Pedestrian Path Prediction using Body Language Traits

R. Quintero[1] J. Almeida[2] D. F. Llorca[1] M. A. Sotelo[1]

[1] Computer Engineering Department. University of Alcalá, Alcalá de Henares, Spain. {raul.quintero,llorca,sotelo}@aut.uah.es.

[2] Department of Mechanical Engineering. University of Aveiro, Aveiro, Portugal. {almeida.j}@ua.pt.

***Abstract*— Driver Assistance Systems have achieved a high level of maturity in the latest years. As an example of that, sophisticated pedestrian protection systems are already available in a number of commercial vehicles from several OEMs. However, accurate pedestrian path prediction is needed in order to go a step further in terms of safety and reliability, since it can make the difference between effective and non-effective intervention. In this paper, we consider the three-dimensional pedestrian body language in order to perform path prediction in a probabilistic framework. For this purpose, the different body parts and joints are detected using stereo vision. We propose the use of GPDM (Gaussian Process Dynamical Models) for reducing the high dimensionality of the input feature vector (composed by joints and displacement vectors) in the 3D pose space and for learning the pedestrian dynamics in a latent space. Experimental results show that accurate path prediction can be achieved at a time horizon of ≈ 0.8 s.**

***Index Terms*— Pedestrian Path Prediction, Prediction of Intentions, Pedestrian Protection Systems, ADAS, Vision.**

I. INTRODUCTION AND RELATED WORK

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, such as robotics, surveillance, human-machine interaction, and intelligent vehicles. As a clear example of that, predicting the path of a pedestrian by means of action classification can lead to further improvement in state-of-the-art driver assistance systems. As a matter of fact, an improvement of 30-50 cm in pedestrians path prediction accuracy may well signify the difference between effective and non-effective intervention in emergency braking systems. Early detection of pedestrians entering a road lane is a current challenge in order to increase traffic safety. Thus, accident analysis [1][2] shows that the ability to initiate emergency braking 0.16 s in advance (with respect to typical human reaction time of 0.5 s) has the potential to reduce the severity of accident injuries up to 50%. Similarly, early recognition of pedestrians stopping actions can lead to much more accurate active interventions in last second automatic maneuvers. As a consequence of that, strong gains are expected to be made in the performance and reliability of pedestrian protection systems.

A large number of works on pedestrian recognition has been published in the past. Remarkable surveys on vision-based state-of-the-art pedestrian detection systems can be found on [3][4][5][6]. All these systems aim at the detection and tracking of the bounding box where the pedestrian body is located in the image plane. In [7] [8], the pedestrian bounding box and the different body parts are detected using discriminatively trained deformable parts-based models, providing an interesting framework for the analysis of pedestrians' gait dynamics. However, while being a remarkable milestone in the domain of pedestrian recognition, the low accuracy and repetitiveness exhibited in the location of the body parts in a sequence of images does not allow to robustly use this method in the tracking of pedestrians joints or limbs. Early approaches for pedestrian tracking use Kalman Filters in a trajectory-based framework, including interacting multiple model filters [9][10] in order to account for different motion dynamics. Nonetheless, the sole consideration of trajectory is clearly insufficient for predicting the pedestrian path in an accurate manner in situations with changing motion dynamics. Thus, empirical studies [11] have demonstrated that when only the trajectory of the pedestrian is available, a higher error rate is produced in drivers judgment regarding the pedestrians intentions.

In contrast to trajectory-based approaches, the consideration of the whole pedestrian body language has the potential to provide early indicators of the pedestrian intentions, much more powerful than those provided by the physical parameters of a trajectory. In this line, vision-based gait recognition has been undertaken in the literature by using motion history images and frame difference [12] in an attempt to analyze human motion. Nonetheless, there is a yawning gap between gait recognition and pedestrians' intention or action classification. In [13], early indicators of the pedestrian's intention to cross the street are divided into those presumably followed by crossing, e.g. turning the head or catching the vehicle driver's eye, and those definitely followed by entering the lane, e.g. bending the upper body. They propose an IMM-EKF algorithm for tracking and for detecting pedestrians' intentions to enter the lane in an intersection monitoring application using a stationary monocular camera. In addition, a HOG-like monocular-video-based descriptor is proposed in combination with SVM classification (MCHOG) in order to speed up the decision on the start of walking (stopping actions and bending in behaviors are not considered). The use of difference of images in their work does not make it directly applicable to moving vehicles.

In [14], a probabilistic approach is proposed for pedestrian action classification (walking vs. stopping; starting-to-walk is not considered) and accurate path prediction from a moving vehicle, at short intervals. They improve traditional trajectory matching approaches by augmenting the underlying features

to include motion cues. Dimensionality reduction of the feature vector is carried out by applying PCA (Principal Component Analysis) on the histograms of Orientation Motion (HoM) features. More recently, in [2] two novel approaches are proposed based on GPDM (Gaussian Process Dynamical Models) and probabilistic hierarchical trajectory matching. Dimensionality reduction of the augmented motion features derived from dense optical flow is carried out using GPDM, as originally proposed in [15][16]. The baseline for comparison are a Kalman filter and its extension to interacting multiple model. While similar performance is attained by the four approaches on walking motion, with near-linear dynamics, during stopping motion the two newly proposed approaches achieve a more accurate position prediction of 10-50 cm at a time horizon of 0-0.77 s.

In this paper, we propose a novel approach for improving the accuracy of pedestrian path prediction in walking and stopping behaviors based on the pedestrian body language. The system description is provided in Section II. The pedestrian body parts are detected using stereo-vision, as described in section II-A, and coded in a low-dimensional embedding as illustrated in section II-B. Predictions are issued on the grounds of statistical formulation, as detailed in section II-C. Experimental results are presented in section III. Finally, we discuss our conclusions and future work in section IV.

II. System Description

We propose to use pedestrians' 3D joints and its displacement vectors as the main clue for predicting the pedestrian path in the short term, given that the pedestrian body pose encodes relevant information regarding the most likely movements in a short time horizon. For this purpose, the pedestrian body parts or joints must be detected in 3D. Point clouds provided by stereo-vision sensors are used as input for that. The proposed system builds on an already existing pedestrian detection function, capable of extracting the pedestrian 3D points from a point cloud. In a second step, the pedestrian 3D joints and displacement vectors are transformed to a low dimensional latent embedding using probabilistic dimensionality reduction techniques. Pedestrian tracking and path prediction is then carried out in the latent space using off-line learned pedestrian behaviors. Finally, the predicted pedestrian 3D pose and global position are recovered from the latent space based on the reverse mean of the trained data. A detailed description of the different parts of the proposed algorithm is provided in the next sections.

A. Stereo Pedestrian Pose Estimation

1) Preprocessing: In the preprocessing step, a point cloud is obtained from the stereo images pair with a subsequent pedestrian extraction step. In our experiments we use the KITTI data-set [17]. This data-set provides left and right images from a stereo setup along with the calibration parameters. Fig. 1 depicts a sequence image example from KITTI in which a pedestrian is walking towards the car's reachable area. A disparity image is calculated using Semi Global Block Matching (SGBM) algorithm modified from

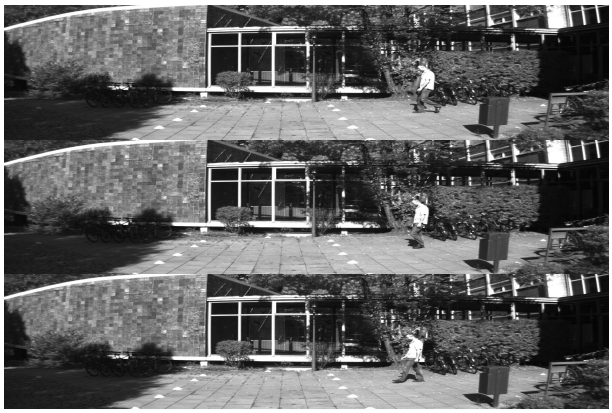


Fig. 1. Sequence example from KITTI in which a pedestrian walks towards the curbstone (artificially delimited by white spots on the pavement).

[18]. The disparity image is used to compute a 3D point cloud. In a first step, a background mask is created using a background subtraction algorithm in order to extract the pedestrian body. In a second step, the ground plane is also detected in the point cloud. These two steps allow to remove most of the points that do not belong to the pedestrian. Euclidean point clustering is applied to the resulting cloud and the largest cluster is assumed to belong to the pedestrian. This pedestrian extraction scheme works well in the KITTI data-set used with standing vehicles, but in a more complex scenario any pedestrian detection algorithm could be used to extract the pedestrian point cloud [6].

2) Pose estimation: The pose estimation algorithm here proposed assumes that a input point cloud is comprised only of points belonging to a single pedestrian that has been previously extracted from the general point cloud provided by the stereo images pair. It is also assumed that the pedestrian is in an upright pose, a common assumption in the pedestrian detection context.

Let $\mathscr{P} = \{p_1, \ldots, p_N\}$ represent the pedestrian point cloud with N points. The overall bounding box of $\mathscr{P}$ provides a rough approximation to the pedestrian height. The height approximation together with the typical human body proportions allows to estimate the size of the body parts. This point cloud is sliced horizontally into overlapping segments corresponding to: the head, shoulders, center torso, lower torso, upper legs and lower legs, respectively. These individual point clouds allow the algorithm to search for each body part in a small subset of $\mathscr{P}$ making the search simpler and faster. The pose estimation algorithm starts by the definition of the head center position as the geometric centroid of all the points in the top sliced point cloud. The head position will be the start for the rest of the body parts. The neck is extracted from the head position. All other body parts are subsequently extracted after that.

The neck position is obtained using a sampling and scoring method reminiscent of the Monte Carlo techniques. A line segment is defined starting at the head position with a predefined length and orientation. This initial line defines the preferential orientation of the neck. With the same starting

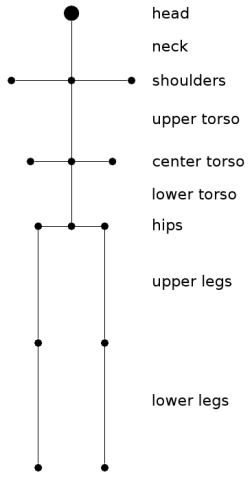


Fig. 2. Body parts detected with the pose estimation algorithm.

point a set of new lines is created. Let us denote these lines as *samples* that correspond to different possible neck positions. All samples are created by reorientation of the preferential sample in two perpendicular directions, with the first one corresponding to the pedestrian main orientation. The samples are uniformly distributed within a boundary that limits the neck movement to account for the human neck relative position limits. In this specific case, the boundary has the shape of a ellipse. Different boundaries are used for different body parts. After creation, the samples are ranked. Let $\mathscr{S} = [S_1, S_2, \ldots, S_K]$ denote all samples. Each sample score is calculated as the sum of a scoring function $f(d(),\lambda,k)$ for all points D, in the specific search point cloud, as expressed in Eq. 1.

$$X_k = \sum_D f\left(d\left(p_d, S_k\right), \lambda, k\right) \tag{1}$$

Function $d(p,S)$ denotes the euclidean distance function from a 3D point p to a line segment, sample S. The scoring function $f()$ provides the individual score for each point based on the euclidean distance of the point to the sample and two parameters. The function is defined as the pdf of the Weibull distribution:

$$f(x,\lambda,k) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} \tag{2}$$

This function provides a degree of control over the location of the maximum score. For instance, the maximum score for each point maybe obtained at a specific distance from the line segment. This allows for the best scoring sample to be placed at a specific distance from the point cloud. This method is used because of the cylindrical nature of body parts. Selecting the sample that best fits the points based on the distance alone would not take this nature into consideration and would yield erroneous results. After all samples are scored the highest scoring sample is selected as the best description of the body part. The next body part will be connected to the terminal position of the previous one and the same method is iteratively applied. Fig. 2 presents all the body parts that are detected using this method. The arms are not extracted because the stereo algorithm does not provide enough resolution for their reliable detection.

The sample and score method is applied to all body parts but with variations. The shoulders preferential position is represented by an ellipse centered at the neck bottom and stretched to best fit the typical human proportions. In this case, samples are created by rotating the initial ellipse using only a vertical axis centered at the ellipse center and the distance function is replaced by a 3D point representing the ellipse distance function. Because of the ellipse representation the shoulders are coupled and cannot move independently. A similar variation is also employed at the hips.

The body orientation can be extracted from the shoulders orientation but with ambiguity. Using just the shoulders, the correct direction cannot be obtained given that both back and front-looking parts provide the same general orientation. To solve this problem motion history is used. By using the consecutive positions of the head, both the direction and velocity of the motion can be extracted. The final pedestrian orientation is obtained from both the shoulders and motion direction. In both the upper and lower legs, the boundary conditions on the reorientation of the preferential sample are modified to best fit the leg motion limitations, for instance: the lower leg may not curve upward at the front.

Due to the fact that both left and right legs use the same slice of the original point cloud, they may converge on the same position, even with different starting positions. This problem is particularly evident when one leg is partially or totally occluded by the other. In order to avoid this problem a sequential search in performed. First, both upper legs samples are scored with the original point cloud. The best overall sample of either the left or right upper legs is selected. All the points that are within a specific range of the selected sample are removed from the point cloud. The opposing upper leg is re-scored on the remaining point cloud. The same procedure is performed in the lower legs.

3) Results on pose estimation: Fig. 3 presents two different pedestrian poses. On the left, the original pedestrian point cloud as extracted by the pre-processing stage is presented. In the middle, the extracted pose with all the created samples and the key positions of the pose is depicted. On the right, the original point cloud is colored based on the corresponding body part; points are classified based on their distance to each body part.

The poses are well estimated especially taking into account the noisy nature of the stereo point cloud. The example poses were obtained at a range of ≈ 14 meters. The recursive nature of the algorithm limits the accuracy of a body part on the accuracy of the previous part. If a part is incorrectly detected all following parts will be affected. The sampling scheme allows to explore the motion space while imposing anthropomorphic limits on the movement of the joints, where a minimization scheme could become stuck in a local minimum. The partial self occlusion of the torso does not affect

Fig. 3. Two different extracted poses. On the left, the segmented pedestrian point cloud. In the middle, the pose extracted with all the samples used. Finally, on the right, the original point cloud colored based on the corresponding body part.

the pose estimation, the head position along with the visible torso side are typically enough for a correct pose estimation. In the case of a severe occlusion of one of the legs, the occluded leg pose cannot be correctly extracted. This case can be detected by identifying an abnormally low maximum score of the occluded leg. The presented method is able to extract correctly the typical pose of a pedestrian walking in any direction. The method does not require multiple initial models or poses and the extraction is based on the simple assumptions of an up-right position and relative body parts size.

B. Dimensionality Reduction

A major goal in statistics modeling and machine learning is to reduce the dimensionality of input data. Several approaches have been followed in the technical literature for this purpose, such as PCA (Principal Components Analysis) [19], SGPLVM (Scaled Gaussian Process Latent Variable Model) [16], and GPDM (Gaussian Process with Dynamic Model) [20]. In our experiments, we use two data-sets that contain the 3D coordinates of body joints and its displacement vectors performing typical pedestrian motions. The first data-set was created at Carnegie Mellon University [21] using a motion capture system (CMU mocap). We will denote it as CMU data-set hereinafter. In CMU data-set each pose is made up of 41 joints along the body. The second data-set was created in our lab by means of the algorithm described in section II-A using the images contained in the KITTI data-set. We will denote it as K-UAH data-set. In K-UAH data-set, a skeleton containing 14 relevant body joints is built from the pedestrian points cloud in 3D. The choice of the 14 joints, as depicted in Fig. 2, has been made based on their discriminating capability for learning pedestrian motion. The use of 3D data and depth information has the potential to significantly augment the performance of pedestrian tracking and prediction. Our goal is to transform the 3D pose data into a low dimensional (3-d) latent space in which tracking and prediction of the pedestrian movements will be carried out based on trained data containing pedestrian motions. Previous works use different kinds of data for learning pedestrian motions. For example, in [20] the CMU data-set is used for introducing the GPDM algorithm. Each pose is defined by 44 Euler angles (joints), three global (torso) pose angles, and three global (torso) translational velocities. In [2], the use of GPDM is also proposed for pedestrian path prediction, although in this case the feature vector contains dense optical flow and disparity information instead of the 3D joints and displacement vectors.

The use of dynamical information in the training stage is useful for time-series data modeling, such as pedestrian motions. GPDM computes the observation and the dynamic mapping separately in a non-linear form. GPDM marginalizes out both mapping parameters and optimizes for the latent variables and the kernel hyper-parameters. The incorporation of dynamics can be used for predicting future data. The definition of the conditional probability of Y given X, θ, and W is provided in Eq. 3.

$$p(Y|X,\theta,W) = \frac{|W|^N}{\sqrt{(2\pi)^{ND}|K_Y|^D}} exp(-\frac{1}{2}tr(K_Y^{-1}YW^2Y^T)) \tag{3}$$

where Y is the centered observed data (3D-pose), 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 dimensions). The elements of the kernel matrix for the observation mapping are computed using Eq. 4.

$$k(x_i,x_j) = \theta_1 exp(\frac{-\theta_2}{2}(x_i - x_j)^T(x_i - x_j)) + \theta_3\delta_{i,j} \tag{4}$$

where $\delta_{i,j}$ is the Kronecker delta function. The dynamic mapping from the latent coordinates is given in Eq. 5.

$$p(X|\beta) = \frac{p(x_1)}{\sqrt{(2\pi)^{(N-1)d}|K_X|^d}} exp(-\frac{1}{2}tr(K_X^{-1}X_{out}X_{out}^T)) \tag{5}$$

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 Eq. 6.

$$k(x_i,x_j) = \beta_1 exp(\frac{-\beta_2}{2}(x_i - x_j)^T(x_i - x_j)) + \beta_3 x_i^T x_j + \beta_4\delta_{i,j} \tag{6}$$

where β_1 to β_4 are kernel hyper-parameters. The goal is to minimize the negative log-likelihood function $-lnp(X,\theta,\beta,W|Y)$ that is given by Eq. 7.

$$\mathscr{L} = \mathscr{L}_Y + \mathscr{L}_X + \sum_j ln\theta_j + \frac{1}{2\kappa^2} tr(W^2) + \sum_j ln\beta_j \quad (7)$$

where

$$\mathscr{L}_Y = \frac{D}{2} ln|K_Y| + \frac{1}{2} tr(K_Y^{-1} Y W^2 Y^T) - N ln|W| \quad (8)$$

$$\mathscr{L}_X = \frac{d}{2} ln|K_X| + \frac{1}{2} tr(K_X^{-1} X_{out} X_{out}^T) + \frac{1}{2} x_1^T x_1 \quad (9)$$

The optimization procedure is carried in two alternative steps. In a first step, $\mathscr{L}$ is optimized with respect to W in closed form. In a second step, $\mathscr{L}$ is optimized with respect to X,θ,β by using SCG (Scaled Conjugate Gradient) [22]. The latent coordinates are initialized by PCA, θ is manually initialized to $[1,1,exp(-1)]^T$, β is set to $[1,1,exp(-1),exp(-1)]^T$, and W is set to an identity diagonal matrix. The details of the learning algorithm are provided in [20].

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 $\mathscr{L}_X$ on the likelihood function by means of a λ element. As proposed in [15], we use a value $\lambda = \frac{D}{d}$. This modification is known as Balanced GPDM.

C. Prediction

GPDM provides a framework for transforming the 3D joints and its displacement vectors into a low dimensional latent space, as described in the previous section, but it also provides the grounds for predicting the next position in the latent space based on the current latent position and the dynamics of the pedestrian motion, as learned during the GPDM training stage. Thus, the latent position in the next frame can be obtained as described in Eq. 10.

$$\mu_X(x) = X_{out}^T K_X^{-1} k_X(x) \quad (10)$$

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 Eq. 6, and $k_X(x)$ is a column vector with elements $k_X(x,x_j)$ for all other latent positions x_j in the model. Eq. 10 can be iteratively used in order to predict the pedestrian position in the latent space a number of frames N ahead in time. The reconstruction of a pedestrian pose and the displacement vectors given the latent position can be obtained from Eq. 11.

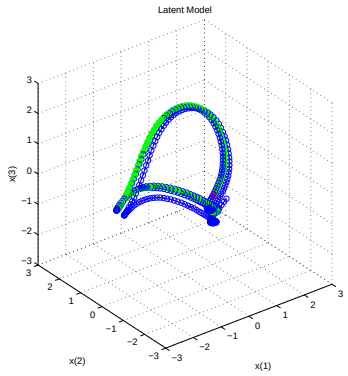
$$\mu = Y^T K_Y^{-1} k_Y(x) \quad (11)$$

where Y is the centered data set, K_Y^{-1} is the inverse matrix of the kernel for the observation mapping (see Eq. 4), and $k_Y(x)$ is a column vector with elements $k_Y(x,x_j)$ for all other latent positions x_j in the model. In our approach, the mean reconstruction errors per joint in the walking test set are 0.60 cm (CMU) and 4.4 cm (K-UAH), while the mean errors per joint in the stopping test set are 0.47 cm (CMU) and 2.22 cm (K-UAH), respectively.

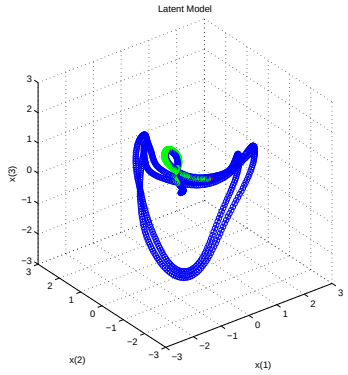
III. Experimental Results

Two systems trained on the CMU data-set (one for a walking trajectory and another for a stopping trajectory) using Balanced GPDM are tested using four test-sets (two sets extracted from the CMU data-set, 120 fps, and two sets extracted from the K-UAH data-set, 10 fps). Comparison between results obtained on the K-UAH data-set and CMU data-set are intended for quantifying the influence of stereo noise, differences in frame rate, and errors committed in the skeleton estimation phase. Two types of pedestrian behaviors are considered. In the first behavior, the pedestrian walks along the lateral direction with respect to the ego-vehicle. This scenario resembles a pedestrian crossing the street. In the second behavior, the person walks and suddenly stops. In this scenario, a pedestrian waits for crossing the street when a vehicle is approaching. Both actions are available in CMU and K-UAH data-sets. The mean squared error between the reconstructed pose (from a latent position) and the test pose (joints and displacement vectors) must be minimized iteratively in order to obtain the most likely latent position for a given test pose. A few issues must be considered before the training stage. The first one is the different number of joints (points) contained in the CMU and K-UAH data-sets. The same number of joints (in the same position) are selected in both data-sets, either in the training and testing steps, aiming at homogenizing the method. For that purpose, only the 14 body joints considered in the K-UAH data-set, as depicted in Fig. 2, are considered in the CMU data-set. As a second consideration, the order of the joints must be the same in the input data for both data-sets, either for training and testing. As a third consideration, the displacement vectors must be scaled due to the different frame-rates in the training and test sets. Finally, the same reference system and movement direction are considered between the training and test sets.

Once the latent position has been estimated, a prediction at a time horizon of N frames ahead can be done using Eq. 10 iteratively. Fig. 4 depicts the trajectory obtained in the latent space after training the system with GPDM (blue), the latent position corresponding to a given test pose (red), and the predicted trajectory in the latent space (green) for a time horizon of 1 s. As can be observed, the predicted trajectory closely resembles the shape of the trajectory used for training both in the walking and stopping cases. The pedestrian global lateral position with respect to the camera can be recovered using the displacement vectors. Tables I and II show the mean lateral prediction error (in cm) for different prediction horizons for the CMU and K-UAH data-sets, respectively. In both cases, the system is trained on the accurate 3D poses contained in the CMU data-set. As can be observed, prediction accuracy is higher when using testing data from the CMU data-set, given that it contains accurate 3D data acquired with a motion capture device. In such a



(a)



(b)

Fig. 4. (a) Balanced GPDM trained on CMU database file 07_01.c3d (walking motions) (blue), the latent position for a given test pose (red dot) and the predicted trajectory on the latent space (green); (b) Balanced GPDM trained on CMU database file 120_19.c3d (stopping motions) (blue), the latent position of a given test pose (red dot) and the predicted trajectory on the latent space (green)

case, the mean lateral error at a time horizon of 0.78 s is around 2 cm, for walking trajectories, and around 3 cm, for stopping trajectories. In contrast, prediction accuracy gets slightly decreased when using the K-UAH data-set, reaching a mean lateral error of 3.82 cm and 12.2 cm in the same conditions, respectively. This is partially due to the fact that the input data for the K-UAH data-set are computed automatically using the vision-based skeleton estimation algorithm previously described. However, the highly accurate results obtained demonstrate the potential of the proposed method for accurate pedestrian path prediction when using accurate 3D data as input and as a model for motion learning. It must be clearly stated that these results have been obtained using a single type of pedestrian dynamics for walking and stopping motions. Accordingly, similar motions have been used for testing in the K-UAH data-set. In any case, the realization of exhaustive experiments for gathering additional results involving many different pedestrians and dynamics would be needed in order to provide the grounds for generalization of the conclusions drawn in this research.

Figs. 5 and 6 show the predicted 3D pose at time horizons of 0, 0.23, 0.5, and 0.78 s for walking trajectories extracted from CMU and K-UAH data-sets, respectively.

TABLE I
LATERAL PREDICTION MEAN ERROR (CM) FOR DIFFERENT PREDICTION HORIZONS (SECONDS) - CMU DATA-SET

	0 sec	0.23 sec	0.5 sec	0.78 sec
Walking	0.23	1.99	2.03	2.10
Stopping	0.10	0.27	0.97	3.10

TABLE II
LATERAL PREDICTION MEAN ERROR (CM) FOR DIFFERENT PREDICTION HORIZONS (SECONDS) - K-UAH DATA-SET

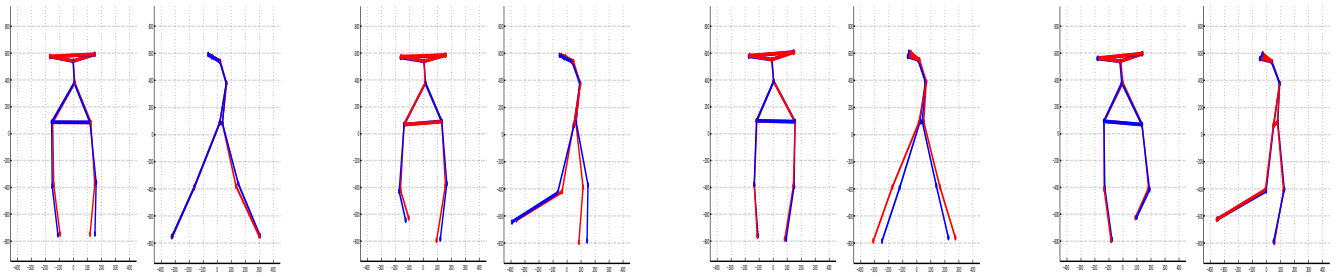
	0 sec	0.23 sec	0.5 sec	0.78 sec
Walking	1.91	2.26	3.44	3.82
Stopping	0.62	4.88	11.39	12.2

Reconstructed poses (depicted in blue) are very similar to the ground-truth poses (depicted in red) when using testing data containing very accurate 3D inputs, as in the case of the CMU data-set. Reconstruction results get a bit worse when using testing data obtained from stereo-vision and skeleton estimation (specially visible in the prediction of the legs), as in the case of the K-UAH data-set, although the general aspect of the body pose is preserved.

IV. CONCLUSIONS AND FUTURE WORK

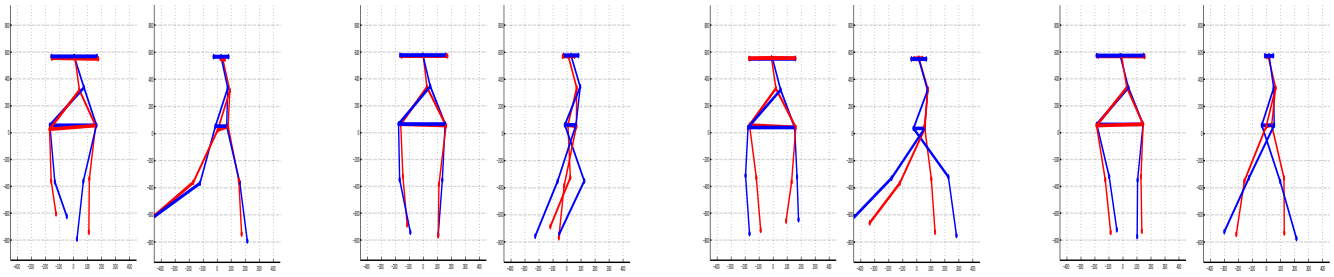
We have developed a system for accurate pedestrian path prediction in a limited time horizon up to ≈ 0.8 s. For such purpose, we propose the use of stereo-vision and probabilistic techniques, namely GPDM, for dimensionality reduction. The 3D structure of the pedestrian joints is built from the point cloud provided by a stereo-vision system and transformed (with displacement vectors) later on into a latent space using GPDM. Predictions are then performed in the latent space using the knowledge learned during the training of the system dynamics. The method has the potential to provide accurate path predictions of 2 cm, for walking trajectories, and 3 cm, for stopping trajectories, at a time horizon of 0.78 s, as demonstrated with the accurate 3D data-set provided by CMU. Experiments with K-UAH data-set, built from 3D data provided by a stereo-vision system, demonstrate that prediction accuracy gets decreased to 3.8 cm and 12 cm for walking and stopping trajectories, respectively at a time horizon of 0.78 s.

As future work, we propose to enhance the method for building the pedestrian skeleton from the point cloud. A more accurate 3D reconstruction of the pedestrian joints would definitely increase the pedestrian path prediction accuracy, as demonstrated in our experiments. In addition, a richer data-set will be created in order to include a representative number of sequences containing pedestrians performing different behaviors, such as walking, stopping, starting and bending-in, with different dynamics. Finally, a decision making system will be developed in order to select the most appropriate tracking system and to provide pedestrian action classification as a function of the pedestrian behavior.



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

Fig. 5. Predicted pose reconstruction (left, frontal view; right, lateral view) at time horizons of (a) 0 s, (b) 0.23 s, (c) 0.5 s, and (d) 0.78 s for walking trajectories extracted from CMU data-set. The reconstruction is depicted in blue, while the ground-truth is in red.



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

Fig. 6. Predicted pose reconstruction (left, frontal view; right, lateral view) at time horizons of (a) 0 s, (b) 0.23 s, (c) 0.5 s, and (d) 0.78 s for walking trajectories extracted from K-UAH data-set. The reconstruction is depicted in red, while the ground-truth is in blue.

V. ACKNOWLEDGMENTS

This work was supported by the Spanish Ministry of Economy under Grant ONDA-FP TRA2011-27712-C02-02 and the Portuguese FCT under Grant SFRH/BD/73181/2010.

REFERENCES

[1] 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.

[2] 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. In Press, 2014.

[3] M. Enzweiler and D. Gavrila, "Pedestrian protection systems: issues, survey, and challenges," *IEEE Transactions on Intelligent Transportation Sysems*, vol. 31, no. 12, pp. 2179–2195, 2009.

[4] D. Gerónimo, A. López, A. Sappa, and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1239–1258, 2010.

[5] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: an evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, 2012.

[6] 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.

[7] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[8] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester, "Discriminatively trained deformable part models, release 5," http://people.cs.uchicago.edu/ rbg/latent-release5/.

[9] M. Farmer, L. H. Rein, and A. Jain, "Interacting multiple model (imm) kalman filters for robust high speed human motion tracking," in *Int. Conf. on Pattern Recognition (ICPR)*, 2002, pp. 20–23.

[10] Y. Boers and J. Driessen, "Interacting multiple model particle filter," *IET Radar, Sonar Navigation*, vol. 150, no. 5, pp. 344–349, 2003.

[11] S. Schmidt and B. Farber, "Recognising the action intentions of humans," *Transportation Research Part F*, vol. 12, pp. 300–310, 2009.

[12] C. C. Lee, C. H. Chuang, J. W. Hsieh, M. X. Wu, and K. C. Fan, "Frame difference history image for gait recognition," in *International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 4, 2011, pp. 1785–1788.

[13] S. Kohler, M. Goldhammer, S. Bauer, S. Zecha, K. Doll, U. Brunsmann, and K. Dietmayer, "Stationary detection of the pedestrianś intention at intersections," *IEEE Intelligent Transportation Systems Magazine*, pp. 87–99, 2013.

[14] C. Keller, C. Hermes, and D. Gavrila, "Will the pedestrian cross? probabilistic path prediction based on learned motion features," *Pattern Recognition. Springer*, vol. 6835, pp. 386–395, 2011.

[15] 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.

[16] N. Lawrence, "Gaussian process latent variable models for visualization of high dimensional data," in *Advanced NIPS*, vol. 16, 2004, pp. 329–336.

[17] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[18] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 30, no. 2, pp. 328–341, 2008.

[19] A. Yao, J. Gall, L. J. V. Gool, and R. Urtasun, "Learning probabilistic non-linear latent variable models for tracking complex activities." in *NIPS*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. C. N. Pereira, and K. Q. Weinberger, Eds., 2011, pp. 1359–1367.

[20] J. M. Wang, D. J. Fleet, and A. Hertzmann, "Gaussian process dynamical models for human motion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 283–298, Feb. 2008.

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

[22] M. F. Moller, "A scaled conjugate gradient algorithm for fast supervised learning," Computer Science Department, University of Aarhus, Aarhus, Denmark, Tech. Rep. PB-339, 1990.