EMIDAS: Explainable Social Interaction-Based Pedestrian Intention Detection Across Street

Nora Muscholl, Matthias Klusch, Patrick Gebhard, Tanja Schneeberger
firstname.lastname@dfki.de
German Research Center for Artificial Intelligence
Saarland Informatics Campus, Saarbrücken, Germany

ABSTRACT

An explainable, accurate, and fast prediction of pedestrian movements in streets is an essential requirement for self-driving cars and remains a daunting challenge. Current algorithmic approaches rely solely on visual information. The information about social interaction between pedestrians across the street is not considered yet. The intention to cross the street can be influenced by social interaction with another pedestrian across the street, which comes with observable social signals such as hand waving. This paper presents EMIDAS, a dynamic Bayesian network model that uses various social signals to predict the intention to meet another pedestrian across the street. For training and evaluating this model, we adopted typical procedures from the area of social signal analysis, which consists of collecting real prototypical scenarios, annotating them concerning the pedestrians' intention to cross the street, and creating scenes from the car's field of view to test the model. This approach's benefit is that it can be employed to explain the reasoning and its underlying knowledge base. Both aspects are essential for future self-driving cars, especially when considering that such future cars have to maintain a level of trust towards the car's passengers.

KEYWORDS

Autonomous driving, pedestrian intention estimation, social interaction

ACM Reference Format:
Nora Muscholl, Matthias Klusch, Patrick Gebhard, Tanja Schneeberger. 2021. EMIDAS: Explainable Social Interaction-Based Pedestrian Intention Detection Across Street. In *The 36th ACM/SIGAPP Symposium on Applied Computing (SAC '21), March 22–26, 2021, Virtual Event, Republic of Korea.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3412841.3441891

1 INTRODUCTION

We live in a social world. Communication and behavior are inherently social. When interacting with others across the streets, we emit social signals as we do in close face-to-face interactions. As shown by researchers, social signals, such as gestures, body posture, or gaze, play a significant role, even more than related verbal counterpart [6, 15]. This observation can be employed for a computational situational assessment of pedestrian behavior in everyday traffic situations.

Traditionally, machine learning approaches are used to predict pedestrian movements within traffic situations (e.g., [11, 13, 22]). Most of them rely on observable pedestrian dynamics features. Few of them use the scene context or the social context to estimate one or more pedestrians' movement trajectories. Albeit it seems obvious, the observation of reciprocal social interaction across the street and related signals have not yet been considered for this task.

As a first approach towards this goal, we present EMIDAS, a computational model for explainable multi-pedestrian interaction estimation across streets. We employ a dynamic Bayesian network (DBN) model that encodes causal-effect relationships between network nodes while including previous observations. DBNs inherently allow retracing decisions, which allows generating explanations on various levels of granularity. A handy feature concerning possible application scenarios, e.g., self-driving cars that are able to explain action decisions. The EMIDAS DBN nodes and their relationships are modeled based on expert knowledge towards the goal to correlate observable social signals with the pedestrian intention to cross the street. EMIDAS is the first approach that represents such correlations in a DBN. The EMIDAS model was successfully integrated into the current state of the art pedestrian path prediction SIMP3 approach, improving the overall prediction performance significantly [16]. SIMP3 is the first socially-aware multi-pedestrian path predictor that takes observed social interaction between pedestrians on opposite sides of the street into account.

In the first step, we identified relevant street scenarios with two pedestrians interacting across the street, showing social signals such as gestures and gaze. A user study checks the street scenario's realism and collects the ground truth of the occurring pedestrians' intention to cross the street to meet the other pedestrian. Within the study, we also query reasons and relation to social signals behind the assessment of a pedestrian crossing a street. The results confirmed our expert-based EMIDAS DBN model.

To get the needed amount of material to train the EMIDAS model parameters, synthetic variations of real scenes are created. Based on them, the model parameters are learned. The model's technical evaluation implies that most scenarios' prediction has a strong positive linear relationship with the collected ground truth.

SAC '21, March 22–26, 2021, Virtual Event, Republic of Korea

ACM ISBN 978-1-4503-8104-8/21/03…$15.00
https://doi.org/10.1145/3412841.3441891

2 BACKGROUND AND RELATED WORK

2.1 Pedestrian Intention Estimation

Pedestrian intention estimation comprises several sub-problems: the prediction of the pedestrian's future path, their goal location, or their most likely next action (walking, standing, crossing). In the context of autonomous driving, an incorrect prediction is particularly critical when the predictor fails to forecast that the pedestrian will step onto the road in front of the vehicle. Current (multi-) pedestrian path prediction methods make use of various features of pedestrian dynamics (e.g., pedestrian position, moving direction, velocity) [2, 3, 11–13, 22], scene context (e.g., distance to curb, traffic light state, crosswalks) [11–13, 22] and social context (e. g., distance to others [2, 13], pedestrian map [3, 11, 12, 22]) in order to learn to estimate the future path of pedestrians. The majority of methods rely on the past trajectory of the pedestrians as a main feature. Therefore, these methods are only effective if a pedestrian is already about to cross the street. Some other methods include the interaction between pedestrians into the feature space to consider behavior that meets social rules, e. g., keeping a comfortable distance from strangers. In addition to path prediction, Rasouli et al. [19] present a method that estimates the intention of pedestrians to cross the street using the pedestrians' appearance, their immediate local surroundings and their motion. However, we have found that a false prediction can occur, especially if a pedestrian *suddenly* crosses the street driven by the intention to meet with others. This intention can be correlated with the pedestrians socially interacting with each other beforehand through exchanging social signals.

2.2 Analysis of Interaction and Social Signals

The analysis of social interaction and social signals have a long tradition in Sociology and Psychology. Since 1995, computer-based social signal analysis has emerged in order to get a better understanding of related internal states of users [18].

Relevant for this work's context is the assessment from Clark and Krych that the observation of human social signals is mandatory for a mutual understanding of a dialog partner [8]. This assessment is not only relevant for face-to-face dialog but for every kind of social interaction, [6, 15]. The relevant dialog situation for this work is the interaction of two pedestrians across the street, which might come with the understanding that one or both has the intention to cross the street.

In the research area Affective Computing and in the research field of socially interactive agents of Human-Computer-Interaction (HCI), the multi-modal analysis and interpretation of social signals [21] provide reliable methods and techniques, even in real-time [1].

3 THE EMIDAS SYSTEM

The foundation of EMIDAS is a DBN that models a dyadic (group of two people) social interaction. The EMIDAS DBN intends to observe two pedestrians being on opposite sides of the street and predicts whether they have the intention to meet the other. This intention is derived by observing social signals that provide information about whether pedestrians interact with each other. The DBN dyadic engagement model from Bauer [4] inspires the EMIDAS intention representation. The EMIDAS DBN represents the dyad by having nodes assigned to each pedestrian (Fig. 1, left pedestrian *A*, right pedestrian *B*). More specifically, each feature that exists to model one of the pedestrian also exists for the second pedestrian.

3.1 Features

The features used to model each pedestrian are their *head orientation*, *body orientation*, currently performed *gesture*, and whether they are *approaching* the other pedestrian. The selection of all features is based on typical categorizations of non-verbal behaviors in the research field of social signal interpretation (Sec. 2.2). With regard to the representation of feature values, DBNs do not consider the logical order. Hence it is preferable not to define too many possible variable values since it would lower the number of observations per value. The *distance* between pedestrians is also used:

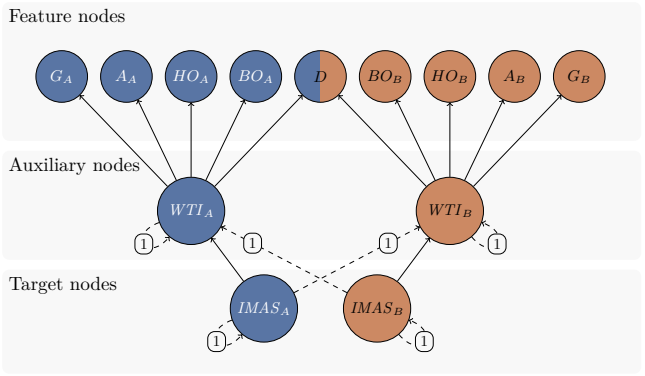


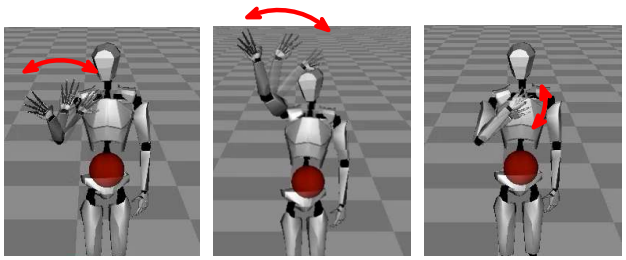
Figure 1: EMIDAS-DBN structure - solid (dashed) edges represent instantaneous (temporal) causal effects.

Head orientation (HO_A, HO_B) describes social signals associated with face and eye behavior and is of great importance since the face, and the eyes are humans' most prevalent communication tools [20]. Since we aim to find social signals that indicate whether two pedestrians are socially interacting, including the head orientation as a feature is indispensable. More precisely, we are interested in whether a pedestrian looks at the other one. It would be nice to include the facial expressions and gaze behavior since they might be correlated with a person's current intention to meet each other. However, from the position of a car-mounted camera, the facial expression of a pedestrian cannot easily be recorded. The values for *head orientation* refers to the other pedestrian B modeled by the DBN by examining the angle between the gaze direction of A and the vector from A to B: *central* or *peripheral* if the angle is at most 5° or 60°, respectively, otherwise: *outside*.

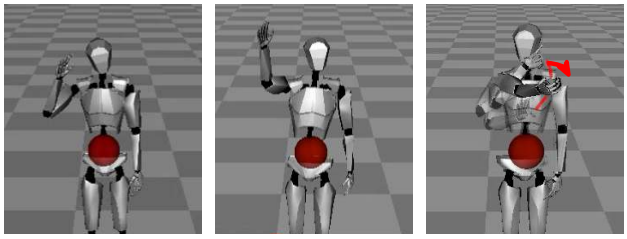
Body orientation (BO_A, BO_B) describes social signals associated with postures. These social signals are considered the most reliable cues about the attitude of people towards others [20]. We consider the body orientation of pedestrian A concerning the position of pedestrian B (and vice versa) by examining the angle between the facing direction of A's body and the vector from A to

B. If the angle is at most 10° or 90°, BO_A takes the value *strongly facing* or *slightly facing*, respectively, otherwise: *turned away*.

Gesture (G_A, G_B) represents social signals associated with gestures that are used to greet someone (Fig. 2). Note that the selected gestures are biased by the European culture. We consider two hand raise gestures, and two other gestures. The four waving gestures are a combination of two different hand levels and two different waving speeds. Figure 2a and 2b show the waving motions with low and high hand levels. The two hand raise gestures (Fig. 2d and 2e) differ in the highest position of the hand. The meaning of the two final gestures goes beyond greeting: signalizing that the other should come (Fig. 2c) or that one will come (Fig. 2f).



(a) Waving with low hand level. **(b) Waving with high hand level.** **(c) "Come here" gesture.**



(d) Arm raise with low hand level. **(e) Arm raise with high hand level.** **(f) "I'm coming" gesture.**

Figure 2: Gestures represented by the variables $G_{A/B}$.

Approaching (A_A, A_B) represents (social) signals associated with any behavior connected to the way people share and organize their spacial surroundings [20]. The feature represents that people currently interacting and want to get closer. The variable A_A for pedestrian A takes on the value *yes* if A is looking in the direction of B and if A walks and contributes to shortening the distance between A and B. If the reduction in distance is too small to be sure that A wants to approach B, *approaching* takes on the value *maybe*. Otherwise: *no*.

Distance (D_A, D_B) is not a social signal. We consider this feature because pedestrians usually don't just start to interact; they wait until they are sure that they know the other and that each other are recognized. The *distance* variable's values are discretized groups, each covering a distance of 3 m.

3.2 Model Prediction

The EMIDAS DBN aims to predict whether one or both pedestrians in the dyad intend to cross the street to meet the other. Since we limit ourselves to scenarios where the pedestrians interact with each other before crossing the street, we are interested in detecting ongoing social interaction. We model the described feature nodes as the effect of the cause *willingness to interact* (*WTI*) of each pedestrians (Fig. 1, center layer). Edges connect the *WTI* nodes with the features associated with the respective pedestrian.

To realize the circumstance that a pedestrian may have the wish to interact with the other, but does not have the intention to cross the street, we introduce the target variable *intention to meet across the street* (*IMAS*) for each pedestrian, which is the cause of the willingness to interact (Fig. 1, lower layer). Both feature variables *WTI* and *IMAS* take on the values *very high, high, medium, low* or *very low*. Both variables and their values are not observable since they represent internal mental states. Thus, there is no apparent algorithm that computes the value of both variables. Only their effects (the EMIDAS DBN features) are perceptible.

A critical advantage (compared to Bayesian Networks) of DBNs is that they all allow modeling influences over time with the help of temporal nodes and temporal edges. Temporal nodes allow modeling a feature variable which value changes over time. A temporal edge will enable us to model the influence of a variable on another variable's future state. In the EMIDAS DBN, all variables change their value over time. Hence, all nodes are temporal nodes.

The EMIDAS DBN temporal edges consists of two groups: *temporal self-loops* and *temporal causal edges*. Temporal self-loops are used for the variables $WTI_{A/B}$ and $IMAS_{A/B}$. Temporal causal edges connect $IMAS_A$ with WTI_B and $IMAS_B$ with WTI_A (Fig. 1).

To represent that previous value influences a variable's value in the present, each feature (variable) node at each time slice has a temporal self-loop that creates a chain connecting each node to its successor in time. This representation is used for the variables $WTI_{A/B}$ and $IMAS_{A/B}$ since these are internal processes that develop and coherently change over time. The temporal causal edges describe that one pedestrian's intention to meet the other (Fig. 1, lower layer) may result in the other pedestrian probably being willing to interact with them at a later stage.

3.3 Explainability of Intention Detection

The EMIDAS system allows for explainability of its pedestrian intention detection by visual means and inference.

Visual means. At each time t, an undirected, bipartite, fully-connected graph G_t can be constructed, called pedestrian interaction graph (PInG). G_t is defined as $G_t = (V_t^{left} \cup V_t^{right}, E_t)$, where $V_t^{left} = \{v_t^i \mid \forall i \in \{1, \ldots, N\}\}$ and $V_t^{right} = \{v_t^j \mid \forall j \in \{1, \ldots, M\}\}$ represent the pedestrians on the left and right side of street, respectively, which are visible from the perspective of the car-mounted camera. The edges $E_t = \{\{v_t^i, v_t^j\} \mid \forall v_t^i \in V_t^{left}, v_t^j \in V_t^{right}\}$ represent all dyads across the street. We use a weighted adjacency matrix A_t to represent the intention each pedestrian has to meet the other from the dyad across the street. Since G_t is a bipartite graph, A_t has the form

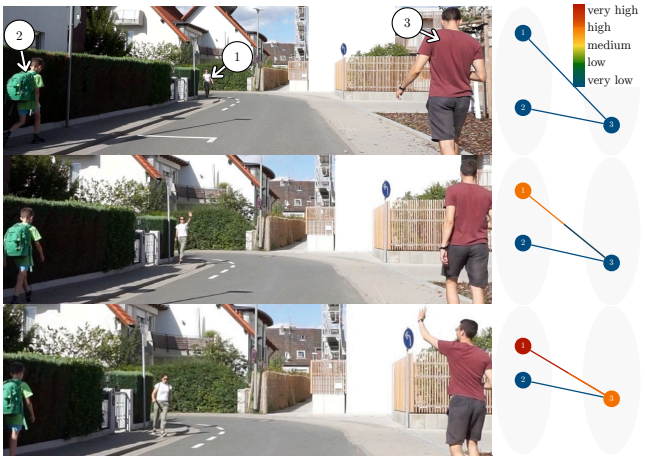
$$A_t = \begin{pmatrix} 0_{n,n} & A_t^{n \to m} \\ A_t^{m \to n} & 0_{m,m} \end{pmatrix},$$

where $A_t^{n \to m}, A_t^{m \to n}$ contain the intentions of the pedestrians on the left sidewalk to meet the pedestrians on the right sidewalk and

vice versa. Each entry of A_t is defined as $a_{i,j} = imas_t^{i\to j}$, where $i \in V_t^{left}, j \in V_t^{right}$. $imas_t^{i\to j}$ denotes the predicted intention of i to meet j across the street at time t. $imas_t^{i\to j}$ is computed by using all available probabilities given by the DBN-node. The values of the variable $IMAS_i[t]$ range from very high to very low using five classes, which we define to be an interval scale. An interval scale, in contrast to nominal or ordinal scales, allows to compute the average of different values. Let $\mathcal{P}_t^{i\to j}$ be the vector symbolizing the probability distribution of the variable $IMAS_i[t]$ given the observed features in the previous $t-1$ time steps w.r.t. pedestrian j. Assigning the number 5 to the value very high and the number 1 to the value very low, we define the *aggregated prediction* $imas_t^{i\to j}$ as follows:

$$imas_t^{i\to j} = \begin{pmatrix}5 & 4 & 3 & 2 & 1\end{pmatrix} \mathcal{P}_t^{i\to j} \tag{1}$$

Figure 3 shows an exemplary development of the PInG in the course of a scene from the view point of a car's camera field of view.



The frames show three pedestrians walking along the street. Pedestrian 1 starts to socially interact with pedestrian 2 and then crosses the street. Each edge $e_t = \{i, j\}$ is colored using the values $a_{i,j}$ and $a_{j,i}$ from A_t and each node i is colored according to the maximal value in the row vector $\begin{pmatrix}a_{i,1} & \dots & a_{i,n+m}\end{pmatrix}$.

Figure 3: Exemplary development of the PInG.

Inference means. Madigan et al. [14] present their evidence balance sheets which graphically display the weight of evidence in BNs and answer the question: "What is the relative importance of each of the features in determining the probability of some given value of the target variable?". Good [10] defines the weight of evidence as a measure for the explanatory importance of a particular finding E for a target hypothesis H. Let $\neg H$ be the negation of the hypothesis H.

$$W(H:E) = \log_{10} \frac{P(E \mid H)}{P(E \mid \neg H)} \tag{2}$$

Good also suggests to multiply the weight of evidence by 100 calling the resulting unit is *centibans*. The weight of evidence is measured using probabilistic inference. We adapt this approach to DBNs to take into account that variables have changing values over time. Several options can be considered to represent to weight of evidence of the dynamic features.

The first option measures the explanatory importance of a variable by considering all its values over time at once (Tab. 1). Changing the location of features in the evidence balance sheet changes their weight of evidence since the weight of evidence is conditioned on the features located above the specific feature. We determine the order in which we consider the feature in the evidence balance sheet using best-first search (BFS) on the WOE value.

Table 1: Evidence balance sheet for $P(IMAS_A[\text{t=9}] = \textit{very high} \mid E[\text{t=0:8}])$ that examines the total development of each feature.

DBN Feature	Feature values [t=0:8]	WOE (centibans)	Target Probability
Initial			10.25%
A_A	$yes_{0:5}$, $maybe_6$, $yes_{7:8}$	93.93	49.84%
D	$[12,15)_{0:3}$, $[9,12)_{4:8}$	95.07	89.87%
BO_A	$strongly\ facing_{0:2}$, $slightly\ facing_3$, $strongly\ facing_{4:8}$	14.84	92.58%
HO_B	$central_{0:8}$	3.26	93.08%
G_A	$None_{0:8}$	1.64	93.32%
HO_A	$central_{0:8}$	0.02	93.33%
BO_B	$strongly\ facing_0$, $slightly\ facing_{1:8}$	-0.16	93.30%
A_B	$yes_{0:5}$, $maybe_6$, $yes_{7:8}$	-23.21	89.09%
G_B	$high\ arm\ raise_{0:8}$	19.33	92.72%

The prior probability of $IMAS_A[\text{t=9}] = \textit{very high}$ is 10.25 % whereas the posterior probability given $E[\text{t=0:8}]$ is 92.72 %. The features A_A and D have the highest importance in determining the posterior probability of $IMAS_A[\text{t=9}] = \textit{very high}$. The order of the evidence in the evidence balance sheet was determined using BFS on the WOE.

An other option for representing the features' weight of evidence would be to measure the explanatory importance of every single variable at each time point independently. However, this is at the expense of understandability since the EMIDAS DBN has nine feature nodes and the evidence balance sheet would contain $9 \cdot T$ rows. Also, this amounts to treat the DBN the same as a BN.

A third option is to measure the explanatory importance of successive time slices having the same value (Tab. 2). By this method, we partition the temporal sequence of feature values to understand which value had a high importance in determining the posterior probability of the chosen value of the target variable.

The first option is able the capture the weight of evidence of the temporal development of each feature. As social interaction is not a static state but the exchange of social signals, the development of the features' value may have contributed to the result. This can be investigated with this presented option. However, the state of variables at a particular point in time may have had an exceptional influence on the result. This can be explained using the third presented option. Because of the aforesaid characteristic of social interaction, the second mentioned option is not useful in this context. Evidence balance sheets can serve as basis for various other options for investigating the features' weight of evidence. The order of the rows in the table as well as the partitioning of the features' values across the rows can be adjusted.

A second type of explaining the EMIDAS DBN applies the most relevant explanation (MRE) method presented by Yuan et al. [23]. The aim of this explanation type is to automatically create questions

Table 2: Evidence balance sheet for $P(IMAS_A[\text{t=9}] = \textit{very high} \mid E[\text{t=0:8}])$ that examines the features partitioned according to their value.

DBN feature	Feature values	WOE (centibans)	Target Probability
Initial			10.25%
A_A	$yes_{0:5}$	80.13	41.97%
G_A	$None_{0:8}$	39.71	64.34%
BO_B	$slightly\ facing_{1:8}$	19.89	74.04%
BO_A	$strongly\ facing_{4:8}$	8.43	77.59%
BO_A	$strongly\ facing_{0:2}$	1.80	78.31%
A_A	$yes_{7:8}$	0.78	78.61%
HO_A	$central_{0:8}$	0.03	78.62%
BO_A	$slightly\ facing_{3}$	0.03	78.63%
A_A	$maybe_{6}$	-0.14	78.58%
A_B	$maybe_{6}$	-5.53	76.36%
A_B	$yes_{7:8}$	-6.04	73.76%
D	$[9,12)_{4:8}$	-1.43	73.12%
D	$[12,15)_{0:3}$	74.85	93.84%
HO_B	$central_{0:8}$	0.61	93.92%
BO_B	$strongly\ facing_{0}$	-2.71	93.56%
G_B	$high\ arm\ raise_{0:8}$	-14.60	91.21%
A_B	$yes_{0:5}$	8.92	92.72%

The evidence $A_A[\text{t=0:5}] = yes$ has the highest importance in determining the posterior probability of $IMAS_A[t = 9] = very\ high$. The order of the evidence in the evidence balance sheet was determined using BFS on the WOE.

that give meaningful insights into the information captured by the DBN's structure and CPTs. Given a target variable $target \in \{IMAS_A, IMAS_B\}$, a selected target value $v_t \in \{very\ high, \ldots, very\ low\}$ and a set of feature nodes M, we use the MRE method to select a subset of features $X \subseteq M$ that conceive a relevant question about the evidence $target = v_t$. This question is then answered using probabilistic inference.

Given a trained EMIDAS DBN that unrolls to $T = 10$ slices, the target variable and its value $IMAS_A[9] = very\ high$ as well as the feature nodes $M = \{HO_A[t], BO_A[t] \mid \forall t \in \{0, \ldots 8\}\}$. We perform $MRE(M : IMAS_A[9] = very\ high)$ and obtain the following most likely explanation:

> ***Question:*** How likely is the prediction that A has a very strong intention to meet B in the next time step when B was in the central field of view of A and A was strongly facing B in the past nine time steps?
>
> ***Answer:*** $P(IMAS_A[9] = very\ high \mid HO_A[0{:}8] = central, BO_A[0{:}8] = strongly\ facing) = \mathbf{45.93}\,\%$

The subset X (in the example above $X = M$) is selected by going through all possible variable assignments and choosing the assignment with the maximal generalized Bayes factor (GBF, see [23]). Some feature assignments do not give a reasonable explanation and have to be pruned. For example, MRE can provide an explanation where pedestrian B is outside the field of view of A but A is strongly facing the pedestrian, which would mean that pedestrian A is turning their head back by at least 125° – but humans can only turn their head back by 90° into one direction. In the context of this work, we limited our search for relevant explanations to constant feature instances (the value of the features does not change over time) to provide simpler question, but let the method choose to which consecutive time steps it assigns the feature values. However, this method can also be used to examine the effect of the temporal development of the features' values.

Given a trained EMIDAS DBN that unrolls to $T = 10$ slices, the target variable and its value $IMAS_A[9] = very\ high$ as well as the feature nodes $M = \{HO_A[t], BO_A[t] \mid \forall t \in \{0, \ldots 9\}\}$. We perform $MRE(M : IMAS_A[9] = very\ high)$ and obtain the following most likely explanation:

> ***Question:*** How likely is the prediction that A has a very strong intention to meet B now when B is in the central field of view of A for the past second, A is approaching B for the past second and A is in the mid peripheral area of B for the past four time steps?
>
> ***Answer:*** $P(IMAS_A[9] = very\ high \mid HO_A[0{:}9] = central, A_A[0{:}9] = y, HO_A[6{:}9] = mid\ peripheral) = \mathbf{77.05}\,\%$

These method to generate explanations allow comprehensible insights into the knowledge base of the DBN. Otherwise, the knowledge base is distributed over all CTPs in the form of conditional probabilities of the child's values given the parents. The explanations can be a support during the conception of DBNs. But more importantly, such explanations provide understandable insights into the model's reasoning and knowledge base to build trust in the model's decisions.

3.4 Implementation

EMIDAS is implemented in Python 3.7 and uses the Python wrapper of the software library SMILE [1] for training the EMIDAS DBN and for performing inference tasks. When a DBN is conceived, it is necessary to define the temporal horizon considered by the DBN. This is necessary, as in practice it is not possible to consider a DBN that unrolls into infinitely many time slices. It is neither possible to train a network with infinitely many nodes nor to perform inference with unlimited data from the past. Let T be the amount of time slices considered by the DBN. To train a DBN with SMILE, the software requires a training dataset that is adapted to the time horizon T. The training data should consist of temporal sequences of data of length T. SMILE uses the EM algorithm to learn the parameters of BNs and DBNs [5].

To predict the intention of pedestrians to meet another pedestrian across the street, we apply the prediction inference query. Given the evidence data $e_{0:t-1}$ from the time slices 0 to $t-1$, SMILE provides the posterior probability distribution of the unobserved variables $IMAS_A[t], IMAS_B[t], WTI_A[t], WTI_B[t]$. SMILE implements various exact inference algorithms as well as various stochastic sampling algorithms. Here, the clustering algorithm was used to obtain the posterior probability distribution of the unobserved variables.

[1] www.bayesfusion.com

4 EVALUATION

To train and evaluate the EMIDAS DBN model, we created a dataset that met our requirements: 1) scenes that capture an interaction between pedestrians across the street followed by a possible road crossing; 2) scene annotations regarding the intention of pedestrians to cross the street to meet some other pedestrian with the necessary ground truth; 3) scenes from ego vision that reproduce possible video footage from the automated vehicles' camera(s).

4.1 Data Set and Scenarios

We identified seven scenarios(S1 - S7) that show two pedestrians interacting across the street for the first step of creating the dataset. Five of them are critical scenarios since they include a street crossing (Fig. 4). The two remaining scenarios are not critical (Fig. 5). Each scenario consists of five consecutive points in time, connected by action transitions (ⓐ - ⓓ), in which the pedestrians perform different social actions together with a walking or a standing action. For each point in time, each pedestrian viewing direction is represented by a gray cone.

Scenario 1 - 6 contains social interaction across the street, consisting of simple greetings or easily understandable gestures (Fig. 2). The scenarios were created with the aim to cover interaction that precedes a sudden change of intention towards crossing the street. We limited ourselves to scenarios where the pedestrian's social interaction is clearly visible, i. e., where social actions beyond directed gaze occur as, for example, by means of gestures. We consider scenarios with two pedestrians only. Moreover, we restricted ourselves to scenarios that primarily occur in residential areas. In these areas, we think it is more likely that a pedestrian will suddenly cross the road without making sure that a car is approaching and without looking out for a crosswalk first.

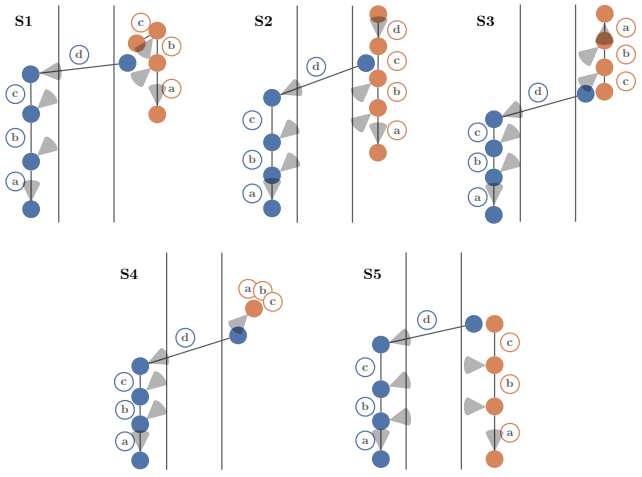


Figure 4: Critical scenarios with pedestrian social actions: ⓐ look straight ahead, ⓑ look towards other pedestrian, ⓒ perform gesture, ⓓ look straight ahead/to other pedestrian.

For obtaining a dataset with the necessary ground truth of the intention to cross the street to meet the other pedestrian, we conducted a distinctive user study. This approach is in accordance with the study conducted for the PIE dataset [19].

4.2 User study

To gather the ground truth of the intention to cross the street to meet the other pedestrian depending on the pedestrians' behaviour, we conducted a user study with 25 participants (14 female, M = 25.36 years, SD = 2.83 years) confronting them with staged scenes showing interactions of two pedestrians over the street. The task of the participants was to label each data point with the willingness to interact and intention to meet across street of each pedestrian.

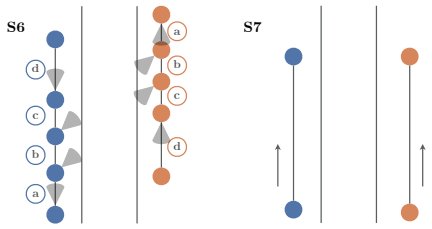


Figure 5: Non-critical scenarios with two pedestrians. In scenario 6, the pedestrians greet each other. Scenario 7 does not contain social interaction.

4.2.1 Procedure. After receiving a link to the online questionnaire via e-mail, participants were given a detailed explanation of the procedure and the consent form. Then, 28 different videos (cf. Sec.4.2.2) showing two pedestrians were presented. The videos were sorted regarding the scenes and presented in a way that the premature end was moved to a later point in time, allowing the video to be viewed progressively and finally showing the video completely. After each partially shown video, participants were asked for their assessment on the level of the pedestrians' intention to meet the other across the street as well as on the level of the pedestrians' willingness to interact with the other without necessarily crossing the street. Participants were asked to base their answers on the premature end of the video. Moreover, they were asked to indicate on what observation they base their decision. Participants were asked whether the scene shown was realistic and how frequently such a scene occurs in real life for the videos that were played entirely. The survey ended with a typical demography questionnaire asking for gender, age, and field of study. All participants received a 15€ voucher for a 60 – 90 min participation (M = 83 min, SD = 38 min, n = 23) via email.

4.2.2 Material. Based on the mentioned scenarios, we created 60 videos capturing different variations of actions in the action transitions (Fig. 4 and Fig. 5 ⓐ - ⓓ) but staying in agreement with the scenario descriptions. We thought that different actions (e.g., gestures) might influence the ground truth. The videos were shot according to the defined scenarios with two instructed persons (a woman and a man). The pedestrians would be walking on opposite sides of the street and possibly cross the street after interacting. The videos were recorded with a static camera to ensure that the

pedestrians are visible for the interaction's entire duration. From the 60 videos, we selected 28 that differ significantly from each other. In line with our goal to examine the pedestrians' intentions to cross the stress, the videos were cut at different times. Depending on the video, between 1 and 5 points in time were selected at which the video was cut ($M = 2.46$, $SD = 0.91$).

4.2.3 Measurements. For the premature videos, *Intention to meet the other across the street (IMAS)* and *Willingness to interact (WTI)* were measured on a five-point Likert scale with construct-specific response choices from 1 (*very low*) to 5 (*very high*). With an open question after the two scales, we collected *explanations* how participants assessed the IMAS and the WTI. For the complete videos, *Realism of the scenes* was assessed on a five-point Likert scale from from 1 (*very unrealistic*) to 5 (*very realistic*). *Frequency of occurrence in reality on the scale* was assessed on a five-point Likert scale from from 1 (*very rarely*) to 5 (*very often*) after the completed videos. We gathered information about the reasons behind the given value for the IMAS and WTI assessments with an open question.

4.2.4 Results. All videos received, on average, a high to very high rating concerning the realism (Tab. 3). The frequency scores are ranging from rarely to very often. As critical scenes might only occur with low probability, we reached the needed variability of frequency. In general, all scenes were generated in OpenDS (Sec. 4.3).

Table 3: Realism and frequency values for each scenario.

Scenario	# scenes	Realism	Frequency
1*	6	$M = 4.03$, $SD = 0.9$	$M = 3.05$, $SD = 0.87$
2*	2	$M = 3.24$, $SD = 1.2$	$M = 2.28$, $SD = 0.9$
3*	10	$M = 4.03$, $SD = 0.89$	$M = 3$, $SD = 0.89$
4*	3	$M = 4.32$, $SD = 0.7$	$M = 3.27$, $SD = 0.89$
5*	4	$M = 3.65$, $SD = 1.13$	$M = 2.73$, $SD = 0.92$
6	1	$M = 4.6$, $SD = 0.58$	$M = 3.92$, $SD = 0.81$
7	2	$M = 4.96$, $SD = 0.2$	$M = 4.8$, $SD = 0.4$

* Critical scenarios. *Note.* Realism measurement: 5-point scale from 1 (*very unrealistic*) to 5 (*very realistic*). Frequency measurement: 5-point scale from 1 (*very rarely*) to 5 (*very often*). Both were measured after the participants saw the full scene.

As assumed, the IMAS of the pedestrian, who eventually crosses the street, increases steadily (Fig. 6). The IMAS of the pedestrian who does not cross the road remains at a low level. The WTI increases steadily for both pedestrians. The WTI of the pedestrian who does not cross the road to interact with the other stays inferior to the WTI of the crossing pedestrian. It can also be observed that the WTI level of the pedestrian who does not cross the road rises even before the time point "p2 gaze". This is because, in five out of the 25 videos showing critical scenes (1-5), the pedestrian does not cross the road that starts the interaction, i.e., who is p1. This occurs in scenario 3, 4, and 5. In the remaining 20 videos, the pedestrian that does not cross the road has not seen the other pedestrian before the relevant time point "p2 gaze". Thus, their WTI with the other is inferior to the willingness of the other during all relevant time points before "p2 gaze".

An analysis of the open questions regarding the reasons behind the IMAS and WTI assessments show that participants paid attention to the following signals: (mutual) gaze, whether a pedestrian was in the field of view of the other, body orientation (towards the other pedestrian or the street), walking direction, distance between both pedestrians, used gestures, whether a pedestrian calls to the other, and whether a pedestrian takes a step towards the street.

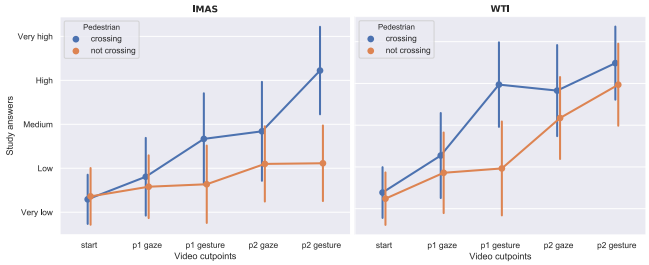


Figure 6: Average and standard deviation of the study participants' assessments about IMAS (left side) and WTI (right side) at each relevant time point over all critical (1-5) scenarios. Each video could not be cut at all five relevant time points. The relevant time points are *start* (before the interaction begins, $n = 75$), *p1 gaze* (p_1 starts the interaction by looking towards p_2, $n = 150$), *p1 gesture* (p_1 makes a gesture, $n = 250$), *p2 gaze* (p_2 looks towards p_1, $n = 425$), *p2 gesture* (p_2 makes a gesture, $n = 650$). The provided n-values represent the number of participants times the amount of videos cut at that time point.

4.3 Benchmark OpenDS-CTS2

The dataset OpenDS-CTS2 consists of 15 949 synthetic scenes created in the 3D driving simulator OpenDS [2] in a three step approach. First, each video used in the study questionnaire was synthetically reproduced (Fig. 7).

Then, each synthetic replication was used to create three additional scenes by mirroring the pedestrians' path along the midline of the road and along the axis perpendicular to the midline of the road. In this way, the EMIDAS DBN will not be biased on the side of the crossing pedestrian. Finally, all obtained scenes were varied by adding small variations in the pedestrians' walking path and altering the pedestrians' walking velocity. The dataset OpenDS-CTS2 contains 2607 scenario-1 scenes, 2134 scenario-2 scenes, 7260 scenario-3 scenes, 375 scenario-4 scenes, 2985 scenario-5 scenes, 196 scenario-6 scenes and 392 scenario-7 scenes. The number of scenes per scenario is not evenly distributed because the different scenario properties allowed to generate an additional amount of scenes.

4.4 Results

We evaluate the prediction performance of the EMIDAS DBN on the *IMAS* variables for different time horizons $T \in \{10, 20, \ldots, 60\}$ of the DBN, where the step size represents 0.1 s. To obtain the *IMAS* predictions of pedestrian A and B at time t, all features in the interval $[t - T + 1, t - 1]$ are inserted in the leaves of the DBN. The upper bound $T \leq 60$ was selected due to the average occurrence of the ad-hoc change of intention at $t = 6.16$ s among all critical scenes in OpenDS-CTS02.

[2] https://opends.dfki.de

Figure 7: Real scene (top) versus related synthetic reproduction (bottom).

We use leave-one-out cross-validation to evaluate the prediction, where each EMIDAS-DBN$_T$ is trained on six scenarios and validated on the seventh one - repeated for each scenario. The evaluation relies on the Pearson correlation coefficient (PCC) [17] between the ground truth and the aggregated prediction value $r_{GT,AGG}$ (c.f. Eq. 1). A value of -1 is interpreted as a perfect negative linear correlation, 0 is interpreted as no linear correlation, and 1 is interpreted as a perfect positive linear correlation.

Table 4: PCC results $r_{GT,AGG}$ for EMIDAS-DBN$_{10}$.

Variable	Test scenario						
	1*	2*	3*	4*	5*	6	7
$IMAS_A$	0.84	0.67	0.59[a]	0.82	0.83	0.23	0.2[c]
$IMAS_B$	0.83	0.69	0.59[b]	0.82	0.83	0.29	0.22[d]

* Critical scenarios
[a] Highest PCC at $T = 60$, $r_{GT,AGG} = 0.66$
[b] Highest PCC at $T = 50$, $r_{GT,AGG} = 0.66$
[c] Highest PCC at $T = 20$, $r_{GT,AGG} = 0.21$
[d] Highest PCC at $T = 30$, $r_{GT,AGG} = 0.24$

Table 4 shows that the prediction on the critical scenarios 1 to 5 has a strong positive linear relationship with the ground truth. This observation does not hold for the non-critical scenarios 6 and 7. A reason might be that the numbers of critical and non-critical scenarios are not equally distributed in the dataset. We observed that having more past data, i. e., a larger T, does not significantly enhance the prediction. To find an explanation for this observation, we examined the temporal development of some feature variables used in the EMIDAS DBN. We found that the value of many feature variables characteristically changed in the time close to the street crossing. For example, the head and the body are oriented towards the other, and a gesture was performed.

5 CONCLUSIONS AND FUTURE WORK

The implications of this work are twofold. First, the hybrid interpretation of social signals and interaction across the street increases the precision of state of the art pedestrian path prediction algorithms [16]. Second, and this is the focus of this paper, cars using this kind of algorithm can explain related decisions and actions. Motivated by the theory of mind of others' approach, the presented EMIDAS cognitive software model interprets social signals concerning one pedestrian's intention to meet another across the street. Based on this, elaborate explanations can be generated.

We see the EMIDAS dynamic Bayesian network (DBN) model as a starting point for hybrid, explainable, accurate, and fast social interaction-based prediction of pedestrian movements in streets. Therefore, we described the necessary modeling process that uses various relevant social signals to predict the intention to meet another pedestrian across the street. The conducted initial user study indirectly validated our designed model since the model could predict the intentions collected in the study well. Nevertheless, an extensive study in the context of traffic psychology is indispensable to refine the presented approach. Using the EMIDAS DBN's structure, we showed ways to explain the model's reasoning and underlying knowledge base.

The EMIDAS approach can be extended in several ways. In future work, it is necessary to consider how to handle pedestrians that are occluded or that a pedestrian involved in social interaction is not visible in the entire scene. Concerning enhancing the explanation of the reasoning process, future work has to examine whether obtained explanations are valuable and contribute to building confidence in the model's decisions. It is then particularly challenging to explain which aspects of social interaction lead to a particular intention prediction. Also, our corpus contains scenes with only a single pedestrian on each sidewalk. It can be extended to more complex scenes with changing numbers of pedestrians on each side of the street. Moreover, the performance of EMIDAS has to be investigated when the scenes are recorded from a moving vehicle.

ACKNOWLEDGMENTS

This work has been funded by the German Ministry for Research and Education (BMBF) in the project REACT under grant 01IW17003 and the German Research Foundation (DFG) within the DEEP project (funding code 392401413).

REFERENCES

[1] Wagner, J., André, E. (2018): Real-time sensing of affect and social signals in a multimodal framework: a practical approach. In The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition-Volume 2 (pp. 227-261).
[2] Abduallah, M.; Kun, Q.; Mohamed, E.; Claudel, C. (2020): Social-STGCNN: A Social Spatio-Temporal Graph Convolutional Neural Network for Human Trajectory Prediction. In Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).
[3] Alahi, A.; Goel, K.; Ramanathan, V.; Robicquet, A.; Fei-Fei, L.; Savarese, S. (2016): Social LSTM: Human trajectory prediction in crowded spaces. In Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).

[4] Baur, T. (2018): Cooperative and transparent machine learning for the context-sensitive analysis of social interactions. Dissertation, CSD, University of Augsburg, Germany.
[5] BayesFusion, LLC (2020): GeNIe Modeler User Manuel (Version 2.5.R4). Retrieved from https://support.bayesfusion.com/docs/GeNIe.pdf. Accessed: June 29, 2020.
[6] Birdwhistell, R. L. (2011): Kinesics and context: Essays on body motion communication. University of Pennsylvania press.
[7] Blaiotta C. (2019): Learning generative socially aware models of pedestrian motion. IEEE Robotics and Automation Letters, 4(4):3433–3440.
[8] Clark , H. H. and Krych, M. A. (2004) Speaking while monitoring addressees for understanding. Journal of memory and language, 50(1):62–81.
[9] Ferrer G.; Sanfeliu A (2014): Behavior estimation for a complete framework for human motion prediction in crowded environments. In Proc. of IEEE Int. Conf. on Robotics and Automation (ICRA).
[10] Good, I. J. (1985): Weight of Evidence: A Brief Survey. Bayesian Statistics, (2), 249-270.
[11] Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A. (2018): Social GAN: Socially Acceptable Trajectories with Generative Adversarial Nnetworks. In Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (pp. 2255-2264).
[12] Lee, N., Choi, W., Vernaza, P., Choy, C. B., Torr, P. H., Chandraker, M. (2017): Desire: Distant Future Prediction in Dynamic Scenes with Interacting Agents. In Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (pp. 336-345).
[13] Liang, J., Jiang, L., Niebles, J. C., Hauptmann, A. G., Fei-Fei, L. (2019): Peeking into the future: Predicting Future Person Activities and Locations in Videos. In Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (pp. 5725-5734).
[14] Madigan, D., Mosurski, K., Almond, R. G. (1997): Graphical explanation in belief networks. Journal of Computational and Graphical Statistics, 6(2), 160-181.
[15] Mehrabian, A. (1980): Public Places and Private Spaces: The Psychology of Work, Play, and Living Environments. Basic Books.
[16] Muscholl, N., Poibrenski, A., Klusch, M. and Gebhard, P. (2020) SIMP3: Social Interaction-Based Multi-Pedestrian Path Prediction By Self-Driving Cars In: Proc. of IEEE International Symp. on Computational Intelligence in Vehicles and Transportation Systems (CIVTS) (pp. t.b.a.).
[17] Pearson, E. S. (1931): The Test of Significance for the Correlation Coefficient. Journal of the American Statistical Association, 26(174), 128-134.
[18] Picard, R. W. (1997): Affective Computing. MIT Press, Cambridge, MA.
[19] Rasouli, A., Kotseruba, I., Kunic, T., Tsotsos, J. K. (2019): PIE: A large-scale Dataset and Models for Pedestrian Intention Estimation and Trajectory Prediction. In Proc. of the IEEE International Conf. on Computer Vision (pp. 6262-6271).
[20] Vinciarelli, A., Salamin, H., Pantic, M. (2009): Social Signal Pprocessing: Understanding Social Interactions Through Nonverbal Behavior Analysis. In Proc. of IEEE Conf. on Computer Vision and Pattern Recognition Workshops (pp. 42-49).
[21] Vinciarelli, A., Esposito, A. (2018): Multimodal Analysis of Social Signals. In The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition-Volume 2 (pp. 203-226).
[22] Xue, H., Huynh, D. Q., Reynolds, M. (2018): SS-LSTM: A hierarchical LSTM model for Pedestrian Trajectory Prediction. In Proc. of IEEE Conf. on Applications of Computer Vision (WACV) (pp. 1186-1194). IEEE.
[23] Yuan, C., Lim, H., Lu, T. C. (2011): Most relevant explanation in Bayesian networks. Journal of Artificial Intelligence Research, 42, 309-352.