# "Seeing is Believing": Pedestrian Trajectory Forecasting Using Visual Frustum of Attention

Irtiza Hasan[1,2], Francesco Setti[1],Theodore Tsesmelis[1,2,3], Alessio Del Bue[3], Marco Cristani[1], Fabio Galasso[2]
University of Verona (UNIVR)[1], OSRAM GmbH[2], Istituto Italiano di Tecnologia (IIT)[3]
irtiza.hasan@univr.it, francesco.setti@univr.it

## Abstract

*In this paper we show the importance of the head pose estimation in the task of trajectory forecasting. This cue, when produced by an oracle and injected in a novel socially-based energy minimization approach, allows to get state-of-the-art performances on four different forecasting benchmarks, without relying on additional information such as expected destination and desired speed, which are supposed to be know beforehand for most of the current forecasting techniques. Our approach uses the head pose estimation for two aims: 1) to define a view frustum of attention, highlighting the people a given subject is more interested about, in order to avoid collisions; 2) to give a short-time estimation of what would be the desired destination point. Moreover, we show that when the head pose estimation is given by a real detector, though the performance decreases, it still remains at the level of the top score forecasting systems.*

## 1. Introduction

Trajectory forecasting stands for predicting where the people would go, *i.e*. estimate the location of pedestrians in the future frames within a surveillance camera video. Usually, these approaches assume to observe the behavior of a pedestrian for some frames, and then predict his/her locations in the next frames. In the last years, some *learning-based* methods have been proposed, mostly based on recurrent neural networks, and in particular on LSTM [3, 4]. Despite performing very well on many public datasets, learning based approaches require a huge amount of annotated data to be trained, which is a time consuming operation that is usually a limit for real life applications. In contrast, *model-based* approaches rely on hand generated models of crowd behavior; to overcome the lack of training data, they usually assume to have in advance rich information describing the future state of a pedestrian, like his desired speed and his destination point [29, 35, 40, 45]. Such a knowledge allows to create a set of hypotheses of trajectories where occlusions are avoided and a social distance among the people is maintained. This strategy brings to a contradiction, since, in particular, the future destination implies to know details about the trajectory that we actually want to forecast.

In this paper, we propose a model-based forecasting approach that discards information coming from the future time steps, advocating the use of the head pose estimation [33, 34, 39] as a way to address this problem. The head pose captures the people focus, also named *Visual Frustum of Attention* (VFOA) [36], as maintained by literature on social psychology and neuroscience [17, 38, 41]. Potentially, the head pose is an agreed proxy for the short-term future prediction, not only in social psychology and neuroscience [8, 9, 12, 26, 28], but also in tracking [29, 36, 6]. Furthermore, knowing the head pose allows describing the social contexts of the moving agents, as people within a cone of visual attention are likely to steer the agents to avoid collisions.

We propose here a new energy-based model to encode the people intention and their social context in a simple and intuitive way. The model simplicity allows us to study the importance of both aspects separately, further to analyzing the influence of the head pose quantization and the aperture of the cone of attention (cf. Fig. 1). Interestingly, we attain best performances when the cone of attention has an aperture of 30 degrees, matching the results of psychological studies [16].

We consider for study four challenging datasets which offer the head pose annotations and which have recently been considered for the trajectory forecasting task: the UCY, Zara01, Zara02 [22] and TownCentre [7]. By only considering the head pose as given by the oracle, *i.e*. causal information, we definitely outperform current state-of-the-art which leverages the future people destination information [29, 35, 40, 45]. The results remain very competitive with state-of-the-art, setting the best score in two cases out of four when we replace the oracle with a per-frame head-pose estimator. We believe that this calls for the importance of head pose for the people trajectory prediction.

We claim two main contributions: 1) the use of head orientation as a novel way to improve the estimation of person's future path, and 2) a novel energy based approach for trajectory forecasting.

The rest of the paper is organized as follows: in Sec. 3 we propose the prediction model, discussing its learning and inference. In Sec. 4 we evaluate against 4 comparative approaches showing also some ablation studies. Finally, we conclude the paper in Sec. 5.

## 2. Related work

A large body of literature have addressed the topic of path prediction, by adopting Kalman filters [18], linear regressions [27], Gaussian regression models [31, 32, 43, 44], autoregressive models [2] and time-series analysis [30]. Our approach departs from these classical approaches because we also consider the human-human interactions and the person intention, expressed by the VFOA.

**Human-human interactions.** The consideration of other pedestrians in the scene and their innate avoidance of collision was first pioneered by [15]. The initial seed was further developed by [22] and [29], which respectively introduced a data-driven and a continuous model. Notably, these approaches remain top performers on modern datasets, as they successfully employ essential cues for track prediction such the human-human interaction and the people intended destination. More recent works encode the human-human interactions into a "social" descriptor [3, 4, 25] or proposes human attributes [46, 24] for the forecasting in crowds. Our work mainly differentiates from [22, 29] because we only consider for interactions those people who are within the cone of interest of the person, which we encode with the VFOA (as also maintained by psychological studies [16]).

**Destination-focused path forecast.** Starting from the seminal work of Kitani *et al*. [19], path forecast has been cast as an inverse optimal control (IOC) problem. Follow-up work has additionally utilized inverse reinforcement learning [1, 47] and dynamic reward functions [21] to address the occurring changes in the environment. We describe these approaches as destination-focused because they all require the end-point of the person track to be known, which later work has relaxed to a set of plausible path end points [10, 23]. We share with these works the importance of the person intention, but we believe that knowing the destination undermines the reason why we may be predicting the trajectories. By contrast, we represent the person intention by their VFOA which, as we show, may be estimated at the current frame.

**VFOA and the social motivation.** The interest into the VFOA stems from sociological studies such as [8, 9, 11, 12, 13, 28, 42], whereby VFOA has been shown to correlate to the person destination, pathway and speed. Interestingly, the correlation is higher in the cases of poor visibility, such as at night time, and in general when the person is being busy with a secondary task (*e.g*. bump avoidance) further to the basic walking. These studies motivate the use of VFOA as a proxy to forecasting trajectories. Using VFOA comes with the further advantage that it can be estimated [5, 34, 37] on a frame basis, thus requiring no oracle information and enabling a real-time system. While our experiment is agnostic about the head pose estimation algorithm, in our experiments we will use an off-the-shelf head pose estimator [14].

## 3. Our Model

We formulate the predictive model as a joint optimization problem, where the position of each individual in the next frame is simultaneously estimated by minimizing an energy function. We gather into the energy three intuitive terms: (1) a collision avoidance term, which accounts for the multi-agent nature of the system, (2) a destination term, which accounts for the goal of each individual behaviour, and (3) a constant velocity term. The general idea behind our model is that, when in an open space, a person walks towards a destination point trying to avoid collisions with other pedestrians and static objects. While doing this, she/he prefers to move smoothly, *i.e*. limiting accelerations both in terms of intensity and direction.

Our cost function has the general form:

$$\mathcal{C} = w_A \cdot E_A + w_V \cdot E_V + w_D \cdot E_D \tag{1}$$

where $w_A$, $w_V$, and $w_D$ are weighting factors, and $E_A$, $E_V$, and $E_D$ are the respective three energy terms discussed in the following.

Let us consider a video sequence of $T$ image frames as $\mathcal{S} = \{I_t\}_{t=1\dots T}$. At each frame $t$, a set of $N$ pedestrians are detected and their position on the ground plane is $P_i(t)\ ,\ i = 1 \dots N$. For each individual, we define his/her head orientation $\theta_i(t)$. Finally, let us indicate with $\hat{P}_i(t+1)$ the predicted location of the individual $i$ at frame $t+1$.

In order to promote smooth trajectories, we define the *velocity term* ($E_V$) as the summation over all the individuals' of the squared $\ell^2$-norm of the acceleration vector:

$$E_V = \sum_{i=1}^{N} \left\| \frac{d^2 P_i(t)}{dt^2} \right\|^2 = \tag{2}$$

$$= \sum_{i=1}^{N} \left\| \hat{P}_i(t+1) + P_i(t-1) - 2P_i(t) \right\|^2 \tag{3}$$

As for the *destination term* ($E_D$), we consider that a person is consistently looking at his/her short-term destination point while walking. Thus, this term is the additive inverse

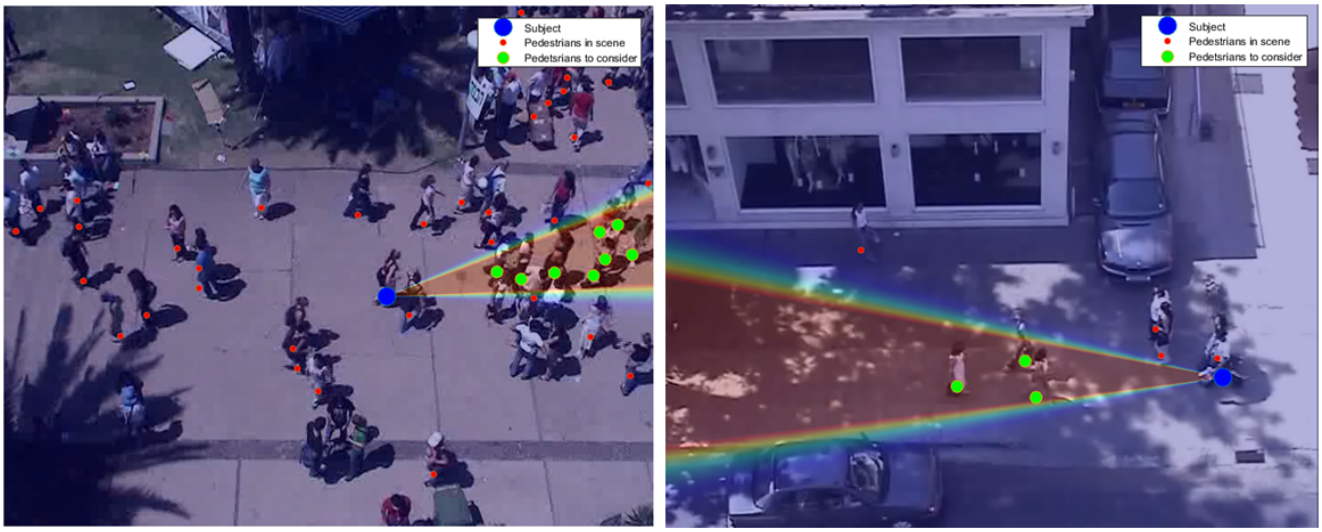


Figure 1. Graphical explanation on the selection of pedestrians to be taken into account for the avoidance term. The large blue dot represents the target pedestrian, the green dots are the pedestrians he/she tries not to collide to, and the small red dots are the pedestrians he/she is not aware of because out of the view frustum. (Best viewed in colors.)

of the cosine of the angle comprised between the gaze direction $\theta_i$, *i.e.* the head pose, and the direction of the predicted velocity:

$$E_D = -\sum_{i=1}^{N} \cos\left(\theta_i(t) - \underline{\angle \hat{P}_i(t+1) - P_i(t)}\right) \qquad (4)$$

where $\underline{\angle \mathbf{v}}$ is the phasor angle of vector $\mathbf{v}$.

For the *avoidance term* ($E_A$), many different models have been proposed in the literature, mostly based on the concept of *social force* [15, 29, 35, 45]. The idea is that a person would not allow another individual to enter his/her personal space; thus, when walking, people adjust their velocity in order to avoid this kind of situations to happen. In this work we model the avoidance potential as a repulsion force that is exponential with respect with the distance between two predicted locations. Unlike many previous works, which consider the repulsion force only when 2 pedestrians are going to be closer than an isotropic comfort area, our method is more biologically motivated, assuming that the pedestrian reacts to what he senses in terms of sight, which is modeled by the VFOA. More formally, this term assumes the summation over all the individuals of the exponential of the minimum distance between the predicted location of the individual itself and the closest predicted location of another individual.

$$E_A = \sum_{i=1}^{N} e^{-\arg\min_j \; d^*_{ij}(t+1)} \;, \text{ with } j \in \mathcal{F}_i(t), j \neq i \qquad (5)$$

where $d^*_{ij} = \|\hat{P}_i(t+1) - \hat{P}_j(t+1)\|^2$, and $\mathcal{F}_i(t)$ is the set of all the individuals inside the VFOA of person $i$ at time $t$. While in theory the view frustum is related to the gaze, we assume that in first approximation, in the scenario we are facing, the gaze is equal to the head orientation. Thus, we model the VFOA as a circular sector of angle 30°, where this last angle has been found experimentally (see in Sec. 4): surprisingly, this angle corresponds to the angle of the human focal attention [16], which can be likened to a "spotlight" in the visual receptive field that triggers higher cognitive processes like object recognition. A graphical explanation of the VFOA is given in Fig. 1.

Thus, the cost function of Equation 1 can be minimized with reference to $\hat{P}_i(t+1), \forall i = 1 \dots N$. So at each step we predict the positions of all the pedestrians in the scene jointly. This optimization problem can be addressed with a *direct search method* for $n$-dimensional unconstrained spaces. The Nelder-Mead simplex method [20], adopted in this work, uses an iterative approach that maintain at each step a non-degenerate simplex of $n+1$ vertices, and updates the simplex according to the function value in the vertices. The method has a very low complexity, since it does not require to compute the gradient (as all the direct search methods) and typically requires the function evaluation on only one or two sample points at each iteration step.

## 4. Experiments

We evaluated our approach on publicly available benchmarks, UCY [22] and TownCentre [7] and compared it against state of the art methods. The benchmark UCY contains three sequences showing two different scenarios. Zara01 and Zara02 sequences show a public street with

Table 1. Dataset Statistics.

| Sequences | # frames | # ped. | # ped. per frame | avg traj. |
|---|---|---|---|---|
| UCY | 5,405 | 434 | 32 | 404 |
| Zara01 | 8,670 | 148 | 6 | 339 |
| Zara02 | 10,513 | 204 | 9 | 467 |
| TownCentre | 4,500 | 230 | 16 | 310 |

shops and cars, the number of pedestrians is quite limited and the trajectories are somehow constrained since entry and exit points are in a limited portion of the image border. UCY sequence is taken in a university campus plaza and it shows a dense crowd moving in several directions without any physical constraint. Similarly, TownCentre dataset portrays a crowded real world city centre scenario. The four datasets have in total of 29,088 frames with 1,016 pedestrians. More details about each sequence are given in Table 1.

The evaluation protocol follows the most recent literature. We first downsample the frame rate of the videos of a factor of 10, resulting in a frame rate of 2.5 fps. Then, for each pedestrian detected, we predict their trajectory for the next 12 frames (4.8 seconds) by considering at every time step the predicted location of the target pedestrian and the ground truth positions of all the others. As for the evaluation metrics, we use the standard *Mean Average Displacement* (MAD) and the *Final Average Displacement* (FAD) error. The MAD metric is given by the average over all the pedestrians and all the frames of the Euclidean distance between the predicted location and the ground truth position. The FAD error is given by the average displacement of the 12-th predicted frame over all the trajectories.

### 4.1. Quantitative results

We compare our method with four state-of-the-art model-based approaches, namely Linear Trajectory Avoidance (LTA) [29], Social Force model (SF) [45], Iterative Gaussian Process (IGP) [40], and multi-class Social Force model (SF-mc) [35]. We also provide results with a baseline method (Lin.) that merely estimates the next locations by using the previous velocity. For a fair evaluation, we need to point out that all the methods use different ground truth data and/or a priori information. All the approaches require the knowledge of the ground truth pedestrian position at each time step. In addition, IGP requires the exact destination point of each pedestrian (*i.e*. the last point of each trajectory, or the point where the pedestrian exits from the scene); LTA, SF and SF-mc require a soft version of the destination point, indeed they only need the direction the individual is pointing (*e.g*. North, South, East or West); SF and SF-mc also require to know which individuals are forming groups.

Differently, our approach does not require the knowledge of destination points or a direction but just the pedestrian position (as the others) and the labelled head orientation of each individual, no group membership is required. The destination point of each pedestrian, as well as other terms in the cost function Eq. 1 are then automatically estimated. We report sample model parameter in Table 4. Since head pose is crucial for forecasting, although people maintain a trajectory to their final destination, there might be the need to take short term deviations in order to avoid collision, obstacles or to engage in human-human interactions (*e.g*. a subject might take few steps in the complete opposite direction of the given destination point). This short term divergence is not addressed in any of the other methods and the head pose seems to be an effective mean towards this end.

Table 2. Mean Average Displacement (MAD) error for all the methods on all the datasets.

| Dataset | Lin. | LTA | SF | IGP | SF-mc | Ours |
|---|---|---|---|---|---|---|
| UCY | 0.57 | 0.51 | 0.48 | 0.61 | 0.45 | 0.38 |
| Zara01 | 0.47 | 0.37 | 0.40 | 0.39 | 0.35 | 0.30 |
| Zara02 | 0.45 | 0.40 | 0.40 | 0.41 | 0.39 | 0.26 |
| Town Centre | 1.3 | 1.8 | 2.1 | – | – | 1.2 |

Table 3. Final Average Displacement (FAD) after 12 frames (4.8 seconds) for all the methods on all the datasets.

| Dataset | Lin. | LTA | SF | IGP | SF-mc | Ours |
|---|---|---|---|---|---|---|
| UCY | 1.14 | 0.95 | 0.78 | 1.82 | 0.76 | 0.78 |
| Zara01 | 0.89 | 0.66 | 0.60 | 0.39 | 0.60 | 0.59 |
| Zara02 | 0.91 | 0.72 | 0.68 | 0.42 | 0.67 | 0.60 |
| Town Centre | 2.7 | 3.67 | 3.8 | – | – | 2.28 |

Table 4. Model parameters obtained from training sequences

| wA | wV | wD |
|---|---|---|
| 0.1 | 1.16 | 1.0184 |

Table 2 and Table 3 show that our method outperforms the state of the art methods in MAD, while it scores worst against the SF-mc on FAD in the UCY sequence. Please note that the comparison with IGP method with the FAD metric is not fair by definition, since it requires the annotation of the final point of each trajectory. Even with the unfair advantage for IGP, in a more densely crowded scenario like UCY, IGP performs poorly, since the short term divergence of a subject is much more prominent and is not addressed by the fixed destination point.

### 4.2. Ablation studies

It is worth noting that all approaches assume that a subject takes the next step accounting for all other pedestrians in the scene. This assumption is far to be true since in normal situations most people are unaware of what is happening behind themselves, and this does not effect their future movements. Thus, to prove the effectiveness of the the view frustum information, we conducted two ablation studies.

First, we turned off the frustum in the avoidance term, taking into account all the pedestrians in the scene. In such a case performances decrease of 2% in MAD and 5% in


Figure 2. Θ angle of the VFOA in relation with the Mean Average Displacement error

Table 5. Mean Average Displacement (MAD) with and without the view frustum condition in the avoidance term.

| Dataset | Ours (no frustum) | Ours |
|---|---|---|
| UCY | 0.41 | 0.38 |
| Zara01 | 0.31 | 0.30 |
| Zara02 | 0.29 | 0.26 |

Table 6. Final Average Displacement (FAD) with and without the view frustum condition in the avoidance term.

| Dataset | Ours (no frustum) | Ours |
|---|---|---|
| UCY | 0.83 | 0.78 |
| Zara01 | 0.65 | 0.59 |
| Zara02 | 0.64 | 0.60 |

Table 7. Mean Average Displacement (MAD) for state of the art methods with destination point estimated from the head orientation.

| Dataset | LTA | SF | Our |
|---|---|---|---|
| UCY | 0.44 | 0.42 | 0.38 |
| Zara01 | 0.33 | 0.32 | 0.30 |
| Zara02 | 0.35 | 0.35 | 0.26 |
| Town Centre | 1.2 | 1.4 | 1.2 |

FAD, showing that the view frustum is beneficial for both metrics in all the sequences. (Table 5 and Table 6)

As a second experiment, we provided to the state-of-the-art approaches the destination points estimated frame-by-frame from the head pose. Results of Table 7, compared with the ones reported in Table 2, demonstrate how the use of head pose is beneficial also for other approaches, improving performances of LTA and SF of 5% and 6% on average respectively.

Fig. 2 shows the study on the span of the Θ angle of the VFOA in relation with the MAD error, when the ground-truth head orientation is known. For this sake, we randomly sample 25 pedestrians per dataset (Zara01, Zara02 and UCY) and we compute the error while modulating Θ from 10 to 75 degrees with a step of 5. As visible in the figure, the range from 10 to 30 gives the best score, with 30 being the best absolute value. Actually, this does correspond to the angle defining the focal attention area [16].

Table 8. Mean Average Displacement error with quantized annotated head pose and with real head pose estimator.

| Dataset | GT | GT(4) | GT(8) | HPE(4) | HPE(8) |
|---|---|---|---|---|---|
| UCY | 0.38 | 0.44 | 0.43 | 0.52 | 0.50 |
| Zara01 | 0.30 | 0.39 | 0.37 | 0.44 | 0.42 |
| Zara02 | 0.26 | 0.35 | 0.34 | 0.39 | 0.38 |
| Town Centre | 1.2 | 1.3 | 1.2 | 1.3 | 1.2 |

### 4.3. Experiments with HPE

Once we have shown the theoretical advantages of our approach, we replace the oracle head orientation with the one estimated from a real head pose estimator [14]. As most of the head pose estimators, the one used in this paper outputs the head pose in a quantized format: dividing the 360° into 4 or 8 classes, thereby we also quantized the ground truth into the same format in order to understand the theoretical bounds that one could reach with the detector.

Looking at the results in table 8 we illustrate that even with the real head pose estimator, we could get competitive results with all the state-of-the-art approaches, which relies on strong ground truth information, highlighting the pragmatism of our approach. Additionally, by quantizating the ground truth we further illustrates that given an accurate pose estimator one could outperform the current state-of-the-art approaches. Moreover, as it can be noticed, finer granularity for head pose estimation proves to be more suitable in trajectory forecasting.

### 4.4. Qualitative results

Besides these quantitative results and ablation studies, we report a qualitative illustration of our predictions in Fig. 3. Along with the proposed approach, we also show trajectories predicted with LTA [29] and SF [45]. Notably, our model is able to better forecast trajectories with highly non-linear avoidance turns, such as to avoid static (3rd row, 3rd column) and moving objects (3rd row, 2nd column), as well as in case a person has to avoid collision with other

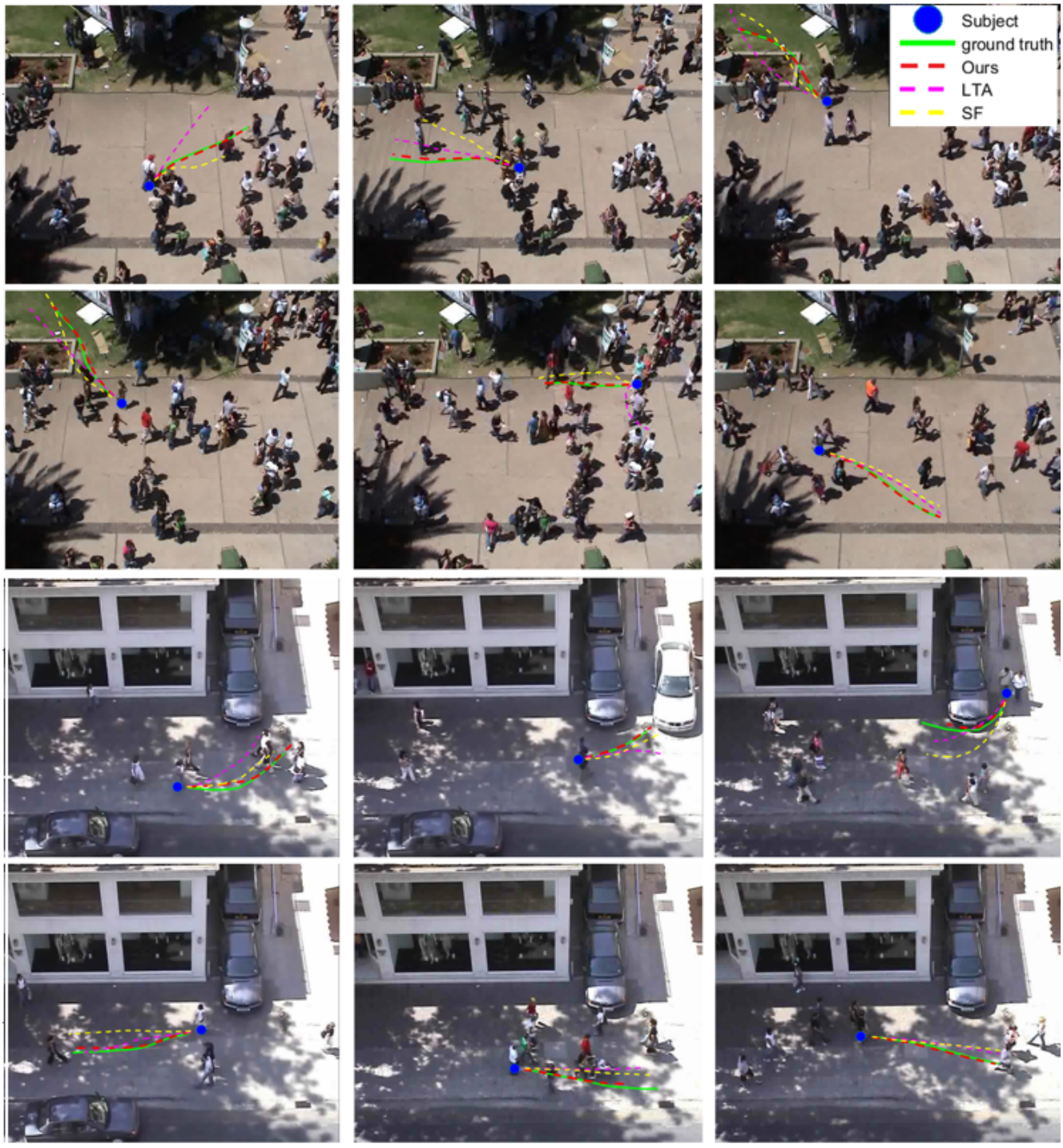


Figure 3. Examples of predicted trajectories on UCY (first two rows) and Zara01, Zara02 (last two rows). Our proposed model is very precise in the prediction of highly non-linear trajectories, where the other approaches such as LTA [29] and SF [45] are less accurate due to the fixed destination points. In particular, our method is able to easily capture short term deviations from the desired path.

pedestrians in the scene (1st, 2nd and 4th rows).

## 5. Conclusion

We have proposed the use of head pose for forecasting the people trajectories, justified the choice from a socio-psychological perspective, introduced a model to exploit the head-pose and a resulted social context, and proved its efficacy with respect to state-of-the-art techniques. The head pose provides the visual frustum of attention (VFOA), which yields the short-term future path of pedestrians, a proxy for their intention.

By means of a simple and intuitive energy-based model, we have proved that having a perfect head pose estimation outperforms the state-of-the-art forecasting performance by in average 8% as (mean average displacement); this can be better appreciated if one considers that, on the same benchmarks, the improvement across 8 years of research has summed to 2.7%. Performance decreases when we adopt real head pose detectors, but it remains at the level of the other forecasting alternatives, which however use ground truth information.

Our proposed system is suitable for real application scenarios, since it does not require information from the future frames. We believe that this direction would potentially motivate more researcher to look into this topic. In addition, we currently ignore grouping activities of the pedestrians, which we would expect to boost performance even more. Further to this topic, we will dedicate future work to evaluating the scaling properties of our system, analyzing scenarios with an increasing crowd density, looking forward to longer time horizons.

**Acknowledgements**: This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement No. 676455.

## References

[1] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *ICML*, 2004.

[2] H. Akaike. Fitting autoregressive models for prediction. *Annals of the institute of Statistical Mathematics*, 21(1):243–247, 1969.

[3] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social LSTM: Human trajectory prediction in crowded spaces. In *CVPR*, 2016.

[4] A. Alahi, V. Ramanathan, and L. Fei-Fei. Socially-aware large-scale crowd forecasting. In *CVPR*, 2014.

[5] S. O. Ba and J.-M. Odobez. A probabilistic framework for joint head tracking and pose estimation. In *ICPR*, 2004.

[6] R. H. Baxter, M. J. Leach, S. S. Mukherjee, and N. M. Robertson. An adaptive motion model for person tracking with instantaneous head-pose features. *IEEE Signal Processing Letters*, 22(5):578–582, 2015.

[7] B. Benfold and I. Reid. Stable multi-target tracking in real-time surveillance video. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3457–3464. IEEE, 2011.

[8] J. F. Caminada and W. J. M. van Bommel. Philips engineering report 43, 1980.

[9] N. Davoudian and P. Raynham. What do pedestrians look at at night? *Lighting Research and Technology*, 44(4):438–448, 2012.

[10] A. D. Dragan, N. D. Ratliff, and S. S. Srinivasa. Manipulation planning with goal sets using constrained trajectory optimization. In *ICRA*, pages 4582–4588, 2011.

[11] S. Fotios, J. Uttley, C. Cheal, and N. Hara. Using eye-tracking to identify pedestrians' critical visual tasks, Part 1. Dual task approach. *Lighting Research & Technology*, 47(2):133–148, 2015.

[12] S. Fotios, J. Uttley, and B. Yang. Using eye-tracking to identify pedestrians' critical visual tasks. part 2. fixation on pedestrians. *Lighting Research & Technology*, 47(2):149–160, 2015.

[13] T. Foulsham, E. Walker, and A. Kingstone. The where, what and when of gaze allocation in the lab and the natural environment. *Vision research*, 51(17):1920–1931, 2011.

[14] I. Hasan, T. Tsesmelis, F. Galasso, A. Del Bue, and M. Cristani. Tiny head pose classification by bodily cues. In *ICIP*, 2017.

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

[16] J. Intriligator and P. Cavanagh. The spatial resolution of visual attention. *Cognitive psychology*, 43(3):171–216, 2001.

[17] J. Jovancevic-Misic and M. Hayhoe. Adaptive gaze control in natural environments. *The Journal of Neuroscience*, 29(19):6234–6238, 2009.

[18] R. E. Kalman et al. A new approach to linear filtering and prediction problems. *ASME Journal of Basic Engineering*, 1960.

[19] K. Kitani, B. Ziebart, J. Bagnell, and M. Hebert. Activity forecasting. In *ECCV*, 2012.

[20] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright. Convergence properties of the nelder–mead simplex method in low dimensions. *SIAM Journal on Optimization*, 9(1):112–147, 1998.

[21] N. Lee and K. M. Kitani. Predicting wide receiver trajectories in american football. In *WACV*, 2016.

[22] A. Lerner, Y. Chrysanthou, and D. Lischinski. Crowds by example. In *Computer Graphics Forum*, volume 26, pages 655–664, 2007.

[23] J. Mainprice, R. Hayne, and D. Berenson. Goal set inverse optimal control and iterative replanning for predicting human reaching motions in shared workspaces. *IEEE Transactions on Robotics*, 32(4):897–908, 2016.

[24] A. Maksai, X. Wang, F. Fleuret, and P. Fua. Non-markovian globally consistent multi-object tracking. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2563–2573. IEEE, 2017.

[25] A. Maksai, X. Wang, and P. Fua. What players do with the ball: A physically constrained interaction modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 972–981, 2016.

[26] D. S. Marigold and A. E. Patla. Gaze fixation patterns for negotiating complex ground terrain. *Neuroscience*, 144(1):302–313, 2007.

[27] P. McCullagh and J. A. Nelder. Generalized linear models, no. 37 in monograph on statistics and applied probability, 1989.

[28] A. E. Patla and J. N. Vickers. How far ahead do we look when required to step on specific locations in the travel path during locomotion? *Experimental brain research*, 148(1):133–138, 2003.

[29] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *ICCV*, 2009.

[30] M. B. Priestley. *Spectral analysis and time series*. Academic press, 1981.

[31] J. Quiñonero-Candela and C. E. Rasmussen. A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research*, 6(12):1939–1959, 2005.

[32] C. E. Rasmussen. Gaussian processes for machine learning. In *Adaptive Computation and Machine Learning*, 2006.

[33] E. Ricci, J. Varadarajan, R. Subramanian, S. Rota Bulò, N. Ahuja, and O. Lanz. Uncovering interactions and interactors: Joint estimation of head, body orientation and F-formations from surveillance videos. In *ICCV*, 2015.

[34] N. M. Robertson and I. D. Reid. Estimating gaze direction from low-resolution faces in video. In *ECCV*, 2006.

[35] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *ECCV*, 2016.

[36] K. Smith, S. O. Ba, J.-M. Odobez, and D. Gatica-Perez. Tracking the visual focus of attention for a varying number of wandering people. *IEEE TPAMI*, 30(7):1212–1229, 2008.

[37] R. Stiefelhagen, M. Finke, J. Yang, and A. Waibel. From gaze to focus of attention. In *VISUAL*, 1999.

[38] T. Taylor, A. K. Pradhan, G. Divekar, M. Romoser, J. Muttart, R. Gomez, A. Pollatsek, and D. L. Fisher. The view from the road: The contribution of on-road glance-monitoring technologies to understanding driver behavior. *Accident Analysis & Prevention*, 58:175–186, 2013.

[39] D. Tosato, M. Spera, M. Cristani, and V. Murino. Characterizing humans on riemannian manifolds. *IEEE TPAMI*, 35(8):1972–1984, 2013.

[40] P. Trautman and A. Krause. Unfreezing the robot: Navigation in dense, interacting crowds. In *IROS*, 2010.

[41] G. Underwood, N. Phelps, C. Wright, E. Van Loon, and A. Galpin. Eye fixation scanpaths of younger and older drivers in a hazard perception task. *Ophthalmic and Physiological Optics*, 25(4):346–356, 2005.

[42] P. Vansteenkiste, G. Cardon, E. D'Hondt, R. Philippaerts, and M. Lenoir. The visual control of bicycle steering: The effects of speed and path width. *Accident Analysis & Prevention*, 51:222–227, 2013.

[43] J. M. Wang, D. J. Fleet, and A. Hertzmann. Gaussian process dynamical models for human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):283–298, 2008.

[44] C. K. I. Williams. Prediction with gaussian processes: From linear regression to linear prediction and beyond. In *Learning in graphical models*, pages 599–621. Springer, 1998.

[45] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg. Who are you with and where are you going? In *CVPR*, 2011.

[46] S. Yi, H. Li, and X. Wang. Understanding pedestrian behaviors from stationary crowd groups. In *CVPR*, 2015.

[47] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. In *AAAI*, 2008.