2018 IEEE Intelligent Vehicles Symposium (IV)
Changshu, Suzhou, China, June 26-30, 2018

Learning to Forecast Pedestrian Intention from Pose Dynamics

Omair Ghori[1,2], Radek Mackowiak[1], Miguel Bautista[2], Niklas Beuter[1], Lucas Drumond[1], Ferran Diego[1] and Björn Ommer[2]

***Abstract*— For an autonomous car, the ability to foresee a humans action is very useful for mitigating the risk of a possible collision. To humans this pedestrian intention foresight comes naturally as they are able to recognize another person's actions just by perceiving subtle changes in posture. Approximating this intention inference ability by directly training a deep neural network is useful but especially challenging. First, sufficiently large datasets for intention recognition with frame–wise human pose and intention annotations are rare and expensive to compile. Second, training on smaller datasets can lead to overfitting and make it difficult to adapt to intra-class variations in action executions. Therefore, in this paper, we propose a real time framework that learns (i) intention recognition using weak-supervision and (ii) locomotion dynamics of intention from pose information using transfer learning. This new formulation is able to tackle the lack of frame-wise annotations and to learn intra-class variation in action executions. We empirically demonstrate that our proposed approach leads to earlier and more stable detection of intention than other state of the art approaches with real time operation and the ability to detect intention one second before the pedestrian reaches the kerb.**

I. Introduction

For modern intelligent systems, enabling a physical autonomous agent to correctly preempt human actions and react appropriately is imperative. In an inner city traffic scenario an autonomous car should be able to foresee a pedestrian's future action and proactively take measures to avoid a potentially dangerous situation. Early intention recognition is therefore key for safe and comfortable driving.

Fig. 1 illustrates an urban traffic scenario where a woman (orange bounding box) is approaching the road. Given a history of visual observations, the task of interest is to predict whether the woman intends to cross the road or stop at the kerb. Based on past experience and anticipation a human observer can anticipate that she is likely to stop at the kerb. This decision is grounded on subtle changes of the observed person [1]. Even if scene context is not visible, as illustrated in the enlarged bounding boxes of Fig. 1, an observer can still infer that she is likely to stop based solely on the subtle changes in posture [2].

Initial works framed intention recognition as a path prediction problem [3], [4], [5], [6], where the focus is to predict the short term path for a pedestrian of interest. Such an approach avoided the difficulties associated with annotating an intention recognition dataset. Annotating each frame in a sequence with intention labels is a challenging, expensive and highly subjective task [1]. As the enlarged bounding boxes of Fig. 1 illustrate, based on a single frame devoid of context, both crossing and stopping intentions look likely. However when viewed as a sequence of images, the true intention becomes clearer. Thus while only a sequential chunk of frames can be annotated for intention, defining the temporal extent of this chunk of frames is a very subjective task. A few earlier methods utilize pedestrian head pose [7], [3] as an additional feature for path prediction. These methods are however, evaluated on staged datasets with actors, where the impact of head pose as a feature for intention prediction cannot be truly gauged. In the wild head pose only implies that a pedestrian is aware of an oncoming car.

Fig. 1: The sequence above illustrates an inner city traffic scene where a person (orange bounding box) approaches the road and eventually stops at the kerb. Our proposed method is able to correctly infer intention one second before the pedestrian reaches the kerb, which for a car traveling at $30kph$ would translate to a braking distance of eight meters. The enlarged bounding boxes focus on the pedestrians without the context of the environment.

In this paper we present a framework for learning pedestrian intention in a weakly supervised manner. Instead of utilizing raw pixel data for inferring intention, we propose to use a high-level structure from a given RGB input frame and subsequently track its temporal dynamics for pedestrian intention recognition. Framing our problem as weakly supervised allows us to overcome the limitations imposed by lack of precise temporal annotations.

[1]The authors are affiliated with with Robert Bosch GmbH, Hildesheim, Germany, (email- Omair.Ghori, Radek.Mackowiak, Niklas.Beuter, Lucas.RegoDrumond, Ferran.DiegoAndilla@de.bosch.com)

[2]The authors are affiliated with Heidelberg University, Heidelberg Germany, (email- miguel.bautista@iwr.uni-heidelberg.de, ommer@uni-heidelberg.de)

Based on the domain knowledge of human posture encoding information about intention, we use human pose as our high-level feature. Although intention recognition using the human skeleton has been studied before [8], [9], [10], these works require detailed 3D pose information for recognition. In contrast we investigate intention recognition in 2D monocular image sequences where groundtruth pose annotations are not available.

The main contributions of this paper are:

- To overcome the lack frame-wise annotations we posit our problem as weakly supervised learning; just a single label per sequence is need for inferring real–time intention recognition per frame.
- We propose human pose as our high-level feature to have a compact feature representation, be interpretable by humans, learn its temporal dynamics, and to have an efficient model for running on embedded hardware.
- In order to enable accurate intention recognition even when using a relatively small dataset for training, we use transfer learning to learn a latent pose representation.

II. Related Work

With the increased focus on achieving fully autonomous driving, pedestrian intention recognition as a means of avoiding possible collisions with pedestrians is very important and has therefore received attention from the research community. Recently Fang et al.[11] utilized pedestrian pose as a feature for intention recognition. In their work they first use a CNN for pedestrian pose estimation. Based on the pose keypoints they compute a total of 396 features based on distances and angles between pairs of keypoints. These features are then used to train a binary classifier for intention recognition. In contrast our approach is completely end-to-end learned with no hand crafted features. Furthermore we utilize the full 14 keypoint skeleton instead of the 9 keypoints use by Fang et al. Most prior works related to pedestrian intention recognition have tended to focus on path prediction [7], [10], [3], [4], [5], [6], [12] instead of explicit intention recognition. Schneider et al. [4] investigated various Kalman filter based models for pedestrian path prediction in a one second future time window. The proposed approach leveraged features extracted from pedestrian motion dynamics only. Kooij et al. [7] extended this initial work by incorporating additional features which approximate the pedestrians behaviour as well as the environmental context. Two non-linear methods for estimating crossing intention of a laterally approaching pedestrian were introduced by Keller et al. [6]. Probabilistic Hierarchical Trajectory Matching (PHTM) tries to find the best match for a partially observed pedestrian motion track among a database of motion snippets. The closest matching pedestrian trajectory is then used as a model for approximating future pedestrian position. For longer term pedestrian path prediction, Rehder et al. [13] frame the problem as one of planning and treat the pedestrians destination as a latent variable. Using inverse reinforcement learning [14] investigates trajectory prediction for multiple people. However, this works only for a fixed surveillance camera setup while we work with a moving camera.

Gaussian Process Dynamical Models (GPDM) have also been demonstrated to be effective using dense optical flow features and pedestrian motion dynamics as input features [6] and body pose as an input feature [10]. Head pose has also been used as a complementary feature for intention recognition [3] [15]. However, both these approaches are only ever evaluated on a dataset of staged scenarios. Hariyono et al. [16] investigated intention recognition by developing a model for a walking human. Unlike our work however, they do not explicitly estimate human pose but instead rely on statistics extracted from a pedestrian bounding box.

III. Proposed Approach

A. Problem Formulation

Having a sequence of images from time 0 to T, $\mathbf{F} = \{F_t : t = 0, ..., T\}$, we are interested in recognizing the intention, y_t, of an observed pedestrian at time t before an action occurs at time T, where $t < T$. Due to the ambiguity of detecting intention from a single frame, we aim to infer the intention class with the maximum probability given a sequence of frames prior to occurrence of action. Thus estimating the intention up to frame F_t is formulated as maximum a posteriori Bayesian inference problem,

$$y_t^{MAP} = \underset{y_t \in \{1,...,M\}}{\arg\max} \; P(y_t|\mathbf{F_{0:t}}), \tag{1}$$

where M is the number of predefined intention categories, and $P(y_t|\mathbf{F_{0:t}})$ is the probability of intention class y_t given all the frames up to frame t. Instead of explicitly computing the probabilities, we take the softmax over the model outputs, which gives us the distribution over the M possible classes as:

$$P(y_t|\mathbf{F_{0:t}}) \approx \frac{\exp h_{y_t}(\mathbf{F_{0:t}})}{\sum_{m=1}^{M} \exp h_m(\mathbf{F_{0:t}})}, \tag{2}$$

where $h_{y_t}(\mathbf{F_{0:t}})$ are the features extracted from frames 0 to t for intention y_t. Due to the recurrent nature of the problem, we are able to estimate the intention class for each frame up to t,

$$Y_t^{MAP} = \underset{Y_t \in \mathcal{M}}{\arg\max} \; P(Y_t|\mathbf{F_{0:t}}), \tag{3}$$

where $\mathcal{M}$ is the set of all possible intention annotations, $Y_t = [y_0, \ldots, y_t]$. The probability of an intention class given a sequence of frames,$P(Y_t|\mathbf{F_{0:t}})$, can be approximated as

$$P(Y_t|\mathbf{F_{0:t}}) \approx \prod_{j=0}^{t} P(y_j|\mathbf{F_{0:j}}), \tag{4}$$

$$\approx P(y_0|F_0) \prod_{j=1}^{t} P(y_j|y_{j-1}, F_j), \tag{5}$$

where $P(0y_j|y_{j-1}, F_t)$ is the probability of the intention class at frame j given the current frame F_j and the previous intention class, y_{j-1}.

The former approximation, Eq.(4), estimates each intention class independently, requiring the computation of all spatio–temporal features $h_{y_t}(\mathbf{F_{0:t}})$ at each timestep before estimating the softmax in Eq.(2). In contrast, the later approximation, Eq.(5), infers the intention class based on the current frame and the previous intention, allowing the extracted features, $h_{y_t}(\mathbf{F_{0:t}})$, to be calculated recurrently, independent of the total number of time steps. Despite reducing the computational cost of the feature extractor, Eq.(5) is not able to learn long–term dynamics of the intention. Therefore, following the idea of computing the feature extractor recurrently, we decouple $P(Y_t|\mathbf{F_{0:t}})$ into two components, a visual and a temporal feature extractor. The former captures spatial frame-wise characteristics of the intention; whereas temporal feature extractor captures long–term dynamics of the intention given the visual features; hence given a sequence of frames before the action actually begins we learn over the temporal dynamics of features indicative of the persons intention. As the classifier starts to see those features occurring during test time, the probability of the correct class will begin to increase.

B. Visual Feature Extraction

As a first step towards intention recognition a feature representation of the visual contents of frame, F_t, needs to be extracted. Normally, both intentions and their associated actions can be well represented by generic visual descriptors [17]. These feature representations work for action classes which differ greatly in execution. In contrast, we focus on human pose as a compact visual feature descriptor. The proposed feature encodes information about a persons intention [2], and helps in identifying subtle differences in motor movements.

The challenge then is to estimate the pose of a pedestrian q in a given frame F_t. In order to do this, first an off-the-shelf pedestrian detector is used to obtain a bounding box image, I_t^q, of pedestrian q. Given this bounding box image as the input of a feature transformation we estimate coordinates of a 14 keypoint human skeleton as defined in [18]. To this end we learn a function, ϕ, parametrized by parameters θ, to regress the real valued vector output, $\hat{Z}$, corresponding to pose of pedestrian q, from the grayscal input image I_t^q,

$$\hat{Z}^q = \phi(I_t^q; \theta). \tag{6}$$

Specifically we train a Convolutional Neural Network (CNN), represented by the transformation ϕ for estimating the pose from the obtained bounding box image. The CNN architecture is similar to the one proposed by Belagiannis et al. [19]. Fig. 2 summarizes the architecture we utilize for pose estimation in the box marked Pose CNN. The top row of Fig. 3 illustrates the pose estimation output of the network on frames of a sequence in our dataset.

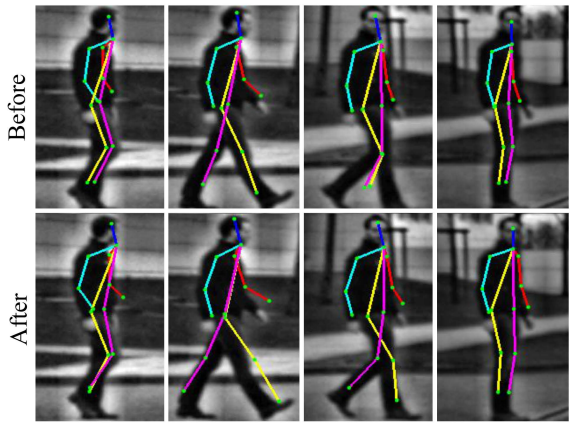


Fig. 3: The top row shows the posture as estimated by our pose CNN after training on a standard pose dataset. The bottom row refinement in pose after end-to-end intention recognition learning. The network is never explicitly trained for pose on the intention dataset.

For parameter learning, Tukey's biweight loss function [19] is minimized. This loss function is defined as:

$$\rho(r_i) = \begin{cases} \frac{c^2}{6}\left[1-\left(1-\left(\frac{r}{c}\right)^2\right)^3\right] & \text{, if } |r| \leq c \\ \frac{c^2}{6} & \text{, otherwise} \end{cases}, \tag{7}$$

where r is the residual, defined as $r = Z - \hat{Z}$ and c is a tuning constant. For regression tasks, Tukey's biweight loss function is more robust to the influence of outliers than the more commonly used $L2$ loss. We set an initial learning rate of 0.01 and use AdaGrad [20] for optimizing the network.

C. Intention Recognition Learning

In the intention recognition step we aim to classify the intention of pedestrian q based on the feature representation, $\mathbf{Z^q_{0:t}}$, extracted from the input frames $\mathbf{F_{0:t}}$ as described in Sec. III-B. For notational simplicity, we denote $\mathbf{Z_{0:t}} = \phi(\mathbf{F}_{0:t}, \theta)$ to represent the features extracted from pedestrian q. Given a sequence of inputs $\mathbf{Z_{0:t}}$, we learn a function, Φ, that maps the temporal dynamics of the input to a

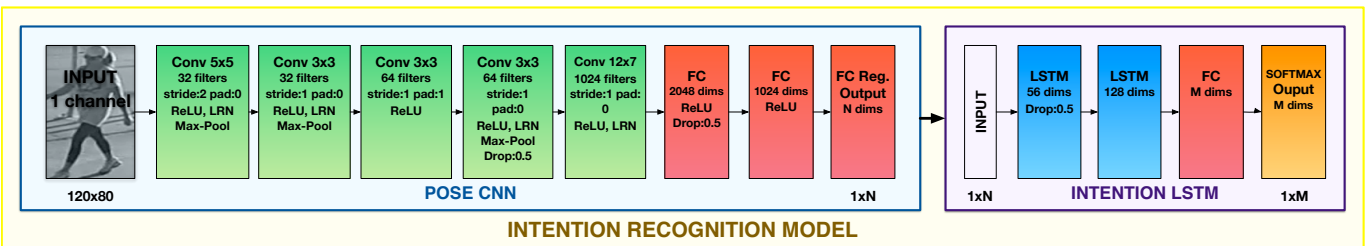


Fig. 2: The network architecture for the pose estimation CNN and the intention recognition network.

Fig. 4: A snapshot showing the variation in appearance of pedestrians in our compiled dataset. Youtube and Self recorded were captured in the wild while Daimler and Hanau are with actors.

M-dimensional output score vector and gives the distribution over the M possible intention classes.Therefore, for intention recognition we can reformulate Eq. (3) as,

$$Y_t^{MAP} = \arg\max_{Y_t \in \mathcal{M}} P(Y_t|\mathbf{F_{0:t}}) = \arg\max_{Y_t \in \mathcal{M}} \prod_{i=0}^{T} \Phi(F_i), \quad (8)$$

where

$$\Phi(F_i) = \frac{\exp h_{y_i}(Z_i, h(Z_{i-1}, \dots); \Omega)}{\sum_{m=1}^{M} \exp h_m(Z_i, h(Z_{i-1}, \dots); \Omega)}, \quad (9)$$

and $h_{y_i}(Z_i, h(Z_{i-1}, \dots); \Omega)$ represents the accumulated spatio–temporal features of the intention y_i up to frame i, and $h(\phi(F_i; \theta), h(\phi(F_{i-1}; \Omega)$ represents the accumulated M-dimensional output score vector. For modeling the required complex long–term temporal dynamics from the input pose, the function $h_{y_i}(Z_i, h(Z_{i-1}, \dots); \Omega)$ is modeled with a shallow LSTM network parametrized by Ω. The LSTM takes as input the estimated human posture, $Z_i = \phi(F_i; \theta)$, at frame i and together with its hidden internal state models the temporal dynamics of the human pose. The network structure is as defined in Fig. 2 in the box marked Intention LSTM. The second LSTM layer is followed by a fully connected layer which outputs a M-dimensional feature vector of scores for all intention classes at each time step.

Before we can start training our network two obstacles must be addressed, namely the lack of posture annotations and frame level intention labels for any of the frames in any of the sequences in the dataset.

We tackle the first issue by utilizing transfer learning. Specifically, the network specified by Eq. (6) is first trained on a standard pose training dataset [18] and then used as an initialization for the Pose CNN in Fig. 2. During end-to-end training the layers in this portion of the overall model receive a smaller learning rate than the recurrent layers, thus the convolutional parameters are fine tuned for intention recognition during training thereby learning more relevant and discriminative features. The effect of this fine tuning on pose estimation is visible in the bottom row of Fig. 3. The lack of frame level intention annotations is dealt with by treating our task as weakly supervised. Only a sequence label, Y_T, is provided which reflects the intention at the final time step, T. We therefore perform back propagation through time only at the final time step of a sequence. Optimizing in this manner ensures that we make no assumptions about when the intention begins to manifest and thereby avoid any bias that might be introduced due to subjective labeling. Therefore, intention recognition is formulated to minimize:

$$\hat{\Omega}, \hat{\theta} = \arg\min_{\Omega, \theta} \mathcal{L}(Y_T, \Phi(F_T; \Omega, \theta)), \quad (10)$$

where $\mathcal{L}$ is the cross-entropy loss between the correct intention class, Y_T, encoded by a one-hot label vector, and the probabilities of the intention class at end of the sequence, $\Phi(F_T; \Omega, \theta)$. For end-to-end training the LSTM network has an initial learning rate set to 0.01 and is optimized with back propagation through time and Adagrad [20].

IV. Experiments

A. Datasets

For evaluating our proposed method we compile a dataset of sequences of pedestrians walking towards the road. It leverages two existing datasets, namely Daimler [4] and Hanau [21] datasets, both of which contain sequences recorded with instructed actors. Moreover, we extend them with sequences of pedestrians in real world traffic scenarios.

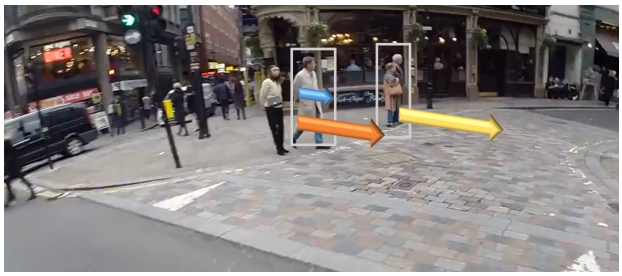
The real world sequences contain recordings from United Kingdom, Canada, United State of America, Germany, Turkey, China and Pakistan. A snapshot of pedestrians in our datasets can be seen in Fig. 4. The diverse appearance and attitude towards road safety make this a particularly difficult dataset. Additionally, the dataset contains sequences recorded at day and night time as well as in cloudy and sunny conditions. In total we have 466 sequences, with 270 sequences reserved for training, 35 for validation and 161 for the test, taking care to preserve any predefined split of the Daimler and Hanau datasets. The real world sequences contain 58 sequences downloaded from YouTube and 315 recorded by us. All sequences were resampled at a frame of 16 frames per second.

In addition to the intention classes introduced by [4], namely: crossing, stopping, starting and turning, we also include the 'walking along' class. This class refers to the case when a pedestrian walks along the road longitudinally. The semantic meaning of each class is illustrated in Fig. 5. We use this dataset for all our evaluations. Reported results are for trainings and evaluations which utilize groundtruth pedestrian bounding boxes as inputs.

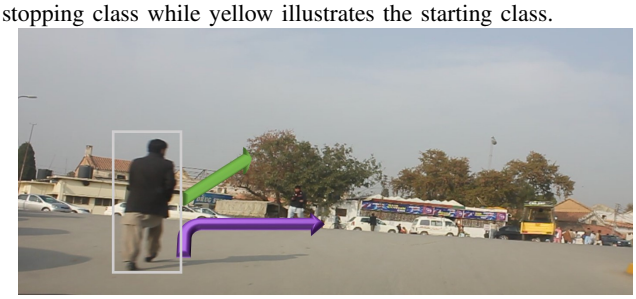
B. Baselines

We contrast our approach with several other methods which do not utilize explicitly explainable features such as pose. The idea is to highlight how domain knowledge, i.e. pose encoding intention, allows us to estimate pedestrian intention more efficiently and robustly.

C3D/SVM: A simple baseline is established by extracting $fc6$ features using the C3D [22] net and training a linear SVM classifier on top. A sliding window of 6 frames, with



(a) The orange arrow represents the crossing class, blue the stopping class while yellow illustrates the starting class.



(b) The green arrow represents the walking-along class while purple shows the turning class.

Fig. 5: The figure shows the five pedestrian intention classes where the arrows only illustrate the direction of pedestrian movement.

a stride of one is run over each sequence in order to extract spatio-temporal features. Wider observation time window widths were also tested but they negatively affected the detection of fast changing intention such as a running person coming to a stop. We therefore settled on a time observation window width of 6.

CNN Slow Fusion (CNN SF) [23]: We train a CNN based on the SF architecture for end-to-end intention recognition. The network takes as input multiple frames and outputs the intention class probabilities. Best results were achieved with a sliding window of 6 frames.

CNN+LSTM variations: This method is similar to our approach except a generic CNN based feature extractor, pre-trained on ImageNet, is used in conjunction with a recurrent network.

Pose Feature/SVM [11]: This method initially uses a state of the art pose estimator [24] trained on the MS-COCO dataset [25] for estimating the human pose. The estimated pose is then used to calculate 396 features (distances and angles between body keypoints). These features are then used to train a SVM for classifying the probability of different intentions.

Probabilistic Hierarchical Trajectory Matching (PHTM) [6]: aims to find the best match for a partially observed pedestrian motion track from among a database of previously observed motion snippets. Then closest matching snippet is then used as a model for extrapolating future pedestrian position for the current observed track. However, this approach requires an offline predestrian trajectory data, being only available for Daimler Dataset.

Latent-dynamic Conditional Random Field (LD-CRF) [3]: also takes as input pedestrian motion dynamics. In addition it also utilizes pedestrian head pose as an input; being a proxy for situational awareness.

C. Intention Recognition Results

From an Advanced Driver Assistance Systems (ADAS) point of view, the two most important classes are a pedestrian crossing the road or stopping at the kerb. All other classes form a subset of these two major classes. Our evaluation is therefore performed at two levels of granularity with respect to the intention classes: a binary classification problem, where the goal is to predict whether an approaching pedestrian plans on stopping or crossing in front of the ego vehicle and a second more in depth multi-class classification scenario with all labeled classes being considered.

Binary classification is evaluated on the basis of the F1-score across three time horizons: one second, half a second and one frame before event. The single frame before event case represents a time interval dependent on the frame rate. The time horizons were chosen keeping in mind realistic urban driving scenarios as well as the erratic nature of pedestrian movement.

Multi-class evaluation analyzes how the probabilities of each intention class varies with respect to Time to Event (TTE) [4]. This metric has previously been employed for measuring performance of pedestrian intention recognition methods [3], [4], [6], [7].

Binary Class Analysis: Table I shows the F1-score across three time horizons for the stopping and crossing classes. It is clear from the results that our proposed model outperforms the baseline methods across all three time horizons. Of particular significance is our superior intention recognition performance one second before the event. The results table reflects that the spatio-temporal features extracted using the C3D net are informative and help in distinguishing between the two intention classes only as the time of event gets closer. One second before the event the C3D based model is biased towards the crossing intention class. As the time of event approaches and pedestrian appearance begins to differ for both classes the accuracy of the C3D based method increases. Furthermore from Table I we can see that the F1 score using C3D/SVM for crossing class fluctuates as time of event approaches. This is a direct consequence of operating on a fixed observation time window. When reasoning about

Method	F1-Score wrt. Time to Event for stopping / crossing		
	1 second	1/2 second	1/16 second
Random	0.14 / 0.11	0.14 / 0.11	0.14 / 0.11
C3D/SVM [22]	0.57 / 0.71	0.60 / 0.59	0.77 / 0.79
CNN SF [23]	0.33 / 0.40	0.47 / 0.52	0.64 / 0.69
VGG-M [26]+LSTM	0.48 / 0.52	0.43 / 0.47	0.43 / 0.46
Latent Pose CNN+LSTM (ours)	**0.71** / **0.72**	**0.72** / **0.73**	**0.87** / **0.85**

TABLE I: F1-score for Pedestrian intention recognition on stopping and crossing classes. Our method outperforms all other evaluated methods with stopping intention being accurately detected one second before event by a significant margin.

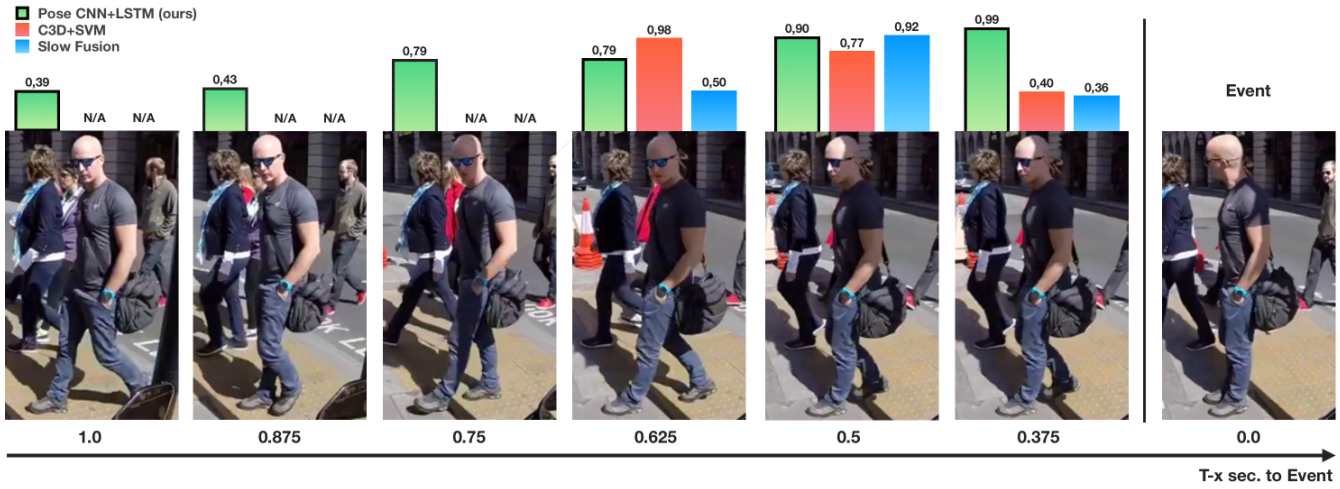


Fig. 6: Variation in stopping probability for a stopping scenario for our method and selected baseline methods. For the other methods an initial time lag, associated with frame accumulation, is also seen before a valid output is available.

intention, frames outside of the observation window are not considered due to which larger variance is observed in output probabilities at each time step. Fig. 6 highlights the probability variation with respect to time for a stopping scenario. Initially there are no visible signs of the pedestrian stopping; our model therefore outputs a higher probability for crossing intention. In contrast, C3D/SVM and Slow Fusion both operate on a chunk of frames and initially their output is not available as the required number of frames are not accumulated. In subsequent frames, our model starts to detect subtle changes of posture like a reduction of step size and leaning back slightly, and hence infers a stopping intention with high probability 0.75 seconds before event.

We extend our baseline comparisons to include methods which utilize pedestrian motion dynamics. Specifically we implemented an approach similar to PHTM [6] and a Latent-dynamic Conditional Random Field (LDCRF) based approach [3]. Also included are results from the Pose feature/SVM [11] approach as it performs evaluation on the same Daimler data subset. This evaluation is performed only for the Daimler [4] subset of the data as it contains pedestrian trajectory information. Fig. 7 shows the variation in average stopping probability for all stopping scenarios. Our posture based approach reacts much more quickly when changes indicative of stopping are seen. From the curve for the LDCRF based approach it can be seen that while head pose plays a part in determining intention, change in head pose occurs much later than changes in posture associated with stopping. The Pose feature/SVM [11] approach performs well on this particular subset, however after time of event the probability of stopping anomalously begins to decrease whereas for the other methods the probability increases.

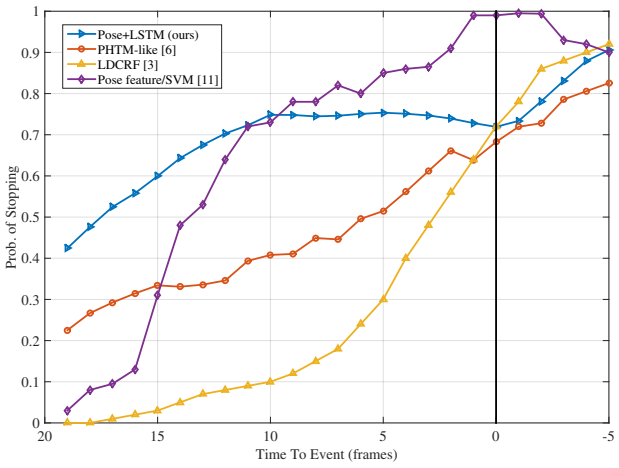
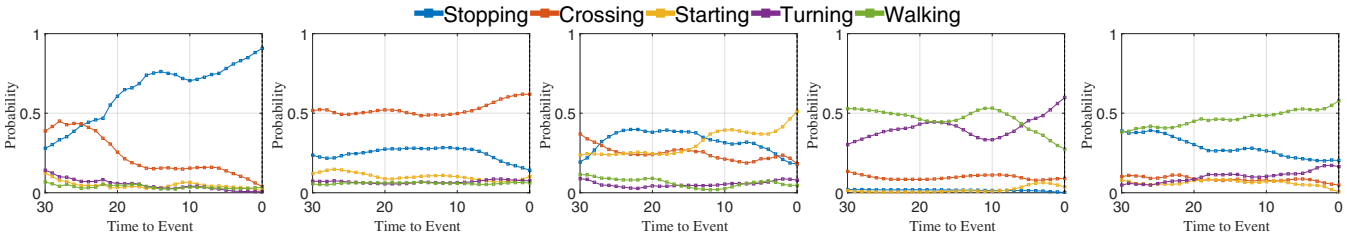


Fig. 7: The curves show the average change in probability of stopping for all Daimler [4] stopping scenarios. Our method reacts more quickly with stopping probability increasing earlier as TTE decreases. Negative TTE values indicate frames after the event occurrence.

Multi-Class Analysis: The results for multi-class analysis are presented in the form of probability graphs as illustrated in Fig. 8. From the graphs it can be seen that as time to event decreases the probability of the correct class increases. Of particular interest are the turning and walking case which have a large semantic overlap. Both actions initially consist of a person walking longitudinally along the road. At a later point in time the person for the turning class suddenly turns on to the road. This case is reflected in Fig. 8d where the initially the probability of walking is higher than the probability of turning but as TTE gets smaller the probability of turning begins to increase quickly. Similar behaviour can be seen for the stopping and starting class.

Run Time Comparison: Table II shows the results for average running time for all the considered methods on a CPU as well as a GPU (Nvidia GTX Titan X). Owing to its tiny size in comparison to the other networks, the slow fusion architecture has the fastest average running time on the CPU. On the GPU our network is the fastest and is able to function in real time for the pedestrian intention recognition use case and fit on embedded hardware. The run time does not include the time needed for pedestrian detection.



(a) Stopping Scenario (b) Crossing Scenarios (c) Starting Scenarios (d) Turning Scenarios (e) Walking Scenarios

Fig. 8: The five graphs above illustrate the variation in average probability for each class for the five different scenarios. As the TTE decreases the probability of the correct class begins to rise.

Method	Running Time in ms	
	CPU	GPU
C3D/SVM [22]	1330 ms	10 ms
CNN SF [23]	**13 ms**	10 ms
VGG-M [26]+LSTM	490 ms	8 ms
Latent Pose CNN+LSTM (ours)	97 ms	**6.1 ms**

TABLE II: Average running times for each method for intention recognition on a per frame basis on both CPU and GPU. As can be clearly seen our network is the fastest on the GPU. These run times do not take into account the time needed for pedestrian detection

Method	F1-Score 1 second before Event
Pose CNN(random init.)+LSTM	Undefined
Pose CNN (fixed)+LSTM	0.64
Pose estimate[24]+LSTM	0.66
VGG-M [26]+LSTM	0.48
Latent Pose CNN+LSTM (ours)	**0.71**

TABLE III: F1-score on the Pedestrian Intention Recognition dataset one second before stopping. The CNN+LSTM trained from random initialization completely overfits on the crossing class and fails to recognize any stopping sequence.

D. Ablation Analysis

To investigate the benefit of using pose as an intermediate feature for intention recognition we extend our experiments. We firstly train the recurrent portion of our network with pose features as predicted by an off the shelf state of the art pose estimation CNN [24]. In addition we also train the recurrent network from a random initialization with the convolutional portion of the network being frozen. Together these experiments allow us to gauge the impact of end to end training, how the quality of predicted pose affects our final intention recognition performance as well as the benefit of allowing the model to abstract the pose feature. For completeness the full model is trained from a random initialization as well. In this case, given the relatively small amount of data and the large number of trainable parameters overfitting is a concern.

The results of our experiments are summarized in Table III. Results for the VGG-M+LSTM network variant are also included for comparison. Our proposed network together with the staggered semi-supervised approach to intention recognition training outperforms other variations to the network structure or training methods. The F1-score for the Pose CNN+LSTM trained from a random initialization illustrates that directly training a deep network using the small available dataset leads to strong overfitting. Furthermore utilizing a pre-trained feature extractor such as VGG-M and fine tuning that still leads to overfitting. It is important to point out that our network has almost eight times fewer parameters than the VGG variant. The LSTM network trained with the output of a state of the art pose estimation network [24] gives competitive results but the achieved performance is still lower than our method.

V. Conclusion and Future Work

In this paper we propose an efficient intention recognition method from pose dynamics. We learn to specifically infer pedestrian intention using weak–supervised learning that bypasses the difficulties of labeling a dataset: (i) the subjective nature of labeling intention sequences and (ii) the cost associated with labeling a sufficiently large dataset. Furthermore, we learn the locomotion dynamics of intention from pose information using transfer learning. Pose offers a compact feature representation which is interpretable by humans and also allows us to have a small model that runs on embedded hardware. Our proposed approach outperforms current state-of-the-art intention recognition systems in terms of accuracy and is able to accurately infer pedestrian intention one second before a pedestrian reaches the road, a capability which is invaluable for urban autonomous driving.

References

[1] S. Schmidt and B. Färber, "Pedestrians at the kerb recognising the action intentions of humans," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 12, no. 4, pp. 300 – 310, 2009.

[2] J. M. Kilner, "More than one pathway to action understanding," *Trends in Cognitive Sciences*, vol. 15, no. 8, p. 352, 2011.

[3] A. Schulz and R. Stiefelhagen, "Pedestrian intention recognition using latent-dynamic conditional random fields," in *Intelligent Vehicles Symposium (IV), 2015 IEEE*, 2015, pp. 622–627.

[4] N. Schneider and D. Gavrila, "Pedestrian path prediction with recursive bayesian filters: A comparative study," in *Pattern Recognition*, 2013, vol. 8142, pp. 174–183.

[5] S. Kohler, M. Goldhammer, S. Bauer, K. Doll, U. Brunsmann, and K. Dietmayer, "Early detection of the pedestrian's intention to cross the street," in *Intelligent Transportation Systems (ITSC), 2012 15th International IEEE Conference on*, Sept 2012, pp. 1759–1764.

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

[7] J. Kooij, N. Schneider, F. Flohr, and D. Gavrila, "Context-based pedestrian path prediction," in *Computer Vision ECCV 2014*. Springer International Publishing, 2014, vol. 8694, pp. 618–633.

[8] Y. Li, C. Lan, J. Xing, W. Zeng, C. Yuan, and J. Liu, *Online Human Action Detection Using Joint Classification-Regression Recurrent Neural Networks*, 2016, pp. 203–220.

[9] M. Zanfir, M. Leordeanu, and C. Sminchisescu, "The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection," in *2013 IEEE International Conference on Computer Vision*, Dec 2013, pp. 2752–2759.

[10] R. Quintero, I. Parra, D. F. Llorca, and M. A. Sotelo, "Pedestrian intention and pose prediction through dynamical models and behaviour classification," in *IEEE 18th International Conference on Intelligent Transportation Systems (ITSC), 2015*.

[11] Z. Fang, D. Vázquez, and A. M. López, "On-board detection of pedestrian intentions," *Sensors*, vol. 17, no. 10, p. 2193, 2017.

[12] H. Kataoka, Y. Aoki, Y. Satoh, S. Oikawa, and Y. Matsui, "Fine-grained walking activity recognition via driving recorder dataset," in *IEEE 18th International Conference on Intelligent Transportation Systems (ITSC), 2015*, 2015.

[13] E. Rehder and H. Kloeden, "Goal-directed pedestrian prediction," in *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, Dec 2015, pp. 139–147.

[14] W.-C. Ma, D.-A. Huang, N. Lee, and K. M. Kitani, "Forecasting interactive dynamics of pedestrians with fictitious play," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[15] F. Flohr, M. Dumitru-Guzu, J. Kooij, and D. Gavrila, "Joint probabilistic pedestrian head and body orientation estimation," in *Intelligent Vehicles Symposium Proceedings, 2014 IEEE*, June 2014, pp. 617–622.

[16] J. Hariyono and K.-H. Jo, "Detection of pedestrian crossing road: A study on pedestrian pose recognition," *Neurocomputing*, vol. 234, pp. 144 – 153, 2017.

[17] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[18] D. Hall and P. Perona, "Fine-grained classification of pedestrians in video: Benchmark and state of the art," in *CVPR*, 2015.

[19] V. Belagiannis, C. Rupprecht, G. Carneiro, and N. Navab, "Robust optimization for deep regression," in *International Conference on Computer Vision (ICCV)*, 2015.

[20] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, 2011.

[21] D. Kondermann, R. Nair, S. Meister, W. Mischler, B. Güssefeld, S. Hofmann, C. Brenner, and B. Jähne, "Stereo ground truth with error bars," in *Asian Conference on Computer Vision, ACCV*, 2014.

[22] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *The IEEE International Conference on Computer Vision (ICCV)*, 2015.

[23] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014.

[24] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.

[25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 740–755.

[26] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *British Machine Vision Conference*, 2014.