Model Selection	42.70
Latent Position Search	3672.54
Path and Pose Prediction	1252.27

IV. CONCLUSIONS AND FUTURE WORKS

We have developed a system for accurate pedestrian path and pose prediction by means of action classification in a limited time horizon up to 1 second. For such purpose, we propose two naïve-Bayes classifiers based on 3D joint positions and joint displacement respectively. This approach allows us to reduce the missclassifications and avoid continuous changes in the transitions between actions, specially from walking and starting to walking and stopping respectively. Once the action has been classified, the most similar pose is found on the 3D space in the sub-set of that action and the latent position on the corresponding B-GPDM model is estimated. Finally, a prediction at a time horizon of 1 second ahead is done. The system provides accurate path predictions with mean errors of 24.4 cm, for walking trajectories, 26.67 cm, for stopping trajectories and 37.36 cm for starting trajectories, at a time horizon of 1 second. These results were obtained using dynamical models created with the high accuracy and high frequency (120 Hz) CMU data-set [10] in which 41 joints are on the pedestrian body. According to our previous results, we believe accuracy can be increased at 1 second time horizons with better performance of the action classifiers. In this line, we plan to introduce contextual information to support the pose information of our classifiers. Our final goal is to develop a pedestrian path and pose prediction system set up in a moving vehicle equipped with stereo cameras and LIDAR. The work presented in this paper can be considered as the best case scenario and further experimentation will be carried out to test how this approach performs with noisy test sequences.

As future work we propose to create a bigger data-set in order to include a signicative number of sequences for the different actions that will help to train definite classifiers. We propose to include sequences where pedestrians are making a turn or even sequences with children. In addition, experiments with pedestrian joint extraction systems in real conditions will be performed to test the real predictive power of the system with noisy samples.

V. ACKNOWLEDGEMENTS

This work was supported by the Research Grants DIS-ADAPT SPIP2014-1300 (General Traffic Division of Spain), IMPROVE DPI2014-59276-R (Spanish Ministry of Economy), and SEGVAUTO-TRIES-CM S2013/MIT-2713 (Community of Madrid).

REFERENCES

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

[2] M. M. Meinecke, M. Obojski, D. M. Gavrila, E. Marc, R. Morris, M. Tons, and L. Lettelier, "Strategies in terms of vulnerable road user protection," EU Project. SAVE-U," Deliverable D6, 2003.

[3] D. F. Llorca, M. A. Sotelo, A. M. Hellín, A. Orellana, M. Gavilán, I. G. Daza, and A. G. Lorente, "Stereo region-of-interest selection for pedestrian protection: A survey," *Transportation Research Part C*, vol. 25, pp. 226–237, 2012.

[4] C. G. Keller, C. Hermes, and D. M. Gavrila, "Will the pedestrian cross? probabilistic path prediction based on learned motion features," in *Pattern Recognition*. Springer, 2011, pp. 386–395.

[5] J. Wang, D. Fleet, and A. Hertzmann, "Gaussian process dynamical models for human motion," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 2, pp. 283–298, Feb 2008.

[6] L. Raskin, M. Rudzsky, and E. Rivlin, "Dimensionality reduction using a gaussian process annealed particle filter for tracking and classification of articulated body motions," *Computer Vision and Image Understanding*, vol. 115, no. 4, pp. 503–519, 2011.

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

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

[9] R. Quintero, J. Almeida, D. F. Llorca, and M. A. Sotelo, "Pedestrian path prediction using body language traits," in *IEEE Intelligent Vehicles Symposium (IV)*, 2014.

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

[11] R. Urtasun, D. J. Fleet, , and P. Fua, "3d people tracking with gaussian process dynamical models," in *IEEE International Conference on Computer vision and pattern recognition (CVPR)*, 2006, pp. 238–245.