IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 30, 2017, accepted November 6, 2017, date of publication November 9, 2017,
date of current version November 28, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2772029

Encoding Motion Cues for Pedestrian Path Prediction in Dense Crowd Scenarios

YUKE LI[1], MOHAMED LAMINE MEKHALFI[1], (Member, IEEE), MOHAMAD MAHMOUD AL RAHHAL[2], (Member, IEEE), ESAM OTHMAN[3], (Student Member, IEEE), AND HABIB DHAHRI[2]

[1]Department of Pattern Analysis and Computer Vision, Istituto Italiano di Tecnologia, 16163 Genova, Italy
[2]Department of Information Science, College of Computer Science and Information Technology, King Saud University (Almuzahmiyah Branch), Riyadh 11451, Saudi Arabia
[3]Department of Computer Engineering, College of Computer Science and Information Technology, King Saud University, Riyadh 11543, Saudi Arabia

Corresponding author: Mohamad Mahmoud Al Rahhal (mmalrahhal@ksu.edu.sa)

This work was supported by the Deanship of Scientific Research at King Saud University through research group under Grant RG-1438-071.

ABSTRACT Pedestrian path prediction is an emerging topic in the crowd visual analysis domain, notwithstanding its practical importance in many respects. To date, the few contributions in the literature proposed quite straightforward approaches, and only a few of them have taken into account the interaction between pedestrians as a paramount cue in forecasting their potential walking preferences in a given scene. Moreover, the typical trend was to evaluate the proposed algorithms on sparse scenarios. To cope with more realistic cases, in this paper, we present an efficient approach for pedestrian path prediction in densely crowded scenes. The proposed approach initiates by extracting motion features related to the target pedestrian and his/her neighbors. Second, in order to further increase the representativeness of the extracted motion cues, an autoencoder feature learning model is considered, whose outcome finally feeds a Gaussian process regression prediction model to infer the potential future trajectories of the target pedestrians given their walking records in the scene. Experimental results demonstrate that our framework scores plausible results and outperforms traditional methods in the literature.

INDEX TERMS Crowd analysis, walking path prediction, motion modeling, computer vision.

I. INTRODUCTION

The mounting availability as well as the affordability of powerful processing facilities has benefited the computer vision field at large. This is manifested by the steady rise of potential research applications that might eventually find way into real scenarios. For instance, computer vision has recently shown that, to some extent, understanding the behavioral patterns of individuals is within reach. It can thus facilitate carrying out a wide range of crowd-related tasks, some instances include:

Crowd management: Crowd modeling and analysis can help comprehending, thus managing, public traffic and gatherings, as well as related events.

Urban planning: Understanding the undergoing interactions in the crowd as well as the behaviors of individuals in common places can assist in designing the structural layout of public spaces in order to accommodate the different crowd mobility flows.

Security and risk management: Monitoring public masses for the aim of security and hazard prevention is one of the top priorities with regards to today's society. The automatization of such process is therefore pivotal to aid ensuring secure and smooth daily activities. Moreover, detecting abnormalities may even help alarming yet preventing potential future threats. In turn, time, cost and human labor can potentially be saved.

In this regard, the relevant literature accumulates a decent amount of papers, addressing several issues related to crowd, such as activity forecasting [1], activity recognition [2], anomaly detection [3], crowd counting and profiling [4]–[7].

Another yet relatively recent crowd behaviour analysis task is pedestrian path prediction, which stands for forecasting the potential future walking route of an individual (or a group of people) provided their prior walking history. With respect to the other crowd analysis and modeling tasks, pedestrian path prediction has been devoted a remarkably scarce attention.

The task of foreseeing the potential walking route of an individual in crowded scenes can be a laborious attempt, given that numerous factors are liable to be treated jointly.

For instance, it has a strong tie with the structural layout of the scene in consideration. In other words, spacious sites can accommodate large masses with high physical freedom of movement, whereas confined spaces can enclose much less people with a limited mobility [8], [9].

Another worth noting fact regards the behavioral tendencies of the person of interest. For instance, impaired people constrained on wheelchairs usually tend to pursue less-crowded paths that can conveniently accommodate them as well as people in their vicinity, besides the way in which they interact globally (i.e., scene dynamics). However, apart from unusual situations, it is socially evident that people normally assumes commonsense social etiquettes (e.g., a typical manner to follow a route that is obstructed by a conversational group of people is, to avoid inconveniences, to detour it rather than crossing it). Such universally recognized behavioral codes suggest that crowd modeling via computer vision is not out of reach.

To date, only several contributions have been reported in the literature regarding pedestrian path prediction. The work proposed in [1], for instance, combines the semantics of the scene in order to derive a trajectory and destination forecasting model based on Markov Decision Processes. However, their approach was evaluated on sparse and slightly crowded scenes, which leaves open the question regarding its performance in dense public scenarios. Inspired by the success of mid-level elements [10], Walker *et al.* [11] suggest a non-parametric approach for visual prediction. The underlying insight is to represent a scene by means of a bunch of mid-level visual elements extracted via a sliding window, and then model the interaction between moving elements (agents) in the scene through a reward function that assigns high values to those agents having the capacity to mobilize in the scene. The main advantage of this method is the unsupervised fashion. However, the definition of agents in this work seems to be subjective given that in heavily crowded scenes, normally large-scale occlusions take place. In [12], a Bayesian cascade model that couples topic mixture model and Gaussian mixtures, fed with an ensemble of words encoding KLT tracklets over a grid of a given scene, is proposed. This is the first attempt that considers the co-occurrence and interaction of objects in the path forecasting context. However, given that the trajectories produced by tracklets are discontinuous, it is hard to assign the predicted outcomes to individuals, which is the main aim of human path prediction. Motivated by the success of Long-Short Term Memory (LSTM) networks in predicting long sequences, particularly in speech and handwriting [13], [14], Alahi *et al.* tailored them to the path prediction issue [15]. The underlying idea is to assign LSTM models to sequences (representing individuals) of the scene at hand, and then top them with a pooling layer to ensure the modeling of the interaction between spatially close sequences. Receiver trajectory prediction in American football games was the subject of study in [16], where a reward function takes jointly static features (relevant to prior knowledge about the game) and dynamic features (short-term prediction of the opposite team players). However, the proposed method remains dependent on offline knowledge about the game. Another LSTM-driven method was proposed in [17], but it was applied on tracklets rather than agents.

In this context, we note that the works mentioned above are jointly characterized by three aspects, namely (i) some of them depend on the scene structural and/or semantic layouts, (ii) only a few attempts consider agent interaction as a key-factor in determining the future walking routes of a given pedestrian, and (iii) even the latter works seem to evaluate their models on datasets characterized by little to moderate pedestrian density, which raises the concern regarding their effectiveness in densely crowded scenes.

To address that, we propose in this work a path forecasting approach, whose main novelty stems from the fact that it takes into account the local agents (pedestrians) in the neighborhood of the target pedestrian in dense scenarios, and demonstrate that an agent-inclusive approach (i.e., the neighbor pedestrians are included in the process) is ought to yield better results than a target-centric approach. To model the interaction between the target pedestrian and its neighbors, we resort to simple but informative features which are further reinforced by a feature learning stage. To strongly consider the interactions of the target with its neighbors, an Auto Encoder (AE) model is applied, which proved to score plausible improvements over the former selected features. Ultimately, such learned features are fed into a prediction model based on Gaussian Mixture Model (GPR) [23]–[25] in order to predict the future walking coordinates of the target agent. We show that our approach can outperform, by far, traditional prediction models in dense scenarios.

The remaining part of the paper is organized as follows. Section 2 elaborates the proposed approach as well as the conceptual backgrounds. Section 3 reports the experimental results and discussions, and Section 4 concludes the paper.

II. THE PROPOSED FRAMEWORK

A. GENERAL PIPELINE

In path prediction, the goal is to discern the potential future walking coordinates across a given scene of an individual by making use of their prior motion records. Precisely, provided the walking history of a pedestrian from time t_0 to time instant t_h, the aim is to infer the spatial walking coordinates from time t_{h+1} to t_{end}. To this end, a typical way to approach this task is to build a model based on the known walking history, and exploit it to produce the future locations of the target pedestrian.

In public places, especially in crowded scenarios, it is valid that people normally follow certain commonsense rules to interact with each other. For instance, etiquettes imply that pedestrians typically tend to detour conversational groups and other walking pedestrians that lie in their way in order to accommodate each other, avoid collisions and inconveniences. Such social observations suggest that modeling the walking routes of pedestrians is within reach if the motion

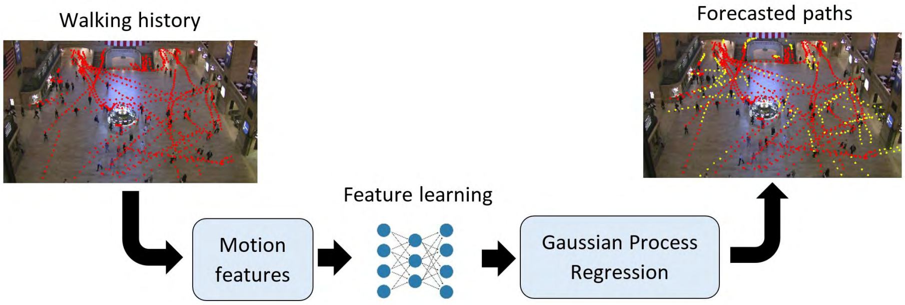


FIGURE 1. Architecture of the proposed path forecasting framework.

patterns characterizing the interaction between pedestrians in the scene are well-capitalized on. In this respect, the essence of our approach is inspired by the earlier fact. Precisely, our framework proceeds by deducing motion features related to (i) the target pedestrian, as well as (ii) his/her nearby pedestrians within a local spatial spot. In order to further improve the representativeness of the extracted motion features, a feature learning block based on AE is appended. Finally, the learned features are fed into learned GPR models to predict the future spatial locations of the target pedestrian. A depiction of our approach is provided in Fig. 1. In the next subsections, details outlining the three building blocks of the framework are conducted.

B. MOTION FEATURES

Let $p_i, i = 1, \ldots, N$ be the target pedestrian, and $F = t_0, \ldots, t_h$ be the uniformly samples time points (i.e., frame indexes) representing the walking records of p_i, we recall that the aim is to make use of the information stored in F to infer the future walking trajectory within the time span t_{h+1} to t_{end}. Thus, in order to learn the prediction models, two information are required, namely (i) features representing the motion patterns of the target p_i with respect to its neighbors, that are deduced from the walking history within F, along with (ii) training labels, which represent the spatial coordinates of p_i within F.

In order to enable the earlier process, the data comprised in F is fragmented into overlapping segments (e.g., a one frame overlap) of k frames each, where the k spatial locations are utilized for feature extraction and subsequently fed into the GPR as input, and their succeeding coordinates at time $k+1$ as output (label) for training to the GPR model. For example, if k equals to five, then the first five spatial points are utilized for motion feature extraction (GPR input) and the sixth point is employed as a label (GPR output), then the same applies for the next overlapping points.

In order to characterize the motion patterns of a target pedestrian $p_i, i = 1, \ldots, N$ as well as those moving in their vicinity, we adopt jointly two distinct but complementary feature modalities. The first one consists in extracting features relative to the pedestrian of interest, where displacement and features are extracted. The displacement features correspond to the Euclidean distances between each two subsequent spatial points of the k frames. Thus, a $(k-1)$-dimensional displacement vector v_d is envisioned, as illustrated in Fig. 2.

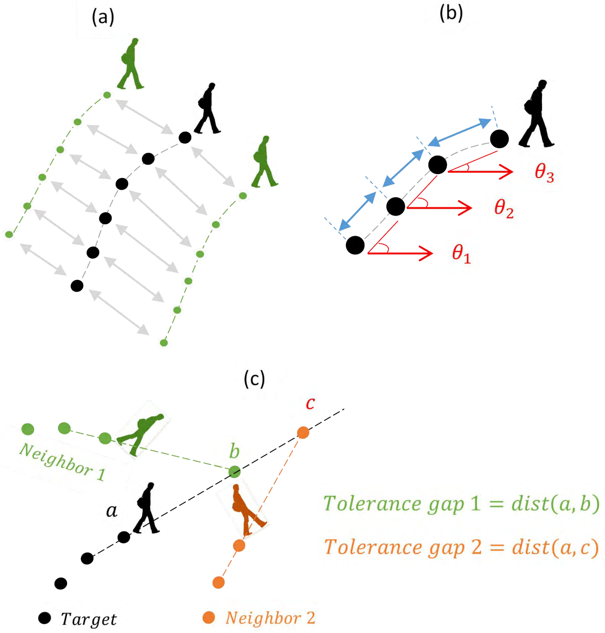


FIGURE 2. Display of the motion feature extraction process. (a) Spatial distances to neighbor pedestrians, black dots refer to the special locations of the target pedestrian and green ones to those of the neighbors. (b) Feature extraction from the walking history of the target pedestrian, blue arrows pertain to spatial displacements, and red ones represent the target's orientation angles. (c) Tolerance gap with respect to nearby pedestrians, it can be noted that the first neighbor is more likely to collide with the target than the second one as its tolerance gap is narrower.

The second modality relates to three features corresponding to the interaction between p_i and its nearby pedestrians that appear within a local sport (e.g., a circle of a one frame radius) at the current time instant.

Therefore, the first one considers Euclidean distance between each of the k spatial points of the target pedestrian and their counterparts (only the ones that occur within the local circle are considered) corresponding to the nearby pedestrians are aggregated in a vector v_N. Thus, the length of v_N depends on how many spatial points of nearby pedestrians coincide with those of the target. For instance, if only two pedestrians appear inside the local area where the first neighbor co-occurs with the target at three time instants, and the second neighbor coincides with the target at two instants, then five Euclidean distance values are gathered in v_N. Fig. 2 illustrates the concept.

While the earlier feature relates to the spatial distances between the target and its neighbors, the second one considers the angular distances between them. Thus, for the target pedestrian and within the k frames, the orientation angles can be extracted from the lines joining each two consecutive points, $k-1$ angular values are thus obtained. Hence, after calculating the individual angular deviations of the target as well as the co-occurring pedestrians inside the local circle, the angular distances can be easily inferred and stored in v_θ.

The third feature expresses the likelihood for the target pedestrian to collide with a nearby pedestrian. For given neighbor pedestrian, that is achievable by virtually extending (linearly) the line joining the last two coordinates of this latter as well as those of the target pedestrian, which eventually leads to a meeting knot along the virtual line of the target (See Fig. 2). Thus, the longer the distance (termed tolerance gap) between the last coordinate of the target and the meeting knot the less likely the collision between the two persons is to take place, and vice versa otherwise. Fig. 2 further demonstrates the concept. Thereupon, the tolerance gap values corresponding to all nearby pedestrians are collected into a vector v_g. Note that this feature is motivated by the assumption that persons tend to pursue a linear line in the short term.

So far v_d, v_N, v_θ and v_g, pertaining to target displacement, spatial neighbor distances, angular neighbor distances, and collision tolerance gaps, are obtained. However, at each time instant, different number of pedestrians are observed, which entails inconsistent feature vector lengths. In order to render all the vectors to a constant dimension, we calculate the first five statistical moments for each of the four vectors, which totals a 20-dim feature vector $x \in R^n$ to be fed into the AE.

C. UNSUPERVISED FEATURE LEARNING

Unsupervised feature learning is one of the paradigms that has been shown to better the underlying structure of features in the recent literature [18]–[21]. One of the successful models in this regard is the Autoencoder, which is characterized by a simple symmetric architecture, a simple learning process performed in an unsupervised fashion, and a good generalization capability.

In its basic form, an AE is composed of three layers, an input layer acquiring the input features of size n, a hidden layer of d nodes, and a reconstruction layer with n nodes (same as the input layer).

Suppose $\mathrm{x} \in \mathcal{R}^{\mathrm{n}}$ be the input vector, $\mathrm{h} \in \mathcal{R}^{\mathrm{d}}$ the output of the hidden layer, the underlying principle of AE is the reproduction, with a certain error, of its input x into an output vector $\hat{\mathrm{x}} \in \mathcal{R}^{\mathrm{n}}$ through h. Thus, a feature reduction is fulfilled if $\mathrm{d} < \mathrm{n}$, whilst a (sparse) over-complete representation is envisioned in a high dimensional manifold if $\mathrm{d} > \mathrm{n}$ [24].

The reproduction of x at he output undergoes two phases, namely an encoding part that maps x into h, and a decoding stage that maps this latter into $\hat{\mathrm{x}}$, through a nonlinear mapping function f as expressed by Eq (1) and Eq (2), respectively:

$$\mathrm{h} = f\,(\mathrm{Wx} + \mathrm{b}) \tag{1}$$

$$\hat{\mathrm{x}} = f\left(\mathrm{W'h} + \mathrm{b'}\right) \tag{2}$$

where $\mathrm{W} \in \mathcal{R}^{d \times n}$ and $\mathrm{b} \in \mathcal{R}^{d}$ represent the weight and the bias of the encoding part, and $\mathrm{W'} \in \mathcal{R}^{n \times d}$ and $\mathrm{b'} \in \mathcal{R}^{n}$ stand for the same of the decoding part. Typically, a sigmoid activation function is employed.

The parameters (W, W′, b and b′) can be estimated by minimizing a cost function $L(\mathrm{x}, \hat{\mathrm{x}})$ expressing the error between the input and its reconstruction. Since the input features are real-valued, a squared error function is adopted:

$$L\left(\mathrm{x}, \hat{\mathrm{x}}\right) = \left\|\mathrm{x} - \hat{\mathrm{x}}\right\|^2 \tag{3}$$

Initially, the weights and the biases are set randomly, and then updated iteratively until a predefined convergence criterion (i.e., maximum number of iterations) is met. The update is achieved through the following equations:

$$\mathrm{W} = \mathrm{W} - \eta \frac{\delta L(\mathrm{x}, \hat{\mathrm{x}})}{\delta \mathrm{W}} \tag{4}$$

$$\mathrm{W'} = \mathrm{W'} - \eta \frac{\delta L(\mathrm{x}, \hat{\mathrm{x}})}{\delta \mathrm{W'}} \tag{5}$$

$$\mathrm{b} = \mathrm{b} - \eta \frac{\delta L(\mathrm{x}, \hat{\mathrm{x}})}{\delta \mathrm{b}} \tag{6}$$

$$\mathrm{b'} = \mathrm{b'} - \eta \frac{\delta L(\mathrm{x}, \hat{\mathrm{x}})}{\delta \mathrm{b'}} \tag{7}$$

where η is the learning rate. Finally, the encoded (learned) features, denoted $\mathrm{l} \in \mathcal{R}^{d}$ are comprised in the hidden layer h, which are then fed into the GPR model briefly outlined below.

D. GAUSSIAN PROCESS PATH FORECASTING

In GP formulation [25]–[27], learning a machine follows a Bayesian estimation problem, where the parameters of the machine are assumed as random variables to be inferred from a Gaussian distribution. Let us consider $\mathrm{L} = \{\mathrm{l_i}\}_{\mathrm{i}=1}^{\mathrm{M}}$ a matrix accommodating the AE-learned feature vectors, and $\mathrm{l_i} \in \Re^{\mathrm{N_c}}$ represents a feature vector of extracted from the K-long frames. Let also $\mathrm{y} = \{\mathrm{y_i}\}_{\mathrm{i}=1}^{\mathrm{M}}$ be the corresponding output target vector, which comprises the spatial coordinate at time $\mathrm{K}+1$ corresponding to their respective vectors in L.

The aim of GP regression is to infer from of training set $\{\mathrm{L}, \mathrm{y}\}$ a function $\psi(\cdot)$ so that $\mathrm{y} = \psi\,(\mathrm{x})$. This can be done by formulating the Bayesian estimation problem directly in the function space view. The observed values y of the function to model are considered as the sum of a latent function following a joint Gaussian distribution and a noise component $\boldsymbol{\varepsilon}$ with zero mean and variance σ_{n}^2:

$$\mathrm{y} \sim \mathrm{GP}\left(0, \mathrm{K}\,(\mathrm{L}, \mathrm{L}) + \sigma_{\mathrm{n}}^2 \mathrm{I}\right) \tag{8}$$

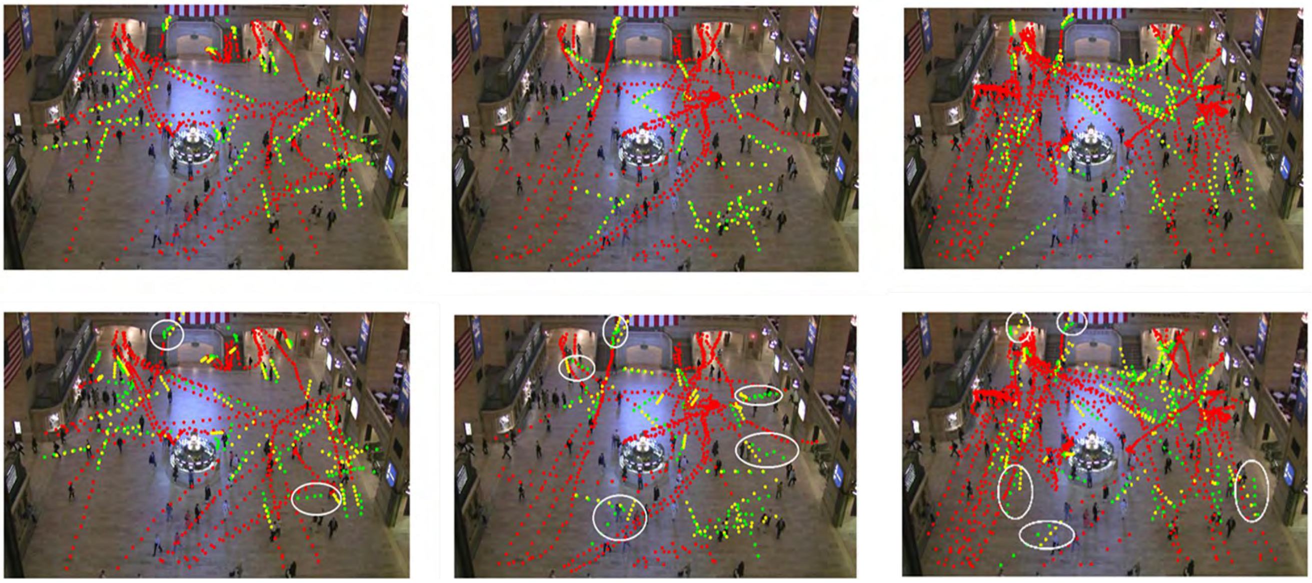


FIGURE 3. Examples of path prediction from three different time slots. First row displays the results of our approach, red dots refer to the training segment, green to actual test segment and yellow to the estimated prediction. Second row represents the results of SF, where circled zones highlight the prediction failures regarding the SF [22], while their counterpart predictions of our approach are properly inferred.

where K(X, X) is the covariance matrix, which is built by means of a covariance (kernel) function computed on all the training sample pairs, and I represents the identity matrix.

A central role in the GP regression model is played by the covariance function $k(x_i, x_j)$ as it embeds the geometrical structure of the training samples. In this paper, we consider the following Matérn covariance function [27]:

$$k\left(l_i, l_j\right) = \theta_0 \left(1 + \frac{\sqrt{3}\left|l_i - l_j\right|}{s}\right) exp\left(-\frac{\sqrt{3}\left|l_i - l_j\right|}{s}\right) \qquad (9)$$

For this covariance function, the hyperparameter vector given by $\odot = [s, \theta 0]$ can be determined empirically by cross-validation. However, the intrinsic nature of GPs allows a Bayesian treatment for the estimation of $\odot$.

These were the fundamental insights outlining GPR. For an in-depth comprehension, the reader is directed to consult [23]–[25].

III. RESULTS AND DISCUSSION

A. DATASET

As highlighted earlier, our approach is meant essentially for densely crowded scenes. In this regard, to the best of our knowledge, the only available dataset that meets our endeavor is the one recently presented in [6], which totals 12684 pedestrians from a one-hour video. The complete trajectory from the time each pedestrian enters the scene up to the exit time is labeled, setting thereby the largest dataset of this kind to date, which we deem as a valuable advantage to realistically assess our approach. For each trajectory, a portion of the available data is employed for training and the remaining part is retained for test. In order to quantify the performance, we utilize the Normalized Mean Squared Error (NMSE) between the predicted trajectory and the actual one, thus the smaller the NMSE the more accurate the prediction, as in [26].

TABLE 1. Influence of the training size.

Setting	Ours (%)	SF [22] (%)	CV [26] (%)	CA [26] (%)
90%	3.06	5.7	5.7	9.2
80%	3.06	5.7	5.9	9.9
67%	4.6	7.6	8.5	10.4
50%	6.2	9.2	9.7	12.4
40%	6.5	9.9	11.9	13.1
20%	7.2	11.5	12.7	14.8

B. EXPERIMENTS

We initiate the experiments by assessing the influence of the training size, where the respective results are summarized in Table 1. Thus, it can be seen that the accuracy increases with the training size, where the best score of 3.06% was obtained for 80% and 90% equally, which is due to the fact that the AE inherently requires a large amount of training data in order to converge faster. Consequently, the former option i.e., 80% for training and 20% for test) is therefore adopted in the remaining experiments.

Second, we study the impact of the local neighbors on the results, as well as the feature learning and path prediction processes (i.e., assigning one GPR/AE model per pedestrian or a global model for all), the results are reported in Table 2. The best result of 3.1% is observed when assigning a single AE and a single GPR model for each pedestrian, while a 3.4% is scored when a global AE is utilized, which is not a big decline given that the AE serves as an unsupervised encoding mechanism. By contrast, the GPR is employed as

TABLE 2. Comparison scores.

Method	NMSE (%)
Ours	3.1
Global AE	3.4
Learning without neighbors	4.6
Without AE	5.6
Global GPR	12.0
SF [22]	5.7
CV [26]	5.9
CA [26]	9.9

a supervised prediction model, hence it cannot be learned globally on a bunch of pedestrians while manifesting a good generalization ability as pedestrians (i) neither exhibit the same walking behaviors, (ii) nor share the same walking context (i.e., they don't necessarily walk across the same spatial spot or encounter the same obstacles), which is confirmed by the quantitative results i.e., a global GPR drops the accuracy largely by almost 9%.

On the other hand, disregarding the neighbors' information incurs roughly a 1.5% drop, which confirms that the interactions between the target and its neighbors plays a pivotal cue in the whole process, which represents one of the strengths of our approach.

TABLE 3. Effect of the size of previous coordinates.

History	1	3	5	8	10
NMSE (%)	4.2	4.2	3.1	3.1	3.4

We further study the impact of k (i.e., the number of preceding frames to be considered) in Table 3. It is clear that a small value (i.e., k = 1 or k = 3) is unreliable, while a high value (i.e., k = 10) slides down the accuracy, leading us to opt for k = 5 as an optimal heuristic. This suggests that a short walking history is insufficient to characterize the future undergoing motion patterns, whilst too much information is not useful, as the future walking behavior of a given pedestrian seems to depend on the near past history.

As for comparison, we implemented three well-known reference methods for crowd analysis, namely Social Force Dynamic (SF) [22], Constant Velocity (CV) [26], and Constant Acceleration (CA) [26]. The results are reported in Table 1, and Table 2. Our method outperforms all three schemes by 2.7%, 2.7%, and 6.2%, respectively. For instance, after our method, the SF yielded 5.7%, ranking second given that it also considers the interaction between the pedestrians into the prediction process, whilst the remaining algorithms do not. Visual examples of path prediction by our method and SF, for different time intervals, are displayed in Fig. 3. It can be seen that, by contrast to our method, the SF can successfully predict trajectories with linear-like nature but fails otherwise (e.g., second row of Fig. 3).

Although our framework is aimed at addressing crowded public settings, we also assess it in a sparse scenario. The ETH HOTEL dataset [8], which contains over 300 different pedestrians is employed. The size of the training and test splits is the same as the PWPD dataset. As per evaluation metric, we use the average MSE as in [1] and [8]. The results are summarized in Table 4. Although our method still outperforms SF [8], CV [21] and CA [21], the gain is not as high as that on the PWPD dataset. For instance, second to ours comes the SF [8] method with a 0.03 decline. This is traced to the fact that our method, although can still outperform in a

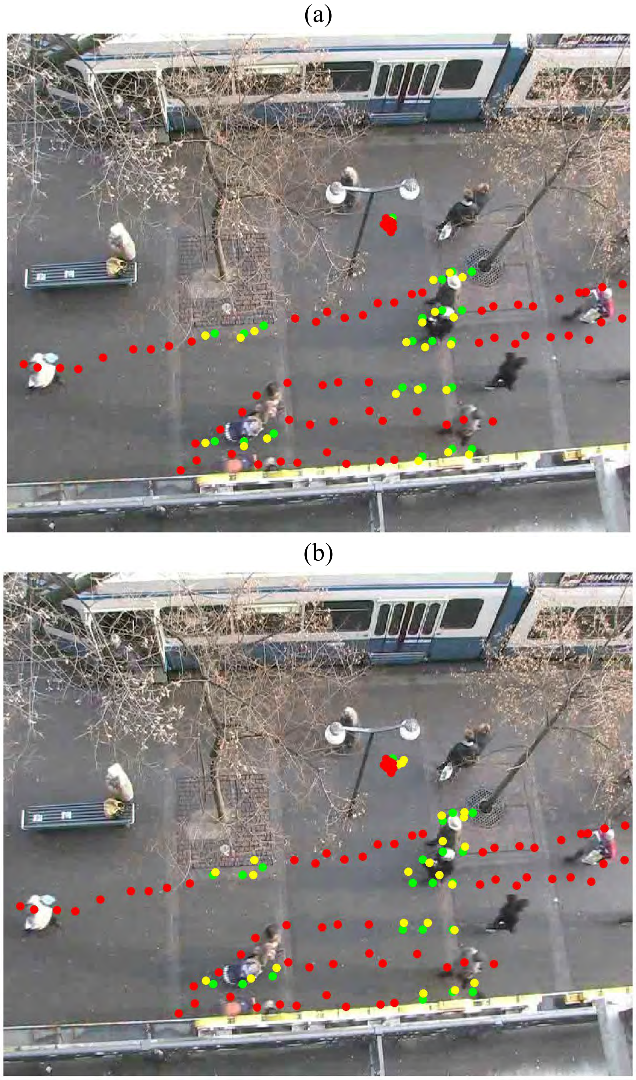


FIGURE 4. Examples of path prediction from the ETH dataset. (a) Our approach (b) Social Force model [22]. It can be noticed that our predictions are more consistent.

TABLE 4. Results on the ETH dataset.

Method	NMSE (%)
Ours	0.22
Global AE	0.29
Learning without neighbors	0.46
Without AE	0.49
Global GPR	0.54
SF [22]	0.25
CV [26]	0.39
CA [26]	0.40

sparse setting, is meant primarily to model the interaction of individuals in a crowded scene. Therefore, sparse scenarios, such as the ETH HOTEL dataset, do not enable to adequately capture the contextual motion features w.r.t a given target pedestrian. Visual examples are illustrated in Fig. 4). It is to observe that our method is better especially in non-linear cases.

For reproduction purposes, it is to note that the above experiments were conducted based on 100 hidden nodes for both AE and GPR models. Although other architectural alternatives (e.g., increasing the number of hidden nodes for instance) might lead to better results, this is out of the scope of the paper.

IV. CONCLUSIONS

This paper proposed a pedestrian walking path prediction approach in crowded scenes. The underlying idea is to characterize the motion patterns of a given target pedestrian as they move in the scene, with respect to their neighbors, and reinforce the inferred features via unsupervised feature learning. Experimental results show that promising scores can be attained versus well-established traditional methods. Nevertheless, potential ameliorations are prone to be achieved by relying for instance on a deeper feature learning architecture, which might introduce notable gains but potentially compromises the processing overheads. Another line of improvement is to assign either AE or GPR models to local regions within the scene, where regions that display similar motion attributes (e.g., pedestrian frequency) are assigned to a single AE/GPR, which we expectedly believe is subject to raise the prediction precision if the region segmentation step is adequately addressed.

REFERENCES

[1] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert, "Activity forecasting," in *Proc. ECCV*, 2012, pp. 201–214.

[2] M. Ziaeefard and R. Bergevin, "Semantic human activity recognition: A literature review," *Pattern Recognit.*, vol. 48, no. 8, pp. 2329–2345, 2015.

[3] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proc. CVPR*, Jun. 2010, pp. 1975–1981.

[4] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proc. CVPR*, Jun. 2015, pp. 833–841.

[5] J. Shao, C. C. Loy, and X. Wang, "Scene-independent group profiling in crowd," in *Proc. CVPR*, Jun. 2014, pp. 2219–2226.

[6] B. Zhou, X. Tang, and X. Wang, "Learning collective crowd behaviors with dynamic pedestrian-agents," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 50–68, 2015.

[7] B. Zhou, X. Tang, H. Zhang, and X. Wang, "Measuring crowd collectiveness," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1586–1599, Aug. 2014.

[8] S. Yi, H. Li, and X. Wang, "Understanding pedestrian behaviors from stationary crowd groups," in *Proc. CVPR*, Jun. 2015, pp. 3488–3496.

[9] S. Yi, H. Li, and X. Wang, "Pedestrian travel time estimation in crowded scenes," in *Proc. ICCV*, 2015, pp. 3137–3145.

[10] S. Singh, A. Gupta, and A. A. Efros, "Unsupervised discovery of mid-level discriminative patches," in *Proc. ECCV*, 2012, pp. 73–86.

[11] J. Walker, A. Gupta, and M. Hebert, "Patch to the future: Unsupervised visual prediction," in *Proc. CVPR*, 2014, pp. 3302–3309.

[12] Y. Yoo, K. Yun, S. Yun, J. Hong, H. Jeong, and J. Y. Choi, "Visual path prediction in complex scenes with crowded moving objects," in *Proc. CVPR*, 2016, pp. 3488–3496.

[13] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proc. ICML*, vol. 14. 2014, pp. 1764–1772.

[14] A. Graves. (2013). "Generating sequences with recurrent neural networks." [Online]. Available: https://arxiv.org/abs/1308.0850

[15] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: Human trajectory prediction in crowded spaces," in *Proc. CVPR*, Jun. 2016, pp. 961–971.

[16] N. Lee and K. M. Kitani, "Predicting wide receiver trajectories in American football," in *Proc. WACV*, Mar. 2016, pp. 1–9.

[17] H. Su *et al.*, "Crowd scene understanding with coherent recurrent neural networks," in *Proc. IJCAI*, 2016, pp. 3469–3476.

[18] A. Coates, H. Lee, and A.Y. Ng, "An analysis of singlelayer networks in unsupervised feature learning," in *Proc. 14th Int. Conf. Artif. Intell. Statist. (AISTATS)*, Ann Arbor, MI, USA, 2011, pp. 1–9.

[19] Y. Wang, Z. Xie, K. Xu, Y. Dou, and Y. Lei, "An efficient and effective convolutional auto-encoder extreme learning machine network for 3d feature learning," *Neurocomputing*, vol. 174, pp. 988–998, Jan. 2016.

[20] S. Rifai, P. Vincent, X. Müller, X. Glorot, and Y. Bengio, "Contractive auto-encoders: Explicit invariance during feature extraction," in *Proc. ICML*, 2011, pp. 833–840.

[21] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, no. 12, pp. 3371–3408, Dec. 2010.

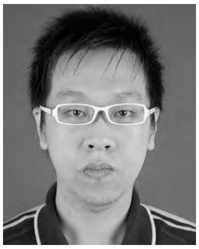
[22] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *Proc. ICCV*, Sep./Oct. 2009, pp. 261–268.

[23] C. K. I. Williams and D. Barber, "Bayesian classification with Gaussian processes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 12, pp. 1342–1351, Dec. 1998.

[24] C. Rasmussen and C. K. I. Williams, *Gaussian Process for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.

[25] J. Quiñonero-Candela and C. E. Rasmussen, "A unifying view of sparse approximate Gaussian process regression," *J. Mach. Learn. Res.*, vol. 6, pp. 1939–1959, Dec. 2005.

[26] S. Yi, H. Li, and X. Wang, "Pedestrian behavior understanding and prediction with deep neural networks," in *Proc. ECCV*, 2016, pp. 263–279.



YUKE LI received the M.Sc. degree in microwave engineering and the Ph.D. degree in computer science from the State Key Laboratory of Information Engineering in Surveying Mapping and Remote Sensing, Wuhan University. He is currently pursuing the Ph.D. degree with the Department of Pattern Analysis and Computer Vision, Istituto Italiano di Tecnologia, under the supervision of Prof. V. Murino. His main interests lie in computer vision and deep learning.

MOHAMED LAMINE MEKHALFI (S'13–M'16) received the State Engineer degree in electronics (with specialization in telecommunications) from the University of Mentouri, Constantine, Algeria, in 2009, the M.Sc. degree in electronics (with specialization in signal processing) from the University of Batna, Algeria, in 2012, under the supervision of Prof. R. Benzid, and the Ph.D. degree in information and communication technology (ICT) from the University of Trento, as part of the ICT International Doctoral School. From 2012 to 2016, he was engaged with the Signal Processing and Recognition Laboratory headed by Prof. F. Melgani with the Department of Information Engineering and Computer Science, University of Trento. Since 2016, he has been a Post-Doctoral Researcher with the Department of Pattern Analysis and Computer Vision, Istituto Italiano di Tecnologia, under the supervision of Prof. V. Murino, where he conducts research mainly in crowd behavior analysis/modeling and person re-identification. His research interests mainly encompass computer vision, pattern recognition, machine learning, and remote sensing.

MOHAMAD MAHMOUD AL RAHHAL (S'14–M'17) received the B.Sc. degree in computer engineering from Aleppo University, Aleppo, Syria, in 2002, the M.Sc. degree from Hamdard University, New Delhi, India, in 2005, and the Ph.D. degree in computer engineering from King Saud University, Riyadh, Saudi Arabia, in 2015.

From 2006 to 2012, he was a Lecturer with Al Jouf University, Sakakah, Saudi Arabia. Since 2015, he has been an Assistant Professor in computer science with King Saud University. His research interests include signal/image medical analysis, remote sensing, and computer vision.

ESAM OTHMAN (S'14) received the B.Sc. (Hons.) degree in computer engineering from Umm Al-Qura University, Mecca, Saudi Arabia, in 2007, and the M.Sc. degree in computer engineering from King Saud University, Riyadh, Saudi Arabia, in 2012, where he is currently pursuing the Ph.D. degree with the Department of Computer Engineering. His research interests include machine learning, pattern recognition, and remote sensing.

HABIB DHAHRI was born in Sidi Bouzid, Tunisia, in 1975. He received the degree in computer science in 2001, and the Ph.D. degree in computer engineering from the National Engineering School of Sfax in 2013. He is currently an Assistant Professor in computer science with King Saud University. His research interest includes computational intelligence: neural network, swarm intelligence, differential evolution, and genetic algorithm.